Эконометрика — Г. Г. Канторович

Последнее обновление имело быть 10 июня 2012 г. в 21:23:55.

Примечание составителя. В соответствии с математической традицией имена векторов указываются либо со знаком вектора наверху, либо жирным символом (так, эквивалентны выражения «$\vec{X}$» и «$\boldsymbol{X}$»). Матрицы также обозначены жирным символом (например, «$\boldsymbol{H}$»).

1 Лекция 1

Специфика эконометрики: народу много, но это первая лекция. К концу курса можно будет сопоставить средний балл, с какого раза сдаётся домашка, опираясь на эти данные.

Курс большой. Во всех мировых планах это один из трёх китов. Это пришло относительно недавно. В СССР курса эконометрики не было ни в одном вузе. Круг вопросов входил в разные предметы. С 1995 года читается этот курс. Нигде в России такого курса не было, а сейчас он стал необходимой частью. Чем занимается эта наука на наших занятиях? Лекции и семинары еженедельные. У кого Мамонтов, у того называется «мамонтометрика». Отличие системы: есть домашки, которые надо сдавать (это самое сложное). У Мамонтова есть упрямость выспрашивать каждого. Каверзные вопросы, чтобы понять, как макрос написан. И это сложно. А кто-то придёт в последний день и будет до 11 вечера под дверью сидеть.

Сложно дать точное определение. «Эконометрика» — слово сложное. Ввёл его Фриш (нобелевский лауреат) как измерения в экономике. Мы изучали микру, знаем классическую микру, и экономические явления описываются в рамках матмоделей и соотношений. Адам Смит ни одной формулы не использовал. А в начале XX века экономика стала эксплуатировать математику. Кроме простых графиков и процентов, там ничего не было. В 30-е годы появился журнал «Econometrica», это журнал №1 сейчас. Это применение математики, выходящее за простые проценты, формулы и графики. Сейчас два русла: матэкономика и эконометрика. Наш курс — последнее в узком смысле. Эконометрика — это раздел экономической науки, изучающий количественные отношения между переменными, выраженными в математической форме, чтобы опровергнуть или подтвердить возможность использования этих количеств. Мы перекладываем понимание экономики так: сначала мы работаем с агентами вербально. Что нам говорит экономическая теория о такой материи, как закон спроса? Обычно приводят по-разному, но в макре пересказывают Кейнса: «Основной экономический закон говорит, что она или он при увеличении дохода в общем и среднем увеличивает своё потребление, но не в той мере, в какой возрастает доход». Этот закон выражен в вербальной форме. Количества есть, меры есть, отношения есть. Как мы работаем с этим законом? Можно работать так, как Смит с Рикардо пошли в общих терминах. Но появляются вещи, которые кажутся нам естественными. Но что такое доход? Что такое потребление? Мы используем категории, потом выяснится, что это располагаемый доход. Первый шаг — выразить то, что мы сказали вербально, в форме математических соотношений. Используем могучий аппарат: если у нас есть расходы (D), то эта переменная есть функция, зависящая от дохода и ещё чего-то (в общем и среднем). $D = f(I, \ldots)$ возрастающая («вторая функция позитивная», — пишут в плохих переводах).

Вместо категорий теория видит переменные (доход, цена). Поэтому мы их намеренно мерим количественно. Матмодель — это ничего хитрого, это совокупность уравнений и неравенств, связывающая экономические переменные. Вот наша модель:

$$\begin{cases} Y = G(X, Z, \ldots); \\ Z = \ldots; \\ X \leqslant S. \end{cases} \tag{1.1}$$

Наше первое ружьё, которое мы вешаем на стену. Но что такое цена? Какая это переменная? Появляются три группы значков. Первая — это **переменные**, которые существуют в модели. Это те экономические величины, которые мы в рамках нашей задачи подразумеваем возможно изменяющимися. Примеры переменных: ВВП (в кризис падает), цена, спрос и предложение (посложнее). Вторая — это **параметр**. Это та переменная, которая не меняется на некотором неизвестном количественном значении. Например, $C = a + bY$ — это маленькая модель. C и Y — переменные, а a и b — параметры, которые зафиксированы. По-другому зададим вопрос, и переменная и параметр могут поменяться. Третья градация — **коэффициент** (параметр с известным численным значением). Если $b = 0{,}8$, то тогда это коэффициент. Переменные меняются, параметры хотя бы постоянные, а коэффициенты мы знаем. Так как матмодель связывает несколько переменных, то мы делим переменные на несколько классов. Самый большой — это независимые и зависимые переменные. Когда переменных много, то мы не можем сказать (как Z в примере), что зависимо, что независимо. Надо рассматривать как неявно заданную вектор-функцию многих переменных. Все входящие переменные делятся на **экзогенные** (это *переменные* (*sic!*), но чьё значение должно стать известным вне рамок модели) и **эндогенные** (относительно которых мы разрешаем систему уравнений $y = \ldots$). Имея систему уравнений, мы можем разрешать её по воле произвола. Однако экономика нам диктует, что надо считать экзогенным, что — эндогенным. Пусть есть рынок товара со спросом и предложением. Пишем: $D = f(P)$, $S = g(P)$. Можем специфицировать вид функций: $f(P) = \alpha - \beta P$, $g(P) = \gamma + \delta P$. Надо добавить ещё одно условие: $D = S$. Получается маленькая модель, которая позволяет скапитанить ещё об одном предмете: среди уравнений мы будем выделять поведенческие (бихевиоральные: спрос есть убывающая функция от цены) и соотношения, являющиеся экономическими тождествами ($R = E + \pi$). Вот так мы их ввели, они просто связаны тождеством. Третий тип — это экономический тип ($D = S$). Равновесие — это третий тип равенства. В модели 3 переменных: D, S, P. Решить модель — это приравнять $f(P) = g(P)$. В

этом уравнении 3 переменных, 3 неизвестных, находятся все численные значения, нет экзогенных переменных. Если что-нибудь ещё добавить, что человек может сберегать ($D = f(P, i)$), то экзогенной станет ставка процента. При наращивании соотношения она может стать и эндогенной. Вот это математическая экономика.

Перейдём к эконометрике. Приходим мы на работу в банк. Вызывает начальник и говорит: «Ты умный, скажи, какой курс устанавливать» (начальник редко говорит более конкретно). Смотрим на сеть обменных пунктов. Вспоминаем, чему учили.

Первый этап любого исследования — это выдвижение гипотезы. Можно придумать самим, можно найти в литературе. Как правило, гипотеза выдвигается вербально. Доллар — это товар, поэтому должен действовать закон спроса. Чем выше цена, тем меньше у нас купят. Объём покупок в долларах убывает при увеличении курса продажи (в общем и среднем при прочих равных условиях, *ceteris paribus* — Кейнс). Вторая ситуация — это мы выдвигаем теорию. Григорий Гельмутович считает, что тем выше финальная оценка, чем чаще человек посещал лекции и семинары. Или: в связи с замедлением темпов роста упадёт спрос и цена на нефть. Мы предполагаем, что объём покупок упадёт, но если мы так скажем, то это будет наш последний день на работе.

Второй шаг: мы будем количественно наполнять модель. Тогда почти понятно: $Q = f(P, \dots)$. И тут есть неопределённость (прелесть исследования). Самая простая убывающая функция — это линейная функция. $Q = \alpha - \beta P$, где P — цена, $\beta > 0$. Можно написать: $\ln Q = a + b \ln P$. Что ещё может влиять на объём продаж? Если пункт расположен в большом универмаге или большом торговом доме, а другой — на улице, то мнения разные. В универмагах курс хуже для нас. Число влияет (пятого зарплата). Дождь на улице влияет? Встаёт одна из важнейших задач: надо отразить только существенные связи между переменными. Тут нужен опыт, научно-исторический опыт, чутьё экономиста. Влияет, если рядом есть другие обменники. Есть принцип бритвы Оккама, и тут надо тоже отобрать существенное. Каким образом отразить в модели то, что не учтено? Эконометрика делает это единообразно: то, что мы явно не учли, тоже как-то влияет. Будет определённо случайная переменная, которой нет в модели. Чаще всего получается аддитивная случайная величина. Неучтённое заставляет нас трактовать это как случайную величину. Детерминированный хаос пытается описать аналогичные величины. $Q = \alpha - \beta P + \varepsilon$. Мы определили зависимость со случайным параметром.

Третий важнейший этап[1] — данные. Мы идём от реальности и пытаемся вывести оттуда количественные соотношения, связывающие переменные. Для продолжения подхода нам нужны данные. Они порождают несколько сложных задач. Мы должны каким-то образом измерить количественно входящие переменные. Если мы продолжаем аналогию с обменным курсом, то знаем: если днём установить курс такой-то, то выручка такая-то. Другой курс — другая величина. Вот сюда подставляем и ищем параметры. Померить мы можем только объём выручки. Значит, должен быть спрос на что-то. Не так очевидно, почему объём выручки — спрос. Когда их много, люди могут удовлетворить свой спрос. А в Белоруссии кто-то принесёт 100 долларов, а следующий их купит. Вспоминаем курс СЭС, и экономическая теория что-то говорит о ВВП, а СЭС говорит: так мы его мерим. Пусть говорят: цена на акцию упала на 2,2 % — и сразу куча сделок. Надо как-то их измерить и оценить. Ещё пример: министерство экономики публикует отчёт, и там будет среднегодовой курс доллара. Но то, что можно измерить, не всегда подразумевается экономической теорией. Спрос невозможно измерить. Есть всегда ошибка. Росстат — это огромная служба, которая вырабатывает культуру правдивой отчётности. Пусть мы работаем в большом банке с десятком курсов. Будут данные за один и тот же период времени по разным объектам. Можно опрашивать о расходах и доходах. Это называется cross-section (идентичные данные по группе объектов в один и тот же момент времени). Очень часто приходится встречать. А если нет возможности сделать cross-section, то тогда мы в маленьких Чебоксарах двигаем курс и смотрим в последовательные моменты за одним и тем же объектом (динамический ряд). В чём разница? В cross-section нет соседей. А в наблюдениях ряда важна последовательность. Если вчера курс был 39, то завтра он не станет 50. Третий, комбинированный тип — pooled. Пусть у нас 10 курсов, и мы за каждым курсом следим. Сочетание и того, и другого. Люди могут договориться в контакте («вы контактные люди»), и один момент: если мы работаем в районном центре, то есть шанс получить информацию по всем курсам во всех обменниках (собрать генеральную совокупность). Эта информация не есть бесплатна. В Москве такие кустарные методы обречены, поэтому применяются выборочные методы, где мы помним об ошибке. Третий пункт — сбор данных — редко сделан за вас.

Четвёртый этап — это в плохих учебниках он отождествляется с эконометрикой. Есть данные, есть модель, надо найти параметры. Если 10 данных, то слишком много уравнений для двух параметров. Поэтому параметры оценивают. Сразу понятно, что получится оценка и вероятностная характеристика. Полученные параметры назовём $\hat{\alpha}$, $\hat{\beta}$.

Пятый этап — диагностика модели. Но те количественные оценки — удовлетворяют ли они разумным требованиям (если $\beta < 0$, то доллар впервые в истории станет товаром Гиффена)? Нет? Значит, что-то недоучли. Если модель проходит диагностику, то всё нормально. Если она не проходит её, то что делать дальше? Мы можем учесть то-то и то-то, сменить МНК на что-то другое. Оказываемся на этапе 4. Можно получить дополнительные данные. Может, неправильно специфицирована модель, что-то не учтено, неправильное выражение. И совсем замечательная ситуация: неверна наша гипотеза. Сегодня теория — набор ветвей, описание, которое можно опровергнуть. Если мы походили по этапам и получили модель, то вопрос не кончается. Начинается вопрос: зачем строили? При выполнении рутины за скобками внимания остаётся что-то для начальника. Если принести

[1] «Они все важнейшие», — сказал Григорий Гельмутович, поэтому запятая не требуется.

оценки коэффициентов и графики, то, может, пройдёт. Выделим две группы. Первая — это что модель адекватна, нормально передаёт реальность, построенная модель не противоречит имеющимся данным (а иногда факты соответствуют модели). Это значит, что с помощью такой модели получили DGP (Data Generated Process).

Популярны исследования: от чего зависит успеваемость детей. Где живёт, где учились, кто мама и папа, и выяснилось, что наличие у мамы высшего образования влияет, а у папы — нет. Образование папы — это ε. А начальник от нас ждёт количественных оценок: если поднимем на 2 копейки, будет так, а если опустим, то так. Можно оценить величину эндогенной переменной по величине экзогенной, не входящей в состав наблюдений (prediction and forecasting). Если ставку экспортной пошлины установить такой, то налоги будут такими. На английском языке море. «Путеводитель в современную эконометрику» не очень хорошо переведён (Verbeek). Неплохой. Хороши старые учебники (Маддалы — Introduction). Excel хорошо помогает всё понять. Хорошо работать в Excel.

2 Лекция 2

Отличительная черта, выделяющая эконометрические исследования, — это *боль*. Если в физике исследователь хочет проверить, как пружина бросает шарик, то он может поставить эксперимент сто раз. В экономике мы лишены направленного эксперимента за редчайшим исключением. В химии можно смешивать воду со спиртом в определённой пропорции, и с вероятностью 0,99 водка получается. А инфляция в январе — одна цифра. Остаётся либо сказаться на описательность, либо работать в единичном пассивном эксперименте.

Пусть кто-то пытается продать или купить доллары в сотне обменных пунктов в городе N, как в прошлом примере. Он собрал данные о цене и объёме. На оси p (горизонтальной, не как у этих экономистов!) будет цена, а зависеть от неё будет количество. Объём спроса падает при падении цены. Но для одного и того же курса есть несколько значений, поэтому это не функция. Приходится по-другому интерпретировать. Эконометрика говорит, что у нас ещё факторы (место, дата, пробка на улице), поэтому мы говорим, что наблюдаемая величина Q является случайной величиной, которая может принимать разные значения. Это приводит нас в мир ТВиМСа. Бессмысленно спрашивать, какой будет значение случайной величины. Если мы пробежали все пункты в городе, то мы собрали генеральную совокупность (population). Значение случайной величины бессмысленно, а характеристика смысл имеет. Например, матожидание. У нас дискретная случайная величина, поэтому матожидание будет каким-то средним. Для каждой позиции по x ищется среднее значение. Это называют $\mathbb{E}(Y \mid X)$ или $\mathbb{E}(Y \mid X_i)$ в дискретном случае. $\mathbb{E}(Y \mid X) = f(X_i)$. Эту функцию мы назовём регрессией Y на X. Это условное математическое ожидание. Но тогда наблюдаемое нами значение Y есть $Y_i = f(X_i) + \varepsilon_i$, где последнее слагаемое — аддитивная случайная составляющая. Есть некоторая детерминированная величина плюс случайная добавка. Это называют уравнением регрессии. Переменная X — регрессор (фактор). Y — regressand, а по-русски — зависимая переменная. Поскольку у нас полная ГС, мы считаем матожидание: $Y_i = \mathbb{E}(Y \mid X) + \varepsilon$. Вот это эконометрическая модель. Population regression curve — это теоретическая регрессия. Что мы вкладываем в эту добавочку ε? Есть разумная интерпретация: переменные, которые не несущественны для Y: расположение, время и неучтённые факторы. Сугубо экономическая вещь — это врождённое непредсказуемое поведение животного. Имманентная неопределённость заставляет агента вести себя по-разному. В матмодели участвуют экономические переменные, но у нас есть только то, что мы смогли померить. Вот есть инфляция, ВВП, а у нас есть только то, что даёт Росстат. Неизвестно, совпадает ли это с теоретическими категориями. Скажем, мы мерим не интеллект, а IQ, балл по диплому. И всё равно у нас измерения с ошибками! Итак, *неучтённые переменные*, *неопределённость поведения*, *расхождения между теорией и данными* и *ошибки* — это 4 компонента ε.

Рассмотрим ситуацию по Москве. Собрать информацию по всем пунктам — дорогая задача. Вместо ГС, мы имеем дело с выборкой. Не population, но sample. Что изменилось? Теперь мы не можем посчитать матожидание так, как тогда. Но то, что нельзя измерить, можно оценить. Мы переходим от точного прогноза к какому-то диапазону. Необходимость оценивания вызвана неполнотой. Это называется выборочной регрессией, по которой мы оцениваем нашу кривую. $\mathbb{E}(Y \mid X) = \alpha + \beta X$. Мы параметризовали модель. $Y = \alpha + \beta X + \varepsilon$. Получили оценку для выборки: $\tilde{Y} = \tilde{\alpha} + \tilde{\beta} X$, $Y_i = \tilde{Y}_i + e_i$. Вот эта e_i — остаток (residual) наблюдения. $Y_i = \tilde{\alpha} + \beta X_i$, и остатки — случайные величины. Остаток — оценка.

Наблюдение можно подогнать под диаграмму разброса, и тогда получится population regression line. Можно оценить Sample regression curve. С точки зрения выборки нам не хватает данных. Поэтому надо суммировать, а можно минимизировать сумму по **методу наименьших квадратов**:

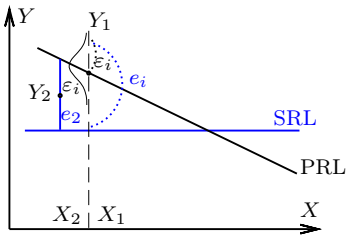

Рис. 1. PRL и SRL

$$\min_{\alpha,\beta} \sum e_i^2 = \min_{\alpha,\beta} \sum (Y_i - \alpha - \beta X_i)^2 \tag{2.1}$$

$$\text{FOC:} \begin{cases} \frac{\partial S}{\partial \alpha} = 0 \\ \frac{\partial S}{\partial \beta} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum (Y_i - \alpha - \beta X_i) = 0 \\ -2\sum (Y_i - \alpha - \beta X_i) X_i = 0 \end{cases} \tag{2.2}$$

$$\sum Y_i - n \cdot \alpha - \beta \sum X_i = 0, \quad \bar{Y} = \frac{1}{n}\sum Y_i; \ \bar{X} = \frac{1}{n}\sum X_i$$

$$\bar{Y} = \alpha + \beta \bar{X}, \quad \alpha = \bar{Y} - \beta \bar{X}, \quad \sum (Y_i - \bar{Y} + \beta \bar{X} - \beta X_i) X_i = 0$$

$y_i = Y_i = \bar{Y}$ зависит от выборки; $\bar{Y}$ — случайная величина. Разница равна $\sum(y_i - \beta x_i)(\bar{X} + x_i) = 0$, т. е. сумма отклонений равна 0. $\sum x_i = \sum(X_i - \bar{X}) = \sum X_i - n\bar{X} = 0$; $\bar{X}\underbrace{\sum Y_i}_{=0} + \sum x_i y_i - \beta\bar{X}\underbrace{\sum x_i}_{=0} - \beta\sum x_i^2 = 0$.

$$\tilde{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2} = \frac{\frac{1}{n-1}\sum x_i y_i}{\frac{1}{n-1}\sum x_i^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \quad \tilde{\alpha} = \bar{Y} - \tilde{\beta}\bar{X} \tag{2.3}$$

$$\sum X_i^2 = \sum x_i^2 + n\bar{X}^2, \quad \sum X_i^2 = \sum(\bar{X} + x_i)^2 = n\bar{X}^2 + 2\bar{X}\sum x_i + \sum x_i^2$$

$$\sum X_i Y_i = \sum x_i y_i + n\bar{X}\bar{Y}$$

Но всё это проходит, если $\sum x_i^2 \neq 0$. Если оно равно 0, то получается несовместная система. Когда все иксы одинаковые, нельзя применить МНК.

Но, как учил Соколов, надо посмотреть условие второго порядка. Матрица Гессе:

$$\boldsymbol{H} = \begin{pmatrix} 2n & \overbrace{2\sum X_i}^{2n\bar{X}} \\ 2\sum X_i & 2\sum X_i^2 \end{pmatrix}$$

Эта функция должна быть хотя бы дважды непрерывной и дифференцируемой. И она не зависит от точки. Функция выпуклая, а минимум глобальный, потому что $\Delta_1 = 2n > 0$, $\Delta_2 = 4n\sum X_i^2 - 4n^2\bar{X}^2 = 4n(\sum X_i^2 - n\bar{X}^2) = 4n\sum x_i^2 > 0$, а $\sum X_i = n\bar{X}$.

Критерий мы написали эвристически, с потолка, просто написали условие максимизации. $e_i = Y_i - \hat{\alpha} - \hat{\beta}X_i$. Иными словами, если $\hat{Y}_i = \hat{\alpha} - \hat{\beta}X_i$, то введём ещё одно обозначение: с волной оценка любая, а с шапочкой (hat) — МНК. Тогда очевидно, что $Y_i = \hat{Y}_i + e_i$. Можно сказать, что $\sum e_i = 0$, что есть первое нормальное уравнение. Это эквивалентно тому, что $\frac{1}{n}\sum e_i = 0$. Это трактуют как отсутствие систематической ошибки. Второе свойство: $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$, SRL $\to Y = \hat{\alpha} + \hat{\beta}X$. Линия выборочной регрессии проходит через точку $(\bar{X}; \bar{Y})$ со средними координатами. Третье свойство: посчитаем $\frac{1}{n}\sum \hat{Y}_i = \bar{\hat{Y}}$, $\frac{1}{n}\sum \hat{Y}_i + \frac{1}{n}\sum e_i$, $\bar{Y} = \bar{\hat{Y}}$. Четвёртое свойство (второе нормальное уравнение): $\sum e_i X_i = 0$. $\sum e_i\hat{Y}_i = \sum e_i(\hat{\alpha} + \hat{\beta}X_i) = \hat{\alpha}\sum e_i + \hat{\beta}\sum e_i X_i = 0 + 0 = 0$. Далее, $\sum Y_i^2 = \sum \hat{Y}_i^2 + 2\sum e_i\hat{Y}_i + \sum e_i^2$, $\sum Y_i^2 = \sum \hat{Y}_i^2 + \sum e_i^2$ — Пифагоровы штаны! $-n\bar{Y}^2 = -n(\bar{\hat{Y}})^2$, $\sum Y_i^2 - n\bar{Y}^2 = \sum Y_i^2$. Важнейшее свойство: $\sum Y_i^2 = \sum \hat{Y}_i^2 + \sum e_i^2 \Rightarrow \mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$ (total sum of squares = residual + explained). У нас в руках есть критерий, насколько хорошо лежат точки вокруг найденной линии. Чем оно меньше, тем оно лучше. Поэтому надо критерий обезразмерить. $R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$ — чем больше это выражение, тем ближе общая сумма квадратов к тому, что мы построили. $1 \geqslant R^2 \geqslant 0$. Это эвристическая мера качества полученной линии.

3 Лекция 3

Мы мерили отклонения от оси y, так как $y = f(x)$, а можно было брать модули, отклонения от икса, отклонения перпендикуляров. Но мы выбрали, потому что нам так удобно.

Мы начали работать в двухмерном пространстве, но в математике можно сменить размерность пространства на количество объектов. И мы рассмотрим эту задачу в другом пространстве — выборочном пространстве (sample space). Вектор $(Y_1, Y_2, \ldots, Y_n) = \boldsymbol{Y}$, $(X_1, \ldots, X_n) = \boldsymbol{X}$. Нарисуем два вектора. Но писать $Y_i = \alpha + \beta X_i$ неудобно, поэтому введём единичный вектор $\vec{\imath} = (1, \ldots, 1)$. Мы хотим сделать, чтобы сумма квадратов отклонения была $e_i = Y_i - \alpha \cdot 1 - \beta X_i$, или $\vec{e} = \vec{Y} - \alpha\vec{\imath} - \beta\vec{X}$. Но сумма альф на единичный вектор плюс бета на вектор $\boldsymbol{X}$ — это линейная комбинация. При разных α, β у нас получается всё пространство. Сумма квадратов отклонений — квадрат длины вектора: $\sum e_i^2 = \|\vec{e}\|^2 = (\vec{e}^T, \vec{e})$. Так наше пространство стало евклидовым. Оси ортогональные. Мы хотим минимизировать длину вектора $\boldsymbol{e}$. По правилу работы с векторами у нас получается рисунок 2.

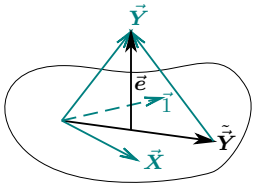

Рис. 2. Векторы

Опустим ортогональную проекцию, и получится минимизирующий вектор. А коэффициенты α и β — разложение вектора $\hat{\vec{Y}}$ по осям. $\sum e_i X_i = 0$, $(\vec{e}^T, \hat{\vec{Y}}) = 0$. Мы выиграли в наглядности, но потеряли смысл. Итак, в случайный член ε мы засунули все отклонения. По этой выборке мы хотим оценить генеральную совокупность: $Y_i = \hat{\alpha} + \hat{\beta}X_i + e_i = \hat{Y}_i + e_i$. Итак,

$$\hat{\beta} = \frac{x_i Y_i}{\sum x_i^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{3.1}$$

X детерминирован, Y случаен. Поэтому все наши оценки — случайные величины, которые нам что-то говорят о теоретических значениях теоретической регрессии. Важны также точечные оценки и их свойства. Какими свойствами обладают оценки как статистические?

$\hat{\beta} = \sum\left(\frac{x_i}{\sum x_i^2}\right) \cdot y_i = \sum k_i y_i$, $\sum k_i = \frac{1}{\sum x_i^2}$, $\sum x_i = 0$. Далее, $\sum k_i y_i = \sum k_i(Y_i - \bar{Y}) = \sum k_i Y_i = \bar{Y}\sum k_i$. Поэтому данная оценка математически называется линейной оценкой. По иксу она явно нелинейная: что-то делить на

что-то. А $\hat\alpha = \frac{1}{n}\sum Y_i - \sum k_i Y_i \bar X = \sum \left(\frac{1}{n} - k_i \bar X\right) Y_i = \sum \theta_i Y_i$. Итак, наши оценки МНК линейны.

Далее, нам нужна несмещённость — unbiasedness. $\mathbb{E}(\tilde\theta) = \theta$. Чтобы установить это для оценок, надо предположить что-то дополнительно: $\mathbb{E}(Y \mid X) = [\alpha + \beta X] + \mathbb{E}(\varepsilon \mid X)$. Нужно: $\mathbb{E}(\varepsilon \mid X) = 0$. Она при каждом X другая; это континуум эпсилонов. Надо, чтобы при любом ε и X была сумма нулевая. Нужна экзогенность X (регрессора) по отношению к ε. $\mathbb{E}(\varepsilon) = 0\ \forall X$ — такое свойство будем считать выполненным. Рано говорить о распределении; скажем же нечто и об ε. В наших выборочных наблюдениях $\mathbb{E}(\varepsilon) = 0\ \forall i = 1; 2; \ldots; n$.

Второе свойство — дисперсия ε_i, она равно σ^2 без индекса. Во всех наших наблюдениях дисперсия одинакова, что обозначается греко-латинским словом *гомоскедастичность*. Иначе называется *гетероскедастичность*. Что бы мы ни взяли, дисперсия этого равна $\mathrm{Var}(\varepsilon) = \sigma^2$.

Кроме того ε_i независимы; выполняется некоррелированность. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \Rightarrow \mathrm{Corr}(\varepsilon_i, \varepsilon_j) = 0$. Если мы изучаем расходы от доходов, то вряд ли наши доходы коррелируют с расходами соседа.

Величина Y_i случайна с матожиданием $\mathbb{E}(Y_i) = \alpha + \beta X_i$ и дисперсией $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$. Наконец, $\mathrm{Cov}(Y_i, Y_j) = \mathbb{E}(\varepsilon_i, \varepsilon_j) = \mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$. Далее, $\hat\beta = \sum k_i Y_i$; $\mathbb{E}(\hat\beta) = \sum k_i \mathbb{E}(Y_i) = \sum k_i(\alpha + \beta X_i) = \alpha \sum k_i + {} + \beta \sum k_i X_i = \beta \sum \frac{x_i}{\sum x_i^2} X_i = \frac{\beta}{\sum x_i^2} \sum x_i(x_i + \bar X) = \beta + \frac{\beta \bar X}{\sum x_i^2} = \beta$, поэтому $\mathbb{E}(\hat\beta) = \beta$. Аналогично, $\mathbb{E}(\hat\alpha) = \alpha$. Мы стали применять МНК не для несмещённости, но получилось всё так хорошо.

Можем ли мы получить ещё важные свойства? Раз $\hat\alpha$ — случайная величина, то есть матожидание, но единственная порождающая случайная величина имеет матожидание и дисперсию.

$$\mathrm{Var}(\hat\beta) = \mathrm{Var} \sum k_i Y_i = \sum_i k_i^2 \mathrm{Var}(Y_i) + 2 \sum_{i<j} k_i k_j \mathrm{Cov}(Y_i, Y_j) = \sigma^2 \sum_i k_i^2 + 0 =$$

$$= \sigma^2 \sum \left(\frac{x_i}{\sum x_i^2}\right)^2 = \frac{\sigma^2}{\left(\sum x_i^2\right)^2} \sum x_i^2 = \frac{\sigma^2}{\sum x_i^2} \quad (3.2)$$

Суммирование было по $i < j$, чтобы дважды не посчитать.

$$\mathrm{Var}(\hat\alpha) = \sigma^2 \sum \theta_i^2 = \sigma^2 \sum \left(\frac{1}{n} - k_i \bar X\right)^2 = \sigma^2 \sum \left[\frac{1}{n^2} + k_i^2 \bar X^2 - 2\frac{k_i}{n} \bar X\right] =$$

$$= \sigma^2 \left[\frac{1}{n} + \bar X^2 \sum k_i^2 - 2\frac{\bar X}{n} \sum k_i\right] = \sigma^2 \left[\frac{1}{n} + \bar X^2 \sum k_i^2 + \frac{\bar X^2}{\sum x_i^2}\right] \quad (3.3)$$

Но так как у нас две случайные величины, то мы хотим посчитать ковариацию двух выражений. Можно найти $\sum \theta_i k_i$, но мы поступим по-другому. $\hat\alpha = \bar Y - \hat\beta \bar X = \frac{1}{n}\sum Y_i - \hat\beta \bar X$. $\bar Y = \hat\alpha + \hat\beta \bar X$. Дисперсии этих двух случайных величин одинаковы: $\mathrm{Var}(\bar Y) = \mathrm{Var}(\hat\alpha) + \bar X^2 \mathrm{Var}(\hat\beta) + 2\bar X \mathrm{Cov}(\hat\alpha, \hat\beta)$. Далее, $\mathrm{Var}(\bar Y) = \mathrm{Var}\left(\sum \frac{1}{n} Y_i\right) = \sigma^2 \sum \frac{1}{n_2} = {} = \frac{\sigma^2}{n} = \sigma^2 \left(\frac{1}{n} + \frac{\bar X^2}{\sum x_i^2}\right) + \bar X^2 \frac{\sigma^2}{\sum x_i^2} + 2\bar X \mathrm{Cov}(\hat\alpha, \hat\beta)$. Ковариация — сумма произведений совместного разброса. Они пропорциональны σ^2, что является разбросом ε. Итак, $\mathrm{Cov}(\hat\alpha, \hat\beta) = -\frac{\sigma^2 \bar X}{\sum x_i^2}$. Чем сильнее разброс объясняющей переменной, тем точнее оценка. И понятно: проводится прямая по точкам. Чем точнее «эпсилен», тем точнее оценки. Сильнее разбрасывайте иксы, а также делайте среднее X равным 0.

Вот сейчас мы подошли к **теореме Гаусса—Маркова**. Пусть что-то выполняется — тогда такие-то свойства. Пусть выполняются следующие условия:

1. $Y = \alpha + \beta X + \varepsilon$;
2. X_i детерминированы и не все равны между собой;
3. $\mathbb{E}(\varepsilon_i) = 0$;
4. $\mathrm{Var}(\varepsilon_i) = \sigma^2$ — гомоскедастичность;
5. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0,\ i \neq j$.

Тогда оценки МНК $\hat\alpha, \hat\beta$ являются линейными, несмещёнными, эффективными (обладают наименьшей дисперсией в классе линейных несмещённых). Оценки OLS суть Best Linear Unbiased Estimators (эффективные линейные несмещённые оценки) — BLUE-оценки.

Начнём доказательство. $\tilde\beta = \sum w_i Y_i$ — класс линейных оценок. Несмещённые: $\mathbb{E}(\tilde\beta) = \beta$. Поэтому у нас появляется ограничение на w_i: $\mathbb{E}(\tilde\beta) = \sum w_i \mathbb{E}(Y_i) = \sum w_i \mathbb{E}(\alpha + \beta X_i + \varepsilon_i) = \sum w_i(\alpha + \beta X_i) = \alpha \sum w_i + \beta \sum w_i X_i$. И это выражение должно быть равно β для любых Y_i. Единственная возможность — это когда $\sum w_i = 0$ и $\sum w_i X_i = 1$. Наша задача — найти минимальную дисперсию этой оценки. $\mathrm{Var}(\tilde\beta) = \sigma^2 \sum w_i^2$, поэтому мы ищем $\min\limits_{w_i} \sigma^2 \sum w_i^2$ subject to $\sum w_i = 0$, $\sum w_i X_i = 1$. Можно решать с лагранжианом. А можно полный квадрат выделить.

4 Лекция 4

Чтобы оценка была несмещённой, мы вывели 2 условия для $\tilde\beta = \sum w_i Y_i$: $\sum w_i = 0$, $\sum w_i X_i = 1$. $\min\limits_{\vec w} \mathrm{Var}(\tilde Y) = {} = \sigma^2 \sum w_i^2$.

$\sigma^2 \sum w_i^2 = \sigma^2 \sum (w_i - k_i + k_i)^2$, где $\hat{Y} = \sum k_i Y_i = \sum \frac{x_i}{\sum x_i^2} Y_i$. Далее, $\sigma^2 \left[\sum (w_i - k_i)^2 + \sum k_i^2 + 2\sum (w_i - k_i)k_i\right]$. Мы хотим увидеть, что последняя сумма равна нулю: $\sum (w_i - k_i)k_i = \sum w_i \frac{x_i}{\sum x_i^2} - \sum k_i^2 = \frac{\sum w_i x_i}{\sum x_i^2} - \frac{\sum x_i x_i}{\left(\sum x_i^2\right)^2} \overset{=}{\bigcirc}$ Поясним: $\sum w_i x_i = \sum w_i (X_i - \bar{X})]1 - \bar{X} \sum w_i = 1$. Продолжим: $\overset{=}{\bigcirc} \frac{1}{\sum x_i^2} - \frac{1}{\sum x_i^2} = 0$. $\sigma^2 \sum w_i^2 = \sigma^2 \left[\sum (w_i - k_i)^2 + \sum k_i^2\right]$. При $w_i = k_i$ верно, что $\sigma^2 \sum k_i^2 = \frac{\sigma^2}{\sum x_i^2}$. Аналогично доказывается для $\hat{\alpha}$.

В итоге мы доказали, что дисперсии $\hat{\beta}$ и $\hat{\alpha}$ суть минимально возможные, если мы рассматриваем линейные несмещённые оценки. Но теорема Гаусса—Маркова говорит только об экстремальных свойствах МНК, но никак не $\mathrm{Cov}(\hat{\alpha}, \hat{\beta})$.

Пусть мы хотим иметь X_0. $Y_0 = \alpha + \beta X_0 + \varepsilon_0$. Это ненаблюдаемая величина, поэтому будем использовать модель $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} X_0$; наш прогноз — случайная величина. Это линейная оценка. Сохраняется ли несмещённость? $\mathbb{E}(Y_0) = \mathbb{E}(\hat{\alpha} + \hat{\beta} X_0) = \alpha + \beta X_0 = \mathbb{E}(Y_0)$. Получается, что матожидание $\hat{Y}_0$ есть матожидание другой случайной величины, а должно быть что-то детерминированное. Что скажем о дисперсии?

$$\mathrm{Var}(\hat{Y}_0) = \mathrm{Var}(\hat{\alpha} + \hat{\beta} X_0) = \mathrm{Var}(\hat{\alpha}) + X_0^2 \mathrm{Var}(\hat{\beta}) + 2X_0 \mathrm{Cov}(\hat{\alpha}, \hat{\beta}) =$$
$$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right) + \sigma^2 \frac{X_0^2}{\sum x_i^2} - 2\sigma^2 X_0 \frac{\bar{X}}{\sum x_i^2} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2 + X_0^2 - 2\bar{X} X_0}{\sum x_i^2}\right) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right) \quad (4.1)$$

Поэтому величина дисперсии минимальна вблизи среднего и быстро нарастает при удалении от неё.

Полученное выражение даёт дисперсию $\hat{Y}_0$. Но иногда говорят: а я хочу найти дисперсию Y_0, то есть не среднего, а индивидуального. Это будет дисперсия такая: $\mathrm{Var}(Y_0) = \mathrm{Var}(\hat{\alpha} + \hat{\beta} X_0 + \varepsilon_0)$. Но альфы и беты линейно зависят от эпсилонов, а ε_0 не из их числа. Свойства выражены только в выборке: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)$, а для теоретической регрессии оно справедливо для всех разных ε. Поэтому можно записать эту дисперсию как сумму двух дисперсий. Она равна $\mathrm{Var}(\hat{\alpha} + \hat{\beta} X_0 + \varepsilon_0) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2} + 1\right)$. Доказывая теорему Гаусса—Маркова, мы нигде не располагали типом распределения. Лишь бы матожидание с дисперсией было. Так мы получили статистические свойства наших оценок. Вспомним второй курс и статистику. То, что мы получили, называют point estimate — точечная оценка.

Можем ли мы построить интервальную оценку для наших случайных величин? Даже если неизвестно распределение, у нас есть **неравенство Чебышёва**:

$$\mathbb{P}\{|\xi - \mathbb{E}(\xi)| > \delta\} \leqslant \frac{\mathrm{Var}(\xi)}{\delta^2} \quad (4.2)$$

А может получиться, что дисперсия, делённая на дельту в квадрате, — двойка. Поэтому нам нужно хоть что-то о распределении. Но у нас ещё один пробел. Мы ввели σ^2 — одинаковую дисперсию всех ε_i. Но без альфы и беты знать сигму довольно странно. Помимо регрессии, нам удобнее использовать ещё один σ^2. Нам надо научиться его оценивать.

$\sigma^2 = \mathrm{Var}(\varepsilon_i)$, $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$, $\sum e_i^2 = \mathrm{RSS}$, $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. $e_i = Y_i - \hat{\alpha} - \hat{\beta} X_i = \alpha + \beta X_i + {} + \varepsilon_i - \hat{\alpha} - \hat{\beta} X_i = \varepsilon_i - (\hat{\alpha} - \alpha) - (\hat{\beta} - \beta) X_i$. Вспомним, чему равны крышки. $\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} = \frac{1}{n} \sum Y_i - \hat{\beta} \bar{X} = {} = \frac{1}{n} \sum (\alpha + \beta X_i + \varepsilon_i) - \hat{\beta} \bar{X} = \alpha + \beta \bar{X} - \hat{\beta} \bar{X} + \bar{\varepsilon}$. Внимание: $\bar{e} = 0$, но $\bar{\varepsilon}$ — это какая-то величина. Имеем: $\bar{\varepsilon} + \alpha - (\hat{\beta} - \beta) \bar{X}$. $e_i = \varepsilon_i - [\bar{\varepsilon} - (\hat{\beta} - \beta) \bar{X}] - (\hat{\beta} - \beta) X_i = (\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta} - \beta) x_i$.

$$\sum e_i^2 = \sum \left[(\varepsilon_i - \bar{\varepsilon}) - (\hat{\beta} - \beta) x_i\right]^2 = \underbrace{\sum (\varepsilon_i - \bar{\varepsilon})^2}_{\mathrm{I}} + \underbrace{(\hat{\beta} - \beta)^2 \sum x_i^2}_{\mathrm{II}} - 2 \underbrace{\sum (\varepsilon_i - \bar{\varepsilon})(\hat{\beta} - \beta) x_i}_{\mathrm{III}} \quad (4.3)$$

Возьмём матожидание от трёх слагаемых. Так как $\sum (\varepsilon_i - \bar{\varepsilon})^2 = \sum \varepsilon_i^2 - n(\bar{\varepsilon})^2$, то верно вот что.

$\mathbb{E}(\mathrm{I}) = n\sigma^2 - n\mathbb{E}\left(\left(\frac{1}{n} \sum \varepsilon_i\right)^2\right) = n\sigma^2 - n \frac{1}{n^2} \cdot n\sigma^2 = (n-1)\sigma^2$.

$\mathbb{E}(\mathrm{II}) = \sum x_i^2 \mathbb{E}(\hat{\beta} - \beta)^2 = \sum x_i^2 \cdot \mathrm{Var}(\hat{\beta}) = \sum x_i^2 \cdot \frac{\sigma^2}{\sum x_i^2} = \sigma^2$.

$\mathbb{E}(\mathrm{III}) = \mathbb{E}\left(\sum (\varepsilon_i - \bar{\varepsilon})(\hat{\beta} - \beta) x_i\right)$ — это что? $\hat{\beta} = \sum k_i Y_i = \sum k_i (\alpha + \beta X_i + \varepsilon_i) = \underbrace{\alpha \sum k_i}_{=0} + \beta \underbrace{\sum k_i X_i}_{=1} + \sum k_i \varepsilon_i$,

поэтому $\mathbb{E}\left(\sum (\varepsilon_i - \bar{\varepsilon})(\hat{\beta} - \beta) x_i\right) = \mathbb{E}\left(\sum (\varepsilon_i - \bar{\varepsilon}) \left(\sum k_j \varepsilon_j\right) x_j\right) \overset{=}{\bigcirc} \sum x_i \mathbb{E}\left[(\varepsilon_i - \bar{\varepsilon}) \left(\sum k_j \varepsilon_j\right)\right] = \sum x_i \mathbb{E}\left(\varepsilon_i \left(\sum k_j \varepsilon_j\right)\right) - {} - \sum x_i \mathbb{E}\left(\bar{\varepsilon} \left(\sum k_j \varepsilon_j\right)\right)$. Но мы слишком рано поменяли матожидание и суммы местами, поэтому продолжим формулу так: $\overset{=}{\bigcirc} \mathbb{E}\left(\sum (\varepsilon_i x_i) \cdot \left(\sum k_j \varepsilon_j\right) - \bar{\varepsilon} \left(\sum x_i\right) \left(x_j \varepsilon_j\right)\right) = \mathbb{E}\left(\sum \varepsilon_i x_i\right) \cdot \left(\sum k_j \varepsilon_j\right) \overset{=}{\bigcirc} \ldots$ У нас стоит сумма произведений одной линейной комбинации на другую. Это сумма всевозможных произведений слагаемого из первой на слагаемое из второй. Это то же, что и $\overset{=}{\bigcirc} \mathbb{E}\left(\sum_{i,j=1}^{n} (\varepsilon_i x_i k_j \varepsilon_j)\right) = \mathbb{E}\left(\sum_{i=1}^{n} x_i k_i \varepsilon_i^2\right) = \sum_{i=1}^{n} x_i k_i \sigma^2 = \sigma^2 \sum_{i=1}^{n} x_i \frac{x_i}{\sum x_j^2} = \sigma^2$.

И тогда у нас окончательно получается: $\mathbb{E}\left(\sum e_i^2\right) = \mathbb{E}(\mathrm{RSS}) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$. Вводим $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\mathrm{RSS}}{n-2}$, и $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$. И она будет несмещённой! Если рассмотреть standard error $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$, то она

смещённая в силу неравенства Йенсена и вогнутости функции: $\mathbb{E}(\hat{\sigma}) \neq \sigma$. Вспомним **неравенство Йенсена**: пусть $(\Omega, \mathcal{F}, \mathbb{P})$ — вероятностное пространство, а $\xi = \xi(\omega)$ — случайная величина с конечным матожиданием. Пусть функция $g = g(x)$ определена на всей вещественной прямой, дифференцируема (для простоты) и выпукла вниз (просто выпуклая). Тогда $g\big(\mathbb{E}(\xi)\big) \leqslant \mathbb{E}\big(g(\xi)\big)$.

Чтобы получить оценку дисперсии, надо написать, что $\widehat{\mathrm{Var}(\hat{\beta})} = \frac{\hat{\sigma}^2}{\sum x_i^2}$. И благодаря несмещённости и детерминированности мы получаем несмещённые оценки характеристик наших коэффициентов (альфа, бета) и их ковариации. Вспомним ЦПТ, которая говорит, что среднее суммы одинаковых случайных величин нормально. Мы сделаем дополнительное предположение, что $\varepsilon_i \sim \mathcal{N}(0; \sigma^2)$. Простая регрессия в условиях нормальности. Это предположение влечёт за собой много приятных для нас последствий. Во-первых, величины ε_i между собой не коррелируют: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ для $i \neq j$. Для нормальных величин они независимы. Во-вторых, $\hat{\alpha}, \hat{\beta}, \hat{Y}_0$ суть линейная комбинация, поэтому они суть BLUE-оценки.

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha; \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)\right), \quad \hat{\beta} \sim \mathcal{N}\left(\beta; \frac{\sigma^2}{\sum x_i^2}\right) \tag{4.4}$$

$$Y_i = \alpha + \beta X_i + \varepsilon_i \sim \mathcal{N}(\alpha + \beta X_i; \sigma^2), \quad \hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \sim \mathcal{N}\left(\alpha + \beta X_i; \sigma^2\left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum x_i^2}\right)\right) \tag{4.5}$$

Оценки $\hat{Y}_i$ выходят гетероскедастичными, поэтому величина остатка e_i, равная $Y_i - \hat{Y}_i$, есть разность двух нормально распределённых случайных величин с матожиданием 0 и выводимой дисперсией, так как тут есть ещё ковариация. Готовый ответ: $e_i \sim \mathcal{N}\left(0; \sigma^2\left(1 - \frac{1}{n} + \frac{x_i^2}{\sum x_i^2}\right)\right)$. Остатки с нулевым матожиданием, но дисперсия разная от точки к точке.

5 Лекция 5

Продолжаем мучить нашу модель $Y = \alpha + \beta X + \varepsilon$. Выполнены условия Гаусса—Маркова, X детерминирован, $\sum x_i^2 > 0$, $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}\,\varepsilon_i = \sigma^2$. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$. А теперь предположение: $\varepsilon \sim \mathcal{N}(0; \sigma^2)$. А если две нормальные некоррелированные, то они независимые. Мы выяснили, что

$$\hat{\alpha} \sim \mathcal{N}\left(\alpha; \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_i^2}\right)\right), \quad \hat{\beta} \sim \mathcal{N}\left(\beta; \sigma^2\left(\frac{\bar{\sigma}^2}{\sum x_i^2}\right)\right) \tag{5.1}$$

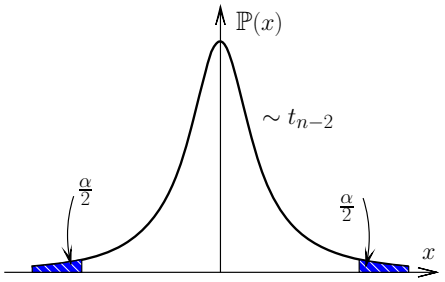

Рис. 3. t-распределение

$$\hat{Y} \sim \mathcal{N}\left(\alpha + \beta X_i; \sigma^2\left(\frac{1}{n} + \frac{\sum (X_i - \bar{X})^2}{\sum x_i^2}\right)\right) \tag{5.2}$$

Поэтому трудно спрогнозировать величины, далёкие от среднего. И ошибки тоже в своей оценке гетероскедастичны:

$$e_i \sim \mathcal{N}\left(0; \sigma^2\left(1 + \frac{1}{n} + \frac{\bar{x}_1^2}{\sum x_i^2}\right)\right) \tag{5.3}$$

То, что наша RSS несмещённая, не так просто: величина, равная $(n-2)\frac{\hat{\sigma}^2}{\sigma^2}$, при нормальности распределения имеет распределение χ^2_{n-2}, где $\sigma^2 = \frac{\mathrm{RSS}}{n-2} = \frac{\sum e_i^2}{n-2}$. Кроме того, статистика χ^2_{n-2} распределена независимо от $\hat{\alpha}$ и $\hat{\beta}$, которые между собой ещё как зависимы. Будет $\widehat{\mathrm{Var}(\hat{\beta})} = \frac{\hat{\sigma}^2}{\sum x_i^2} = \hat{\sigma}_\beta^2$. И она несмещённая: $\mathbb{E}(\hat{\sigma}_\beta^2) = \sigma_\beta^2$. Тогда естественно, что $(n-2)\frac{\hat{\sigma}_\beta^2}{\sigma_\beta^2} \sim \chi^2_{n-2}$. Аналогично, $(n-2)\frac{\hat{\sigma}_\alpha^2}{\sigma_\beta^2} \sim \chi^2_{n-2}$ и тоже не зависит от альфы и беты. Стандартизуем величину: $Z = \frac{\hat{\beta} - \beta}{\sigma_\beta} \sim \mathcal{N}(0; 1)$. Мы хотим оценить β, поэтому нам нужно построить такую статистику, которая не содержит этого параметра и имеет известное распределение:

$$\frac{\frac{\hat{\beta} - \beta}{\sigma_\beta}}{\sqrt{\frac{\hat{\sigma}_\beta^2 / \sigma_\beta^2 (n-2)}{n-2}}} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_\beta} = \frac{\mathcal{N}(0; 1)}{\sqrt{\frac{1}{n-2}\chi^2_{n-2}}} \sim t_{n-2} \tag{5.4}$$

Доверительный интервал — это такое множество, где с доверительностью вероятностью находится наша величина. $\mathbb{P}\{w_l \leqslant w\theta \leqslant w_n\} = 1 - \alpha$. θ — параметр, для которого мы строим интервал. Это интервал со случайными концами, и концы зависят от выборки, а параметр θ у нас известный. Пусть у нас генерируется θ, и границы (случайные концы) зависят от выборки; они с заданной вероятностью содержат внутри себя значение параметра θ. Так как наша статистика имеет t-распределение, то $\mathbb{P}\{\hat{\beta} - t_{n-2;\frac{\alpha}{2}}\hat{\sigma}_\beta < \beta < \hat{\beta} + t_{n-2;\frac{\alpha}{2}}\hat{\sigma}_\beta\} = 1 - \alpha$. Это то же, что и $\mathbb{P}\left\{-t_{n-2;\frac{\alpha}{2}} \leqslant \frac{\hat{\beta} - \beta}{\hat{\sigma}_\beta} \leqslant t_{n-2;\frac{\alpha}{2}}\right\}$.

Получим:

$$\mathbb{P}\left\{\chi^2_{n-2;1-\frac{\alpha}{2}} \leqslant (n-2)\frac{\hat\sigma^2}{\sigma^2} \leqslant \chi^2_{n-2;\frac{\alpha}{2}}\right\} = 1-\alpha \Leftrightarrow \mathbb{P}\left\{\frac{\hat\sigma^2}{\chi^2_{n-1;\frac{\alpha}{2}}} \leqslant \sigma^2 \leqslant \frac{\hat\sigma^2}{\chi^2_{n-1;1-\frac{\alpha}{2}}}\right\} = 1-\alpha \tag{5.5}$$

Но на практике такое даётся строить редко. Куда чаще проходит…

Проверка гипотез. Она связана с проверкой имени Пирсона. Гипотеза касается одного или нескольких параметров. Один параметр — одна гипотеза. $\mathcal{H}_0\colon \beta = \beta_0$, $\mathcal{H}_1\colon \beta \neq \beta_0$. Проверим самую первую: $\beta = 0$. Это значит, что фактор X незначим и вообще не влияет на Y. Противоположная гипотеза означает, что Y хоть как-то зависит от X. $t_\beta = \frac{\hat\beta}{\hat\sigma_\beta}$. Это называется t-ratio. $|t_\beta| < t_{n-2;\frac{\alpha}{2}}$ — это значит, что зависимости нет. Ошибка первого рода — это отвергнуть верную $\mathcal{H}_0$. Обычно проверяют для $\alpha = 10\,\%; 5\,\%; 1\,\%$. Если мы хотим отвергнуть связь, то приходится брать небольшие α. Thumb rule: если $5\,\%$ и 30 наблюдений, то t-распределение уже стандартное нормальное. На пяти процентах критическое значение примерно равно 2. Тогда не нужны таблицы. Если мало наблюдений, то надо помнить, что у t более толстые хвосты, поэтому надо увеличивать значение до 2,3. Надо не отвергать гипотезы в пользу альтернативной. Если взять более крупное α, то повышается риск ошибиться по-второму. И есть вероятность того, что α_0 совпадёт с границей доверительного интервала. Такое можно посчитать только при помощи компьютера. α^* — это p-value. Это критическое значение вероятности. Это значение позволяет проверять гипотезу на любом возможном уровне значимости. Чем больше α, тем проще отвергнем.

В программе EViews эта штука называется `Prob`. Проверка односторонней гипотезы — вещь хорошая, например, когда речь идёт о предельной склонности к потреблению, потому что она не может быть меньше 0. Если $\mathcal{H}_0\colon \sigma^2 = \sigma_0^2$, то $\mathcal{H}_1\colon \sigma^2 \neq \sigma_0^2$. Берётся Стьюдент.

Вернёмся к проверке значимости коэффициентов. Почти все программы выдают следующую табличку:

Параметр	Расчётная оценка	Стд. ошибка оценки	t	p-value	Дов. инт.
α	$\hat\alpha$	$\hat\sigma_\alpha$	$\hat\alpha/\hat\sigma_\alpha$		
β	$\hat\beta$	$\hat\sigma_\beta$	$\hat\beta/\hat\sigma_\beta$		

Свободные члены он просто называет intercept, так как при проведении прямой линии свободный член отсекается на прямой y. Гипотеза о значимости коэффициентов — это наличие линейной связи.

Следующая гипотеза — это понять, разумная гипотеза или нет. У нас есть разброс относительно регрессии, и если разброс большой, то модель неадекватна данным (наоборот говорить хуже). Проверяется адекватность модели. Индикатор качества модели — это fitting. По-русски — критерий подгонки модели. $R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$.

Мы любим простую гипотезу. Поэтому $\mathcal{H}_0\colon R^2 = 0$ (регрессия неадекватна), $\mathcal{H}_1\colon R^2 > 0$ (регрессия адекватна). Модель хороша, если она что-то объясняет. Процедура проста: $\frac{\hat\sigma^2}{\sigma^2}(n-2) \sim \chi^2_{n-2}$, $\frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-2}$; TSS =? $\mathrm{TSS} = \sum(Y_i - \bar Y)^2 = \sum y_i^2$. $\hat Y = \hat\alpha + \hat\beta X$. $\bar Y = \hat\alpha + \hat\beta \bar X$, $\hat y_i = \hat\beta x_i$ (вычли). $\sum \hat y_i^2 = \hat\beta^2 \sum x_i^2$, $R^2 = 0 \Leftrightarrow \beta = 0$ (модель ничего не объясняет). $\frac{\hat\beta\sqrt{\sum x_i^2}}{\sigma} = \frac{\hat\beta}{\hat\sigma_\beta} \sim \mathcal{N}(0;1)$, $\frac{\mathrm{ESS}}{\sigma^2} = \frac{\hat\beta^2 \sum x_i^2}{\sigma^2} \sim \chi^2_1$. Если рассчитать $\frac{\mathrm{ESS}/1}{\mathrm{RSS}/(n-2)}$ при $R^2 = 0$, то это то же, что и $\frac{\frac{1}{1}\chi^2_1}{\frac{1}{n-2}\chi^2_{n-2}} \sim \mathrm{F}_{1;n-2}$. Это известное распределение Фишера—Снедекора. И квантиль распределения $\mathrm{F}_{1;(n-2);\alpha}$ таков по смыслу: если мы попали в критическую область, что ESS довольно велико, мы отвергаем нулевую гипотезу, наши данные противоречат гипотезе, что нет связи (мы этого хотим), и модель адекватна. Поэтому F-статистика велика для правдивой гипотезы.

Аббревиатура ANOVA — дисперсионный анализ. Источник — SS (1), регрессия — ESS $(n-2)$, остатки — RSS $(n-1)$, общая — TSS. В следующем столбце показывают степени свободы. Потом идёт MSS (делённое на число свободы).

Источник	`SS`	`df`	`MSS`	`F`
Регрессия	ESS	1	ESS/1	$\mathrm{F} = \frac{\mathrm{ESS}/1}{\mathrm{RSS}/(n-2)}$
Остатки	RSS	$n-2$	$\mathrm{RSS}/(n-2)$	
Общие	TSS	$n-1$	$\mathrm{TSS}/(n-1)$	

На всё больших уровнях значимости наша модель адекватнее.

6 Лекция 6

В прошлый раз мы проговорили статистический анализ — диагностику.

Как правильно представлять результаты работ? Пусть мы строили линейную регрессию, получили $\hat Y = 1{,}62 + 0{,}58X$. Принято сохранять одинаковое количество значащих цифр или знаков после запятой. Если не хотим словесно описывать, то надо писать: t-отн (0,85), (2,13), s.e.(…), $t = \frac{\hat\beta}{\mathrm{s.e.}(\hat\beta)}$, приводится R^2, а F-тест как-то сегодня не пишут. Сегодня просто говорим p-value (α_{crit}). Можно это запихнуть в сноску. Пусть кто-то пишет, что $R^2 = 0{,}03$, но это не очень хорошо.

$\mathrm{F} = \frac{\mathrm{ESS}/1}{\mathrm{RSS}/(n-2)} \sim \mathrm{F}_{1;n-2}$. $\mathcal{H}_0\colon R^2 = 0$, $\mathcal{H}_1\colon R^2 > 0$. Выведем простое соотношение между R^2 и F. $R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}}$; $\mathrm{F} = \frac{\mathrm{ESS}}{\mathrm{RSS}}(n-2) = \frac{R^2(n-2)}{1-R^2}$, так как $\mathrm{TSS} = \mathrm{ESS} + \mathrm{RSS}$, $1 = R^2 + \frac{\mathrm{RSS}}{\mathrm{TSS}}$, $\mathrm{RSS} = (1-R^2)\mathrm{TSS}$. Поэтому $\mathcal{H}_0\colon \mathrm{F} = 0$, $\mathcal{H}_1\colon \mathrm{F} > 0$. Выразим по-другому: $R^2(n-2+\mathrm{F}) = \mathrm{F}$, $R^2 = \frac{\mathrm{F}}{\mathrm{F}+n-2} = 1 - \frac{n-2}{\mathrm{F}+n-2}$. И $\mathrm{ESS} = \hat\beta^2 \sum x_i^2$. Тогда мы

помним, что t-статистика $\hat\beta$ распределена, как t_{n-1}, и $\frac{\mathcal{N}(0;1)}{\sqrt{\frac{1}{k}\chi^2_k}} = t_k$, поэтому $\mathrm{F}_{1;n-2;\alpha} = t^2_{n-1;\frac{\alpha}{2}}$. Поэтому оба метода проверки в этом простом случае с одной регрессией можно применять любой из этих тестов, результат-то одинаков.

Рассмотрим ещё одно следствие. $R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\hat\beta^2 \sum x_i^2}{\sum y_i^2} = \frac{\hat\beta^2 \frac{1}{n}\sum x_i^2}{\frac{1}{n}\sum y_i^2} = \beta \frac{S_x^2}{S_y^2} = \hat\beta^2 \frac{s_x^2}{s_y^2}$. Однако мы можем обратить внимание на следующее:

$$R^2 = \frac{\left(\sum x_i y_i\right)^2 \sum x_i^2}{\left(\sum x_i^2\right)^2 \sum y_i^2} = \left(\frac{x_i y_i}{\sqrt{\sum x_i^2 \cdot \sum y_i^2}}\right)^2 = r^2 \tag{6.1}$$

— выборочный коэффициент корреляции. Легко проверить, что такой же коэффициент корреляции будет здесь: $r^2 = \big(\mathrm{Corr}(Y,X)\big)^2 = \big(\mathrm{Corr}(\hat Y, X)\big)^2$. Сейчас кажется, что это одинаково — ан нет.

Рассмотрим регрессию без свободного члена — regression through the origin. В ряде содержательных задач наличие свободного члена не очень объяснимо и не очень желательно. $Y = \alpha + \beta X + \varepsilon$. Мы моделируем инфляцию, и она зависит от экспансии денежной базы. Но нет общепринятой точки зрения на инфляцию. В общем, $\alpha = 0$, так как без расширения монетарной политики источника инфляции нет, как сказали бы монетаристишки. А вот так: $\hat Y = \hat\beta X$. В этом случае мы легко повторим всё, что было ранее. $\min\limits_\beta \sum e_i = \sum (Y_i - \beta X_i)^2 = S$, $\frac{dS}{d\beta} = -2\sum (Y_i - \beta X_i) X_i = 0$. Отсюда $\hat\beta = \frac{\sum X_i Y_i}{\sum X_i^2}$, $\frac{d^2 S}{d\beta^2} = 2\sum X_i^2$, вторая положительная производная — минимум!

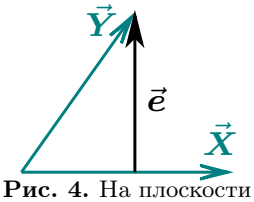

Рис. 4. На плоскости

$\mathrm{Var}(\hat\beta)\frac{\sigma^2}{\sum X_i^2}$. В EViews надо перечислить все регрессоры. При нулевом свободном члене исчез вектор $\boldsymbol{Y}$. Значит, $\sum e_i = \sum 1 \cdot e_1 \neq 0$. И это свойство не выполняется. Это плохо: когда мы писали, что $\hat Y_i = Y_i + e_i$, то это ничего, но $\hat{\bar Y} = \bar Y + \bar e$, а последнее не равно нулю. Мы не сможем вычесть кое-что, что позволяло совершить переход к TSS и иже с ними. Имеем: $\mathrm{TSS} \neq \mathrm{ESS} + \mathrm{RSS}$. $R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} \neq R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$. Люди пытались придумать замену этому коэффициенту. В старых книжках писалось, что надо заменить X на x, а пакеты не говорят ничего о R^2 (хотя и считают его как единица минус …).

После Нового года будем изучать статистику Дарбина—Уотсона. И она тоже не работает в регрессиях без свободного члена.

Раньше не было компов, поэтому нужны были простые вычисления. Пусть у нас есть свободный член. Наш $\boldsymbol{e}$ ортогонален плоскости, порождённой векторами $\boldsymbol{X}$ и $\boldsymbol{1}$. А $\hat{\boldsymbol{Y}}$ и $\boldsymbol{e}$ определяются подпространством любых других векторов; коэффициенты будут в разложении другими, поэтому $\hat{\boldsymbol{Y}}$ и $\boldsymbol{e}$ определяются линейной оболочкой независимо от тех векторов. Поэтому если мы сделаем простое преобразование с точностью до константы, то… Пусть энергозатраты зависят от температуры (Цельсий или Фаренгейт). $Y = \alpha + \beta X$ линейно. Если построим такую регрессию, то оценки $\hat\alpha$ и $\hat\beta$ будут разными, но остатки и качество измерения будут одинаковыми. Когда переменная одна, а вторая из единичек, то линейные преобразования другие.

Изменение масштаба. $X \to \gamma X$ (строили в миллионах — перешли к миллиардам). Тогда $\hat\beta = \frac{\sum x_i y_i}{\sum x_i^2}$, и масштаб не меняется, а $\hat\alpha$ возрастает во столько же раз.

Нас может волновать центрируемость или нецентрируемость переменных. Центрирование: $X \to x = X - \bar X$. Нормирование: $\to \frac{X - \bar X}{\mathrm{s.e.}(x)}$. Матожидание 0, ошибка стандартная равна единице. Тогда $\hat\beta = \frac{\sum x_i y_i}{\sum x_i^2} \Rightarrow \tilde x_i = \frac{x_i}{\sqrt{\frac{1}{n}\sum x_i^2}}$, $\tilde y_i = \frac{y_i}{\sqrt{\frac{1}{n} y_i^2}}$. Тогда

$$\hat\gamma = \frac{\sum \tilde x_i \tilde y_i}{\sum \tilde x_i^2} = \frac{x_i y_i \frac{1}{\mathrm{s.e.}(X)\cdot \mathrm{s.e.}(Y)}}{\sum x_i^2 \frac{1}{\left(\mathrm{s.e.}(X)\right)^2}} = \frac{\sum x_i y_i \cdot \mathrm{s.e.}(X)}{\sum x_i^2 \cdot \mathrm{s.e.}(Y)} = \frac{\sum x_i y_i \sqrt{\sum x_i^2}}{\sum x_i^2 \sqrt{\sum y_i^2}} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}} = r_{xy} \tag{6.2}$$

— выборочный коэффициент корреляции. К тому же Y и X могут быть измерены в разных единицах. А теперь мы их привели к одному масштабу.

Множественная линейная регрессия.

Нам стало неудобно таскать α, поэтому $Y = \beta_0 + \beta_1 + \ldots + \beta_k X_k + \varepsilon$, тут $k+1$ факторов. Многофакторная модель! Пишется так: $Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$. Тут k коэффициентов, и принимается, что $X_1 \equiv 1$. Запишем условное матожидание: $\mathbb{E}(Y \mid X_1, \ldots, X_k) = f(X_1, \ldots, X_k)$. Рассмотрим естественное обобщение: $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ \ldots \\ Y_n \end{pmatrix}$. Составим матрицу: $\boldsymbol{X} = \begin{pmatrix} X_1^1 & X_2^1 & \ldots & X_k^1 \\ \vdots & \vdots & \ddots & \vdots \\ X_1^n & X_2^n & \ldots & X_k^n \end{pmatrix}$. Во второй модели первый столбец меняется на единицы. $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.

Да и естественно, что $k < n$, чтобы не приходилось через точку две чёрт-те что. $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \Rightarrow \boldsymbol{Y} = \hat{\boldsymbol{Y}} + \boldsymbol{e}$.

Оценим эту красоту по МНК:

$$\min_{\vec{\beta}} \sum e_i^2 = \min_{\beta_1;\ldots;\beta_k} \sum (Y_i - \beta_1 X_1^i - \ldots - \beta_k X_k^i)^2 = \min_{\beta_1;\ldots;\beta_k} \sum (Y_i - \beta_1 - \beta_2 X_2^i - \ldots - \beta_k X_k^i)^2 = \min_{\vec{\beta}} S \tag{6.3}$$

$\frac{\partial S}{\partial \beta_i} = -2\sum_j (Y_i - \beta_1 - \ldots - \beta_k X_k^j) X_i^j = 0 = -2\sum_j e_j X_i^j$, $i = 1; 2; \ldots; k$. Имеем СЛУ, k штук. Это система нормальных уравнений.

$\sum_j Y_j X_i^j = \left(\sum X_i^j\right)\beta_1 + \beta_2 \sum X_2^j X_i^j + \ldots + \beta_k \sum X_k^j X_i^j$. Чтобы записать это в матричном виде, нам нужно записать столбец на столбец, но так нельзя. Поэтому транспонируем:

$$\underbrace{(\boldsymbol{X}^T \boldsymbol{X})}_{k\times k} \underbrace{\boldsymbol{\beta}}_{k\times 1} = \underbrace{\boldsymbol{X}^T}_{k\times n} \underbrace{\boldsymbol{Y}}_{n\times 1} \tag{6.4}$$

Имеем: $\det(\boldsymbol{X}^T\boldsymbol{X}) \neq 0$, $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Эту формулу невозможно записать не в матричной форме.

7 Лекция 7

Андрей Александрович не понижает планки от пересдачи к пересдаче. Сдаёте домашки — экзамен напишете.

Была получена формула для оценки, получаемой по МНК. $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. А чем хорош именно этот метод? Оценки обладают рядом замечательных свойств, которые можно только пожелать, — несмещённость и эффективность. Эффективность подразумевается минимумом дисперсии. Если мы выбор ограничим линейными несмещёнными, то оценки МНК самые хорошие в своём классе. Нелинейная оценка — это $\hat{\sigma}^2 = \frac{\text{RSS}}{n-k}$, но об этом позже.

Рассмотрим матрицу $\boldsymbol{X}$: $\begin{pmatrix} 1 & X_{21} & \ldots & X_{k1} \\ 1 & \vdots & \vdots & \vdots \\ 1 & X_{2n} & \ldots & X_{kn} \end{pmatrix}$. Первый столбец единичный ($\boldsymbol{S} = (1; \ldots; 1)$), так как $Y_i = \beta_1 \cdot 1 + \beta_2 \cdot X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$, $X_1 \equiv 1$, $i = \overline{1;n}$. На векторы $\boldsymbol{X}_i$ натянута гиперплоскость, но есть ещё и случайная ошибка, поэтому вектор $\boldsymbol{Y}$ торчит в пространство. МНК ищет такой вектор $\hat{\boldsymbol{Y}}$, лежащий в этой плоскости (должен быть линейной комбинацией этих переменных), который будет ближе всего к вектору $\boldsymbol{Y}$. А мы минимизируем $\sum (Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})^T(\boldsymbol{Y} - \hat{\boldsymbol{Y}})$.

Свойства МНК:

1. $\sum e_i = 0$, $(e, S) = 0$;
2. Вектор остатков ортогонален любой объясняющей переменной;
3. $\sum e_i \hat{Y}_i = 0$.
4. Вектор проходит через среднюю точку: $\bar{\boldsymbol{Y}} = \hat{\beta}_1 + \ldots + \hat{\beta}_k \bar{X}_k$, откуда следует, что $\bar{\hat{\boldsymbol{Y}}} = \bar{\boldsymbol{Y}}$.

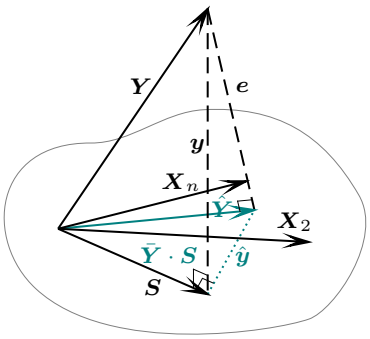

Рис. 5. В пространстве

Имеет место теорема Пифагора. Все эти замечательные свойства имеют место только тогда, когда в регрессии присутствует константа, иначе теорема Пифагора неверна. Даже при прохождении через 0 используется наша константа. Если это 0, то это ухудшает точность прогноза. Хорошо. При экстраполяции к нулю даже не всегда корректно поведение функции, поэтому говорят, что глубина прогноза небольшая.

$y_i = Y_i - \bar{Y}$, $\hat{y}_i = \hat{Y}_i - \bar{Y}$. $\bar{\boldsymbol{Y}} \cdot \boldsymbol{S}$ — то проекция $\hat{\boldsymbol{Y}}$ на $\boldsymbol{S}$. Но и проекция $\boldsymbol{Y}$ точно такая же! $(\hat{\boldsymbol{Y}} - \bar{\boldsymbol{Y}} \cdot \boldsymbol{S}, \boldsymbol{S}) = 0$, $(\boldsymbol{Y} - \bar{\boldsymbol{Y}} \cdot \boldsymbol{S}, \boldsymbol{S}) = 0$. Рассмотрим треугольник, который образован $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, $\boldsymbol{e}$: $(\boldsymbol{e}, \hat{\boldsymbol{y}}) = 0$.

> «На сайте компании, которая производит громоотводы, написано: „равносторонний прямоугольный треугольник“. Почему так? Риманова метрика? Отнюдь, просто простой народ не знает слова „равнобедренный“, а так всем вроде понято».

$\boldsymbol{y}^t\boldsymbol{y} = \hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}} + \boldsymbol{e}^T\boldsymbol{e}$. Отсюда следует, что TSS = ESS + RSS. Внимание читающим Магнуса, Катышева, Пересецкого: у них ESS — это error, $R^2 = \frac{\text{RSS}}{\text{TSS}}$, RSS — regression.

$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = \text{Corr}(\boldsymbol{Y}, \hat{\boldsymbol{Y}})$. Свойства такие же, как у коэффициента корреляции: если индексы трансформировать линейным преобразованием, то тогда на R^2 это не скажется. То есть эта оценка инвариантна относительно линейного преобразования. Ещё чудесное свойство: есть такая ситуация, когда он всегда растёт. Сколько надо включить переменных, чтобы вырос R^2? Все. Свойство не очень хорошее: если мы напихаем в регрессию фигни, то у нас R^2 будет хорошим. Поэтому применяют ещё скорректированный коэффициент детерминации. Кстати, чем больше гиперплоскость, тем короче перпендикуляр, поэтому разница будет меньше. Вводится корректировка на число степеней свободы: $\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$. Этот коэффициент уже не тот. Когда добавляем переменную, то тогда эффект аналогичен уменьшению RSS, поэтому при большом количестве переменных он не вырастет.

Переходим к теореме Гаусса—Маркова. Григорий Гельмутович сегодня не смог провести лекцию, а Елена Владимировна скучала по эконометрике, сейчас даже обрадовалась, что вновь ей займётся. Елена Владимировна даже не доказывала Гаусса—Маркова без матриц, так как всегда в процессе что-то терялось. Когда на комиссии позволяют заменить вопрос на тот, который знает студент, то тогда выбирают именно эту теорему.

1. Если модель правильно специфицирована (постоянная эластичность рассматривается как постоянная, или нет лишних переменных, или есть все важные переменные),
2. Если $\operatorname{rank}(\boldsymbol{X}) = k$,
3. Если $\mathbb{E}(\boldsymbol{\varepsilon}_i) = 0$,
4. Если $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$ (что влечёт гомоскедастичность),

то тогда $\hat{\boldsymbol{\beta}}$ — BLUE НЛО.

Доказательства.

$\hat{\boldsymbol{\beta}} = \underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T}\boldsymbol{Y}$ — линейное преобразование. $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$? Правда ли это? Несмещённость доказывается так:

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \Leftrightarrow \boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} + \boldsymbol{X}^T\boldsymbol{\varepsilon} \Leftrightarrow \underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T}_{\hat{\boldsymbol{\beta}}}\boldsymbol{Y} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}$. Итак, $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}$, поэтому $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underbrace{\mathbb{E}(\boldsymbol{\varepsilon})}_{=0} \Rightarrow \mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Для доказательства эффективности запишем ковариационную матрицу:

$\operatorname{Cov}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\operatorname{Cov}(\boldsymbol{Z})\boldsymbol{A}^T$, $\operatorname{Cov}\boldsymbol{\beta} = \mathbb{E}\underbrace{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}_{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T =$ переставим $= \mathbb{E}\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big) =$

$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underbrace{\mathbb{E}(\boldsymbol{\varepsilon}\cdot\boldsymbol{\varepsilon}^T)}_{\operatorname{Cov}(\boldsymbol{\varepsilon})}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Итак, на главной диагонали у этой матрицы стоят минимально возможные значения: $\operatorname{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, поэтому наша оценка есть the best.

Нам для принятия гипотез и совершения выводов нужно знать распределение эпсилонов.

Нормальная регрессия. $\boldsymbol{\varepsilon} \sim \mathcal{N}(0; \sigma^2\boldsymbol{I})$, величины некоррелированные, оттого независимые, ибо нормальные. Сейчас опять будет много матриц. Если $\boldsymbol{\varepsilon}$ нормальный, то $\boldsymbol{Y}$ тоже нормальный: $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}; \sigma^2\boldsymbol{I})$. Далее, $\hat{\boldsymbol{\beta}} \sim \mathcal{N}\big(\boldsymbol{\beta}; \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big)$, $\hat{\boldsymbol{Y}} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \ldots)$ — сейчас посчитаем. $\hat{\boldsymbol{Y}} = \big\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big\}\boldsymbol{Y}$, где выражение $\big\{\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big\}$ — идемпотентная матрица. Определение идемпотентности: $\boldsymbol{P} = \boldsymbol{P}^T$, $\boldsymbol{P}^2 = \boldsymbol{P}$. Да и в квадрате наша матрица даст $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\cancel{\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}}\boldsymbol{X}^T = \boldsymbol{P}$. Наконец, $\operatorname{Cov}(\hat{\boldsymbol{Y}}) = \boldsymbol{P}\underbrace{\operatorname{Cov}(\boldsymbol{Y})}_{\sigma^2\boldsymbol{I}}\boldsymbol{P}^T = \sigma^2\boldsymbol{P}\cdot\boldsymbol{P} = \sigma^2\boldsymbol{P}$.

Поэтому $\hat{\boldsymbol{Y}} \sim \mathcal{N}\big(\boldsymbol{X}\boldsymbol{\beta}; \sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big)$. А, и остались остатки: $\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}$, $(\boldsymbol{I} - \boldsymbol{P})^T = (\boldsymbol{I} - \boldsymbol{P})$ (идемпотентность), $(\boldsymbol{I} - \boldsymbol{P})(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{I}\boldsymbol{I} - \boldsymbol{P}\boldsymbol{I} - \boldsymbol{I}\boldsymbol{P} + \boldsymbol{P}\boldsymbol{P} = \boldsymbol{I} - \boldsymbol{P} - \boldsymbol{P} + \boldsymbol{P} = (\boldsymbol{I} - \boldsymbol{P})$. $\operatorname{Cov}(\boldsymbol{e}) = (\boldsymbol{I} - \boldsymbol{P})\operatorname{Cov}(\boldsymbol{Y})(\boldsymbol{I} - \boldsymbol{P}) =$ $= \sigma^2(\boldsymbol{I} - \boldsymbol{P})$. Главная проблема: остатки зависимые. Есть масса тестов, которые строят остатки от чего-то. Но это нарушение условий теоремы, так как у нас остатки должны быть независимыми. Итак, $\boldsymbol{e} \sim \big(0, \sigma^2(\boldsymbol{I} - \boldsymbol{P})\big)$. Остатки не всегда суть нормальные наблюдения, а мы по остаткам всё проверяем. Какие бонусы у нас есть при наличии вектора ошибок? Нам понадобится ещё $\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-k}$, $\mathbb{E}(\hat{\sigma}^2) = \sigma^2$.

Общая линейная гипотеза. Это какие-то линейные ограничения на коэффициенты: $\boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{q}$. Это СЛУ относительно $\boldsymbol{\beta}$, поэтому $\boldsymbol{H} = r \times k$, где r — число уравнений связи, $r < k$. $\mathcal{H}_0 \colon \boldsymbol{H}\boldsymbol{\beta} = \boldsymbol{q}$, $\mathcal{H}_1 \colon \boldsymbol{H}\boldsymbol{\beta} \neq \boldsymbol{q}$. Есть гипотеза: $\boldsymbol{H} = (0; \ldots; 0; \underbrace{1}_{s}; 0; \ldots; 0)$, $q = 0$. Гипотеза о значимости. $\beta_s = 0 \Leftrightarrow \boldsymbol{H}\boldsymbol{\beta} = 0$. Основная гипотеза — модель неадекватна. Альтернативная гипотеза — хотя бы один из бета значим. $\mathcal{H}_0 \colon \beta_2 = \ldots = \beta_k = 0$ (модель неадекватна), $\mathcal{H}_1 \colon \beta_1^2 + \ldots + \beta_k^2 > 0$. $\boldsymbol{H} = \begin{pmatrix} 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \ldots & 1 \end{pmatrix}$, $\boldsymbol{q} = 0$.

Нельзя переменные выкидывать по одной. На этот случай существует *гипотеза о наличии группы лишних переменных*: $\beta_s = \beta_{s+1} = \ldots = \beta_k = 0$. Это записывается в матричной форме:

$$\boldsymbol{H} = \begin{matrix} \\ \frac{s-1}{} \\ s \end{matrix} \left[\begin{array}{ccc|cccc} 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ \hline 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \vdots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & 0 & \ldots & 1 \end{array}\right]$$

s

Гипотеза о постоянной отдаче от масштаба: $\ln Y = a + \alpha \ln K + \beta \ln L + \varepsilon$, $\mathcal{H}_0\colon \alpha + \beta = 1 \Leftrightarrow \boldsymbol{H}\tilde{\boldsymbol{\beta}} = q$, $H = (0;1;1)$, $q = 1$, $\hat{\boldsymbol{\beta}} = \begin{pmatrix} a \\ \alpha \\ \beta \end{pmatrix}$. Как проверять? $\mathrm{F} = \frac{(\mathrm{RSS}_{\mathrm{restr}} - \mathrm{RSS}_{\mathrm{unrestr}})/r}{\mathrm{RSS}_{\mathrm{unrestr}}/(n-k)} \sim \mathrm{F}_{r;n-k}$. Модель без ограничений даёт RSS, что unrestricted. Потом насильно впихнули в модель: restricted RSS должна подурнеть качеством. Restricted должно резко вырасти в необъяснённой части. Если ухудшение незначительно, то число маленькое. Значительность определяется критическим значением $\mathrm{F}_{\alpha;r;n-k}$. Попадание в правый хвост даёт нам лишнюю сущность ограничения.

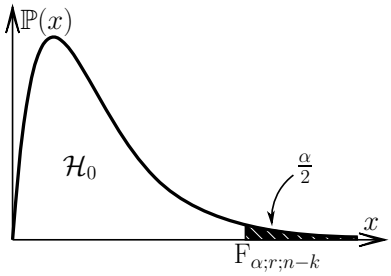

Рис. 6. Фишер

8 Лекция 8

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Это unrestricted регрессия, нет ограничений на параметры. $\mathcal{H}_0\colon \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q}$, $\mathcal{H}_1\colon \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{q}$. Всего имеется r ограничений: $\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q}$. Обычно разрешают систему, подставляют в уравнение, и остаётся unrestricted с меньшим количеством переменных. Ищем $\min \sum e_i^2$. В результате после построения регрессии в unrestricted мы получили $\mathrm{RSS}_{\mathrm{unrestr}}$, во втором — $\mathrm{RSS}_{\mathrm{restr}}$. У restricted жёстче и неудобнее ограничения, поэтому $\mathrm{RSS}_{\mathrm{restr}} \geqslant \geqslant \mathrm{RSS}_{\mathrm{unrestr}}$. В статистике принято смотреть, большой или небольшой прирост неоценённости даёт наложение ограничений. Рассматривается так: $\mathrm{F} = \frac{(\mathrm{RSS}_{\mathrm{restr}} - \mathrm{RSS}_{\mathrm{unrestr}})/r}{\mathrm{RSS}_{\mathrm{unrestr}}/(n-k)} \sim \mathrm{F}_{r;n-k}$. $\mathrm{TSS} = \sum y_i^2$, $\mathrm{TSS} = \mathrm{RSS}_{\mathrm{restr}} + \mathrm{ESS}_{\mathrm{restr}} = \mathrm{RSS}_{\mathrm{unrestr}} + \mathrm{ESS}_{\mathrm{unrestr}}$, $\mathrm{RSS} = (1 - R^2)\mathrm{TSS}$. Упростим:

$$\mathrm{F} = \frac{(1 - R^2_{\mathrm{restr}} - 1 + R^2_{\mathrm{unrestr}})/r}{(1 - R^2_{\mathrm{unrestr}})/(n-k)} = \frac{(R^2_{\mathrm{restr}} + R^2_{\mathrm{unrestr}})/r}{(1 - R^2_{\mathrm{unrestr}})/(n-k)} \tag{8.1}$$

Будет проверяться гипотеза: $\mathcal{H}_0\colon \beta_{e+1} = \ldots = \beta_k = 0$, $\mathcal{H}_1\colon \beta_{e+1}^2 + \ldots + \beta_k^2 > 0$. Сравниваются более короткая и более длинная регрессии, чтобы понять, значат ли вышеуказанные факторы или нет. Но! Если отдельно каждый коэффициент незначим, то это не значит, что они все в сумме также незначимы. Это объясняется наличием ковариации между ними. Может быть и наоборот: все незначимы, а по отдельности они значимы.

В жизни бывает масса ситуаций, в которых, например, мы моделируем потребителя. Смысл говорит: как он там максимизирует функцию полезности? Если он девушка, то он выберет один набор товаров, если мужчина, то второй. Наоборот ситуация: некоторые получили диплом, а некоторые занимались бизнесом. Есть-ли отдача отъ высшаго образованія?[1] Собираются данные о выпускниках, и утверждается: стартовая зарплата зависит от наличия диплома бакалавра — так ли это? Попытаемся проверить. $D = \begin{cases} 1, & \exists \text{ diploma;} \\ 0, & \nexists \text{ diploma.} \end{cases}$ Строим простую модель: $Y = \alpha + \beta D + \varepsilon$. Переменная D получилась искусственной (dummy), булевской. Но теорема Гаусса—Маркова не запрещает нам оценить параметры и утверждать: а можно во всех этих случаях посчитать среднее (Вася вылетает из Вышки на половине и идёт работать). $\hat{Y}_0 = \hat{\alpha} + \hat{\beta} D_0$, и получается: для вылетевших $\hat{Y}_0 = \hat{\alpha}$, для получивших диплом $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}$. $\mathcal{H}_0\colon \beta = 0$, $\mathcal{H}_1\colon \beta > 0$. Причём несущественно, какому признаку давать ноль, какому — единицу. Где ноль — там признак базовый. Если взять в системе D 2 и 0, то просто коэффициент увеличивается. Это значит, что нет смысла бороться и крушить черепа за что-либо иное, кроме 1 и 0.

Ситуацию можно усложнить: нет диплома, диплом бакалавра и диплом магистра. Причём тогда нельзя смешивать признаки. Нам предлагается довольно естественное решение: $D_i = \begin{cases} 2, & \text{magister;} \\ 1, & \text{graduate;} \\ 0, & \nexists\text{diploma.} \end{cases}$ Но это неверно! Нет возможности точно измерить разницу между магистром и бакалавром или бакалавром и неудачником. Получится тогда $2\hat{\beta}$, $\hat{\beta}$, 0, но это может быть и не так! Вводятся две дамми-переменные: $D_1 = \begin{cases} 1, & \exists\delta; \\ 0, & \text{otherwise.} \end{cases}$ $D_2 = \begin{cases} 1, & \exists\delta\mu; \\ 0, & \text{otherwise.} \end{cases}$ У нас можно рассчитать разницу между дипломами в модели $Y = \alpha + \beta D_1 + \gamma D_2 + \varepsilon$: без диплома имеем $\hat{Y}_0 = \hat{\alpha}$, с бакалавром — $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}$, с магистром — $\hat{Y}_0 = \hat{\alpha} + \hat{\gamma}$, $\mathcal{H}_0\colon \gamma - \beta = 0$, $\mathcal{H}_1\colon \gamma - \beta \neq 0$.

Можем ли мы приять полную симметрию? $D_1 = \begin{cases} 1, & \nexists\delta; \\ 0, & \text{otherwise.} \end{cases}$ $D_2 = \begin{cases} 1, & \exists\delta; \\ 0, & \text{otherwise.} \end{cases}$ $D_3 = \begin{cases} 1, & \nexists\delta\mu; \\ 0, & \text{otherwise.} \end{cases}$ В модели получается вот что: $Y = \alpha + \beta D_1 + \gamma D_2 + \delta D_3 + \varepsilon$, и получится, что если в одном случае единица, то в остальных — нули ($D_1 + D_2 + D_3 = 1$). Но тогда будет линейная зависимость, и мы не сможем применить МНК: $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0$. Ошибкой является введение дамми-переменных для всех факторов плюс свободный член. Это называется *dummy trap*.

[1] Въ старой орѳографіи частицы «бы», «ли», «же» имѣли дефисное написаніе.

Пусть карьерный рост зависит от стажа: $Y_0 = \alpha + \beta X + \varepsilon$. Тогда α — стартовая зарплата, β — склонность к карьерному росту. Тогда $D = \begin{cases} 1, & \exists\delta; \\ 0, & \text{otherwise.} \end{cases}$ Тогда γ показывает человеческую оценку карьерного роста. Усложним модель: $Y_i = \alpha + \beta X + \varepsilon + \gamma D + \delta(D\boldsymbol{X})$, где в матрице $\boldsymbol{X}$ на некоторых столбцах поставлены нули. Это называется mixed dummy. Когда нет диплома, то $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. С дипломом начинает работать формула $\hat{Y} = (\hat{\alpha} + \hat{\gamma}) + (\hat{\beta} + \hat{\delta})X$. Если мы ставим гипотезу, что образование не нужно, то гипотеза будет такой: $\mathcal{H}_0 \colon \gamma = \delta = 0$, $\mathcal{H}_1 \colon \gamma^2 + \delta^2 > 0$. Воспользуемся уже использованной F-статистикой: $\mathrm{F} = \frac{(R^2_{\text{unrestr}} - R^2_{\text{restr}})/2}{(1 - R^2_{\text{unrestr}})/(n-4)} \sim \mathrm{F}_{2;n-4}$. Мы окрасили выборку: красненькие — с дипломом, коричневые — все остальные. Мы ответим на вопрос: нужно ли для каждого подмножества строить отдельную регрессию, или we can pool the data? Это важный, но игнорируемый вопрос.

Важно помнить, что, например, может быть такая ситуация по выборкам: выборки могут отличаться или сдвигом, или коэффициентом, или и тем, и другим. А могут они и совпасть. Поэтому значимость коэффициента γ даёт нам существенность сдвига, δ — наклона. Просто англичане слишком ленивы, чтобы правильно произносить слова…

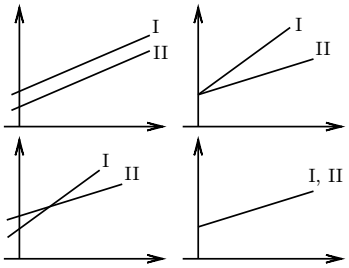

Рис. 7. Влияние γ и δ

Строим RSS_1, RSS_2, $\mathrm{RSS}_{\text{pooled}}$. Тогда $\frac{(\mathrm{RSS}_{\text{pooled}} - \mathrm{RSS}_1 - \mathrm{RSS}_2)/2}{\mathrm{RSS}_{\text{pooled}}/(n-k)} \sim \mathrm{F}_{2;n-k}$. Мы строим регрессию по одной выборке, по второй, потом совместной. Этот тест Чоу (Chow) — та же формула, что и для дамми-переменных. Техника дамми гибче, а Чоу только говорит, равны они или нет. Дамми говорит, за счёт чего они рознятся: содержательный член или коэффициент наклона.

В регрессии с дамми-переменными мы проверяем гипотезу о том, что две разные регрессии, но одинаковый σ^2. Поэтому тут неявно заложено, что σ^2 одинаковые. Это очень нехорошо! Как проверить дисперсию в двух регрессиях? Это довольно сложно.

Не надо быть шибко умным и слушать курс экиста, чтобы понять, что было в США в 1929–33 году. Поэтому возникает гипотеза: модель развития экономики могла поменяться. Она могла смениться и при Ельцине, и при Путине. Поэтому дамми такая: $D = \begin{cases} 1, & t < t^*; \\ 0, & t \geqslant t^*. \end{cases}$ Это помогает в одну модель запихнуть разные значения коэффициентов, поэтому говорят: structural break. Есть тест структурной однородности модели. Филлипс исследовал темп роста денежных доходов населения от уровня безработицы. Введём: $Y = \alpha + \beta\frac{1}{X} = \varepsilon$. Рассмотрим индекс повременной оплаты. Кроме обратной зависимости, есть нормальный уровень безработицы — NAIRU. Поэтому должна быть асимптота при $X \to \infty$.

В разных периодах динамика могла быть разной: $d = \begin{cases} 0, & 1958\text{–}1969; \\ 1, & \text{otherwise} \end{cases}$ $\hat{Y} = 10{,}1 - 10{,}3d - 17{,}5\frac{1}{X} + 38{,}1\frac{D}{X}$. Соответствующие t: 7,2; −6,1; −2,1, 4,1. $R^2 = 0{,}838\,7$, 69–77: $\hat{Y} = 10{,}1 - 17{,}5\frac{1}{X}$, 58–69: $\hat{Y} = 0{,}259 + 20{,}56\frac{1}{X}$.

Есть ещё сезонность. Приходится производить deseasoning, устранить сезонную составляющую. Мы можем не устранять сезонности, но учесть оную при построении модели. $d_1 = \begin{cases} 1, & \text{I квартал}; \\ 0, & \text{otherwise.} \end{cases}$ $d_2 = \begin{cases} 1, & \text{II квартал}; \\ 0, & \text{otherwise.} \end{cases}$ Итак, сколько кварталов, столько и d_i. Получаем: $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \gamma_1 d_1 + \ldots + \gamma_3 d_3 + \delta_1(X_1 d_1) + \delta_2(X_2 d_2) + \ldots$. В общем случае не надо учитывать сезонность. На третьей пересдаче у него спросят: как выглядят данные? — «1 0 0 0»… При перекрёстных данных везде сохраняются единички. Тогда $\mathcal{H}_0$: нет сезонности, $\mathcal{H}_1$: $\exists$ сезонность.

9 Лекция 9

$\vec{\boldsymbol{Y}} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$

Нам многое не нравится в теореме Гаусса—Маркова: не все предположения естественны. Мы начнём ослаблять условия и рассматривать интересные ситуации. Откажемся от не очень реалистичного предположения, что $\boldsymbol{X}$ детерминированы. $\boldsymbol{X}$ — стохастический регрессор. Неприятная ситуация: теперь два источника стохастичности: $\boldsymbol{X}$ и $\boldsymbol{\varepsilon}$. Поэтому возникает вопрос: если мы рассматриваем случайные величины $\boldsymbol{X}$ и $\boldsymbol{\varepsilon}$, то говорить что-то можно только тогда, когда мы знаем их совместное распределение.

Теперь всё случайное: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\boldsymbol{X}$ стохастический. $Y_i = \alpha + \beta X_i + \varepsilon_i$, $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \sum k_i y_i$. $\mathbb{E}(\hat{\beta}) = \mathbb{E}\left(\sum \frac{x_i y_i}{\sum x_i^2}\right)$. Если подставить y, то выражение примет вид: $\hat{\beta} = \beta + \frac{\sum x_i\varepsilon_i}{\sum x_i^2}$, $\mathbb{E}(\hat{\beta}) = \int\cdots\int \frac{\sum x_i\varepsilon_i}{\sum x_i^2}\mathbb{P}(X, \vec{\varepsilon})\,dx\,de$. Если регрессор стохастический, то оценка становится смещённой. Поэтому мы должны внести какие-то упрощения.

Пусть $\boldsymbol{X}$ случаен и не зависит от $\boldsymbol{\varepsilon}$. Это возможно тогда и только тогда, когда их совместное распределение раскладывается в распределение двух плотностей (factorising). $\boldsymbol{X}$ и $\boldsymbol{\varepsilon}$ входят только в произведение: $\mathbb{P}(\boldsymbol{X}, \boldsymbol{\varepsilon}) = \mathbb{P}_X(\boldsymbol{X}) \cdot \mathbb{P}_\varepsilon(\boldsymbol{\varepsilon})$. $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \sum k_i(x)\varepsilon_i$, $\mathbb{E}(\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \sum_i \mathbb{E}(k_i X \cdot e_i)\ominus$

$$\text{Вспомним: } \int g(x)f(y)\underbrace{\mathbb{P}_X(X)\mathbb{P}_Y(Y)}_{\mathbb{P}_2(X,Y)}\,dx\,dy = \underbrace{\int g(x)\mathbb{P}_X(X)\,dx}_{\mathbb{E}\left(g(x)\right)} + \underbrace{\int f(y)\mathbb{P}_Y(Y)\,dy}_{\mathbb{E}\left(f(y)\right)}$$

$\ominus\boldsymbol{\beta} + \sum \mathbb{E}_X(k_i) \cdot \underbrace{\mathbb{E}_\varepsilon(\vec{\varepsilon}_i)}_{0}$.

1. Пункт 1 я либо не отметил на лекции, либо пропустил.
2. $\boldsymbol{X}$ стохастический. А что такое полный ранг? $\boldsymbol{X}$ — стохастическая случайная матрица, и неясно, что такое $\operatorname{rank}(\boldsymbol{X})$. Мы не знаем, что это такое.
3. $\mathbb{E}(\vec{\varepsilon}) = 0$;
4. $\operatorname{Var}(\vec{\varepsilon}) = \sigma^2$;
5. $\operatorname{Cov}(\vec{\varepsilon}) = \sigma^2 I$.

Нам важно это свойство, чтобы $\det(\boldsymbol{X}^T\boldsymbol{X}) \neq 0$. В выборке это должно быть выполнено. Но $\boldsymbol{X}$ — статистическая матрица, поэтому хочется сказать что-то общее. Выпишем, чего мы хотим от статистических свойств матрицы $\boldsymbol{X}$. Может вдруг оказаться, что $\det = 0$. Мы интуитивно хотим, чтобы такие случаи были редкими. Пока у нас конечная выборка, сформулировать это трудно. Рассмотрим $\operatorname*{plim}_{n\to\infty}(\boldsymbol{X}^T\boldsymbol{X})$ — предел по вероятности. Очевидно, что эта матрица будет размерности $(k \times n) \cdot (n \times k) = k \times k$. Может получиться дурная бесконечность, поэтому надо поделить на n: $\operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right) = \boldsymbol{S}_{\boldsymbol{X}^T\boldsymbol{X}}$. Это этакая предельная ковариационная матрица, и она пусть будет невырожденной положительно определённой. Надо, чтобы $\boldsymbol{S}_{\boldsymbol{X}^T\boldsymbol{X}}$ было только положительно определённой (этого достаточно). И даже в этом простом случае придётся написать: должен существовать предел по вероятности, и он должен быть невырожденной матрицей. Если выполнена замена ранга на теоретическое условие, то оценка остаётся BLUE-оценкой. Это свойство сохранится в своём классе.

Чтобы было гарантировано некоторое хорошее свойство, мы должны ослабить одно условие. И ослабим непривычным образом: мы начинаем заменять третье условие. $\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = 0$. Условное матожидание ошибки по отношению к $\boldsymbol{X}$ равно 0, каким бы $\boldsymbol{X}$ ни был. Это более слабое условие, чем $\mathbb{E}(\boldsymbol{e}) = 0$. Запишем безусловное через условное: $\mathbb{E}(\boldsymbol{e}) = \mathbb{E}_X\big(\underbrace{\mathbb{E}_\varepsilon(\boldsymbol{e} \mid \boldsymbol{X})}\big) = 0$. Это называется условием экзогенности $\boldsymbol{X}$ по отношению к $\boldsymbol{\varepsilon}$.

Скажем так: в терминах условной по $\boldsymbol{X}$ ковариационной матрицы $\operatorname{Cov}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \sigma^2\boldsymbol{I}$. Для разнообразия посмотрим также общий случай с матрицами. $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + \underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T}\boldsymbol{\varepsilon}$. Какое смешное выражение же получилось: $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathbb{E}_X\left(\mathbb{E}_\varepsilon(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\varepsilon} \mid \boldsymbol{X}\right) = \mathbb{E}_X\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\underbrace{\mathbb{E}(\boldsymbol{\varepsilon} \mid \boldsymbol{X})}_{=0}\big) = \hat{\boldsymbol{\beta}}$.
Вывод: если $\boldsymbol{X}$ экзогенен, то сохраняется несмещённость оценки МНК. Но экзогенность — это сильное условие. Получается, что $\boldsymbol{X}$ экзогенные ко всем $\boldsymbol{\varepsilon}$, и это значит, что все $\boldsymbol{X}$ не влияют ни на какие $\boldsymbol{\varepsilon}$. Это разумно там, где нет времени.

У нас $\mathbb{E}\left(\begin{pmatrix}\varepsilon_1\\ \vdots \\ \varepsilon_n\end{pmatrix} \,\middle|\, \boldsymbol{X}\right)$. Поэтому остались не все $\boldsymbol{X}$, а только соответствующие наблюдению. Потребуем: $\mathbb{E}(\varepsilon_i \mid X_{i,j\colon j=1;\dots;k}) = 0$. Последнее значит, что строчка — i — та же самая, а j — все столбцы. $\mathbb{E}(\boldsymbol{\varepsilon}_t \mid \boldsymbol{X}_t) = 0$ — условие предопределённости (predetermined). Увы, при нём мы не можем доказать несмещённость МНК. Зато остаётся возможность получить состоятельную оценку. В этом случае происходит довольно простая вещь: $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^T\vec{\varepsilon}\right)$. Если нет экзогенности, не докажется несмещённость. Несмещённости нет, но, может, есть состоятельность. Это означает $\operatorname{plim}(\hat{\boldsymbol{\beta}})$, и Слуцкий говорит: если функция непрерывна, то plim равна произведению, сумме и т. д. $\operatorname{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \underbrace{\operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}}_{\boldsymbol{S}^{-1}_{\boldsymbol{X}^T\boldsymbol{X}}} \cdot \operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}^T\vec{\varepsilon}\right)$. Значит, если $\operatorname{plim}\left(\frac{1}{n}\boldsymbol{X}^T\vec{\varepsilon}\right) = 0$, то всё хорошо. Как это назвать? $\boldsymbol{\varepsilon}$ асимптотически некоррелированные с $\boldsymbol{X}$. Это гарантирует, что оценки МНК будут состоятельными, а если хотя бы для одного $\boldsymbol{X}$ такой предел не равен нулю или не существует, то $\operatorname{plim}\hat{\boldsymbol{\beta}} \neq \operatorname{plim}\boldsymbol{\beta}$. Иначе будет несостоятельность! Поэтому-то корреляция $\boldsymbol{X}$ и $\boldsymbol{\varepsilon}$ так страшна!

Рассмотрим $\frac{1}{n}\sum_{i=1}^{n} X_i^j\varepsilon_i$ — матожидание этого чудесного выражения мы можем посчитать, ведь это есть $\mathbb{E}_X\left(\frac{1}{n}\sum X_i^j\mathbb{E}(\varepsilon_i \mid X_i^j)\right)$. При условии предопределённости это равно нулю. По Чебышёву, оценка состоятельна, если $\mathbb{E}(\hat{\cdot}) \to \mathbb{E}(\cdot)$, $\operatorname{Var}(\cdot) \to 0$. Вектор длины k: $\operatorname{Cov}\left(\frac{1}{n}\sum \boldsymbol{X}^T\vec{\varepsilon}\right)$. $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)$ размерности $k \times k$. Рассчитаем: $\operatorname{Cov}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\varepsilon}\right) = \mathbb{E}\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{\varepsilon} \cdot \frac{1}{n}\boldsymbol{\varepsilon}^T\boldsymbol{X}\right) = \frac{1}{n^2}\mathbb{E}(\boldsymbol{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{X}) = \frac{1}{n^2}\mathbb{E}_x\left(\mathbb{E}_\varepsilon(\boldsymbol{X}^T\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\boldsymbol{X} \mid \boldsymbol{X})\right) = \frac{1}{n}\mathbb{E}_x\big(\frac{1}{n}\boldsymbol{X}^T\underbrace{\mathbb{E}_\varepsilon(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T \mid \boldsymbol{X})}_{\sigma^2\boldsymbol{X}}X\big) =$
$= \frac{\sigma^2}{n}\mathbb{E}_X\left(\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}\right)$. Теперь можно сказать: эта матрица имеет свой предел по вероятности, поэтому её матожидание — какой-то вектор. Но при $n \to \infty$ всё выражение стремится к 0. Мы помним, что к 0 стремится ковариационная матрица, а plim сходится.

Если $\boldsymbol{X}$ и $\boldsymbol{\varepsilon}$ независимы, то мы имеем BLUE-оценки. Если $\boldsymbol{X}$ экзогенный, то выполняется. Если $\boldsymbol{X}$ предопределённый, то есть состоятельность. Но как только нарушено слабое условие предопределённости, т. е. регрессор коррелирован с $\boldsymbol{\varepsilon}$, то тогда оценка становится несостоятельной и смещённой. К сожалению, не так часто возникает некоррелированность.

Первый случай — ошибка в измерениях $\boldsymbol{X}$. Второй случай — оцениваемое нами уравнение неединично, что говорит о том, что оно часть системы. Объясним второй случай. $\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t; \\ Y_t = C_t + I_t. \end{cases}$ Предположим, что ε удовлетворяют всем условиям теоремы Гаусса—Маркова. Как пятиклассник, подставим: $C_t = \alpha + \beta(C_t + I_t) + \varepsilon_t \Rightarrow$ $\Rightarrow C_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} I_t + \frac{1}{1-\beta}\varepsilon_t$. Система с тремя переменными: C, Y, I. Мы должны выбрать эндогенные и экзогенные переменные. Конечно, потребление эндогенно: $Y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{1}{1-\beta}\varepsilon_t$. В чём ужас: у нас X — объясняющая переменная — коррелирует с ε. И оценки МНК получатся смещёнными и несостоятельными: simultaneously biased estimates. Мы никогда не уверены: будь у нас любая система соотношений, где среди объясняющих переменных есть объясняемая переменная из другого уравнения, — у нас возникнет корреляция, и оценка МНК смещённая и несостоятельная.

Вроде бы Фридман не говорил об ошибках в измерениях переменных. Каждый потребитель знает некоторый уровень перманентного дохода. Стипендия — перманентный доход, премия — неперманентный. И потребительское поведение разложено на перманентное и случайное. Отношение такого типа связывает предсказуемые величины. Сейчас мы откажемся от больших букв. Большими будем называть измерения, а маленькими — истинные величины. $y = \alpha + \beta x + \varepsilon$, $X = x + u$, $Y = y + v$, где u и v суть transitory. У нас целых три случайные величины. Предположим, что всё-всё независимо из трёх величин: $\operatorname{Cov}(\varepsilon, u) = \operatorname{Cov}(\varepsilon, v) = \operatorname{Cov}(u, v) = 0$. Пусть все случайные величины — u, v, ε — удовлетворяют условиям теоремы Гаусса—Маркова. Строим регрессию: $Y = A + BX + \varepsilon$. Тогда $\hat{B} = \frac{\sum\limits_i \varepsilon_i}{\sum x_i^2}$. Мы хотим получить эту оценку и сравнить её с тем β, который был бы в другой оценке. Выяснится, что x коррелирует с ε.

10 Лекция 10

Проблем нет, когда регрессор X и ошибка ε независимы. Мы же для облегчения жизни ослабили условие $\mathbb{E}(\vec{\varepsilon} \mid X) = 0$ следующим: эпсилон должен иметь нулевое матожидание относительно только своего наблюдения: $\mathbb{E}(\varepsilon_i \mid \vec{X}_i^T) = 0$, и тогда оценки будут состоятельными. Если регрессор не коррелирует с ошибкой, то оценки состоятельные. Если оценка состоятельная, то достаточное условие — асимптотическая несмещённость, то есть оценка наблюдения стремится к оцениваемому параметру, а дисперсия стремится к нулю. Самый же неприятный момент — это ~~$\operatorname{Cov}(X^T, \vec{\varepsilon}) \neq 0$~~ $\mathbb{E}(X^T\vec{\varepsilon}) \neq 0$.

Рассмотрим пример ошибки в переменных. Классический пример — доход в модели Милтона Фридмана. Наша предпосылка: переменные отцентрированные. $Y_t^0 = \beta_1 + \beta_2 x_t^0 + u_t^0$. Перманентный доход — это ненаблюдаемая переменная. Мы наблюдаем $x_t = x_t^0 + v_{1t}$, $y_t = y_t^0 + v_{2t}$, где v — transitory income, случайное потребление, а у нас есть только x и y. $u_t^0 \sim \text{i.i.d.}(0, \sigma_u^2)$, $v_{1t} = \text{i.i.d.}(0, w_1^2)$, $v_{2t} = \text{i.i.d.}(0, w_2^2)$. Природа ошибок разная. Ошибка измерений — это u, и $\operatorname{Cov}(u_t^0, v_{1t}) = \operatorname{Cov}(u_t^0, v_{2t}) = \operatorname{Cov}(v_{1t}, v_{2t}) = \operatorname{Cov}(x_t^0, u_t^0) = \ldots = 0$. Однако у нас есть только x и y, поэтому искомое соотношение имеет следующий вид: $y_t - v_{2t} = \beta_1 + \beta_2(x_t - v_{1t}) + u_t^0$. $y_t = \beta_1 + \beta_2 x_t + \underbrace{(v_{2t} + u_t^0 - \beta_2 v_{1t})}_{u_t}$ — это та же регрессия, но более сложная структура ошибки, в которой 3 компонента. Гомоскедастичность сохраняется: $\operatorname{Var}(u_t) = w_2^2 + \sigma_u^2 + \beta_2^2 w_1^2$, $x_t = x_t^0 + v_{1t}$, $u_t = u_t^0 + v_{2t} - \beta_2 v_{1t}$. И ошибки ни экзогенные, ни предопределённые, ни некоррелированные, так как верно, что $\operatorname{Cov}(x_t, u_t) = -\beta_2 w_1^2 \neq 0$, и корреляция появляется из-за ошибки в регрессоре; она увеличивает дисперсию, но не влияет на корреляцию регрессора с ошибкой.

$\hat{\beta}_2 = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{(x_t^0 + v_{1t})(y_t^0 + v_{2t})}{\sum (x_t^0 + v_{1t})^2}$, $\operatorname{plim}\hat{\beta}_2 = \frac{\operatorname{plim}\frac{1}{n}\sum(x_t^2 y_t^2 + x_t^0 v_{2t} + v_{1t} y_t^0 + v_{1t} v_{2t})}{\operatorname{plim}\frac{1}{n}\sum(x_0^2 + v_{1t}^2 + 2x_t^0 v_{1t})} \Ominus$ в числителе только одна ковариация ненулевая, да и в знаменателе останутся только ошибки: $\Ominus \frac{\operatorname{Cov}(x_t^0, y_t^0)}{\sigma_x^2 + w_1^2}$; $\sigma_x^2 = \operatorname{Var}(x_t^0)$, $\beta_2 = \frac{\operatorname{Cov}(x_t^0, y_t^0)}{\sigma_x^2}$, и получилась несостоятельность, хоть мы и записали всё в явном виде: $\operatorname{plim}\hat{\beta}_2 = \beta_2 \frac{\sigma_x^2}{\sigma_x^2 + w_1^2} \neq \beta_2$. Вывод: за счёт ошибки в измерениях оценка у нас заниженная. Это именно та ситуация, с которой столкнулись экономисты, когда строили зависимость потребления от дохода, и mpc у них было меньше, чем есть. Фридман предложил разложить на transitory и permanent и тем самым объяснил, почему в простой модели занижается предельная склонность к потреблению.

Вся наша ошибка зависит от $\frac{\sigma_x^2}{\sigma_x^2 + w_1^2} = \frac{1}{1 + \frac{w_1^2}{\sigma_x^2}}$, или соотношения полезного сигнала и случайного шума. Итак, существуют другие соотношения, кроме нашего случая. МНК получается смещённым, и нам остаётся либо надеяться на то, что ошибки маленькие, либо использовать…

Метод инструментальных переменных. Метод устроен так, чтобы проверить, ходит студент на лекции или нет. Те, кто не ходит, говорят, что β_{IV} — это «бета-4» (на самом деле «ай-ви»).

$\tilde{\beta}_{IV}$. Условия: $\operatorname{Cov}(X, \vec{\varepsilon}) \neq 0$, $\mathbb{E}(\vec{\varepsilon} \mid X) \neq 0$. Будем считать, что у нас есть, кроме наших ненадёжных наблюдений, ещё некоторые переменные — инструментальные (инструмент — Z). Relevant variable: $\operatorname{Cov}(X_i, Z_i) \neq 0$, то есть она имеет отношение к X. Важно, что $\operatorname{Cov}(Z_i, \varepsilon_i) = 0$. Инструмент лучше, чем X, он не должен страдать от того, от чего болеет X.

Где взять инструмент — это отдельная проблема. По-прежнему действует наше соотношение о больших и маленьких буквах: $\alpha + \beta X + \varepsilon$. $\tilde{\beta}_{\text{IV}} = \frac{\sum z_i y_i}{\sum x_i z_i}$. Но это не то же, что подставить X. Инструментальные переменные —

это не замена регрессора на Z! Далее, $\beta_{\text{IV}} = \frac{\sum z_i(\beta x_i + \varepsilon_i)}{\sum x_i z_i} = \beta + \frac{\sum z_i \varepsilon_i}{\sum x_i z_i}$. z — стохастический регрессор, через матожидание мы не прорвёмся. $\operatorname{plim} \beta_{\text{IV}} = \beta + \frac{\operatorname{plim} \frac{1}{n} \sum z_i \varepsilon_i}{\operatorname{plim} \frac{1}{n} \sum x_i z_i} = \beta$, так как числитель равен нулю, знаменатель не равен нулю. Поэтому оценка состоятельная! Итак, если мы имеем инструмент, то мы можем получить несмещённую оценку даже при коррелированности регрессора и случайной ошибки.

Итак, у нас есть линейная регрессия: $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$, $\mathbb{E}(\boldsymbol{X}^T\vec{\varepsilon}) \neq 0$. Договоримся, что все регрессии, которые не коррелируют с $\boldsymbol{X}$, мы выбираем как инструменты, и мы будем заменять те переменные из матрицы $\boldsymbol{X}$, которые коррелированные с ошибкой, инструментами. $\boldsymbol{Z}$ — матрица инструментов размера $m \times k$, и столбцы этой матрицы включают все столбцы из матрицы $\boldsymbol{X}$, которые не коррелируют с $\boldsymbol{\varepsilon}$. Ковариация всех столбцов $\boldsymbol{Z}$ с $\boldsymbol{\varepsilon}$ равна 0; ради этого мы и работаем. Тогда оценкой МИП будет аналогичная по структуре формула. $\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{Y}}$, $\tilde{\vec{\beta}}_{\text{IV}} = (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\vec{Y}$. $\det \boldsymbol{Z}^T\boldsymbol{X} \neq 0$, и это не означает, чтобы прямо все ячейки матрицы были ненулевыми. МИП даёт линейную по Y оценку. Доказать её несмещённость — гиблое дело. А благодаря Слуцкому мы находим plim (главное, чтобы предел только существовал!):

$$\tilde{\vec{\beta}}_{\text{IV}} = (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T(\boldsymbol{X}\vec{\beta} + \vec{\varepsilon})) = \vec{\beta} + (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\vec{\varepsilon}, \quad \operatorname{plim} \tilde{\vec{\beta}}_{\text{IV}} = \boldsymbol{\beta} + \operatorname{plim} \frac{1}{n}\left(\boldsymbol{Z}^T\boldsymbol{X}\right)^{-1} \cdot \operatorname{plim} \underbrace{\left(\frac{1}{n}\boldsymbol{Z}^T\vec{\varepsilon}\right)}_{=\vec{0}} \tag{10.1}$$

$\operatorname{plim} \frac{1}{n}\boldsymbol{Z}^T\boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{Z}^T\boldsymbol{X}}$, $\det \boldsymbol{S}_{\boldsymbol{Z}^T\boldsymbol{X}} \neq 0$. Так как каждый элемент этой матрицы — скалярное произведение строчки инструмента на строчку $\boldsymbol{X}$, то есть простая ковариация $\boldsymbol{Z}$ с $\boldsymbol{X}$. Кроме того, мы хотим лучшей оценки. Тут у нас появляется интуитивное понимание того, что нам нужен валидный инструмент, который не коррелирует с $\boldsymbol{\varepsilon}$. А что мы хотим сказать о корреляции $\boldsymbol{Z}$ с $\boldsymbol{X}$? Чем лучше коррелирован $\boldsymbol{Z}$ с $\boldsymbol{X}$, то тем больше информации он получит. Но если $\boldsymbol{Z}$ будет близок к $\boldsymbol{X}$, то он будет коррелировать с $\boldsymbol{\varepsilon}$. Очень сильная корреляция может нам вредить. Итак, если инструмент слабо коррелирует с $\boldsymbol{X}$, то он слабый. А если сильный, то никак об этом не говорится. А что произойдёт, если мы поменяем столбцы в матрице $\boldsymbol{X}$ или $\boldsymbol{Z}$? Содержательно ничего. А чуть более тщательное рассмотрение приведёт нас вот к какому выводу: если у нас есть некоторый набор иксов, то бессмысленно задавать вопрос: «Для какого $\boldsymbol{X}$?» Ответ: вся совокупность инструментов для всей совокупности $\boldsymbol{X}$, то есть вся линейная оболочка инструмента заменяет линейную оболочку $\boldsymbol{X}$. Просто меняется группа на группу.

Есть **двухшаговый метод наименьших квадратов** — Two-Stage Least Squares (TSLS). И во многих случаях этот метод работает.

Рассмотрим следующую ситуацию: есть линейная регрессия, $Y = \beta_1 + \beta_2 X^2 + \ldots + \beta_k X^k + \gamma_1 Z^1 + \ldots + \gamma_l Z^l + \varepsilon$. Итак, Z экзогенны для ε, а X коррелированные и портят жизнь. TSLS говорит, что на первом шаге строятся вспомогательные регрессии всех эндогенных (коррелированных) на все экзогенные (некоррелированные). Итак, построили регрессию на все хорошие переменные:

$$X^2 = \alpha_1 + \alpha_2 Z^1 + \ldots + \alpha_{l+1} Z^l + w; \quad \ldots; \quad X^k = \delta_1 + \delta_2 Z^1 + \ldots + \delta_{l+1} Z^l + w_k, \tag{10.2}$$

что даёт нам $\hat{X}^2; \ldots; \hat{X}^k$. Тогда по МНК получается: $Y + \beta_1 + \beta_2 \hat{X}^2 + \ldots + \beta_k X^k + \gamma_1 Z^1 + \ldots + \gamma_l Z^l + \varepsilon$, и оценки TSLS являются оценками МИП, где инструмент — это Z, полученный из МНК. $\hat{\beta}_{\text{TSLS}}; \hat{\gamma}_{\text{TSLS}} \rightleftharpoons \tilde{\vec{\beta}}_{\text{IV}}; \tilde{\vec{\gamma}}_{\text{IV}}$. Просто столбик каждой коррелированной с ε величины X меняем на оценку его из регрессии на первом шаге. Тогда получим состоятельные оценки: $C_t = a + bY_t + \varepsilon_t$, $Y_t = C_t + I_t$, $Y_t = A + BI_t + wt$, $\hat{Y}_t = \hat{A} + \hat{B}I_i$, $C_t = a + b\hat{Y}_t\varepsilon_t$.

Показано, что двухшаговый метод наименьших квадратов даёт самые эффективные из метода инструментальных переменных. $\tilde{\vec{\beta}}_{\text{IV}} = \boldsymbol{\beta} + (\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\vec{\varepsilon}$, $\operatorname{Cov}(\tilde{\vec{\beta}}_{\text{IV}}) = \mathbb{E}\big((\tilde{\vec{\beta}}_{\text{IV}} - \hat{\vec{\beta}})(\tilde{\vec{\beta}}_{\text{IV}} - \vec{\beta})^T\big) = \mathbb{E}\big((\boldsymbol{Z}^T\boldsymbol{X})^{-1}\boldsymbol{Z}^T\vec{\varepsilon}\vec{\varepsilon}^T[(\boldsymbol{Z}^T\boldsymbol{X})^{-1}]^T\big)$. Но математическое ожидание взять сложно, поэтому берём условное матожидание: $\mathbb{E}_{\boldsymbol{Z},\boldsymbol{X}}\big(\mathbb{E}(\ldots \mid \boldsymbol{X}, \boldsymbol{Z})\big) = \mathbb{E}_{\boldsymbol{Z},\boldsymbol{X}}\big(\sigma^2(\boldsymbol{Z}^T\boldsymbol{X})^{-1} \times (\boldsymbol{Z}^T\boldsymbol{Z}) \times (\boldsymbol{X}^T\boldsymbol{Z})^{-1}\big)$. Всё стандартно: линейная оценка с оценкой её ковариационной матрицы. Так как оценка линейна по $\boldsymbol{Y}$, то условно по $\boldsymbol{Z}$ и $\boldsymbol{X}$ оценка будет асимптотически нормально распределённой. Но за нас ковариационную матрицу считает софт.

Ещё момент: k штук коррелированных переменных и l некоррелированных. Какое между ними отношение? Мы об этом не написали. Но каждый коррелированный столбец мы меняем на инструмент. Нужно, чтобы была обратная матрица, и она должна быть квадратной. А в TSLS строим регрессию каждого $\boldsymbol{X}$ на все $\boldsymbol{Z}$, поэтому инструментов может быть больше, чем плохих регрессоров. МНК совершенно от этого не страдает; в исходной формуле мы меняем всё на оценки $\boldsymbol{X}$. Меньше инструментов быть не может, а больше — пожалуйста.

11 Лекция 11

Мы рассмотрели $Y = \alpha + \beta_1 X^1 + \ldots + \beta_2 X^2 + \gamma_1 Z^1 + \ldots + \gamma_e Z^e + \varepsilon$. Введём $W^1, \ldots, W^m$ — инструменты, причём их необязательно искать в нашем уравнении. Инвестиции не входят в уравнение. На первом шаге строим вспомогательные регрессии $\hat{X}^i = \hat{a}_1 W^1 + \ldots + \hat{a}_m W^m$, $i = \overline{1;k}$, то есть для каждого плохого регрессора строим его регрессию от хороших, а затем строим регрессию с инструментами: $Y = \alpha + \beta_1 \hat{X}^1 + \ldots + \beta_k X^k + \gamma_1 Z^1 + \ldots + \gamma_e Z^e + \varepsilon$, $m \geqslant k$, поскольку плохих переменных — k штук — должно быть меньше. Двухшаговый МНК из имеющегося

набора инструментов отбирает их оптимальные с точки зрения эффективности линейной комбинации. Это просто даже в Excel: заменили истинные на столбцы оценок и прогнали через Excel.

Начали мы с OLS (МНК), обобщение — GLS (ОМНК)… ба, мы же ещё не говорили о нём! И потом пришли к IV. У нас эвристичность: мы что-то придумали, а затем доказали его хорошесть. Потом мы увидели в IV точно так же, что у МНК несостоятельные оценки — а давайте рассмотрим такую оценку. Но надо же подвести общую базу.

Метод моментов. В литературе и статьях мы могли видеть GMM — обобщённый метод моментов. Наши требования к модели формулируются на теоретическом языке: это, это, это — а теоретические свойства не очень понятно как проверить[1]. ММ говорит: потребуем аналога теоретического свойства для выборочных характеристик. На выборочном уровне должно быть выполнение желаемых теоретических свойств. $\vec{Y} = \boldsymbol{X}\vec{\beta} + \varepsilon$, а потом появлялось, что $\mathbb{E}(\boldsymbol{X}^T\vec{\varepsilon}) = 0$. Заменим теоретическое условие $\operatorname{Cov}(X^j, \varepsilon) = 0$ выборочным аналогом $\frac{1}{n}\sum X_i^j e_i = 0$, $j = \overline{1;k}$. Далее, $\boldsymbol{X}^T\vec{Y} = (\boldsymbol{X}^T\boldsymbol{X})\vec{\beta} + \underbrace{\boldsymbol{X}^T\vec{\varepsilon}}$, где $\boldsymbol{X}^T\vec{\varepsilon} = \boldsymbol{X}^T(\vec{Y} - \boldsymbol{X}\tilde{\vec{\beta}}) = 0$, это можно поделить на $\frac{1}{n}$ и закончить:

$$\boldsymbol{X}^T\vec{Y} - (\boldsymbol{X}^T\boldsymbol{X})\tilde{\vec{\beta}} = 0 \Rightarrow \tilde{\vec{\beta}}_{\text{MM}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{Y} \tag{11.1}$$

Пусть теперь у нас есть корреляция: $\operatorname{Cov}(\boldsymbol{X}, \vec{\varepsilon}) \neq 0$, $\operatorname{Cov}(\boldsymbol{W}, \vec{\varepsilon}) = 0$. Заменим на выборочный аналог: $\frac{1}{n}\sum W_i^j e_i = 0$. Тогда $W^T(\vec{Y} - X\tilde{\vec{\beta}}) = 0$, $W^T\vec{Y} - (W^TX)\tilde{\vec{\beta}} = 0$. Но это только тогда, когда число регрессоров равно числу плохих стохастических параметров. В данных случаях мы свойства полученных оценок доказывали для МНК и для МИП. Можно ли что-то утверждать для любого метода моментов? Мы приравняли второй смешанный момент к нулю. Но можно же было взять и другой момент! Что-то хорошее будет и для общего подхода метода моментов.

Перейдём к **методу максимального правдоподобия**. Его почему-то не любят студенты. Он был предложен в начале XX века Фишером, хотя его ещё Гаусс применял. ML — maximal likelihood. Идея философская: то, что мы вводим интерпретацию расхождения между моделью и наблюдениями, переносит нас в случайный мир. Случайность порождают ε, X и Y. Поэтому нам иногда трудно различить возможные распределения ε. Наблюдаемое нами событие может осуществиться с ненулевой плотностью вероятности для любых законов распределения и параметров. Поэтому MLM говорит: пусть есть выборка объёма M. Мы предполагаем, что существует совместная функция распределения. Она может быть распределением Пуассона, равномерным, t, и у них есть параметры. Поэтому плотность совместного распределения зависит и от параметров: $\mathbb{P}_n(x_1;\dots;x_n;\theta_1;\dots;\theta_m) = \mathbb{P}_n(\vec{x} \mid \vec{\theta})$, $\vec{\theta} \in \mathbb{R}^m$.

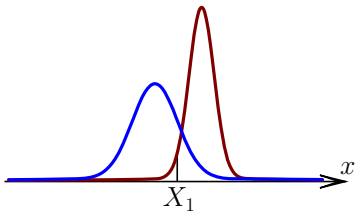

Рис. 8. Ненулевая вероятность любого распределения

$$\mathbb{P}_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \mathbb{P}_1(x \mid \mu, \sigma^2) \tag{11.2}$$

Но у нас есть ненулевая вероятность, что наше наблюдение принадлежит какому-то любому распределению! Метод максимального правдоподобия: более правдоподобно, что были такие параметры, при которых эта вероятность максимальна по μ и σ^2: $\mathbb{P}_1(X \mid \mu, \sigma^2)\,dx = \mathbb{P}\big(\{X_1 \in (x, x+dx)\}\big) \to \max\limits_{\mu,\sigma^2}$. Поэтому надо рассматривать функцию наоборот — функция от параметров θ при фиксированных наблюдениях x:

$$\max_{\vec{\theta}} \mathbb{P}_n(\vec{x} \mid \vec{\theta}) = \max_{\vec{\theta}} L(\vec{\theta} \mid \vec{X}) \to \tilde{\vec{\theta}}_{\text{ML}} = \vec{\theta}_{\text{ML}} \tag{11.3}$$

Функция правдоподобия — это многомерная плотность, рассматриваемая от параметров, когда наблюдения x — параметры нашей функции. Наша экспонента — это на самом деле $L(\mu, \sigma^2 \mid X_1)$. Но всё-таки μ и σ^2 — это не случайные величины.

Пусть $X \sim \mathrm{U}[a,b]$, $p(x) = \begin{cases} \frac{1}{b-a}, \\ 0, \end{cases}$ $x \in (-\infty;a) \cup (b;+\infty)$. В точках a и b плотность выколота.

$\mathbb{P} = \{X_1;\dots;X_n;a;b\} = \prod\limits_{i=1}^{n} \mathbb{P}\{X_i \mid a;b\}$. $L(a,b \mid \vec{X}) = p_n(x) = \begin{cases} \frac{1}{(b-a)^n}, & a \leqslant X_i \leqslant b; i = \overline{1;n}; \\ 0, & \text{otherwise.} \end{cases}$ Для нахождения оценок правдоподобия надо промаксимизировать по a и b по здравому смыслу: чем меньше знаменатель, тем больше правдоподобие. Поэтому $a \leqslant X_i \to \forall i$, $\tilde{a}_{\text{ML}} = \min\limits_i X_i$, $\tilde{b}_{\text{ML}} = \max\limits_i X_i$. Просто у нас сталось так, что функция недифференцируема.

Если $L(\vec{\theta} \mid \vec{X})$ дифференцируема, то FOC — это $\frac{\partial L(\vec{\theta}|\vec{X})}{\partial\theta_j} = 0$, $j = \overline{1;m}$. Уравнение правдоподобия:

$$\frac{\partial}{\partial\vec{\theta}} L(\vec{\theta} \mid \vec{X}) = 0, \quad \frac{\partial L}{\partial\vec{\theta}} = 0, \quad \nabla L = 0 \tag{11.4}$$

[1] Когда главное предложение окаймлено придаточным, запятые не требуются. Подобно данному предложению, в нижеследующем запятой не требуется: «**Вдеть нитку в игольное ушко** никто не знал **как проще**» (сравните: никто не знал, **как проще вдеть нитку в игольное ушко**).

Мы возьмём логарифм функции правдоподобия, или log-likelihood. $\ln L(\vec\theta \mid \vec X) = l(\vec\theta \mid \vec X)$; $\frac{\partial l}{\partial \theta_j} = \frac{1}{L}\frac{\partial L}{\partial \theta_j} = 0$, и технически потом будет сумма, которая уйдёт. Придумано ещё одно название: score-function (счёт, запас). $\mathrm{SF} = \frac{\partial l}{\partial \theta_j}$ — градиент этой функции.

Условия первого порядка необходимы, но недостаточны. Мы же ищем глобальный максимум, а FOC даёт локальные экстремальные точки, причём все, поэтому нужна какая-то локально-глобальная теорема. Для того чтобы решение данной системы FOC являлось $\tilde{\vec\theta}_{\mathrm{ML}}$, надо ввести условие на матрицу Гессе в точке решения. Это гессиан.

$$\text{Hesse-matrix: } \left\| \frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j} \right\|\Big|_{\tilde{\vec\theta}_{\mathrm{ML}}} \quad \text{при } i,j = \overline{1;m} \tag{11.5}$$

Минор — это определитель подматрицы с вычеркнутыми строками. Главный (principal) минор — это определитель с вычеркнутыми одномерными строками и столбцами. Угловой минор (leading principal minor) — это главный минор с вычеркнутыми последними строками и столбцами. $\operatorname{sgn}(\Delta_n) = (-1)^n$. Когда появляются нули, то тогда недостаточно смотреть все leading principal minors. Надо смотреть все главные миноры всякого порядка. В матрице $\underbrace{\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & t \end{pmatrix}}_{3\times 3}$ $\tilde\Delta_1 = \{a, e, t\} \leqslant 0$, $\tilde\Delta_1 = \{\det(abde), \det(efht), \det(acgt)\} \geqslant 0$, $\tilde\Delta_3 = \{\det abcdefght\} \leqslant 0$.

Ладно, мы отобрали локальные максимумы. Но MLF-функция может быть полимодальной, тогда MLM будет неудобно использовать. А на компьютерах сразу легко стало просчитывать MLF. Есть процедура Ньютона—Рафсона…

Пусть $Y_i = \beta_1 + \beta_2 X_i^2 + \ldots + \beta_k X_i^k + \varepsilon_i$, $\varepsilon_i \to$ Gauss, Markov. $\varepsilon_i \sim \mathcal{N}(0;\sigma^2)$. У нас случайная величина Y, мы любим Y. $\mathbb{E}(Y_i) = \beta_1 + \beta_2 X_i^2 + \ldots + \beta_k X_i^k$ — уравнение регрессии, $\operatorname{Var}(Y_i) = \sigma^2$, $Y_i \sim \mathcal{N}(\ldots;\sigma^2)$. Плотность случайных величин есть произведение одномерных плотностей:

$$\mathbb{P}_n(Y_1, \ldots, Y_n) = \prod_{i=1}^n p_i(Y_i) \tag{11.6}$$

$$\begin{aligned} l(\mu, \sigma^2 \mid Y_1, \ldots, Y_n) &= \sum_{i=1}^n = \ln p_i(Y_i) = \\ &= \sum \ln \underbrace{[(2\pi)^{-0,5}(\sigma^2)^{-0,5}]}_{-\frac12 \ln 2\pi - \frac12 \ln \sigma^2} + \left(-\frac{1}{2\sigma^2}(Y_i - \ldots)^2 \right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum \underbrace{\mathrm{RSS}}_{e_i^2} \end{aligned} \tag{11.7}$$

$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum e_i^2 = 0$, $\frac{\partial \ln L}{\partial \beta_j} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta_i}\sum e_i^2 = 0$, поэтому $\tilde{\vec\beta}_{\mathrm{ML}} = \hat{\vec\beta}$. МНК не требует для этого нормальности, а МП требует. Раз уж оценки совпали, то $\sum e_i^2 = \mathrm{RSS}$, откуда $\tilde\sigma^2_{\mathrm{ML}} = \frac{\mathrm{RSS}}{n} \neq \hat\sigma^2$. Одинаковые оценки коэффициентов, но разные оценки параметра σ^2, причём в МНК оценка несмещённая, следовательно, эта оценка смещённая.

ММП не гарантирует несмещённости, однако она асимптотически несмещённая. И ММП всегда даёт состоятельные оценки.

12 Лекция 12

$\vec Y = \boldsymbol{X}\vec\beta + \vec\varepsilon$, $\vec\varepsilon \sim \mathcal{N}(\vec 0; \sigma^2 I_n)$, $\tilde{\vec\beta}_{\mathrm{ML}} = \hat{\vec\beta}$, $\tilde\sigma^2_{\mathrm{ML}} = \frac{\mathrm{RSS}}{n} \neq \frac{\mathrm{RSS}}{n-k} = \hat\sigma^2$.

Любая формула и любая статистика может оценивать любой параметр. Вопрос заключается в точности оценки. Наш контрпример показал, что хорошего свойства МНК — несмещённости — может не быть: $\mathbb{E}(\tilde{\vec\beta}_{\mathrm{ML}}) = \vec\beta$, $\mathbb{E}(\tilde\sigma^2_{\mathrm{ML}}) \neq \sigma^2$. $\max L(\vec\theta \mid \boldsymbol{X}) \Leftrightarrow \max \ln L(\vec\theta \mid \boldsymbol{X})$. Условие правдоподобия: $\frac{\partial}{\partial\vec\theta} L = 0$. Пусть у нас есть какое-то преобразования параметра (перепараметризация).

$\vec g(\vec\theta) = \vec\varkappa$, $\vec\theta \in \mathbb{R}^m$, $\vec g \in \mathbb{R}^m$. Условие существование обратной функции — это необращаемость в ноль производной. Нужно, чтобы матрица Якоби не равнялась нулю в каждой точке области определения.

$$\mathcal{I} = \begin{pmatrix} \frac{\partial g_1}{\partial\theta_1} & \cdots & \frac{\partial g_1}{\partial\theta_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial\theta_1} & \cdots & \frac{\partial g_m}{\partial\theta_m} \end{pmatrix}, \quad \det\mathcal{I} \neq 0 \tag{12.1}$$

Мы хотим найти оценки ММП для новых параметров. Тогда функция правдоподобия зависит от $\varkappa$, то есть от обратной функции: $\tilde{\vec\varkappa}_{\mathrm{ML}}$; $L\big(\vec\theta(\vec\varkappa) \mid X\big) = L_1(\vec\varkappa \mid X)$, $\frac{\partial L_1}{\partial\vec\varkappa} = 0$. Мы хотим найти $\tilde{\vec\varkappa}_{\mathrm{ML}}$. Пример обратной функции: $\vec\theta = \binom{\mu}{\sigma^2}$.

$$\frac{\partial L_1}{\partial \vec{\varkappa}} = 0 = \frac{\partial L}{\partial \vec{\theta}} \times \frac{\partial \vec{\theta}}{\partial \vec{\varkappa}} = 0 \Rightarrow \frac{\partial L}{\partial \vec{\theta}} \quad \text{ой, я забыл дописать?} \tag{12.2}$$

У нас записана СЛУ с невырожденной матрицей, которая имеет единственное ненулевое решение. Градиент вектор-строка записана не $\boldsymbol{A}\vec{x} = 0$, а $\vec{x}^T\boldsymbol{A}^T = 0$, но ведь это одно и то же. Условия первого порядка будут эквивалентны. Свойство инвариантности: если у нас есть другая параметризация нашей плотности вероятности или функции вероятности с невырожденной матрицей, то тогда $\widetilde{[\vec{g}(\vec{\theta})]}_{\text{ML}} = \tilde{\vec{\varkappa}}_{\text{ML}} = \vec{g}(\tilde{\vec{\theta}}_{\text{ML}})$. Мы получили то замечательное свойство, что $\tilde{\sigma}_{\text{ML}} = \sqrt{\tilde{\sigma}^2_{\text{ML}}}$. То есть $\widetilde{\frac{a+b}{2}} = \frac{\tilde{a}+\tilde{b}}{2} = \frac{X_{\min}+X_{\max}}{2}$. Функция от оценок сама является оценкой максимального правдоподобия, поэтому мы параметризуем так, как это удобно.

На свойстве инвариантности хорошие свойства для конечных выборок кончаются. Остальные свойства получаются асимптотически.

Оценка ММП асимптотически состоятельна: $\operatorname{plim}\tilde{\vec{\theta}}_{\text{ML}} = \vec{\theta}_{\text{ML}}$. Для МНК результат о состоятельности следует из рассмотрения неравенства Чебышёва. Оценки несмещённые, матожидание постоянно, а дисперсия МНК при $n \to \infty$ стремится к нулю. Что бы мы ни оценивали ММП, оценки состоятельны. А что такое состоятельность? По мере увеличения размера выборки статистика $\tilde{\vec{\theta}}_{\text{ML}}$, будучи случайной величиной, сходится к детерминированной. Но есть ещё одно свойство: случайная величина при увеличении объёма выборки может стремиться к другой случайной величине. У детерминированной случайной величины функция плотности — это δ-функция: «Ничего… Ничего… Ничего… Ничего себе!!!»

Последовательность случайных величин X_n стремится по распределению к другой случайной величине Y, если разница кумулятивных функций распределения стремится к 0 для всех точек непрерывности:

$$X_n \xrightarrow{d} X \Leftrightarrow \left|F_n(x) - F(x)\right| \to 0 \quad \forall x\colon x \text{ — точка непрерывности } F(x) \tag{12.3}$$

$F(x) = P\big(\{X \leqslant x\}\big)$. Это всегда неубывающая функция, стремится к 0 и 1 в бесконечностях, её производная — плотность распределения. Свойство сходимости: разница при стремлении в бесконечность в каждой точке стремится к нулю. Сходимость по распределению: $\xrightarrow{d}$, где d — distribution. Можно написать и так: $X_n \overset{\text{as}}{\sim} X$. Смысл таков: $X_1; \ldots; X_n \sim$ i.i.d. $\Rightarrow \bar{X}_n = \frac{\sum X_i}{n} \overset{\text{as}}{\sim} \mathcal{N}(\mu, \sigma^2)$; предельное распределение не может зависеть от n. Но мы только что сделали неправильную запись. $\mu = \mathbb{E}(X_i)$, $\operatorname{Var}(X_i) = \sigma^2$. Сформулируем же ЦПТ в форме Линдберга—Леви:

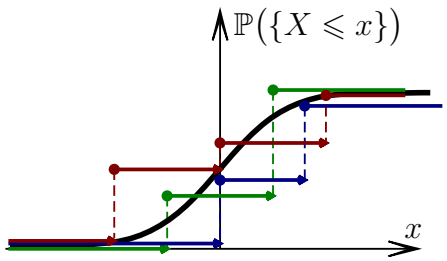

Рис. 9. Сходимость по распределению

> Если $X_1; \ldots; X_n$ является выборкой некоего распределения с конечным матожиданием μ и дисперсией σ^2, то $\sqrt{n}(\bar{X}_n - \mu) \overset{\text{as}}{\sim} \mathcal{N}(0; \sigma^2)$.

Можно снизить требование к случайной выборке из одинакового распределения: $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0; \sigma^2)$. Но распределение этой случайной величины так само по себе будет очень сложным, а при больших выборках асимптотически оно ведёт себя вполне понятно безо всякого n, и мы можем его выразить как результат. Нам неизвестно распределение в конечных выборках, а в пределе имеем невырожденную случайную величину. Если бы не написали $\sqrt{n}$, то было бы $\frac{\sigma^2}{n}$, и в бесконечности случайная величина вырождалась бы в детерминированную! $\sqrt{n}$ нам не даёт получить детерминированную величину. Предельное распределение можно рассматривать как аппроксимацию распределения. Смысл: мы будем неизвестное распределение заменять предельным. Это значит, что для левой величины можно написать приблизительный доверительный интервал, квантили будем ему писать, гипотезы проверять с помощью предельного распределения. Никогда не спрашивайте: «Асимптотически — это сколько?» Иногда полезно оценить скорость предельного распределения, но иногда это бессмысленно.

Ослабим ЦПТ и запишем её в форме Линдберга—Феллера: в случае разных дисперсий и разных матожиданий мы не требуем одинакового распределения. Только сохраним то, что есть случайная выборка.

> Пусть $X_1; \ldots; X_n$ — случайные величины, и у каждой своё распределение: $\mathbb{E}(X_i) = \mu_i$, $\operatorname{Var}(X_i) = \sigma_i^2$ — и все случайные величины независимы в любом количестве. Введём обозначение: $\bar{\mu}_n = \frac{1}{n}\sum\limits_{i=1}^{n}\mu_i$, $\bar{\sigma}_n^2 = \frac{1}{n}\sum\limits_{i=1}^{n}\sigma_i^2$. Ни одна случайная величина X_n не должна по дисперсии доминировать над другими, т. е. все величины должны быть сопоставимы, бишь $\lim\limits_{n\to\infty}\dfrac{\max\{\sigma_i^2\}}{n\bar{\sigma}_n^2} = 0$. Пусть также верно, что $\lim\limits_{n\to\infty}\bar{\sigma}_n^2 = \bar{\sigma}^2 > 0$. Если у нас есть числовая последовательность из ненулевых членов, то тогда предел не должен быть нулевым, как в случае $\left\{\frac{1}{n}\right\} \to 0$. Итак, дисперсия невырожденная. Тогда $\sqrt{n}(\bar{X}_n - \bar{\mu}_n) \sim \mathcal{N}(0; \bar{\sigma}_n^2)$.

Предельное матожидание и дисперсия — это не предел матожиданий и дисперсий, а матожидание и дисперсия предельного распределения. В многомерном случае тоже верно: пусть есть многомерная случайная величина с

одинаковыми ковариационными матрицами, и тогда предел будет многомерной нормальной случайной величиной. Если $X \sim t_n$, то при $n \to \infty$ $X \sim \mathcal{N}(0;1)$, то есть $t_n \xrightarrow{d} \mathcal{N}(0;1)$. С χ^2 не так всё просто: $\chi^2_n \xrightarrow{d} \mathcal{N}(n;2n)$. Но n убегает в бесконечность, поэтому нормальность будет далеко справа.

Сходимость по вероятности — это сходимость к детерминированной величине, а по распределению — это к случайной величине. К разным объектам они сходятся, посему несопоставимы.

Пусть $X_n \xrightarrow{d} X$, а $\operatorname{plim} Y_n = \mathbb{C}$. Тогда

- $X_n Y_n \xrightarrow{d} \mathbb{C}X$;
- $X_n + Y_n \xrightarrow{d} X + \mathbb{C}$;
- $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{\mathbb{C}}$ $(\mathbb{C} \neq 0)$;
- Теорема Слуцкого. Если g — непрерывная функция, то $g(Y_n) \xrightarrow{\mathbb{P}} g(\mathbb{C})$; $g(X_n) \xrightarrow{d} g(X)$;
- Если $\operatorname{plim}(X_n - Z_n) = 0$ и $X_n \xrightarrow{d} X$, то $Z_n \xrightarrow{d} X$.

Сложные выражения мы отныне станем разбивать на кусочки, которые сходятся либо по распределению, либо по вероятности. Теперь вернёмся к ММП.

Оценки ММП являются асимптотически нормальными. Доказательство этого опирается на ЦПТ, из того же корня растёт. Если любая оценка МП является асимптотически нормальной, то ясно, что в конечных выборках мы будем использовать это предельное распределение. Да, есть ошибка в замене нашей выборки на гипотетическую. Можно сказать вот что о параметрах распределения. Так как оценка является состоятельной, то plim этой оценки равен оценке: $\tilde{\vec{\theta}}_{\text{ML}} \sim \mathcal{N}(\vec{\theta}; \operatorname{Cov}(\ldots))$. Но дисперсию надо оговорить отдельно. Дисперсия связана с матрицей Фишера. В классе всех состоятельных асимптотически нормальных оценок оценка ММП наиболее эффективна (неточный термин). Внимание, нет слова «линейных». Аккуратность требуется к тому, что у нас есть вектор. Поэтому рассмотрим $\operatorname{Cov}(\tilde{\vec{\theta}}_{\text{ML}})$ и другую оценку — $\tilde{\vec{\varkappa}}$ — со своей ковариационной матрицей. $\tilde{\vec{\varkappa}}$ состоятельна и асимптотически нормальна. Тогда разность ковариационных матриц есть симметричная отрицательная полуопределённая матрица. Оценка МНК для нормальной регрессии совпадает с оценкой ММП. Это говорит, что она имеет минимальную дисперсию среди всех состоятельных асимптотически нормальных оценок (слово *линейных* не требуется). Теперь оценка ММП есть ~~BLUE~~ BUE-оценка.

13 Лекция 13

Итак, $\tilde{\vec{\theta}}_{\text{ML}} \overset{\text{as}}{\sim} \mathcal{N}(\vec{\theta}; \boldsymbol{I}^{-1}(\vec{\theta}))$, где $\boldsymbol{I}(\vec{\theta})$ — информационная матрица Фишера. Вспомним о границе Крамера—Рао—Фреше (minimum variance bound). Любая выборка несёт какую-то информацию о генеральной совокупности. Благодаря ей мы делаем ту или иную оценку. Так как выборка не несёт полной информации, мы делаем неточную оценку. Максимальная точность, основанная на выборке, говорит: это та минимальная дисперсия, которая вообще доступна по информации, содержащейся в выборке, каким бы ни было распределение с параметрами. Существование такой границы — это достаточно естественно. Оценка может достигать этой границы, а может и не достигать. Поэтому если она и достижима, то только для ММП. Оно может даже само или не сесть, или сесть на неё асимптотически. В многомерном случае это ковариационная матрица, или информационная матрица Фишера. Два эквивалентных результата выведены Рао на основании Крамера, а Фреше сам когда-то получил похожие результаты. Эта матрица связана с матрицей Гессе log-likelihood function. $I(\theta) = \mathbb{E}\left(-\left\|\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right\|\right)$. Эта матрица есть матрица Гессе для этой задачи. $\left\|\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right\| = \boldsymbol{H}$. Кроме того, эта матрица — случайная величина при условии X. Поэтому нам нужно ещё и матожидание по всем X. Рао доказал:

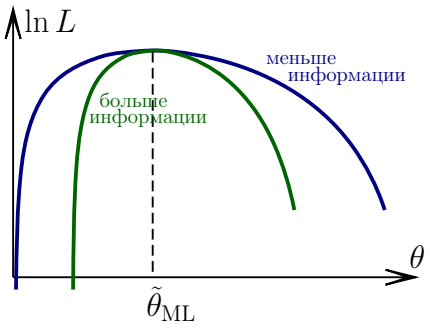

Рис. 10. Качество функции правдоподобия

$$\mathbb{E}\left(-\left\|\frac{\partial^2 \ln L}{\partial\theta_i \partial\theta_j}\right\|\right) = \mathbb{E}\left(\left(\frac{\partial \ln L}{\partial \vec{\theta}}\right)^T \left(\frac{\partial \ln L}{\partial \vec{\theta}}\right)\right) \tag{13.1}$$

Это верно при некоторых условиях регулярности. Однако в большинстве случаев посчитать это матожидание совсем не просто. Появляется два метода оценки неизвестной информационной матрицы в зависимости от того, что проще. Но от методов оценки зависит и качество полученной матрицы. Проблема «хоть как-то оценить» стала несложной. Напомним, что градиент функции правдоподобия — это score function.

Как ММП проверять гипотезы? В регрессии более-менее понятно. Приходится пользоваться приёмами для проверки гипотез о нормально распределённых векторах. Если надо проверить, что $\theta_j = 0$ против $\mathcal{H}_1 : \theta_j \neq 0$, то представим, что каким-то способом мы через ММП мы оценили целый вектор и получили оценку $(\tilde{\theta}_j)_{\text{ML}} \overset{\text{as}}{\sim} \overset{\text{as}}{\sim} \mathcal{N}(\theta_j; \operatorname{Var}(\theta_j))$. Так как это нормальное распределение, то $t = \frac{(\tilde{\theta}_j)_{\text{ML}}}{\text{s.e.}(\tilde{\theta}_j)_{\text{ML}}} \overset{\text{as}}{\sim} \mathcal{N}(0;1)$, так как это t-распределение.

Существует триада асимптотических тестов. Пусть мы оценили $\tilde{\vec{\theta}}_{\text{ML}} \overset{\text{as}}{\sim} \mathcal{N}(\vec{\theta}; \boldsymbol{I}^{-1}(\vec{\theta}))$. Гипотеза у нас может быть сумасшедшим соотношением: $\sin\theta_1 + \operatorname{tg}\theta_2 = 30$ (вот мы сумасшедшие). В общем виде: $\mathcal{H}_0\colon \vec{F}(\vec{\theta}) = 0$, $\mathcal{H}_1\colon \vec{F}(\vec{\theta}) \neq 0$. Если у нас есть две функции, то более правдоподобна более крутая функция правдоподобия (информации больше). $\mathcal{H}_0\colon \theta = \theta_0$; $g(\theta) = 0$; $\mathcal{H}_1\colon g(\theta) \neq 0$. Если справедлива нулевая гипотеза, то тогда θ_0 должна быть близка к $\tilde{\theta}_{\text{ML}}$.

Способ проверки номер раз. Если оценка близка к истинному значению, то тогда значения функции правдоподобия должны быть близки. $\ln L(\tilde{\theta}_{\text{ML}}) - \ln L(\theta_n) = -\ln\left(\frac{L(\theta_n)}{L(\tilde{\theta}_{\text{ML}})}\right) = -\ln\lambda$, и эта оценка есть **Likelihood Ratio Test**, который делает вот что: строится LR, и имеется свойство: $-2\ln\lambda \overset{\text{as}}{\sim} \chi^2_1$, а если $\vec{F} \in \mathbb{R}^l$, то $-2\ln\lambda \overset{\text{as}}{\sim} \chi^2_l$. Чтобы не путались, ещё раз: $\lambda = \frac{L(\theta_0)}{L(\tilde{\theta}_{\text{ML}})}$.

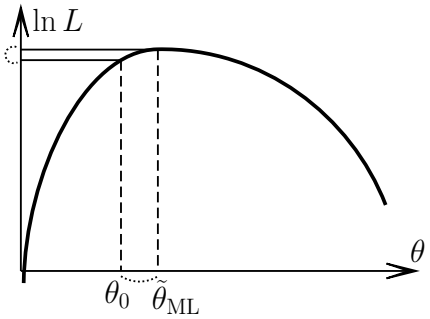

Рис. 11. Likelihood Ratio Test

В доказательстве асимптотического распределения используется то, что $\tilde{\theta}_{\text{ML}}$ асимптотически нормальна. Ошибка: нельзя проверять гипотезу $\theta = 2$ против $\theta = 3$. Мы можем сравнивать только оценку ММП и другую, но не две оценки между собой. Глобальный максимум — это безусловный максимум, то есть unrestricted. Когда гипотеза, то тогда restricted. Если близки между собой, то тогда дробь равна единице, логарифм равен нулю. Оценка держится на том, что мы максимальное значение делим на более мелкое, поэтому $\lambda \leqslant 1$. Критически для нас попасть в правый хвост распределения. Но данный подход не говорит ничего о распределении ошибок. Вторая неприятная ситуация: а что вы так любите вертикальную ось? Можно же по горизонтали мерить. Поэтому ставим вопрос о $g(\tilde{\theta}_{\text{ML}})$, и так появился тест Вальда (Wald-test).

Вальд-тест в общем случае: $\vec{F}(\tilde{\vec{\theta}}_{\text{ML}}) = \vec{0}$? В случае линейности F распределение асимптотически нормально, считается матожидание. А если нелинейная g, то мы раскладываем точку в окрестности θ_{ML} по Тейлору и оставляем только первый член. В эконометрике это получило название Δ-метода, так как $\Delta g \approx g'(\theta_{\text{ML}})\Delta\theta$, как $\Delta y \approx f'(x_0)\Delta x$. В многомерном случае верно: $\Delta\vec{F} = \nabla\vec{F}\cdot\Delta\vec{x}$, или $\Delta\vec{F} = \mathcal{I}(\tilde{\theta}_{\text{ML}})\Delta\vec{\theta}$. Можно даже написать распределение нормального вектора $\vec{F}(\tilde{\theta}_{\text{ML}})$ и формулы для W-теста. Итак, для теста Вальда надо посчитать $g(\tilde{\theta}_{\text{ML}})$. Но оказалось, что не надо считать правдоподобие для θ_0, что есть restricted. Раньше мы оценивали две модели, а теперь делаем только unrestricted model и строим тест. Конечно, этот тест асимптотический. В конечных выборках мы много сказать не можем. В одномерном случае мы проверяем гипотезу, например, $\mathcal{H}_0\colon \theta_1+\theta_2 = 4 \Leftrightarrow$ $\Leftrightarrow \mathcal{H}_0\colon \frac{4}{\theta_1+\theta_2} = 1 \Leftrightarrow \mathcal{H}_0\colon (\theta_1+\theta_2)^2 = 16$. Но у нас могут получиться несколько разные результаты, так как метод не инвариантен по способу задания зависимости.

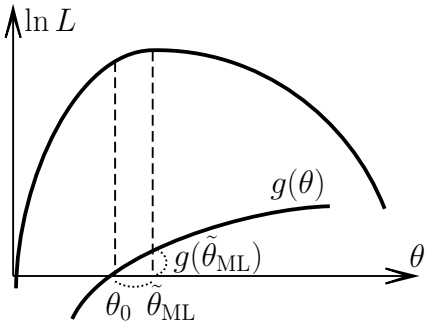

Рис. 12. Wald test

В двух предыдущих тестах мы проверяли близость θ_0 к $\tilde{\theta}_{\text{ML}}$ либо по одной оси, либо по другой. Но есть ещё один способ: если точка графика от θ_0 близка к максимуму, то производная в точке θ_0 по степени отличия от 0 говорит о близости точки к максимуму. Этот тест — это **тест множителей Лагранжа** (score test, Lagrange multipliers test).

Фактически мы делаем $\max L(\vec{\theta})$ subject to $\vec{F}(\vec{\theta}) = \vec{0}$. Мы делаем сначала restricted, затем unrestricted:

$$\mathcal{L} = \ln L(\vec{\theta}) - \vec{\lambda}^T\vec{F}(\vec{\theta}) \tag{13.2}$$

$$\text{FOC: } \frac{\partial\mathcal{L}}{\partial\theta_j} = 0 \Rightarrow \underbrace{\frac{\partial \ln L}{\partial\theta_j}}_{=0} - \underbrace{\vec{\lambda}^T\frac{\partial F(\vec{\theta})}{\partial\theta_j}}_{=0} = 0;\ \frac{\partial\mathcal{L}}{\partial\lambda_l} = 0 \tag{13.3}$$

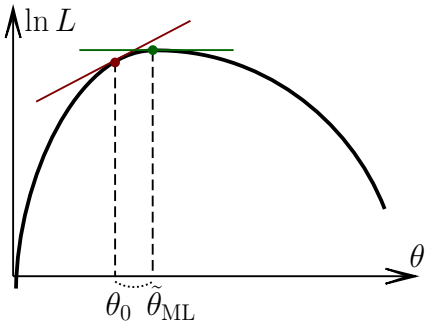

Рис. 13. Score Test

$\mathcal{H}_0\colon \lambda^0 = 0 \Leftrightarrow \mathcal{H}_0\colon \frac{\partial \ln L}{\partial\theta} = 0$. С точки зрения эконометрики замечательная особенность третьего подхода: не надо считать оценку ММП. В restricted model просто смотрится характеристика. Эта функция будет в точке решения. Для применения теста множителей Лагранжа надо подсчитать только restricted. В случае unrestricted у нас разные дисперсии, сложно всё оценивать. Поэтому считается функция правдоподобия в нужной точке, функция правдоподобия раскладывается в ряд Тейлора, и потом оценивается многомерный вектор. Оба теста сводятся к тому, что есть некоторый нормально распределённый нормальный вектор с определённым матожиданием и оценённой ковариационной матрицей. Проверяется, что параметры равны нулю.

Распределение специальной квадратичной формы от нормально распределённого вектора. Пусть у нас есть вектор $\vec{X} \sim \mathcal{N}(0; \boldsymbol{\Sigma})$, где $\boldsymbol{\Sigma} = \operatorname{Cov}(\vec{X}) = \mathbb{E}\left(\vec{X}\vec{X}^t\right)$. Договоримся, что $\vec{X} \in \mathbb{R}^k$. Если мы организуем матрицу квадратичной формы $\vec{X}^T\boldsymbol{\Sigma}^{-1}\vec{X}$, то она будет иметь распределение χ^2_k. Так как написана обратная матрица, то предполагается невырожденность.

Любая ковариационная матрица положительно полуопределённая: $\vec{z}^T\mathbf{\Sigma}\vec{z} \geqslant 0\ \forall \vec{z} \in \mathbb{R}^k$. Все собственные числа положительно полуопределённой матрицы неотрицательны: $\lambda_1; \ldots; \lambda_k \geqslant 0$. Пусть $\boldsymbol{P}$ — матрица из собственных векторов $\boldsymbol{P} = (\vec{P}_1 \ldots \vec{P}_k)$. Далее, собственные векторы, соответствующие различным собственным числам, ортогональны. Если собственные векторы соответствуют одинаковым числам, то тогда ортогонализация проводится по методу Грама—Шмидта. Разные собственные числа — там автоматически ортогональны. Если в матрице $\boldsymbol{P}$ ортогональные векторы суть столбцы, то $\boldsymbol{P}^T\boldsymbol{P} = \begin{pmatrix} 1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 1 \end{pmatrix} = \boldsymbol{I}_k$. Если матрица приводится к диагональному виду, то матрица перехода от базиса к базису — это $\mathbf{\Sigma} = \boldsymbol{P}^{-1}\mathbf{\Lambda}\boldsymbol{P}$. Кстати, $\begin{pmatrix} \vec{P}_1^T \\ \vdots \\ \vec{P}_k^T \end{pmatrix} \underbrace{\left(\mathbf{\Sigma}\right)\left(\vec{P}_1 \quad \cdots \quad \vec{P}_k\right)}_{(\lambda_1\vec{P}_1 \quad \cdots \quad \lambda_k\vec{P}_k)} = \begin{pmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_k \end{pmatrix}$, раз $\mathbf{\Sigma}\vec{P}_i = \lambda_i\vec{P}_i$ (свойство собственных значений), поэтому $\mathbf{\Lambda} = \boldsymbol{P}^T\mathbf{\Sigma}\boldsymbol{P}$, или $\mathbf{\Sigma} = \boldsymbol{P}\mathbf{\Lambda}\boldsymbol{P}^T$. Так, $\mathbf{\Sigma}^{-1} = (\boldsymbol{P}\mathbf{\Lambda}\boldsymbol{P}^T)^{-1} = \boldsymbol{P}\mathbf{\Lambda}^{-1}\boldsymbol{P}^T$, следовательно, у матрицы $\mathbf{\Sigma}^{-1}$ те же самые собственные векторы и обратные собственные числа $\left(\frac{1}{\lambda_i} > 0\right)$.

Теперь начинаем маленькие хитрости. В силу фантазии нам хочется представить $\mathbf{\Sigma}^{-1}$ как произведение одинаковых матриц, что есть $\boldsymbol{P}\mathbf{\Lambda}^{-0,5}\mathbf{\Lambda}^{-0,5}\boldsymbol{P}^T$, где $\mathbf{\Lambda}^{-0,5} = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \frac{1}{\sqrt{\lambda_k}} \end{pmatrix}$, поэтому $\mathbf{\Sigma}^{-1} = (\boldsymbol{P}\mathbf{\Lambda}^{-0,5})(\boldsymbol{P}\mathbf{\Lambda}^{-0,5})^T$. Мы разложили матрицу на произведение двух транспонированных матриц. Это называется разложением квадратичной положительно определённой матрицы. Поэтому $\vec{X}^T\mathbf{\Sigma}^{-1}\vec{X} = \vec{X}^T\underbrace{(\boldsymbol{P}\mathbf{\Lambda}^{-0,5})}_{[(\boldsymbol{P}\mathbf{\Lambda}^{-0,5})^T\vec{X}]^T}(\boldsymbol{P}\mathbf{\Lambda}^{-0,5})^T\vec{X}$. Пусть $\vec{Y} = \mathbf{\Lambda}^{-0,5}\boldsymbol{P}^T\vec{X}$. Тогда $\vec{X}^T\mathbf{\Sigma}^{-1}\vec{X} = \vec{Y}^T\vec{Y}$.

Вернёмся к случайным величинам. $\mathrm{Cov}(\vec{Y}) = \mathbb{E}(\vec{Y}\vec{Y}^T) = \mathbb{E}\left(\lambda^{-\frac{1}{2}}\boldsymbol{P}^T\vec{X}\vec{X}^T\boldsymbol{P}\lambda^{-\frac{1}{2}}\right) = \lambda^{-\frac{1}{2}}\boldsymbol{P}^T\mathbb{E}(\vec{X}\vec{X}^T)\boldsymbol{P}\lambda^{-\frac{1}{2}} = $
$= \lambda^{-\frac{1}{2}}\boldsymbol{P}^T\mathbf{\Sigma}\boldsymbol{P}\lambda^{-\frac{1}{2}} = \lambda^{-\frac{1}{2}}\mathbf{\Lambda}\lambda^{-\frac{1}{2}} = \boldsymbol{I}_k$. $\boldsymbol{Y}$ нормально распределён, нулевое матожидание, дисперсия 1, некоррелированность, поэтому $\sum\limits_{i=1}^{k} Y_i^2 \sim \chi_k^2$. Все сидят на диагонали, имеем сумму нормальных величин в квадрате. Поэтому для любого нормально распределённого вектора с невырожденной ковариационной матрицей квадратичная форма с матрицей, обратной ковариационной матрице, имеет распределение χ_k^2. Оценка ММП асимптотически нормальна, ковариационная матрица асимптотически равна обратной информационной, поэтому $(\tilde{\vec{\theta}}_{\mathrm{ML}} - \vec{\theta}) \overset{\mathrm{as}}{\approx} \mathcal{N}$ и

$$(\tilde{\vec{\theta}}_{\mathrm{ML}} - \vec{\theta})^T[\boldsymbol{I}(\theta)](\tilde{\vec{\theta}}_{\mathrm{ML}} - \vec{\theta}) \sim \chi_k^2 \tag{13.4}$$

14 Лекция 14

$$Y = X\vec{\beta} + \vec{\varepsilon}, \quad \vec{\varepsilon} \sim \mathcal{N}(0; \sigma^2\boldsymbol{I}_n), \quad l = \ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{\sum\limits_{i=1}^{N} e_i^2}{2\sigma^2}, \quad \vec{\beta}, \quad \vec{\sigma}^2, \quad \vec{e} = \vec{Y} - \boldsymbol{X}\vec{\beta} \tag{14.1}$$

Уравнения правдоподобия: $\frac{\partial l}{\partial\vec{\beta}} = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\vec{\beta}}\sum e_i^2(\vec{\beta}) = 0$; $\frac{\partial l}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum(\ldots)}{2\sigma^4} = 0$.

Мы должны получить такие же оценки МП, как и для НК: $\tilde{\vec{\beta}}_{\mathrm{ML}} = \hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. $\tilde{\sigma}^2_{\mathrm{ML}} = \frac{\mathrm{RSS}}{n}$. Мы должны просчитать:

$$\max_{\vec{\beta},\sigma^2}\ln L = \mathbb{C} - \frac{n}{2}\ln\tilde{\sigma}^2 - \frac{1}{2\tilde{\sigma}^2}\mathrm{RSS} = \mathbb{C} - \frac{n}{2}\ln\left(\frac{\mathrm{RSS}}{n}\right) - \frac{1\cdot n\cdot\mathrm{RSS}}{2\mathrm{RSS}} = \mathbb{C} - \frac{n}{2} =$$
$$= \left(\mathbb{C} - \frac{n}{2} + \frac{n}{2}\ln n\right) - \frac{n}{2}\ln\mathrm{RSS} = \mathrm{const} - \frac{n}{2}\ln\mathrm{RSS} \tag{14.2}$$

Это даёт нам то, что $\max\limits_{\vec{\beta},\sigma^2} L = \mathrm{const}(\mathrm{RSS})^{-\frac{n}{2}}$. Рассмотрим эту матрицу. Матрица Фишера зависит от оценок. Это $\boldsymbol{I}(\vec{\beta},\sigma^2) = \mathbb{E}\left(-\frac{\partial^2\ln L}{\partial\vec{\theta}\partial\vec{\theta}^T}\right)$.

Напомним: градиенты функции правдоподобия — это score function. $\frac{\partial^2\ln L}{\partial\vec{\beta}\partial\vec{\beta}^T}$ — это просто гессиан для нашей системы, и он будет следующим: $\frac{\partial^2 l}{\partial\vec{\beta}\partial\vec{\beta}^T} = \frac{\partial}{\partial\vec{\beta}}\left(\frac{\partial l}{\partial\vec{\beta}^T}\right) = \frac{\partial}{\partial\vec{\beta}}S(\vec{\beta})$, где $S(\vec{\beta})$ — score function. Далее, это $-\frac{1}{2\sigma^2}\frac{\partial}{\partial\vec{\beta}}\underbrace{\sum e_i^2(\vec{\beta})}_{-2\sum e_iX_i}$, а выражение $\left\|\frac{\partial^2 l}{\partial\beta_j\partial\beta_j}\right\| = -\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$ не зависит от Y, ε и β. Она совпадает с гессианом для

матрицы МНК, только появился множитель $-\frac{1}{2\sigma^2}$. Матрица сменит знакоопределённость. $\frac{\partial^2 l}{\partial(\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{\mathrm{RSS}}{\sigma^6}$. Кстати, $\frac{\mathrm{RSS}}{\sigma^2} = n$, поэтому последнее выражение равно $\frac{n}{2\sigma^4} - \frac{\mathrm{RSS}}{\sigma^6} \overset{\tilde\sigma^2}{\Rightarrow} \frac{n}{2\sigma^2} - \frac{n}{\sigma^2} == -\frac{n}{2\sigma^4}$. Поэтому $\frac{\partial^2 l}{\partial\vec\beta\partial\sigma^2} = \frac{1}{2\sigma^4}\frac{\partial}{\partial\vec\beta}(\mathrm{RSS}) = \frac{1}{2\sigma^4}(-2\sum e_i X_i)$. Вот такое нехорошее выражение появилось.

Вернёмся к нашей информационной матрице. То, что мы получились, берём со знаком минус и матожидаем. Мы всё писали в предположении детерминированных $\boldsymbol{X}$, поэтому матрица $-\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}$ не есть случайна. Смотрим матрицу:

$$\boldsymbol{I} = \begin{pmatrix} \sigma^2\boldsymbol{X}^T\boldsymbol{X} & 0 \\ 0 & \frac{n}{\sigma^4} \end{pmatrix} \tag{14.3}$$

Матожидание e_i равно нулю. Далее, $\binom{\tilde{\vec\beta}_{\mathrm{ML}}}{\tilde\sigma^2_{\mathrm{ML}}} \overset{\mathrm{as}}{\sim} \mathcal{N}\left(\binom{\vec\beta}{\sigma^2}; \boldsymbol{I}^{-1}(\tilde{\vec\beta}_{\mathrm{ML}}, \tilde\sigma^2_{\mathrm{ML}})\right)$. А $\mathrm{Cov}(\ldots) = \begin{pmatrix} \frac{1}{\sigma^2}(\boldsymbol{X}^T\boldsymbol{X})^{-1} & 0 \\ 0 & \frac{\sigma^4}{\sigma^n} \end{pmatrix}$ (внимание, ковариационная матрица в целом положительно определённая, а она есть матожидание гессиана); гессиан отрицательно знакоопределён, поэтому найденная точка — максимум. Рассмотрим субботние асимптотические тесты.

На этом моменте надо перечитать материал о триаде асимптотических тестов (хотя ниже он будет пересказан).

Likelihood Ratio Test. Давайте проверим $\mathcal{H}_0\colon \boldsymbol{R}\vec\beta = \vec q$ при $\mathcal{H}_1\colon \boldsymbol{R}\vec\beta \neq \vec q$, $\beta_i = 0$, ограничений l штук, ограничений меньше, чем параметров, поэтому $\operatorname{rank}\boldsymbol{R} = p$. $\vec q(\vec\beta) = 0$, $g(\beta) = 0$. Первый классический асимптотический тест говорит, что $\mathcal{H}_0\colon \theta = \theta_0$, $\mathcal{H}_1\colon \theta \neq \theta_0$. Разность логарифмов — это логарифм частного. Тест Likelihood Ratio придуман Нейманом и выглядит так: $0 \leqslant \lambda = \frac{L(\vec\theta_0)}{L(\vec\theta_{\mathrm{ML}})} \leqslant 1$, и $-2\ln\lambda \overset{\mathrm{as}}{\sim} \chi^2_p$, $L(\tilde\theta_{\mathrm{ML}}) = \mathrm{const}(\mathrm{RSS})^{-\frac{n}{2}}$. Поэтому наш максимум $l(\tilde\theta_{\mathrm{ML}})$ — это unrestricted, а $t(\vec\theta_0)$ — это restricted при условии нулевой гипотезы. Поэтому $\frac{L(\vec\theta_0)\to\text{restricted}}{L(\vec\theta_{\mathrm{ML}})\to\text{unrestricted}}$. $L(\theta_0) = \mathrm{const}(\mathrm{RSS}_{\mathrm{restr}})^{-\frac{n}{2}}$. Получается остаточная сумма. Она больше, чем введено ограничений. Поэтому константы сокращаются, и получается:

$$\lambda = \left(\frac{\mathrm{RSS}_{\mathrm{restr}}}{\mathrm{RSS}_{\mathrm{unrestr}}}\right)^{-\frac{n}{2}}; \quad -2\ln\lambda = n(\ln\mathrm{RSS}_{\mathrm{restr}} - \ln\mathrm{RSS}_{\mathrm{unrestr}}) \overset{\mathrm{as}}{\sim} \chi^2_p \tag{14.4}$$

Для проверки гипотезы нам надо построить две модели — с ограничениями и без — посчитать логарифмы.

Пусть мы хотим проверить гипотезу об адекватности регрессии. Надо представить: $\mathcal{H}_0\colon \beta_2 = \ldots = \beta_k = 0$, $\mathcal{H}_1\colon$ не так. В случае верности нулевой гипотезы имеем: $Y_i = \beta_1 + \varepsilon_i$, $\hat\beta_1 = \bar Y$. Тогда $\mathrm{RSS}_{\mathrm{restr}} = \sum(Y_i - \bar Y)^2 = \mathrm{TSS}$, $\mathrm{RSS}_{\mathrm{unrestr}} = \mathrm{RSS}$. Тогда отношение правдоподобий даёт:

$$-2\ln\lambda = n\ln\frac{\mathrm{TSS}}{\mathrm{RSS}} = n\ln\frac{\mathrm{TSS}}{\mathrm{TSS}(1-R^2)} = n\ln\frac{1}{1-R^2} = -n\ln(1-R^2) \tag{14.5}$$

При этом мы сравнивали значения функции правдоподобия в максимуме без ограничений и с ограничениями. Близость между ограничениями проверяется по значению функции.

Wald Test. Способ проверять по значению аргумента сама напрашивается, но только в 1943 году Абрахам Вальд сообразил, что можно проверять $\mathcal{H}_0\colon g(\tilde\theta_{\mathrm{ML}}) = 0$. Идейно лучше этот способ тем, что нам не надо знать θ_0. Зная ограничение, мы посчитаем ограничение на функцию МП, поэтому считается только unrestricted model. Для большей понятности рассмотрим случай с одной переменною.

Если есть модель $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, то $\mathcal{H}_0\colon \beta_2 = 0$, $\mathcal{H}_1\colon \beta_2 \neq 0$. Тогда $\frac{\hat\beta_2}{\mathrm{s.e.}(\hat\beta_2)} \sim t_{n-2} \Leftrightarrow \frac{\overbrace{\hat\beta_2^2}^{\tilde\beta_2^{\mathrm{ML}}}}{\text{оценка }\mathrm{Var}(\hat\beta_2)} \sim \chi^2_1$.

Далее, $\tilde\sigma^2_{\mathrm{ML}} = \frac{\mathrm{TSS}(1-R^2)}{n}$, $\hat\beta_2 = \frac{\sum x_i y_i}{\sum x_i^2}$, $\mathrm{Var}(\hat\beta_2) = \frac{\overbrace{\sigma^2}^{\to\tilde\sigma^2_{\mathrm{ML}}}}{\sum x_i}$. Мы получили выражение для теста Вальда:

$$\frac{\left(\sum x_i y_i\right)^2 n\left(\sum x_i^2\right)}{\left(\sum x_i^2\right)^2 \cdot \mathrm{TSS}(1-R^2)} = \frac{nR^2}{1-R^2} \overset{\mathrm{as}}{\sim} \chi^2_1 \tag{14.6}$$

Lagrange multipliers test. В 1948 году, после Второй мировой, тот самый Рао придумал ещё один тест близости оценок. Он состоит в подсчёте угла наклона градиента в ограниченной модели и сравнении оного с нулём. Это Rao score test, но в 1959 некто Силвей назвал его методом множителей Лагранжа, так как нулевое значение производной эквивалентно тому, что множитель Лагранжа равен нулю. Lagrange multiplier-статистика имеет очень простой вид: $LM = nR^2$.

Три этих теста имеют очень важные последствия. Любая из этих трёх тестовых статистик имеет распределение χ^2_p. В больших выборках эти тесты эквивалентны по своей статистической силе. Но для теста отношения правдоподобий нужно две модели, для Вальда — unrestricted, для Лагранжа — restricted. Конечно, restricted проще, там много множителей равно нулю. А в конечных выборках дать ответ не так просто.

Мы получили, что тестовые статистики следующие: LR: $-n\ln(1-R^2)$, W: $\frac{nR^2}{1-R^2}$, LM: nR^2. Упорядочим эти выражения: $W \geqslant LR \geqslant LM$. Вспомним матан: $x \geqslant \ln(1+x) \geqslant \frac{x}{1+x}$. Всё справедливо при малых x, а второе и третье выражения имеют радиус сходимости, равный единице. Пусть же ныне $x = \frac{W}{n}$. Тогда $\frac{W}{n} \geqslant \ln\left(1+\frac{W}{n}\right) \geqslant \frac{\frac{W}{n}}{1+\frac{W}{n}}$, откуда и получено исходное неравенство. Это соотношение выполняется всегда. С точки зрения проверки гипотезы это означает, что если значение очень большое, то мы отвергаем нулевую гипотезу. Если отвергнем Вальда, то она и по LM отвергнута. Если по LM не отвергнута, то тогда все три не отвергнуты. Если значения разные, то это нам простор для фантазии — понаходиться на разных границах.

Ныне оставим случай линейной регрессии и вернёмся к исходной ситуации. Мы договорились, что ММП состоит в том, что мы максимизируем функцию правдоподобия или её логарифм. Тогда уравнение правдоподобия имеет следующий вид: $\frac{\partial \ln L}{\partial \vec{\theta}} = 0 \Leftrightarrow \vec{S}(\vec{\theta}) = 0$. Данные уравнения суть нелинейные, и без компьютеров это было только для простых случаев. Но есть несколько трудностей. Во-первых, надо считать численно. Во-вторых, данная система нелинейных уравнений может иметь не одно решение. Поэтому может быть найден локальный максимум. Глобальный максимум ищут процедурой Ньютона—Рафсона (процедура градиентного спуска). Пусть $\vec{\theta}_0$ — начальное значение. Мы получаем не ноль: $\frac{\partial l}{\partial \vec{\theta}_0} \neq 0$. Но у нас есть ряд Тейлора, поэтому

$$\vec{S}(\vec{\theta}) = \vec{S}(\vec{\theta}_0) + \underbrace{\mathcal{I}}_{\frac{\partial^2 l}{\partial \vec{\theta}_0 \partial \vec{\theta}_0^T}} \times (\vec{\theta}_1 - \vec{\theta}_2) \rightsquigarrow \vec{\theta}_1 = \vec{\theta}_0 = -\left(\frac{\partial^2 L}{\partial \tilde{\theta}_0 \partial \vec{\theta}_0}\right)^{-1} S(\vec{\theta}_0) \tag{14.7}$$

Это процедура Ньютона—Рафсона. Если мы далеко, то можем не туда сойтись, может плохо приближать линейное представление ряда Тейлора, но мы на каждом шаге просчитываем матрицу вторых производных.

Альтернативный подход — это score method. Он считает матрицу один раз, то есть берёт $\boldsymbol{I}^{-1}(\theta)$. Там стоит минус матожидание, можно посчитать один раз, и будет так. Но не всегда мы можем посчитать матожидание.

15 Лекция 15

Ранее мы разобрали первое неприличное ругательство — гетероскедастичность, а ныне рассмотрим другое — **мультиколлинеарность**. Итак, продолжим ослаблять условия теоремы Гаусса—Маркова. Пусть $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$, модель правильно специфицирована, $\boldsymbol{X}$ детерминирован (от этого мы отказывались), $\operatorname{rank}\boldsymbol{X} = k$, $\mathbb{E}(\vec{\varepsilon}) = 0$, $\operatorname{Cov}(\vec{\varepsilon}) = \sigma^2\boldsymbol{I}_n$. Мы рассмотрели в таких случаях, что нужно сделать, и записали МИП и ММП. Надо только сделать предположения о $\boldsymbol{X}$ и $\vec{\varepsilon}$.

Сегодня мы поничтожим свойство $\operatorname{rank}\boldsymbol{X} = k$. Он будет меньше, но не настолько, чтобы по одному наблюдению оценить три параметра, то есть верно, что $n > k$. $\hat{\vec{\beta}} = \underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}}_{k\times k}\boldsymbol{X}^T Y$. Линейная зависимость столбцов означает компланарность, копространственность — множественную параллельность, или мультиколлинеарность. Если $\operatorname{rank}\boldsymbol{X} < k$ — это perfect, абсолютная, теоретическая мультиколлинеарность. Если это так, то $(\boldsymbol{X}^T\boldsymbol{X})$ необратима, и МНК неприменим.

Рассмотрим обычную регрессию: $Y_i = \beta_1 + \beta_2 X_2^i + \beta_3 X_e^i + \varepsilon_i$. Предположим, что $X_2^i = 3X_3^i$, и тогда верно: $Y_i = \beta_1 + (3\beta_2 + \beta_3)X_3^i + \varepsilon_i$. Тогда оценивать будем коэффициенты $\hat{\beta}_1$ и $\widehat{(3\beta_2 + \beta_3)}$. Оценить у нас получится только совместное влияние второго и третьего коэффициентов. Они показывают изменение величины при прочих равных, но имеющееся линейное соотношение не даёт возможности оценки индивидуального вклада переменных. Но суммарный вклад легко оценится. Если прогнозировать, то нормально, а ставить вопрос об отдельном влиянии X_2 и X_3 нельзя. Получается будто бы dummy trap. Вторая опасность — вводить переменные, связанные экономическим тождеством: потребление, ВВП, инвестиции; прибыль, оборот, издержки. Если есть perfect collinearity, то blind alley — тупик из-за существующих связей. Надо переписать модель, что-то исключив.

При наличии мультиколлинеарности $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0$, а ВВП и инфляция известны нам неточно и с ошибкой. Можно ожидать, что что-то нехорошее будет тогда, когда определитель близок к нулю. Imperfect multicollinearity.

Рассмотрим пример с двумя переменными: $Y_i = \beta_1 X_1^i + \beta_2 X_2^i + \varepsilon_i$. Напишем центрированные переменные: $y_i = \beta_1 x_1^i + \beta_2 x_2^i + \varepsilon_i$. Запишем систему нормальных уравнений I порядка: $\begin{cases} \beta_1 \sum\limits_{i=1}^{n} x_1^2 + \beta_2 \sum x_1 x_2 = \sum x_1 y; \\ \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 = \sum x_2 y. \end{cases}$

После подстановки данных по 38 точкам получили: $200\beta_1 + 150\beta_2 = 350$, $150\beta_1 + 113\beta_2 = 263$, откуда $\hat{\beta}_1 = \hat{\beta}_2 = 1$, $\det = 200 \cdot 113 - 150 \cdot 150 = 100$. Теперь выкинули 2 наблюдения, стало их 26. n нигде не присутствует, поэтому просто как-то изменятся коэффициенты: $199\beta_1 + 149\beta_2 = 347{,}5$, $149\beta_1 + 112\beta_2 = 261{,}5$, $\hat{\beta}_1 = -0{,}5$, $\hat{\beta}_2 = 3$. Разительно неприятная ситуация! Опа! Если бы мы приняли верным первое решение, то $mpc = 1$, а во втором случае зависимость другого знака, а другой коэффициент вырос в несколько раз. Решение неустойчиво по отношению к нашим данным. Чуть-чуть поменяли данные — определитель, будучи около ста, ушёл. Но мы не сказали, в каких единицах.

Повторите такой простой эксперимент. Возьмите данные, например, курс евро к доллару в зависимости от цены на нефть или золото. Получили 3 столбика, а потом говорим: а давайте учтём ещё одну переменную: $2P_1 + 3P_2 + \varepsilon$. Это линейная комбинация, но почти. При этом коэффициенты начнут прыгать. Это и есть

квазимультиколлинеарность. Сегодня такой эксперимент очень легко провести. Проще всего пошатать данные. Можно искусственно. Можно убрать 2–3 наблюдения и проверить устойчивость робастно.

Выпустим кошку из мешка: r^2 между переменными равен 0,95, переменные были почти линейными. Квазимультиколлинеарность (неточная коллинеарность, практическая мультиколлинеарность). Коварство природы вот в чём: если теоретическая мультиколлинеарность, то это сигнал, что надо что-то посмотреть. Некоторые программы делают псевдообращение по Пенроузу, и для вырожденной матрицы считается обобщённый определитель. Программа же ругается, что на ноль деление. А с квазимультиколлинеарностью мы не можем полагаться на оценки, даже если с определителем всё хорошо. И коэффициенты, и оценки прыгают, и t-отношения, и F-отношения прыгают. При наличии подозрения на квазимультиколлинеарность (для обозначения этого длинного и неудобного слова я даже ввёл LaTeX-команду `\qmk`) необходимо проверить несколько признаков. Но даже они ненадёжны, потому что квазимультиколлинеарность — это почти линейная зависимость, и не будет статистического теста на оную. Нельзя так проверить, как с гетероскедастичностью. Мультиколлинеарность есть всегда, но в разной мере. И лишь в редкой ситуации столбцы матрицы $\boldsymbol{X}$ ортогональны.

Итак, каковы же признаки мультиколлинеарности?

1. Высокая значимость модели (высокий R^2) и очень много незначимых по t коэффициентов. Пример: $R^2 = 0{,}99$ по четырём переменным, из которых три незначимы. Из этого простого признака можно сделать более надёжный сигнал: если $R^2 \approx 1$ и все коэффициенты значимые, то мультиколлинеарности нет. Это, конечно, не стопроцентный признак, так как мы набрали плохих переменных в неадекватную модель. Но обратное даёт одно из подозрений на мультиколлинеарность.
2. Большие по величине элементы корреляционной матрицы r_{ij}^2;
3. Высокие значения VIF (далее) в методе вспомогательных регрессий.

Что делать? Известно что: так же, как и в теории, надо выбросить одну, как раньше — дамми, переменную.

Появление мультиколлинеарности обязано не данным, а генеральной совокупности. Если ещё раз взять данные, например, для издержек, прибыли, оборота, то соотношение не изменится, а выборка будет другой. Проблема экономиста — это работа с тем, что есть. Пусть мы моделируем стоимость квадратного метра в Москве, а цена сделки нам редко известна. Хотим построить зависимость цены м^2 от чего-то. Пусть мы уже выбрали квартиры одинакового ранга и качества. Первая переменная — это площадь. Вторая переменная — располагаемый доход и спрос на жильё. Теоретически может быть всё: и хороший доход с маленькой хибаркой, и бабушка на семи комнатах. Практически мы почти никогда с этим не сталкиваемся: данные устроены так, что люди с бóльшим доходом будут иметь больше площади, и у нас получается нормальное размазанное облако. Поэтому это и означает квазимультиколлинеарность. Это проблема выборки, которую мы имеем (at hand).

Рассмотрим чудной способ поправить данные. При наличии довольно однонаправленного облака нам нужно найти данные, которых пока нет в наших пустых областях. Если бы мы нашли бабульку с низким доходом и большой площадью, то мы бы нарушили мультиколлинеарность. Второй тип разрушителя мультиколлинеарности — это Мавроди, который богат, но живёт непонятно где. Поиск данных, удалённых от зависимости (на рис. 14), может помочь решить проблему. Вывод: квазимультиколлинеарность связана с недостаточной информативностью данных и неинформативностью выборки. Наверно, квазимультиколлинеарность связана с коррелированностью данных, что есть почти линейная зависимость. Парная корреляция — это не очень хорошая ситуация. π, TC, TR сильно коррелированы, поэтому в давние времена придумали считать корреляции r_{ij} по всем i и j. Далее матрица всевозможных корреляций $\{r_{ij}\}$ входила в расчёт.

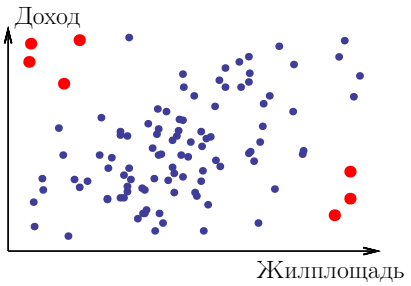

Рис. 14. Новые данные

Рассмотрим общую модель: $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$. Если записать СЛУ и решить её, то можно решить в общем виде. Запишем оценку коэффициента β_2, но не всего, а дисперсии: $\beta_2 = \frac{\sigma^2}{\sum x_2^2 (1 - r_{23}^2)}$, и это один из элементов матрицы $\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}$. После представления коэффициента $\widehat{\mathrm{Var}(\hat{\beta}_2)} = \frac{\hat{\sigma}^2}{\sum x_2^2} \frac{1}{1 - r_{23}^2}$ получим, что изначально было так: $\widehat{\mathrm{Var}(\hat{\beta}_2')} = \frac{\hat{\sigma}^2}{\sum x_2^2}$. Полученный коэффициент назвали $\frac{1}{1 - r_{23}^2} = VIF$, то есть коэффициент вздутия инфляции, variance inflation factor. Поэтому второй способ судить о мультиколлинеарности — это коэффициент VIF. Есть *эвристическое правило* (= с потолка). Если $R^2 < 0{,}8$, то мультиколлинеарности, скорее всего, нет. Если $R^2 > 0{,}8$, то, может быть, есть. Мультиколлинеарность вздувает дисперсию незначимых коэффициентов, а высокая дисперсия зависит не только от этого; ещё от оценки $\hat{\sigma}^2$ и меры разброса величины X. Если он сильно разбросан, он сильно гасит плохие качества. Если хорошая модель, то тогда сами остатки маленькие, поэтому даже высокий VIF тогда не очень напакостит.

Рассмотрим многомерный случай: пусть модель равна $Y = \sum\limits_{i=1}^{k} \beta_i X_i + \varepsilon$. Тогда $\mathrm{Var}(\hat{\beta}_i) = \frac{\sigma^2}{\sum x_i^2} \frac{1}{1 - R_i^2}$, где R^2 берётся из регрессии i-го X на все остальные. Строится вспомогательная регрессия: $X_i = \alpha_1 X_1 + \ldots + \alpha_{i-1} X_{i-1} + \\ + \alpha_{i+1} X_{i+1} + \ldots + \alpha_k X_k + \varepsilon \to R_i^2$. Этот способ лучше, чем попарная корреляция, так как мы смотрим попарную мультиколлинеарность. Посчитать VIF просто самому, а в старых программах это называется auxiliary regressions. Такое эвристическое правило: R^2 и VIF не должны превышать 0,8.

Понятие мультиколлинеарности ввёл Рагнар Фриш. Есть чисто математической способ описать ситуацию: не удаётся точно найти обратную матрицу. Есть причина, по которой вычисления становятся неустойчивой. Задачу обращения назвали некорректной задачей, матрицу — сингулярной, а затем появился класс методов укрощения задач с сингулярными матрицами. Метод для Bad Conditioned Matrix — метод борьбы с плохо обусловленными матрицами. Совсем плохо обусловленная матрица — это $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0$, но это даже не так, потому что $\det(\varepsilon\boldsymbol{I}) = \det\begin{pmatrix}\varepsilon & & \\ & \ddots & \\ & & \varepsilon\end{pmatrix} = \varepsilon^n$, а обратная равна $\frac{1}{\varepsilon^n}\boldsymbol{I}$. Проводятся преобразования Жордана. Если какие-то данные являются машинным нулём, то возникают проблемы. Вот в этом и кроется проблема: числа сильно рознятся по порядкам. Матрица $\boldsymbol{X}^T\boldsymbol{X}$ симметрична, и она приводится нами в вид диагональный, где на диагоналях стоят собственные числа $\lambda_i > 0$, поскольку матрица положительно определена. Но при приведении к диагональному виду можно получить большие и очень маленькие λ_i, поэтому могут появляться нулевые столбцы. Есть такие меры: $BCI = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$, или даже иногда $\frac{\lambda_{\max}}{\lambda_{\min}}$. Некоторые программы имеют встроенную рутину подсчёта собственных чисел. Так и получаем Bad Condition Index. Если $BCI > 30$ (то, что с корнем), то матрица плохо обусловлена. $\frac{\lambda_{\max}}{\lambda_{\min}} > 1\,000$ — это уже надо понимать, что здесь плохо, это точно плохая степень.

Но не так плоха мультиколлинеарность. Может, у нас R^2 высокий, то бог с ней, с мультиколлинеарностью. Просто показатели плохонькие. Наличие связи не так мешает прогнозировать. Нельзя выделить каждую переменную, но прогнозы точные, поэтому на финансовых рынках квазимультиколлинеарность — меньшее зло.

Сейчас ножки Буша уже неактуальны, поэтому рассмотрим спрос на куриное мясо. $Y = \beta_0 + \beta_1 X_1 + {} + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$. Y — спрос на куриное мясо per capita в фунтах (нестоимостная характеристика). X_2 — очищенная от инфляции цена куря за фунт. X_1 — располагаемый доход на душу населения в реальных терминах (в долларах на фунт). X_3 — цена свинины, X_4 — цена говядины. Данные строились за 1960–82 гг. В 1960 году фунт курятины стоил 42 цента, свинина — около 50, говядина — около 72. В 1982 г. фунт курицы стоил 0,7 \$, свинины — 1,68 \$, говядины — 2,32 \$. Модель: $\widehat{\ln Y} = 2{,}19 + 0{,}34\ln X_1 - 0{,}50X_2 + 0{,}15\ln X_3 + 0{,}09\ln X_4$. $t_0 = 14{,}06$, $t_1 = 4{,}1$, $t_2 = -4{,}6$, $t_3 = 1{,}5$, $t_4 = 0{,}9$. Так последние коэффициенты незначимые! $R^2 = 0{,}982\,3$, $\bar{R}^2 = 0{,}978\,4$. Дополнительные регрессии: $R_1^2 = 0{,}984\,6$, $R_2^2 = 0{,}942\,8$, $R_3^2 = 0{,}975\,9$, $R_4^2 = 0{,}976\,4$. Нельзя выбросить подоходный налог, поэтому выбросить то, что сильнее всего связано, — это лекарство горше болезни.

16 Лекция 16

В субботу будет Рождество.

VIF — индекс обусловленности матрицы. Сегодняшняя техника позволяет проверить модель малым изменением данных. Если данные прыгают, то доверия к модели нет. Итак, если есть мультиколлинеарность, то надо бороться с ней. И хорошего лекарства против неё нет.

Метод первый — «устранение переменных». Пример с куриным мясом. Мы разрываем квазилинейную связь, выбрасывая переменную. Судя по R^2, надо было выбрасывать располагаемый доход, но это ущербно: если должен быть располагаемый доход, то не выполняется первое условие теоремы Гаусса—Маркова. Поэтому получается тришкин кафтан. Надо по возможности не выбрасывать содержательных переменных.

Метод второй. Если данные плохие, то надо оные пополнить. Кажется, что это казуистика, но часть данных — это госстатистика, а часть собирается Росстатом и социологами. Пополнение данные требует затрат, а если нет охвата всего разброса данных, то нам известно, что искать. Рассмотрим пример платы за электричество. Это зависит от площади и от насыщенности техникой. У обеспеченности техники есть прокси-переменная — располагаемый доход. Почти несомненно, что данные будут иметь положительный наклон, потому что недвижимость — часто накопленного богатства. А если мы добудем три дополнительные точки там, где нет облака (рис. 14), то квазимультиколлинеарность резко уходит: $\mathrm{Var}(\hat{\beta}_i) = \frac{\hat{\sigma}^2}{\sum x_i^2(1-R_i^2)}$. Каждое наблюдение сумму увеличивает. Но дополнительные наблюдения в стороне, уменьшающие R_i^2, — это метод добычи дополнительной информации.

Пусть у нас построена следующая зависимость: $\widehat{\ln Y} = 24{,}3 + 0{,}87X_1 - 0{,}03X_2$, $R^2 = 0{,}968\,2$, $t_0 = 3{,}85$, $t_1 = 2{,}77$, $t_2 = -1{,}16$. X_2 — накопленное богатство, X_1 — располагаемый доход, Y — потребление. А, ещё интересно: $n = 10$. Как-то странно, что зависимость потребления отрицательная от накопленного богатства. X_2 кажется незначимым, но налицо сильная мультиколлинеарность. Мы добыли ещё 30 точек и получили уравнение: $\widehat{\ln Y} = 2{,}09 + 0{,}73X_1 + 0{,}06X_2$, $R^2 = 0{,}969\,7$, $t_0 = 0{,}87$, $t_1 = 6{,}00$, $t_2 = 2{,}06$. Оба коэффициента значимы, R^2 высокий, опасности мультиколлинеарности нет, поэтому модель хороша. И даже R^2 повысился! Действительно, причиной была мультиколлинеарность. Правда, свободный член незначим, но без необходимости его не стоит убирать.

Метод третий — переосмысление модели. Переспецификация или ещё какое-то переосмысление. Мы рассматривали логлинейное куриное мясо, но почему оно не линейное? Если мы тот же пример посчитаем без логарифмов и построим линейную модель, то $\hat{Y} = 37{,}23 + 0{,}05X_1 - 0{,}61X_2 + 0{,}20X_3 + 0{,}07X_4$. X_1 — располагаемый доход ($t = 10{,}00$), X^2 — цена куриного мяса ($t = -3{,}75$), X_3 ($t = 3{,}11$) X_4 ($t = 1{,}36$). Стала незначимой цена говядины. $R^2 = 0{,}942\,6$, $\bar{R}^2 = 0{,}929\,8$. Переход к логарифмам может повлиять на статистическую зависимость. Кроме теста Бокса—Кокса и преобразований Зарембки, ничего сделать нельзя.

Метод четвёртый — учёт априорной информации о модели. Пусть мы хотим оценить простую функцию Кобба—Дугласа, это функция с постоянной эластичностью, то есть $\ln Q = A + \alpha \ln K + \beta \ln L + \varepsilon$. Если это государство, то данные страдают от инфляции, поэтому K и L растут. Пусть мы избавились от инфляции, но в корреляции труда и капитала нам даже сомневаться не надо. Но, например, из предыдущих исследований нам известно, что в экономике была постоянная отдача от масштаба. Это означает, что $\alpha + \beta = 1$. Тогда мы инкорпорируем эту априорность — let it be! — и запишем модель так: $A + \alpha \ln K + (1 - \alpha) \ln L + \varepsilon$. Поэтому $\ln \frac{Q}{L} = A + \alpha \left(\ln \frac{K}{L}\right) + \varepsilon$. Объясняющая переменная — трудовооружённость выпуска, объяснение — отношение капитала к одному работнику. Априорная информация — связь таким отношением. Поэтому определяется только часть коэффициентов, а остальные заменяются известными соотношениями и этими коэффициентами.

Тобин много чего наделал, и он предложил следующую ситуацию: pooling the data. Рассмотрим регрессию $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$ от цены и располагаемого дохода. Данные были не по стране, а по отдельным штатам. 48 наблюдений за средней ценой покупки. Но цена, уплаченная за автомобиль, и доход связаны. Данные устроены так: в каждый момент времени есть несколько рядов cross-section. Тобин предложил вот что: бóльшая динамика идёт от года к году, а внутри одного года средние цены меняются мало. Поэтому он взял отдельные cross-sections, цена там почти не меняется, поэтому он построил отдельную регрессию: $Y = \alpha_0 + \alpha_1 X_3 + w$. Он взял данные только одного cross-section, так как данные медленнее меняются. Потом α_1 поставил в уравнение и теперь использовал все данные, разорвав мультиколлинеарность. Итак, две стадии: сначала оценивается один коэффициент, а потом подставляется и помогает вычислить остальные. И не всегда это работает.

Метод пятый — преобразования переменных. Его трудно отличить от переспецификации, но некоторая разница есть. Экономисты построили функцию зависимости спроса на импорт от ВВП и потребительских цен. $Y = -108{,}2 + 0{,}045 X_1 + 0{,}931 X_2$, $t_1 = 1{,}23$, $t_2 = 1{,}844$, $R^2 = 0{,}989\,4$. Ни один содержательный фактор не значим, R^2 более чем хорош, налицо мультиколлинеарность. ВВП и CPI, и построение в номинальных переменных страдает от инфляции. Поэтому коэффициент 0,045 не говорит, каких рублей произойдёт изменение в Y — начала или конца периода — и поэтому надо убрать мультиколлинеарность, вызванную инфляцией. Сложность вот в чём: есть у нас CPI, а как дефлировать ВВП? Дефлятор ВВП очень мало общего имеет с CPI. Когда это 2 %, то нормально. Сегодня более половины ВВП — это услуги, а CPI маленький. Дефлятор по ВВП рассчитать трудно из-за громадного серого сектора (досчёт). И есть импорт. Его-то как дефлировать? Надо дефлировать своим дефлятором. Но переход к $\frac{Y}{X_2}$ и $\frac{X_1}{X_2}$ — это не дефлирование, а прокси для реальных переменных, поэтому $Y^* = -1{,}39 + 0{,}202 X^*$, $t = 12{,}22$, $R^2 = 0{,}914\,2$. Коэффициент 0,202 — это мультипликатор ВВП в неизменных ценах, Y — импорт в неизменных ценах. Значит, на импорт идёт 20 % экономики. И это отличается от того 0,045.

Метод шестой, изощрённо-математический. Есть ridge regression. Гребневая регрессия. Проблема возникала с тем, что $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ — вырожденная матрица, и это эквивалентно делению на ноль. Такие вопросы возникали в 1940–50-х годах. Возник вопрос регуляризации у академика Тихонова. Идея ridge-регрессии: у матрицы плохой индекс обусловленности $\frac{\lambda_{\max}}{\lambda_{\min}}$. Рассмотрим матрицу $\boldsymbol{X}^T \boldsymbol{X} + \rho \boldsymbol{I}$), в которой из-за прибавления какой-то диагональной матрицы уменьшится $\frac{\lambda_{\max}}{\lambda_{\min}}$ (вообще в единицу может прийти). Рассмотрим прогрессию $\rho_k \to 0$. Тогда модифицированная матрица будет стремиться к исходной, которую нельзя посчитать. Считается ρ_1, ρ_2 до ρ_k, и где-то мы остановимся. Мы $\rho \boldsymbol{I}$ ввели, как будто гребень.

Метод седьмой. Он имеет самостоятельное применение, но хорош для уменьшения мультиколлинеарности. Principal components **(главные компоненты)**. Любят его социологи и трудовики. Рассмотрим его общо. У нас есть проблема: в k-мерном пространстве[1] они почти зависимы, и размерность пространства менее k. Наш метод — это снизить размерность. Это метод многомерного статистического анализа.

Пусть есть объясняющие переменные $x_1; \ldots; x_k$. Давайте от этих переменных перейдём к линейным комбинациям переменных. Рассмотрим $z = \alpha_1 x_1 + \ldots + \alpha_k x_k$. Очевидная вещь: если потом перейти к новым координатам z, то как трактовать линейную комбинацию? Что такое два ВВП плюс три CPI? Да не везде это важно. Пусть x — ставки процента по инструментам или доходность акций. Тогда z — доходность портфеля. Не всегда переход даёт что-то интерпретировать. Я хочу снизить размерность и говорю: надо будет меньше z, так как эффективная размерность меньше. Но хочется сохранить максимум информации, которая сидит в x. Чем мерить информацию? Разумно эвристически захотеть большой изменчивости z в виде дисперсии: $\operatorname{Var} Z \to \max\limits_{\alpha_{1;\ldots;k}}$. Тогда надо ещё что-то сказать: если увеличить все α в 2 раза, то дисперсия подскочит в 4 раза. Проведём нормировку, чтобы длина вектора равнялась единице: условие $\sum \alpha_i^2 = 1$ нам поможет. То есть $\vec{\alpha}^T \vec{\alpha} = 1$. Моё упрощение: z — центрированная величина. Тогда максимизируется выборочная дисперсия при условии нормировки:

$$\begin{cases} \max\limits_{\vec{\alpha}} \operatorname{Var}(z) = \frac{1}{n} \sum z_i^2 \sim \max\limits_{\vec{\alpha}} \sum z_i^2; \\ \sum \alpha_i^2 = 1. \end{cases} \tag{16.1}$$

[1] Приехал как-то провинциальный математик в Москву. Идёт по улице и видит афишу на столбе: «*Камерный* оркестр». Покупает билет, заходит, внимательно слушает, уходит, на улице пожимает плечами и говорит: «Вырожденный случай. Ка равно трём!»

$\sum z_i^2 = \vec{z}^T\vec{z}$, $z = \vec{x}\vec{\alpha}$. Запишем функцию Лагранжа:

$$\mathcal{L} = \vec{\alpha}^T(\vec{x}^T\vec{x})\vec{\alpha} + \lambda\left(1 - \sum \alpha_i^2\right) \tag{16.2}$$

$$\text{FOC: } \frac{\partial\mathcal{L}}{\partial\alpha_i} = 0, \quad \frac{\partial\mathcal{L}}{\partial\lambda} = 0 \Leftrightarrow \vec{\alpha}^T\vec{\alpha} = 1 \tag{16.3}$$

Преобразуем градиент квадратичной формы: $\alpha^T(\vec{x}^T\vec{x})\vec{\alpha} = (\boldsymbol{x}^T\vec{\alpha})^T(\boldsymbol{x}^T\vec{\alpha})$. Тогда $\frac{\partial\mathcal{L}}{\partial\vec{\alpha}} = 2(\boldsymbol{x}^T\boldsymbol{x})\vec{\alpha} - \lambda\cdot 2\vec{\alpha} = 0$. Это первое соотношение. Перепишем его: $(\boldsymbol{x}^T\boldsymbol{x})\vec{\alpha} = \lambda\vec{\alpha}$. Задача фантастически известная, поэтому ищем собственный вектор $\vec{\alpha}$ и собственное число λ. Собственный вектор определяется с точностью только до сомножителя, поэтому мы должны вспомнить условие и найти оный единичной длины. Матрица $\boldsymbol{x}^T\boldsymbol{x}$ — симметричная матрица, все собственные числа действительные и положительные. k собственных чисел, поэтому нам нужен глобальный максимум нашего условия. Даже не надо писать условие второго порядка: $\vec{\alpha}_i^T(\boldsymbol{x}^T\boldsymbol{x})\vec{\alpha}_i = \vec{\alpha}_i^T\lambda_i\vec{\alpha}_i = \lambda_i(\vec{\alpha}_i^T\vec{\alpha}_i) = \lambda_i$, поэтому нам нужен $\lambda_{\max}$, и ему соответствует некоторый собственный вектор. Поэтому теперь у нас есть линейная комбинация — *первая главная компонента*. У неё самая большая выборочная дисперсия. Она перехватывает больше всего информации из $\vec{x}$. Можем рассмотреть следующее по величине λ_i, и значение целевой функции будет поменьше, но дисперсия будет тоже большой. Имеется вектор $(z_1; z_2; \ldots; z_k)$ — если они не равны по величине, а матрица симметрична, то собственные векторы ортогональны. Они ортогональны друг к другу. Поскольку z есть линейное преобразование, то мы $\boldsymbol{x}^T\boldsymbol{x}$ переводим в $\boldsymbol{z}^T\boldsymbol{z}$, и, оказывается, матрица принимает диагональный вид. Это матрица собственных векторов. Матрица $\boldsymbol{x}^T\boldsymbol{x}$ — исходная — содержит с точностью до $\frac{1}{n}$ дисперсию $\boldsymbol{x}$ по диагонали. Но $\mathrm{Var}(x_1) + \ldots + \mathrm{Var}(x_k)$ — это след матрицы, а при преобразовании след матрицы сохраняется, поэтому это то же самое, что и $\mathrm{Var}(z_1) + \ldots + \mathrm{Var}(z_k) = \lambda_1 + \ldots + \lambda_k$, а эта сумма — след матрицы — равна сумме собственных чисел. Это линейное преобразование. Поэтому если вместо исходной регрессии строить регрессию вида $\vec{Y} = \boldsymbol{Z}\vec{\beta} + \vec{\varepsilon}$. Если Z ортогональны, то будем иметь диагональную матрицу в произведении:

$$\hat{\vec{\beta}} = (\boldsymbol{z}^T\boldsymbol{z})^{-1}\boldsymbol{z}^T\boldsymbol{Y} \tag{16.4}$$

Вывод: можно ортогональные компоненты выбросить, и тогда оценки коэффициентов не изменятся. Это теорема Фриша—Воу—Ловелла (Frisch—Waugh—Lovell). Мешает только то, что $\lambda_{\min} \ll \lambda_{\max}$, выкидываем последние члены, и первые 3–4 компоненты в основном тащат почти всю дисперсию $\boldsymbol{X}$, которая равна следу матрицы: $\frac{\lambda_1}{\sum\lambda_i}; \frac{\lambda_1+\lambda_2}{\sum\lambda_i}; \ldots; 1$. На рис. 15 показано, как кумулятивная дисперсия зависит от количества включённых лямбд. Получается, что первые 3–4 компоненты несут 95 % дисперсии $\boldsymbol{X}$, поэтому можно строить регрессию не на исходные X, а на главные компоненты в неполном количестве. Тогда после построения регрессии на первые две штуки Z можно вернуться к $\boldsymbol{X}$. Выбрасываются последние λ, чтобы потерять как можно меньше информации.

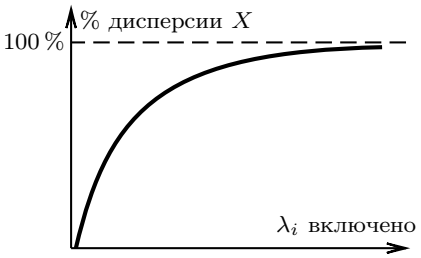

Рис. 15. Содержание Var

Пример. В 1982 году один выпускник уехал в США и стал работать в компании «Линч». Его группа играла на Нью-Йоркской бирже. Идея очень простая: продавать то, что завтра понизится, и покупать сегодня то, что завтра повысится. Акций много, считать надо быстро. Понятно, набирается динамика цен всех акций, но это 2 000 переменных, и ежу понятно, что есть мультиколлинеарность. Они по динамике цен акций считали огромную матрицу $\boldsymbol{x}^T\boldsymbol{x}$ и считали главные компоненты. И 95 % дисперсии описывают 7 из 2 000 компонентов! Это очень мало! Они стали прогнозировать 7 главных компонент. Первая главная компонента: у 9 акций большие α, а у остальных почти нет. Первая главная компонента — это нефтяной портфель. И так реально работала система. Вот точно линейная комбинация акций интерпретируется как портфель. Если x разной природы, каждый измеряется в своих единицах, и неравномерное изменение компонентов приводит к неинвариантности и изменению нагрузки. В EViews его нет, а в SPSS и STATA он есть. Но если поменять переменные, пойдут другие главные компоненты. Применять его надо разумно и с головой.

Всем хорошего Нового года, и чтобы первые главные компоненты были по сумме близки к дисперсии. И до 18 января.

Я хотел бы показать на конкретном примере смысл метода главных компонент. Рассмотрим следующий набор данных:

№ набл.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	Y
1	17,001	16,083	−10,581	18,434	3,832	21,114	4,957	121,231	182,874
2	4,543	3,737	0,502	9,740	2,285	−0,016	0,448	7,413	37,015
3	15,558	15,423	−11,125	11,239	3,607	17,708	5,271	67,397	170,632
4	10,935	8,031	−4,838	16,622	3,452	10,826	3,471	36,185	110,967
5	18,129	15,540	−11,321	25,761	4,047	20,402	15,048	111,681	213,679
6	13,380	5,237	−6,285	16,993	3,796	16,121	2,278	69,005	130,501

7	29,453	36,021	−20,579	25,192	4,888	44,929	17,545	275,658	346,661
8	14,000	19,785	−8,177	22,104	3,826	16,235	11,991	121,546	179,252
9	27,836	28,421	−17,718	31,769	4,075	38,782	4,559	352,499	314,755
10	6,163	7,970	−0,423	1,072	2,663	14,230	2,081	7,232	60,360
11	11,332	7,587	−7,043	13,116	3,007	15,295	1,905	56,590	123,335
12	12,553	10,970	−6,270	12,672	3,218	13,593	2,332	62,455	122,697
13	17,538	20,420	−10,580	16,828	3,649	26,657	2,653	147,626	191,088
14	13,314	14,829	−8,376	19,704	3,335	13,395	4,069	64,210	158,637
15	12,603	13,744	−7,435	13,733	3,863	17,441	2,504	65,287	130,018
16	18,900	16,197	−11,944	15,076	4,079	22,361	28,658	159,467	209,791
17	12,351	7,866	−7,069	11,976	3,840	16,436	4,822	41,166	121,883
18	24,338	29,080	−17,778	23,719	4,380	34,405	7,942	320,357	289,133
19	21,633	19,849	−12,561	20,942	4,057	31,112	3,072	228,862	231,303
20	21,917	15,409	−10,907	20,983	4,449	29,470	10,392	133,745	213,256

Уравнение настоящей зависимости: $Y = 3X_1 + 2X_2 - 5X_3 + 1{,}5X_4 + 0{,}3X_5 + 0{,}7X_6 + 0{,}5X_7 + \ln X_8 + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0; 4)$. Результаты регрессии таковы (заодно пошатаем данные): $R^2_{20} = 0{,}998$, $\bar{R}^2_{20} = 0{,}997$, $R^2_{17} = 0{,}998$, $\bar{R}^2_{17} = 0{,}997$.

β	3	2	−5	1,5	0,3	0,7	0,5	0,14÷0,24
$\hat{\beta}$	2,357	1,340	−5,352	1,975	−8,329	1,455	0,943	0,008
t-статистика	1,750	3,585	−5,365	5,404	−1,875	2,689	4,771	0,221
$\hat{\beta}$ по 17 набл.	3,643	1,082	-5,964	1,972	−4,507	0,222	0,687	0,033
t-стат. по 17 набл.	2,865	3,074	−5,780	6,135	−1,082	0,337	3,632	0,852

Сразу видно, что X_5 даёт большие сбои. Кроме того, первый коэффициент после пошатывания данных резко возрос и стал более значимым. Между X заложена очень высокая корреляция. Рассмотрим матрицу ковариаций (заодно убедимся по корреляционной матрице, что почти все величины очень сильно зависят от X_1):

Ковариационная матрица:

42,290	48,908	−33,362	36,266	3,494	65,612	21,374	601,995
48,908	70,300	−40,624	41,489	3,812	79,148	24,838	756,903
−33,362	−40,624	27,819	−28,429	−2,776	−51,274	−17,683	−478,085
36,266	41,489	−28,429	46,537	3,012	49,505	15,963	539,148
3,494	3,812	−2,776	3,012	0,371	5,358	2,411	44,040
65,612	79,148	−51,274	49,505	5,358	110,728	28,117	954,723
21,374	24,838	−17,683	15,963	2,411	28,117	48,008	245,117
601,995	756,903	−478,085	539,148	44,040	954,723	245,117	10009,323

Корреляционная матрица:

1,000	0,897	−0,973	0,817	0,882	0,959	0,474	0,925
0,897	1,000	−0,919	0,725	0,747	0,897	0,428	0,902
−0,973	−0,919	1,000	−0,790	−0,864	−0,924	−0,484	−0,906
0,817	0,725	−0,790	1,000	0,725	0,690	0,338	0,790
0,882	0,747	−0,864	0,725	1,000	0,836	0,571	0,723
0,959	0,897	−0,924	0,690	0,836	1,000	0,386	0,907
0,474	0,428	−0,484	0,338	0,571	0,386	1,000	0,354
0,925	0,902	−0,906	0,790	0,723	0,907	0,354	1,000

Найдём собственные значения и собственные векторы для ковариационной матрицы (за нас это сделает Wolfram Mathematica 8):

Собственные значения

10252,7	48,848	24,232	18,302	8,799	2,097	0,353	0,047

Собственные векторы

0,060	0,233	−0,243	−0,209	−0,233	0,351	0,813	0,015
0,075	0,247	−0,336	−0,050	0,887	−0,126	0,119	−0,018
−0,047	−0,205	0,190	0,155	−0,040	−0,803	0,496	−0,054
0,053	0,103	0,162	−0,937	−0,043	−0,252	−0,130	0,028
0,004	0,041	−0,027	−0,030	−0,036	0,026	−0,031	−0,997
0,095	0,283	−0,745	0,061	−0,386	−0,376	−0,244	0,042
0,024	0,861	0,446	0,208	−0,069	−0,098	−0,039	0,018
0,988	−0,097	0,101	0,064	−0,014	0,002	−0,003	−0,004

Теперь получим координаты центрированных наблюдений ($x_i = X_i - \bar{X}$) в базисе из собственных векторов. Для этого умножим матрицу наблюдений на матрицу из собственных векторов и получим следующий результат:

№ набл.	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8
1	−1,080	−0,801	−1,408	−1,806	0,331	0,961	0,120	−0,050
2	−118,305	−8,833	8,762	1,867	2,324	−0,421	0,845	0,143
3	−55,078	2,577	−4,854	1,625	2,467	3,937	0,512	0,053
4	−87,451	−1,546	2,005	−3,877	−0,271	−0,456	0,135	−0,123
5	−9,883	9,650	3,610	−7,547	−0,991	−0,553	−0,543	0,189
6	−54,527	−4,021	0,959	−2,669	−5,713	0,014	−0,323	−0,211
7	157,132	12,447	−8,524	0,566	2,967	−0,737	0,009	0,109
8	−0,878	3,940	5,934	−3,241	5,455	−2,265	−0,251	−0,261
9	231,699	−10,102	2,604	−2,599	−1,580	−0,607	0,162	0,292
10	−117,101	−2,640	−4,557	10,478	0,492	−2,971	−0,216	0,151
11	−67,001	−3,546	−0,888	0,262	−2,409	0,889	−1,307	0,441
12	−61,090	−3,304	−0,203	0,723	0,835	0,998	0,904	0,049
13	25,737	−2,746	−5,748	0,910	1,779	−0,812	−0,051	0,114
14	−58,522	0,255	0,151	−6,086	3,794	0,610	−0,002	0,199
15	−57,598	−1,270	−3,720	−0,170	1,727	−0,109	−0,413	−0,396
16	37,404	16,668	10,780	8,165	−2,476	0,848	0,088	0,047
17	−82,035	1,009	−2,549	0,770	−2,802	0,800	−0,683	−0,253
18	199,027	−6,775	3,435	4,015	2,061	1,426	−0,573	−0,300
19	106,948	−7,325	−1,206	1,433	−2,684	−0,309	0,345	−0,102
20	12,604	6,363	−4,583	−2,818	−5,303	−1,244	1,242	−0,091
Var	10252,7	48,847	24,232	18,302	8,799	2,097	0,353	0,047
Var, %	99,008 %	0,472 %	0,234 %	0,177 %	0,085 %	0,020 %	0,003 %	0,000 %

Комментарии здесь излишни.

17 Лекция 17

Мы рассмотрели два жизненных отклонения от теоремы Гаусса—Маркова. Не все цели достижимы на наших данных. Когда мы занимаемся реальным исследованием, то нам недостаточно голой теории; нам нужны адекватные данные, и мы будем или себя обманывать, или, что чаще встречается, других.

Начинаем следующее нарушение — **гетероскедастичность**.

$\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Была дисперсия однородная, а стала $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$ — неоднородной. В векторном виде мы писали так: $\mathrm{Cov}(\vec{\varepsilon}) = \sigma^2\boldsymbol{I}$, а теперь это нарушается, теперь $\mathrm{Cov}(\vec{\varepsilon}) \neq \sigma^2\boldsymbol{I}$. Это называлось сферичностью нарушений (возмущения в многомерном пространстве походили на шарики). Сферичность нарушений — это когда в трёхмерном пространстве линиями уровня являются окружности, в четырёхмерном — сферы, далее — многомерные сферы. В двухмерном пространстве мы закручивали шапочку и получали шляпу с бесконечными полями. Гетероскедастичность — это когда шляпу сплющивают и она становится овальной. А если ещё и оси наклоняют, то это ковариация. Давайте определимся, когда это естественно, а когда нет.

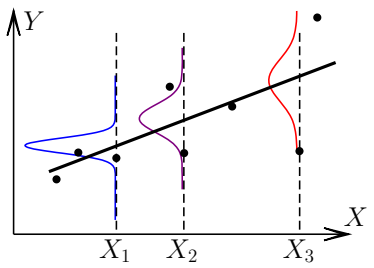

Рис. 16. Разная мера разброса

В одномерном случае имеем разброс точек; при каждом X можем построить распределение Y. Когда была гомоскедастичность, то для разных X будет нулевое условное матожидание (центр в пересечении линии

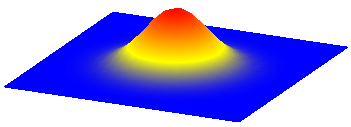

Рис. 17.1. Гомоскедастичность

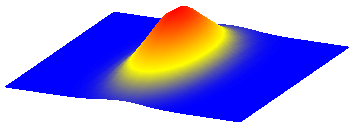

Рис. 17.2. Гетероскедастичность

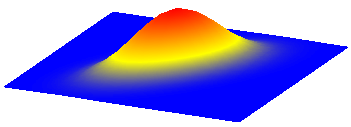

Рис. 17.3. Гетероскедастичность и ковариация ε

регрессии и фиксированного X) и такая же дисперсия. Расширение (мера разброса) одинаковое, даже если распределение какое-то ненормальное. Гетероскедастичность — это когда дисперсия разная.

Пример: personal disposal income, по вертикали — consumption, по горизонтали — PDI. Гомоскедастичность — это когда разброс потребления у разных домохозяйств одинаков. Гетероскедастичность — разный. Низкодоходное население имеет почти нулевой разброс (еда, одежда, обязательные платежи, обязательное потребление). Чем выше доход, тем больше разброс потребления, так как экономические агенты разные. У них нет однородности по одному параметру, поэтому будет неоднородность и по другому параметру.

Реальная ситуация на Нью-Йоркской бирже. Это одна из самых либерализованных торговых площадок, но некоторые ограничения там действуют (вроде нашего ФАС). Покупают и продают акции брокеры, лицензированные деятели, это какая-то мера защиты. А за услуги брокера назначаются комиссионные. До апреля 1973 года услуги по продаже и покупке регулировались; была security and exchange commission, возникло естественное движение: надо либерализовать цены на комиссионные услуги. Сторонники меры говорили, что ставка снизится. Электронного доступа не было, у них там были пальцы, они кричали, победили сторонники либерализации, и либерализовали цены на брокерские услуги. Условно изобразим, что произошло после этого. Рассмотрим среднюю ставку комиссионных за одну акцию в зависимости от времени. Стоимость акции для удобства стандартизована. Это почти как процент сделки. На бирже торгуются пакеты акций разной величины. Кто-то покупает три акции моего кроссовка Майкрософта, а кто-то берёт сто Норникеля. Есть помесячные усреднённые данные: цена брокериджа за акцию среди тех, которые торгуются от 1 до 199, снизилась с 56 до 41 цента за акцию. Вторая группа: 200–999 акций; снизилась с 40 до 25 центов за акции. Третья группа: 1000–9999 акций; снижение с 25 до 12 центов за акцию. Четвёртая группа: $> 10\,000$ акций; снижение с 15 до 5,5 центов за акцию.

Наверно, либерализация с точки зрения клиентов привела к снижению, а брокерские компании, ясно, недовольны. Но и колебания цены меняются: они меньше всего у самой большой группы и больше всего у самой маленькой группы. Почему на малых пакетах больше разброс? Кто продаёт и покупает крупные пакеты акций? Крупные финансовые организации: mutual funds, страховые фонды, инвестбанки. Стоимость пакетов очень высока, и каждая брокерская компания готова за это бороться. Это высококонкурентный рынок, брокеры только на объёме выезжают. А индивид даст заказ знакомому брокеру, конкуренция меньше, поэтому разброс больше. Экономические агенты разные, поэтому разброс разный. Очень часто разброс выражается через степень конкуренции.

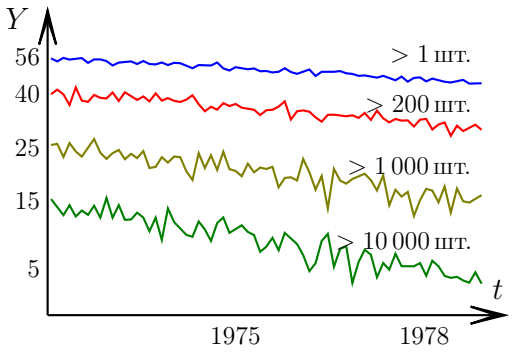

Рис. 17. Комиссионные на бирже

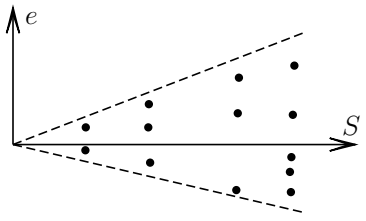

Рис. 18. Разброс по отраслям

Неоднородность агентов — частая причина гетероскедастичности. Если в одной выборке и мелкие фирмы, и «Камаз», то можно её ожидать. Надо показать начальнику, который не понимает слова «гетероскедастичность», что надо разбить выборку на группы.

Более чистый с точки зрения экономики пример. Экономика США, данные cross-section. Рассмотрим R&D, или НИОКР; затраты на них в США в разных отраслях. Мы хотим построить что-то следующее: $R\&D = a + b(SALES) + \varepsilon$. У маленькой отрасли меньше расходов на R&D. Построим маленький график. Самая маленькая у Containers and Packing, а самая большая в автомобильной. Non-bank Finance, Metals and Mining рассмотрели. Получили: $R = 193 + 0{,}031S$. t-отношения: $t_1 = 0{,}194\,8$, $t_2 = 2{,}77$, $R^2 = 0{,}478\,3$. Изобразим остатки; они лежат в уголке. Чем выше объём продаж, тем больше отклонение и разброс. Конечно, чистая величина — это качественно связанный с дисперсией показатель. Есть гетероскедастичность. Чем более крупная отрасль, тем больше разброс в расходах на R&D. Самая крупная отрасль — это автомобильные гиганты. У них больше разброс, потому что они индивидуалы, у них нет этой конкуренции. В отраслях с маленьким составом предприятий расходы на R&D более или менее стандартизованы. Расхождения в крупности экономических агентов — намёк на существование гетероскедастичности.

$\mathrm{Var}(\varepsilon_i) = \sigma_i^2$, $\mathrm{Cov}(\vec{\varepsilon}) = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_n^2 \end{pmatrix} = \mathrm{diag}\{\sigma_1^2, \dots, \sigma_n^2\} = \boldsymbol{\Omega}$. Поскольку нарушено условие теоремы Гаусса—Маркова, то оценки BLUE $\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{Y}$ какие будут? Оценки линейные; несмещённые, так как $\mathbb{E}(\hat{\vec{\beta}}) = \mathbb{E}\big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{Y}\big) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbb{E}(\boldsymbol{X}\vec{\beta} + \vec{\varepsilon}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\vec{\beta} = \vec{\beta}$. Осталось поничтожить эффективность: если $Y = \alpha + \beta X + \varepsilon$, то $\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum x_i^2}$ при гомоскедастичности. А теперь мы вообще не знаем, что такое σ^2. $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i(\beta x_i + \varepsilon_i)}{\sum x_i^2} = \beta + \frac{\sum x_i \varepsilon_i}{\sum x_i^2}$. Если $k_i = \frac{x_i}{\sum x_i^2}$, то $\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\sum k_i\varepsilon_i) = \sum k_i^2 \mathrm{Var}(\varepsilon_i) = \sum k_i^2\sigma_i^2$. Теперь эта сумма не свернётся, и получится в итоге: $\frac{\sum x_i^2\sigma_i^2}{\left(\sum x_i^2\right)^2}$. Дисперсия смещённая. А мы же смотрим по t-статистике и делаем выводы. Поэтому оценка дисперсии из-за смещения может плохо повлиять на выводы. В зависимости от σ_i^2 наша оценка МНК может быть и overestimate, и underestimate. Если есть данные с гетероскедастичностью, то мы используем обычную формулу МНК, получаем несмещённые оценки коэффициентов, но неправильные оценки дисперсии и неправильные выводы о значимости коэффициентов. Получается развилка: понятно, что программные пакеты посчитают оценку рутинно как $\hat{\sigma}^2 = \frac{\mathrm{RSS}}{n-k}$, но это вообще ни о чём не говорит. Первая возможность — использовать не МНК. Вторая возможность — придумать, как оценивать дисперсию в методах гетероскедастичности, и видоизменить формулу МНК. Третий путь — придумать формулу, которая бы работала и там, и там, но тогда мы потеряем в эффективности и ещё кое в чём.

Запишем уравнение: $Y_i = \alpha + \beta X_i + \varepsilon_i$ при условии $\mathrm{Var}(\varepsilon_i) = \sigma_i^2$. Чтобы получить BLUE-оценку, надо как-то поработать с дисперсией. Например, она поменяется квадратично, если мы случайную величину изменим в несколько раз. Поэтому поделим на σ_i: $\frac{1}{\sigma_i}Y_i = \frac{\alpha}{\sigma_i} + \beta\left(\frac{X_i}{\sigma_i}\right) + \frac{\varepsilon_i}{\sigma_i}$, и это то же самое. Условие таково: $\mathrm{Var}\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{\mathrm{Var}(\varepsilon_i)}{\sigma_i^2} = 1$. Имеем обычное уравнение регрессии: $Y_i^* = \alpha\left(\frac{1}{\sigma_i}\right) + \beta X_i^* + \varepsilon_i^*$, корреляции нет. Альфа перестал быть свободным членом, и он вылетает. Надо искусственно создать переменную $\frac{1}{\sigma_i}$, а потом строим регрессию с двумя переменными, но без свободного члена. Получаем $\hat{\beta}$, $\hat{\alpha}$, но из другого уравнения, и задача свелась к предыдущей. Посмотрим же: при МНК мы минимизируем сумму квадратов отклонений, здесь сделали то же самое. Раньше мы строили $\min\limits_{\alpha,\beta}\sum\left(Y_i^* - \alpha\frac{1}{\sigma_i} - \beta X_i^*\right)^2 = \min\limits_{\alpha,\beta}\sum\frac{1}{\sigma_i^2}\underbrace{(Y_i - \alpha - \beta X_i)^2}_{e_i^2} = \min\limits_{\alpha,\beta}\sum\frac{1}{\sigma_i^2}e_i^2 = \min\limits_{\alpha,\beta}\sum w_i e_i^2$. Это называется Weighted Least Squares — взвешенный МНК, мы придали отклонениям разный вес. Получается так, как на рисунке 16: имеется набор разных отклонений, а при большой дисперсии маленькое отклонение более вероятно, а при маленькой менее вероятно.

Итак, надо каждую переменную поделить на σ_i, которых мы пока не знаем.

$$\begin{pmatrix} Y_1^* \\ \vdots \\ Y_n^* \end{pmatrix} = \begin{pmatrix} 1/\sigma_1 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_n \end{pmatrix}\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} Y_1/\sigma_1 \\ \vdots \\ Y_n/\sigma_n \end{pmatrix}$$

То же самое сделаем со столбцами матрицы $\boldsymbol{X}$. $\boldsymbol{\Omega}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & & 0 \\ & \ddots & \\ 0 & & 1/\sigma_n^2 \end{pmatrix}$. А матрицу, на которую мы домножаем, логично назвать $\boldsymbol{\Omega}^{-0,5}$, тогда $\boldsymbol{X}^* = \boldsymbol{\Omega}^{-0,5}\boldsymbol{X}$, $\vec{Y}^* = \boldsymbol{\Omega}^{-0,5}\vec{Y}$, $\vec{\varepsilon}^* = \boldsymbol{\Omega}^{-0,5}\vec{\varepsilon}$. Далее, формула МНК такова:

$$\hat{\vec{\beta}} = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\vec{Y}^* = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-0,5}\boldsymbol{\Omega}^{-0,5}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}^{-0,5}\boldsymbol{\Omega}^{-0,5}\vec{Y} \tag{17.1}$$

Нужна некоторая осторожность в обосновании действий, поэтому останется:

$$\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\vec{Y} \tag{17.2}$$

Эта матрица возникла в двух местах. Это линейная формула, так как мы что-то одно домножаем на $\vec{Y}$. Оценки несмещённые, эффективные, линейные, то есть BLUE. Эта формула — частный случай формулы GLS — Generalized Least Squares. Обобщённость за счёт гетероскедастичности. WLS ничем не отличается от GLS. GLS мы можем обобщить ещё на более широкий случай, когда $\boldsymbol{\Omega}$ не диагональна, т. е. есть корреляция.

Ничего нового мы не сделали: преобразовали данные, применили ТГМ и получили НЛО-оценки. В некоторой литературе это называется теоремой Эйткена (Aitken). Остался небольшой пробел; оценки BLUE, и надо только выписать ковариационную матрицу оценок GLS.

$$\mathrm{Cov}(\tilde{\vec{\beta}}_{\mathrm{GLS}}) = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1} = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-0,5}\boldsymbol{\Omega}^{-0,5}\boldsymbol{X})^{-1} = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1} \tag{17.3}$$

Иногда исследователи предпочитают записывать ковариационную матрицу по-другому:

$$\mathrm{Cov}(\vec{\varepsilon}) = \sigma^2 \boldsymbol{V} = \boldsymbol{\Omega} \Rightarrow \mathrm{Cov}(\tilde{\vec{\beta}}_{\mathrm{GLS}}) = \left(\boldsymbol{X}^T \frac{1}{\sigma^2} \boldsymbol{V}^{-1} \boldsymbol{X} \right)^{-1} = \sigma^2 (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \tag{17.4}$$

Ясно, что эта формула тоже работает. Если мы встречаемся с гетероскедастичностью, то мы можем либо преобразовать данные и пользовать МНК, либо использовать формулу GLS. Если $\varepsilon_i \sim \mathcal{N}(0; \sigma_i^2)$, то в силу линейности оценка $\hat{\vec{\beta}}$ будет нормально распределённой, и можно будет применять всякие t- и F-статистики. Ружьё вперёд на одной или двух лекциях: то, что мы получили, имеет куда более общий характер. Есть ковариационная матрица $\vec{\varepsilon}$, это $\boldsymbol{\Omega}$ — полностью определённые матрицы. Обобщённый МНК обобщает на любые отклонения от сферичности возмущения, и матрица должна быть положительно определена. Слабое место: мы якобы знаем σ_i^2. Это основная наша проблема.

18 Лекция 18

Итак, при гетероскедастичности у нас $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$, $\mathrm{Cov}(\vec{\varepsilon}) = \boldsymbol{\Omega} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$.

Оценки МНК по-прежнему линейны и несмещённы, но они теряют эффективность, а по рутинной формуле они вообще сомнительны, то есть коэффициенты могут со значимостью ссориться. Выход двухсторонний: преобразовываем данные. $\vec{Y}^* = \boldsymbol{\Omega}^{-0,5}\vec{Y}$, $\boldsymbol{X}^* = \boldsymbol{\Omega}^{-0,5}\boldsymbol{X}$, $\vec{Y}^* = \boldsymbol{X}^*\vec{\beta} + \vec{v}$, то есть веса обратно пропорциональны дисперсиям. Мы получаем $\hat{\vec{\beta}}$, а $\tilde{\vec{\beta}}_{\mathrm{GLS}} = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\vec{Y}$, $\mathrm{Cov}(\tilde{\vec{\beta}}_{\mathrm{GLS}}) = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}$. Эта оценка является BLUE. Поэтому теорема Эйткена: в условиях гетероскедастичности вместо OLS несмещённые оценки даёт GLS. Мы вывели более общую формулу: этот обобщённый метод годится не только для разных корреляций. Это годно и для случая, когда $\mathrm{Cov}(\vec{\varepsilon}) = \boldsymbol{\Omega}$ — это не просто диагональная, а вообще симметричная и положительно определённая. Кстати, всякая симметричная матрица приводится к диагональному виду из собственных векторов. $\exists \lambda_1, \ldots, \lambda_n \in \mathbb{R}$. $\exists \vec{p}_1, \ldots, \vec{p}_n$ — собственные векторы. Они ортогональны друг другу: $\vec{p}_j^T \vec{p}_i = 0$ при $i \neq j$. Просто $\boldsymbol{p}^T\boldsymbol{p} = (\cdots)\begin{pmatrix}\vdots\end{pmatrix} = \boldsymbol{I}$, $\boldsymbol{p}^{-1} = \boldsymbol{p}^T$. Раз матрица приводится к диагональному виду, то это означает, что $\boldsymbol{\Omega} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}$, где $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$. Для нашего случая это ещё проще: $\boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P} = \boldsymbol{P}^T\boldsymbol{\Lambda}\boldsymbol{P}$. Далее, $\boldsymbol{\Omega}^{-1} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}^{-1}(\boldsymbol{P}^T)^{-1} = \boldsymbol{P}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{P} = \boldsymbol{P}^T\boldsymbol{\Lambda}^{-0,5}\boldsymbol{\Lambda}^{-0,5}\boldsymbol{P} = (\boldsymbol{\Lambda}^{-0,5}\boldsymbol{P})^T(\boldsymbol{\Lambda}^{-0,5}\boldsymbol{P}) = \boldsymbol{C}^T\boldsymbol{C}$.

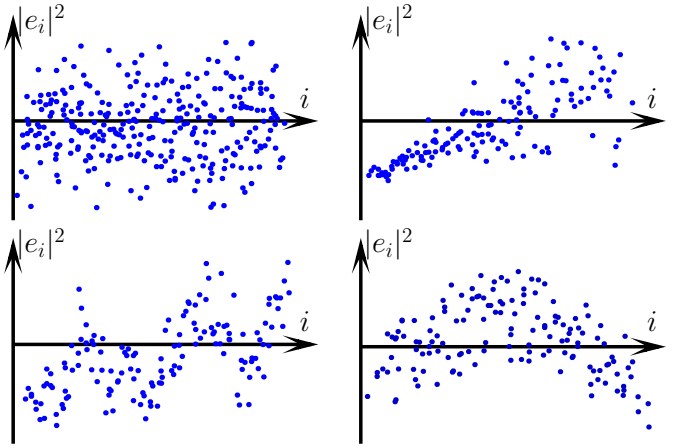

Рис. 19. Гетероскедастичности нет на первой диаграмме, на остальных — есть

Посмотрим на полученную матрицу как на матрицу некоторого линейного преобразования: $\vec{Z} = \boldsymbol{C}\vec{Y}$. $\boldsymbol{C}\vec{Y} = \boldsymbol{C}\boldsymbol{X}\vec{\beta} + \boldsymbol{C}^T\vec{\varepsilon}$. Матрица невырожденная, поэтому мы что-то там линейно преобразовали и ничего не потеряли. Ковариационная матрица — это матожидание произведения вектора на его транспонирование. $\mathrm{Cov}(\vec{\varepsilon}) = \mathbb{E}(\vec{\varepsilon}\vec{\varepsilon}^T) = = \boldsymbol{\Omega}$ — так было раньше. Отныне $\mathrm{Cov}(\vec{v}) = \mathbb{E}(\boldsymbol{C}\vec{\varepsilon}\vec{\varepsilon}^T\boldsymbol{C}^T) = \boldsymbol{C}\mathbb{E}(\vec{\varepsilon}\vec{\varepsilon}^T)\boldsymbol{C}^T = \boldsymbol{C}\boldsymbol{\Omega}\boldsymbol{C}^T = \boldsymbol{\Lambda}^{-0,5}\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{\Lambda}\boldsymbol{P}\boldsymbol{P}^T\boldsymbol{\Lambda}^{-0,5} = = \boldsymbol{\Lambda}^{-0,5}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-0,5} = \boldsymbol{I}$. Поэтому задача сводится вот таким образом: $\vec{Y}^* = (\boldsymbol{\Lambda}^{-0,5}\boldsymbol{P})\vec{Y}$, $\boldsymbol{X}^* = (\boldsymbol{\Lambda}^{-0,5}\boldsymbol{P})\boldsymbol{X}$, $\vec{Y}^* = \boldsymbol{X}^*\vec{\beta} + \vec{v}$,

$$\hat{\vec{\beta}} = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^*\vec{Y}^* = (\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}^{-1}\vec{Y} = \tilde{\vec{\beta}}_{\mathrm{GLS}}$$

Оказывается, для случая чистой гетероскедастичности мы получили статистику, которая верна для гораздо общего случая, и поэтому это тоже называется $\tilde{\vec{\beta}}_{\mathrm{GLS}}$, и она тоже BLUE. Это очень удобно, лишь бы $\boldsymbol{\Omega}$ была симметричной положительной. А если даже $\boldsymbol{\Omega}$ единична, то мы должны подставить её, и тогда у нас получается частная формула МНК. Поэтому последний даёт частный и очень важный случай.

Тогда перед нами стоит понятная дилемма: когда мы строим эконометрическую модель, мы не знаем, какое предположение о матрице $\boldsymbol{\Omega}$ будет более разумной. Когда данные собраны по cross-section, то дисперсии будут разными. Поэтому развилка: если одинаковость, то OLS, если гетерость, то GLS. Надо распознать то, что нужно, учитывая, что мы матрицы $\boldsymbol{\Omega}$ не знаем. Если мы чего-то не знаем в эконометрике, то мы должны оценить, но это не вполне возможно, так как у нас k регрессоров и n сигм. Приходится ввести предположение: σ_i связана с номером переменной. Поэтому встаёт задача выявить связь между σ_i и объясняющими факторами. Но нечего нам взять, кроме как квадраты остатков. Если бы у нас для каждого X была генеральная совокупность или хотя бы выборка для каждого фиксированного X, то тогда было бы хорошо. Но у нас только одно наблюдение инфляции в январе, только один выпуск фирмы «Ромашка», поэтому мы вводим оценку $\tilde{\boldsymbol{\Omega}}$. $\frac{\mathrm{RSS}}{n-k}$ — это несмещённая матрица у нас была. Но в этом случае мы должны отказаться от несмещённости и требовать хотя бы состоятельности. Слуцкий говорит: если мы найдём состоятельную оценку матрицы, то оценка $\tilde{\vec{\beta}}_{\mathrm{GLS}}$ будет состоятельной. Пределы по вероятностью равны произведению пределов, требуется только

$\det(\boldsymbol{X}^T\boldsymbol{X}) \neq 0 \neq \det \boldsymbol{\Omega}$.

Когда мы ставили оценку матрицы вместо истинной $\boldsymbol{\Omega}$, то это называют feasible GLS, а мы по-русски как-то странно называем. Теперь осталось посмотреть по данным, надо ли применять feasible GLS (FGLS) или OLS. $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon} - OLS \Rightarrow e_i$. Далее следуем органолептическим путём (пробуем на вкус, цвет, запах). Теперь рассмотрим пары «номер наблюдения — e_i» Диапазон может быть одинаковым, может расширяться, может идти по диагонали или лежать в области любой прелестной формы. Так вот, только на первом рисунке гомоскедастичность. Если рисунок первый, то всё нормально, да и то надо мерить, а все остальные надо уточнять. Жаль, мы не можем оценить $n + k$ параметров по n наблюдениям, поэтому сокращаем наблюдения и предполагаем, что есть зависимость квадратов остатков от какой-то переменной. В предыдущем случае изменчивость e_i могла быть объяснена доходом семьи. Проведём тест.

Тест Парка (Park). Он устроен следующим образом: $\ln e_i^2 = a + b \ln X_i + w_i$. $\mathcal{H}_0$: homo, $\mathcal{H}_1$: hetero. Когда гомоскедастичность, то тогда $b = 0$, то есть гипотеза о значимости коэффициента. $\mathcal{H}_0\colon b = 0$ против $\mathcal{H}_1\colon b \neq 0$. Если мы отвергаем нулевую гипотезу, то есть гетероскедастичность, и мы находим тип зависимости $\sigma(X_i)$. Если не нашли, то на уровне 5 % гипотеза о гетероскедастичности такого вида отвергается. Если не нашли зависимости, то можно брать другую переменную (не располагаемый доход, а затраты на отдых). Проверяя p-value, мы предполагаем, что либо w_i мало и они нормальны, либо их много и они асимптотически нормальны. Оценка: $e_i^2 = e^{\hat{a}} \cdot X_i^{\hat{\beta}}$.

Glejser test. $|e_i| = \alpha + \beta \cdot g(X_i) + w_i$. В своей работе Глейзер придумал 5 основных функций, но нам и трёх хватит: X_i, $\frac{1}{X_i}$, $\sqrt{X_i}$. Те же гипотезы: $\mathcal{H}_0\colon \beta = 0$ против $\mathcal{H}_1\colon \beta \neq 0$. Для Глейзера есть удобный приём: если у нас много объясняющих переменных (5, 6) и мы не знаем, от какой из них искать, то тогда есть понятный приём: а возьму-ка я их линейную комбинацию, и от неё будет зависимость, раз от одного чего-то есть. Удобнее взять ту, которую мы уже построили: $e_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 X_2^i - \ldots - \hat{\beta}_k X_k^i$. Тогда $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_2^i + \ldots + \hat{\beta}_k X_k^i$. В качестве переменной, от которой пытаемся построить, можно брать $\hat{Y}$. Пусть мы нашли гетероскедастичность, а дальше нам говорят: мы хотим оценку сигм. Тогда $\tilde{\sigma}_i = |\hat{e}_i| = \hat{\alpha} + \hat{\beta} g(X_i)$. Тогда мы знаем преобразование, и мы можем собрать матрицу для FGLS. Мы имеем конструктивный способ, как использовать все известные данные.

Всё, строим регрессию: $\frac{Y_i}{|\hat{e}_i|} = Y_i^*$. Одна из практических сложностей заключается в том, что $|\hat{e}_i| = \hat{\alpha} + \hat{\beta} X_i + w_i$. Но чем это плохо? Получается, что оценка $\hat{e}_i$ может убежать в отрицательную область, и это часто встречается при анализе финансовых рядов. Хорошего выхода из этого нет. Некоторые авторы предлагают наблюдения, а не оценку. Некоторые предлагают взять модуль оценки. Борьба с ними привела в конце 80-х к появлению GARCH-моделей, так как ARCH-модели приводили к отрицательным оценкам. В GARCH оценка никогда не отрицательна. Когда мы делим на σ_i, у нас получается возмущение с одинаковой и даже единичной дисперсией. Но может быть и такое, что α незначима, что $\tilde{\sigma}_i = \tilde{\beta} X_i$. При выполнении преобразования мы пишем: $\tilde{\sigma}_i \cong \beta X_i$, $Y_i^* = \frac{Y_i}{\cancel{\beta} X_i}$, $Y_i = \alpha + \beta X_i + \varepsilon_i$, или $\frac{Y_i}{X_i} = \alpha \frac{1}{X_i} + \beta + \frac{e_i}{X_i}$. Это забавный случай, когда два коэффициента меняются местами. И не всегда обязательно находить оценку этого коэффициента.

Эти тесты называются конструктивными, так как они дают инструмент для получения состоятельной оценки из имеющейся информации. Но за это приходится платить конкретной спецификацией типа зависимости. Есть более общие тесты, которые не предполагают точной спецификации остатков. Нам придётся платить: вот обнаружим мы гетероскедастичность, а конкретных действий никаких не предвидится.

Тест Голдфелда—Квандта. Если у нас есть гетероскедастичность, которая меняется от какой-то переменной, то пусть $\sigma_i = f(X_i)$, обязательно монотонная зависимость. Тогда переупорядочим наблюдения в сторону возрастания X_i. Упорядочили наблюдения. Тогда, если есть гомоскедастичность, при маленьких и больших X дисперсия одинакова, и выборочные оценки будут мало отличаться, а если они отличаются, то это признак гетероскедастичности: стандартное отклонение растёт монотонно с изменением этой величины. Тест предлагает разбить все наблюдения на 3 группы: n_1, d и n_2 наблюдений (в сумме n). Теперь найдём RSS_1 и RSS_2. В силу предположения о нормальности $\frac{\mathrm{RSS}_1}{\sigma_1^2(n-n_1)} \sim \chi^2_{n-n_1}$, $\frac{\mathrm{RSS}_2}{\sigma_2^2(n-n_2)} \sim \chi^2_{n-n_2}$. И тогда $\mathcal{H}_0\colon \sigma_1^2 = \sigma_2^2$ (гомо), $\mathcal{H}_1\colon \sigma_1^2 \neq \sigma_2^2$ (гетеро). Тогда

$$\frac{\mathrm{RSS}_1(n-n_2)}{(n-n_1)\mathrm{RSS}_2} \sim \mathrm{F}_{n-n_1;n-n_2}$$

Очень удобно, когда $n_1 = n_2$, и это рекомендация от авторов. А какое взять d? Он влияет на две вещи: большое d теряет информацию, но зато средние точки с большой вероятностью относятся как туда, так и туда, поэтому чем больше мы выкинем d, то тем чётче будет разница между RSS. Рекомендация авторов: от $\frac{1}{4}$ до $\frac{1}{3}$. Если у нас 30 точек и 5 или 6 объясняющих переменных, то мы оставили 10 слева и справа, но не очень хорошо оценивать. После того как мы обнаружили гетероскедастичность по Голдфелду—Квандту, то мы должны что-то делать.

Тест Бройша—Пейгана (Бреуша—Пагана), Breusch—Pagan. Он не требует выбора одной переменной и не требует точной спецификации соотношения, как другие тесты. Он предполагает, что $\sigma_1^2 = h(\alpha_1 Z_1^i + \ldots + \alpha_m Z_n^i)$ — некоторая функция. В Магнусе есть ошибка: на самом деле это не определённая, а произвольная функция! Это тест асимптотический.

19 Лекция 19

Итак, мы затронули **тест Бройша—Пейгана**. Это тестирование гетероскедастичности при помощи множителей Лагранжа. Сила этого теста — в его универсальности. Менее хорошее качество — худшая конструктивность. Мы проверяем score test: логарифм функции правдоподобия равен нулю. Этот тест — асимптотический тест, а не конечный. Когда Cross-section огромное, а не 20–30 точек, то тогда хорошо.

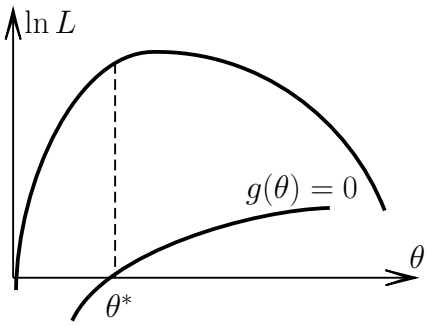

Рис. 20. Тест Бройша—Пейгана

$\sigma_i^2 = h(\alpha_0 + \alpha_1 Z_1^i + \ldots + \alpha_p Z_p^i)$, $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$. В учебнике Магнуса, Катышева, Пересецкого пропущена буква h: это общая функция от линейной комбинации. Тогда $\mathcal{H}_0$: есть гомоскедастичность, $\alpha_1 = \ldots = \alpha_p = 0$, $\mathcal{H}_1$: гетероскедастичность, $\alpha_1^2 + \ldots + \alpha_p^2 > 0$. Тест множителей Лагранжа строит только модель с ограничениями.

Первый этап теста стандартен: строим остатки. $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{e} \Rightarrow \vec{e}$. Далее, $\tilde{\sigma}^2 = \frac{\sum e_i^2}{n} = \frac{\mathrm{RSS}}{n}$. Далее, $g_i = \frac{e_i^2}{\tilde{\sigma}^2}$, который крутится вокруг единицы. Потом строим по OLS $g_i = \gamma_0 + \gamma_1 Z_1^i + \ldots + \gamma_p Z_p^i + w_i$. Далее, TSS = ESS+RSS. Если RSS большой, а ESS маленький, то такой зависимости почти нет. ESS маленький — гомоскедастичность. Если R^2 большой, то гетероскедастичность. Бройш и Пейган доказали, что $\frac{\mathrm{ESS}}{2} \overset{\text{as}}{\sim} \chi_p^2$. Компьютер возвращает p-value, или `prob`. Если он большой, то p-value малый, гетероскедастичность, гипотеза о гетероскедастичности не отвергается. Этот тест есть везде, но работает только асимптотически, не требуется предположения о нормальности ошибки, но работает не всегда.

Второй пример — расходы на R&D по отраслям. Данные cross-section, но их мало, всего 18 наблюдений. Тогда $\widehat{R\&D} = 192{,}99 + 0{,}0319 \cdot S$ (Sales). Стандартные отклонения: 990,99 у свободного и 0,008 3 для коэффициента, $r^2 = 0{,}478\,3$. Коэффициент значим, но как-то всё равно мало… Применили тест Парка к остаткам и получили: $\widehat{\ln e_i^2} = 5{,}687\,7 + 0{,}714 \ln S$, t-статистики 0,66 и 1,162 6, $r^2 = 0{,}077\,9$. Гетероскедастичности вроде нет.

А применим к этим же остаткам тесты Глейзера: $\widehat{|e_i|} = 578{,}57 + 0{,}011\,9S$, t: 0,85, 2,093 1, $r^2 = 0{,}215$. Другая модель: $\widehat{|e_i|} = -507{,}0 + 7{,}97\sqrt{S}$, t: $-0{,}5$, 2,27, $r^2 = 0{,}26$. Третья: $\widehat{|e_i|} = 2\,273{,}7 - 19\,925\,000\frac{1}{S}$, t: 3,76, $-1{,}02$, $r^2 = 0{,}14$.

Первое: третий вариант столь же плох, как и с логарифмами. Возьмём-ка ту, где больше r^2. Надо поделить на e_i. Если посчитать регрессию без свободного члена, то будет значимый коэффициент, и нужное преобразование для единичной регрессии таково: $Y_i^* = \frac{Y_i}{\sqrt{S_i}}$, $S_i^* = \frac{S_i}{\sqrt{S_i}} = \sqrt{S_i}$, и надо строить регрессию от корня. Регрессия: $\frac{\hat{Y}_i}{\sqrt{S_i}} = \alpha\frac{1}{\sqrt{S_i}} + \beta\sqrt{S_i} + u_i$. Тогда $\frac{\hat{Y}_i}{\sqrt{S_i}} = -246{,}68\frac{1}{\sqrt{S_i}} + 0{,}036\,8\sqrt{S_i}$, t: $-0{,}55$, 5,17. $r^2 = 0{,}63$. Да, эта регрессия применена в условиях, когда теорема Гаусса—Маркова гарантирует BLUE оценки. Мы собираемся использовать эти оценки в исходной регрессии. Поэтому регрессия для объёма продаж будет не с 192,99 и 0,031 9, а $-246{,}68$ и 0,036 8. Вот этот способ — преобразование данных по тесту Глейзера. Но зачастую и просто логарифмирование снижает остроту гетероскедастичности. Если мы построим такую регрессию, то $\widehat{\ln(R\&D)} = -7{,}36 + 1{,}322 \ln S$, t: $-3{,}98$, 7,87, $R^2 = 0{,}799\,4$. После логарифмирования ни один из трёх тестов не обнаружил гетероскедастичности. Поэтому просто логарифмирование тоже помогает бороться с гетероскедастичностью.

На Нью-Йоркской фондовой бирже были дебаты между юстицией и биржевиками по поводу комиссии. Мы смотрели данные, и там была гетероскедастичность. Но это можно обнаружить только тогда, когда мы приняли решение. Они исследовали пока в нерегулированной экономике, и противники либерализации посчитали вот какую регрессию: $\hat{Y}_i = 476\,000 + 31\,348X_i - (1\,083 \cdot 10^{-6})X_i^2$, t: 2,98, 40,39, $-6{,}54$, $R^2 = 0{,}934$. Замечательное уравнение, всё значимо. Входит и X^2 со значимым отрицательным коэффициентом. Это эффект масштаба — признак естественной монополии. Получается, это монополия, надо регулировать. Противники монополии не стали потрясать кулаками и говорить о невидимой руке. Они проанализировали остатки и сказали, что гетероскедастичность и что t-статистики могут быть другими. Они показали, что есть гетероскедастичность, поэтому надо подправить $\sqrt{X}$. Они построил уравнение такое: $\hat{Y}_i = 342\,000 + 25{,}77X_i + (4\,34 \cdot 10^{-6})X_i^2$, t: 32,3, 7,07, 0,503. И вот тут уже из-за учёта гетероскедастичности X^2 незначим, и коэффициент положительный. Поэтому неучёт гетероскедастичности привёл к качественно неправильному решению.

Есть и другие тесты: Вайди, Рисе-тест. Многие построены на идее а-ля Глейзер: зависимость модуля остатков от чего-то. А последнее время стал очень популярным третий путь. Чем МНК плох? Рутинная процедура неправильно считает дисперсию, и оценка его неэффективная. Идея Уайта: да и чёрт с ним. Давайте попробуем правильно записать оценки ковариационной матрицы МНК, забыв GLS и попробовав применять OLS, правильно посчитать ковариационную матрицу. Поэтому мы жертвуем эффективностью оценки, зато мы можем не проверять наличие гетероскедастичности. Мы должны быть спокойны, что мы не очень сильно ошибаемся. Уайт сказал, что его тест асимптотический, но работает и при гомо-, и при гетероскедастичности. Как он работает?

Тест Уайта. Рассмотрим ситуацию с гетероскедастичностью. $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$, $\mathrm{Cov}(\vec{\varepsilon}) = \boldsymbol{\Omega} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_n^2\}$, $\hat{\vec{\beta}} = (X^TX)^{-1}X^T\vec{Y} = \vec{\beta} + (X^TX)^{-1}X^T\vec{\varepsilon}$. $\mathrm{Cov}(\hat{\vec{\beta}}) = \mathbb{E}\big((\hat{\vec{\beta}} - \vec{\beta})(\hat{\vec{\beta}} - \vec{\beta})^T\big) = \mathbb{E}\big((X^TX)^{-1}X^T\vec{\varepsilon}\vec{\varepsilon}^TX(X^TX)^{-1}\big)$. Когда X детерминированы, то матожидание действует только на внутреннюю скобку. Это равно $(X^TX)^{-1}(X^T\boldsymbol{\Omega}X)(X^TX)^{-1}$, что есть сэндвичная формула: один кусочек мяса в центре, два кусочка хлеба. $\Omega = \sigma^2\boldsymbol{I} \Rightarrow \sigma^2(X^TX)^{-1}$ в нормальном простом случае. Какая размерность у серединочки? $(X^T\boldsymbol{\Omega}X) = (k \times n)(n \times n)(n \times k) = (k \times k)$, и при

увеличении числе наблюдений она не меняется, только каждый элемент чуточку меняется. А дальше Уайт доказал, что то, то он предложил, — состоятельная оценка средней матрицы и всей линейной комбинации. Это верно, даже если $\boldsymbol{X}$ — случайная матрица. Лишь бы $\mathbb{E}(\vec{\varepsilon} \mid \boldsymbol{X}) = 0$. Что же предложил Уайт? Способ оценки элементов ковариационной матрицы. Оценки Уайта для дисперсии.

При $n \to \infty$ получается дурная бесконечность в матрице $(X^T X)^{-1}$, поэтому надо обеспечить сходимость. Оценка Уайта: $n \cdot \widetilde{\mathrm{Cov}(\hat{\vec{\beta}})} = \left(\frac{1}{n} X^T X\right)^{-1} \times \left(\frac{1}{n} \sum\limits_{s=1}^{n} e_s^2 (x_s^T x_s)\right) \times \left(\frac{1}{n} X^T X\right)^{-1}$. x_s^T — это строчка под номером s, это s-е наблюдение. В произведении получается столбец на строку, то есть матрица. Она умножается на квадрат остатков. Так получается и вторая матрица. Все матрицы складываются с весами, равными квадрату остатков. Полученная вот так оценкой является состоятельной оценкой матрицы. И стали применять оценки Уайта для борьбы с возможной гетероскедастичностью. HCE — оценка, совместимая с гетероскедастичностью (heteroscedasticity-consistent estimator). Если мы просим использовать оценки Уайта, то матрица берётся по такой сложной оценке, а дальше даются обычные статистики. Изменятся только стандартные ошибка каждого коэффициента. Уравнение одно и то же, а сопровождающая статистика будет другой. Способ становится всё более популярным. Косвенно тесть есть: если оценки сильно расходятся с МНК, то была гетероскедастичность. Если коэффициент на границе области значимости, то результат должен быть с гетероскедастичности.

Переходим к борьбе с **автокорреляцией**. Это нарушение условий теоремы Гаусса—Маркова. Отныне $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) \neq 0$ для $i \neq j$. Начнём с простой ситуации, гетероскедастичности нет. Имеем матрицу: $\mathrm{Cov}(\vec{\varepsilon}) = \begin{pmatrix} \sigma^2 & & * \\ & \ddots & \\ * & & \sigma^2 \end{pmatrix}$. Если нам известна матрица $\boldsymbol{\Omega}$, то BLUE оценку даёт обобщённый МНК: тогда $\tilde{\vec{\beta}}_{\mathrm{GLS}} = (X^T \Omega^{-1} X)^{-1} (X^T \Omega^{-1} \vec{Y})$. Но мы можем и не знать $\boldsymbol{\Omega}$. Мы знаем и общий результат из feasible GLS. Если мы хотим оценить матрицу $\boldsymbol{\Omega}$, то будет неизвестных параметров 1, если нет гетероскедастичности, и $1 + \frac{n(n-1)}{2} > n$. Мы ещё вернёмся к общему случаю сферичности возмущений, но пока рассмотрим простую ситуацию: появляется корреляционная связь между соседними наблюдениями. $\mathrm{Corr}(\varepsilon_i, \varepsilon_{i+1}) \neq 0$, появляется понятие соседа. Когда имеет значение сосед, это анализ временных рядов, хотя пространственная корреляция тоже имеет право на рассмотрение. Чтобы это подчеркнуть, у нас индекс наблюдения будет не $_i$, а $_t$.

Пусть $\varepsilon_{t+1} = \rho \varepsilon_t + v_{t+1}$. Это называется AR(1), авторегрессия первого порядка. Определение: $AR(p)$: $e_t = \rho_1 \varepsilon_{t-1} + \rho_2 \varepsilon_{t-2} + \ldots + \rho_p \varepsilon_{t-p} + v_t$, и это уже не марковский процесс. v_t — это white noise (белый шум, что есть i.i.d.). А ε — это ошибка в регрессии $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$, то есть значение в следующий момент имеет пропорциональность ошибке в прошлый момент плюс случайная добавка. Первый вопрос: $\mathbb{E}(\varepsilon_{t+1}) = \rho \mathbb{E}(\varepsilon_t)$, и может статься, что матожидания разные или одинаковые. Потребуем, чтобы $\mathbb{E}(\varepsilon_t) = \mathrm{const}$, не меняется со временем. При $\rho \neq 1$ $\mathbb{E}(\varepsilon_t) = 0$. Дисперсия: $\mathrm{Var}(\varepsilon_{t+1}) = \rho^2 \mathrm{Var}(\varepsilon_t) + \mathrm{Var}(v_{t+1}) + 2\rho \mathrm{Cov}(\varepsilon_t, v_{t+1})$. Так как мы вводим v, то договорились, что $v(t)$ обладает марковским свойством. Оно более широко, чем мы тут применяем. Мы сделали качественный сдвиг: у нас в соотношения проникли наблюдения с разные моменты времени. Раньше состояние Y полностью определялось полностью другими переменными, а теперь оно зависит от $t-1$ и от траектории движения, по которой оно пришло. Мы перешли от функции к функционалу. Состояние системы зависит от состояния системы в предыдущий момент времени и не зависит от того, что было раньше. Это названо марковским свойством. Это зависит от текущего состояния и предыдущего, а от предыстории не зависит. Марковское свойство: $\boldsymbol{\Omega}_t = f(\boldsymbol{\Omega}_{t-1})$. Марковское свойство. Иногда белый шум называется innovations: это то новое, что пришло в момент t. Если это так, то v_t не зависит ни от чего, что приходит раньше, и $\mathrm{Cov}(\varepsilon_t, v_{t+1}) = 0$. Тогда $\sigma_\varepsilon^2 = \rho^2 \sigma_\varepsilon^2 + \sigma_v^2$, откуда следует, что $\sigma_\varepsilon^2 = \frac{\sigma_v^2}{1-\rho^2}$. Имеем: $\rho^2 \neq 1$, $|\rho^2| < 1$. Посчитаем смысл параметра ρ.

$\mathbb{E}(\varepsilon_{t-1} \mid \varepsilon_t) = \rho \varepsilon_{t-1} + v_t$, $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \rho \cdot \sigma_\varepsilon^2 + \underbrace{\mathbb{E}(\varepsilon_{t-1}, v_t)}_{=0}$. $\rho = \frac{\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1})}{\sqrt{\sigma_\varepsilon^2}\sqrt{\sigma_\varepsilon^2}}$ — коэффициент корреляции между текущим и предыдущим моментом времени.

20 Лекция 20

Более коротко рассмотренная нами проблема называется автокорреляцией: $\varepsilon_t = \rho \varepsilon_{t-1} + v_t$, $v_t \sim$ White Noise. Мы рассматриваем общую модель с $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$, $\mathbb{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \frac{\sigma_v^2}{1-\rho^2}$,f $\mathrm{Corr}(\varepsilon_t, \varepsilon_{t-1}) = \rho$, $|\rho| < 1$. Мы получили, что коэффициент корреляции между двумя моментами. Если сдвинуться на шаг, то инфляция между апрелем и мартом или августом и сентябрём имеет одинаковую корреляционную связь.

Откуда содержательно появляется марковская схема? Очень много экономических явлений обладает инерционную. Если в январе инфляция 10 %, то в феврале навряд ли будет дефляция. Если в этом месяце инфляция высока, то в следующем месяце, скорее всего, он будет высок, нежели низок, поэтому коэффициент корреляции одинаковый. Бывает отрицательная корреляционная связь. Если в этом месяце положительный шок, то в следующем отрицательный. В микре был cobweb effect: если там цена выше равновесия, то в следующий период она будет ниже. Это подсказывает нам визуальный критерий. Если после построения регрессии мы строим график просто остатков и они некоррелированные, то после любых e будет примерно поровну положительных и отрицательных. А если $\rho > 0$, то у нас будут длинные волны, то есть число пересечений оси гораздо меньше.

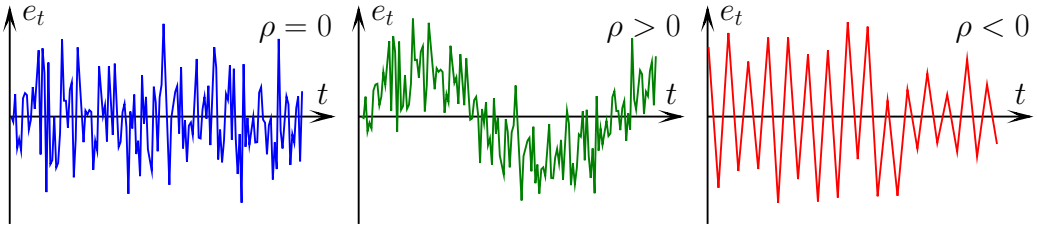

Рис. 21. Параметры автокорреляции

Если же $\rho < 0$, то будет, как на финансовых рынках. Пугливый инвестор: все побежали туда, а он цены сбил. Так что полезно смотреть в Excel графики остатков регрессии.

Мы взяли такую схему, и возникает вопрос: параметр ρ — это коэффициент корреляции между одним шагом. А будет ли корреляция между ε_t и ε_{t-2}? Здравый смысл подсказывает, что да. $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-2})$ — надо умножить: $\varepsilon_t = \rho\varepsilon_{t-1} + v_t = \rho(\rho\varepsilon_{t-2} + v_{t-1}) + v_t = \rho^2\varepsilon_{t-2} + \rho v_{t-1} + v_t$. Далее, $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-2}) = \rho^2\sigma_\varepsilon^2 + 0$. $\mathrm{Corr}(\varepsilon_t, \varepsilon_{t-2}) = \frac{\mathrm{Cov}}{\sigma^2} = \rho^2$. Корреляция между двумя будет ρ^2 от единого параметра. Более того, $\mathrm{Corr}(\varepsilon_t, \varepsilon_{t-\tau}) = \rho^\tau$. Обращаем внимание на две вещи: корреляция между значениями ε, разнесёнными между τ шагами, описывается всевозможными произведениями промежуточных ρ, а всевозможный набор корреляций связан только со сдвигом во времени. Это удобное свойство, но тем самым мы полностью определили ковариационную матрицу:

$$\mathrm{Cov}(\vec{\varepsilon}) = \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon^2 & \rho^2\sigma_\varepsilon^2 & \ddots & \rho^{n-1}\sigma_\varepsilon^2 \\ \rho\sigma_\varepsilon^2 & \sigma_\varepsilon^2 & \rho\sigma_\varepsilon^2 & \ddots & \rho^{n-2}\sigma_\varepsilon^2 \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \rho^{n-2}\sigma_\varepsilon^2 & \ddots & \ddots & \ddots & \rho\sigma_\varepsilon^2 \\ \rho^{n-1}\sigma_\varepsilon^2 & \rho^{n-2}\sigma_\varepsilon^2 & \ddots & \rho\sigma_\varepsilon^2 & \sigma_\varepsilon^2 \end{pmatrix} = \sigma_\varepsilon^2 \begin{pmatrix} 1 & \rho & \rho^2 & \ddots & \rho^{n-1} \\ \rho & 1 & \rho & \ddots & \rho^{n-2} \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \rho^{n-2} & \ddots & \ddots & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \ddots & \rho & 1 \end{pmatrix} \circledast$$

Все $\rho\sigma_\varepsilon^2$ расположены на диагоналях, ближайших к главной, и так далее; пойдут $\rho^2, \ldots, \rho^{n-1}\sigma_\varepsilon^2$. По свойствам определителя Вандермонда это равно

$$\circledast \frac{\sigma_v^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \ddots & \rho^{n-1} \\ \rho & 1 & \rho & \ddots & \rho^{n-2} \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ \rho^{n-2} & \ddots & \ddots & \ddots & \rho \\ \rho^{n-1} & \rho^{n-2} & \ddots & \rho & 1 \end{pmatrix}$$

Вся матрица выражается через один параметр. Если мы знаем параметры, то мы даже знаем все последствия. Представим себе, что это матрица $\boldsymbol{\Omega} = \mathrm{Cov}(\vec{\varepsilon})$. Если мы, несмотря на наличие такой матрицы, отличной от $\sigma^2\boldsymbol{I}$, применяем OLS, то оценка остаётся линейной, несмещённой, а вот с эффективностью явно нет, так как надо применять теорему Эйткена.

Если мы применим рутинную оценку дисперсии одномерной регрессии $Y_t = \alpha + \beta X_t + \varepsilon_t$, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, то $\hat{\beta} = \frac{\sum x_t y_t}{\sum x_i^2} = \beta + \frac{\sum x_t\varepsilon_t}{\sum x_i^2} = \beta + \sum k_i\varepsilon_i$. $\mathrm{Var}(\hat{\beta}) = \sigma^2\varepsilon \sum k_i^2 = \frac{\sigma_\varepsilon^2}{\sum x_t^2} + 2\mathbb{E}\left(2\sum_{i<j} k_i k_j \varepsilon_i \varepsilon_j\right)$ из-за автокорреляции. Но здесь, в отличие от гетероскедастичности, мы можем сказать больше. Если у нас стандартный случай вроде инерции, то у нас ковариации положительного знака. Поэтому, не принимая во внимание слагаемого ковариаций, мы занижаем оценку. Нам кажется, что оценка t, делённая на дисперсию, выше, коэффициент значимее, а он на самом деле незначимый. Применяя МНК к автокорреляции, мы увидим, что коэффициенты позначимее. Если есть автокорреляция и коэффициент незначимый, то ничего страшного (если, конечно, нет cobweb-эффекта), а если наоборот, то плохо.

Мы начнём с нереалистичной процедуры, зная ρ. Если мы знаем ρ и σ_v^2, то выписываем сразу матрицу: $\tilde{\vec{\beta}}_{\mathrm{GLS}} = (X^T\Omega^{-1}X)X^T\Omega^{-1}Y$, $\mathrm{Cov}(\tilde{\vec{\beta}}_{\mathrm{GLS}}) = (X^T\Omega^{-1}X)^{-1}$. Считать обратную матрицу мы не хотим, лень — двигатель прогресса, и тут мы вспоминаем, что обобщённый МНК эквивалентен обычному, если предварительно преобразовать данные так, чтобы сделать одинаковую дисперсию. Если мы преобразуем данные так, чтобы можно было применить условия ТГМ, то мы получим удобную матрицу.

Начнём с простой одномерной регрессии: $Y_t = \alpha + \beta X_t + \varepsilon_t$, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$. v_t у нас статистически хороший, а нам надо исключить ε. Попробуем записать, как в прогрессии в школе: $Y_t = \alpha + \beta X_{t-1} + \varepsilon_{t-1}$, домножим на ρ и вычтем из изначального выражения: $\underbrace{Y_t - \rho Y_{t-1}}_{Y_t^*} = \alpha\underbrace{(1-\rho)}_{1^*} + \beta\underbrace{(X_t - \rho X_{t-1})}_{X_t^*} + v_t \Rightarrow$ OLS. Проблема: мы не знаем

t_1. $t = 2, \dots, T$. Выкидываем одно наблюдение и используем МНК. Дальше люди сказали: мы теряем одну точку, и надо бы правильно сделать обобщённый МНК для эффективности. Нужно что-то для первого наблюдения. Поэтому ввели **поправку Прейса—Уинстона**. Довольно просто её установить: Y_1^* зависит не от будущих, а только от первого наблюдения, да ещё так, чтобы в результате получилось так: $\vec{Y}^* = \Omega^{-0,5}\vec{Y}$. Поправка: $Y^* = \sqrt{1-\rho^2}Y_1$, $X_1^* = \sqrt{1-\rho^2}X_1$, $1^* = \sqrt{1-\rho^2}\cdot 1$. Если мы соберём всё вместе, то какое преобразование мы должны делать? Это явное линейное преобразование: $\vec{Y}^* = T\vec{Y}$. Матрица T должна быть чем-то, что мы умножаем на матрицу-столбец $\vec{Y}$. Поэтому

$$T = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & \ddots & 0 \\ -\rho & 1 & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots \\ 0 & \ddots & -\rho & 1 \end{pmatrix} = \Omega^{-1/2}.$$

Нижняя поддиагональ равна $-\rho$, главная — все единицы, кроме первого корня. Такое преобразование очень похоже на $Y_t - \rho Y_{t-1}$. Поэтому это называется **квазиразностью**. Вычитаем не всё, а чуть-чуть. И это наш выбор — выкидывать наблюдение или делать преобразование. Если же мы возведём $\Omega^{-1/2}$ в квадрат, то мы сможем применять GLS.

Конечно, вопрос — откуда взять ρ. Сначала надо посмотреть, есть ли автокорреляция. Потом, если нет, OLS нам всё даёт. Если есть, то надо оценить ρ, провести преобразования и применить МНК к преобразованным данным. Визуально мы уже видим, на зуб пробовать тоже хорошо, но лучше использовать статистическую процедуру. Её вывести очень нелегко. Первый статистический тест был разработан в 1953 году Д$_{\text{э}}$рбином и Уотсоном, а потом пошли Дарбин, Ватсон, Дурбин…

Что предложили **Дарбин и Уотсон**? dw-тест и его статистика. На первом шаге применяется МНК, и мы получаем e_t — остатки. Дарбин—Уотсон предложили рассмотреть следующую статистику:

$$\mathrm{dw} = \frac{\sum\limits_{t=2}^{T}(e_t - e_{t-1})^2}{\sum\limits_{t=1}^{T} e_t^2}$$

Нетрудно заметить, что эта статистика неотрицательна: $\mathrm{dw} \geqslant 0$. Если раскрыть квадраты сверху, то получится одна квадратичная форма от остатков, делённая на другую. Мы должны рассмотреть отношение двух квадратичный форм. Раскроем и получим:

$$\mathrm{dw} = \frac{\sum\limits_{t=2}^{T} e_t^2 + \sum\limits_{t=2}^{T} e_{t-1}^2 - 2\sum\limits_{t=2}^{T} e_t e_{t-1}}{\sum\limits_{t=1}^{T} e_t^2} = 2 - 2\frac{\sum\limits_{t=2}^{T} e_t e_{t-1}}{\sum\limits_{t=1}^{T} e_t^2}$$

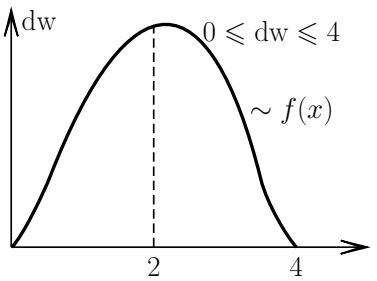

Рис. 22. Статистика Дарбина—Уотсона

Если наблюдений много, то слагаемые почти равны примерно, поэтому верно приближённое равенство, что напоминает выборочный коэффициент корреляции с точностью до концевых моментов. Выборочный коэффициент корреляции почти лежит между -1 и 1, и если он равен -1, то разность равна 4, а если равна 1, то в итоге 0. Имеем вывод: $0 \leqslant \mathrm{dw} < 4$. Тогда мы получаем, что вся статистика сосредоточена на интервале $[0; 4)$. Она имеет какой-то холмоподобный вид. Беда в следующем: статистика эта, в отличие от известных нам распределений, зависит от иксов, то есть она не инвариантна от наблюдений.

Эта функция зависит от матрицы X. Сегодня на некоторых программах нет проблемы построить распределение. А Дарбин и Уотсон — блестящие умы. В зависимости от икса это распределение бегает в некотором коридоре, сидит в некотором коридорчике, каким бы X ни был. В учебниках не отмечается или отмечается неправильно: это распределение несимметричное. Его мода больше, чем 2, и это важно.

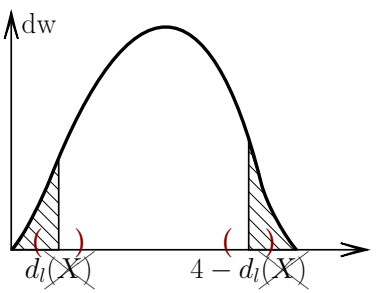

Рис. 23. Не зависящая от X зона

Должна быть некоторая критическая область для проверки гипотез: $\mathcal{H}_0$: нет автокорреляции, $\mathcal{H}_1$: есть автокорреляция. Это в марковской схеме эквивалентно: $\mathcal{H}_0\colon \rho = 0$, $\mathcal{H}_1\colon \rho \neq 0$. Дарбин и Уотсон рассматривали одностороннюю гипотезу: $\mathcal{H}_2\colon \rho > 0$, $\mathcal{H}_3\colon \rho < 0$. Получается так потому, что распределение несимметричное. Поэтому область $\mathcal{H}_2$ соответствует большим значениям dw, $\mathcal{H}_3$ — меньшим dw. А нолик где-то посередине. Критическая область — это квантиль от нуля до чего-то. Назвали его $d_l(X)$ (Дарбин-нижний, зависящий от X). И всё-таки эти значения

имеют какой-то интервальчик. Поэтому рассматриваются не зависящие от X границы критической области (только от α, T, k'). Поэтому надо ввести критическую область, в которой мы не принимаем решения. Поэтому три области: d_l, d_u, и наличествует следующее правило. Если выборочная статистика меньше, чем d_l, то отвергается нулевая гипотеза в пользу положительной автокорреляции. Если больше, чем d_u, то тогда нет автокорреляции. А есть зона отказа от решения. Граница критической области разъезжается на две подграницы, и они подсчитали эти границы в зависимости от числа наблюдений, доверительной вероятности и числа оцениваемых параметров. У них в таблицах $k-1=k'$ — коэффициенты без свободного члена. Если мы попали в $0 \leqslant d \leqslant d_l$, то положительная автокорреляция.

А что делать со второй группой гипотез? Там всё хуже, распределение асимметричное, но надо взять симметричную границу критической области: $4-d_l$ и $4-d_u$, то есть новых таблиц не делается. Общее правило: $d_u \leqslant d \leqslant 4-d_u$ — нет автокорреляции, потом зона неопределённости, а потом снова есть автокорреляция.

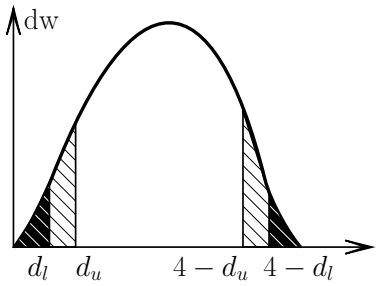

Рис. 24. 5 зон Дарбина—Уотсона

Есть незримые assumptions. Эти таблицы построены в предположении нормальности v. Более существенно другое: при выводе этого распределения явно используется предположение, что модель содержит свободный член. Нельзя применять dw-статистику, если в модели нет свободного члена, и это большая неприятность, потому что элементарное уравнение инфляции часто пишется без свободного члена. В этом случаем имеем огромное количество случаев неправильного применения данной статистики. Второе: dw-статистика годится только для марковской схемы — авторегрессии первого порядка. Статистика эта не работает, если у нас есть GARCH-эффект или даже ARCH-эффект, то есть гетероскедастичность условной дисперсии.

Если мы по dw-статистике определили, что автокорреляция есть, то мы должны применять ОМНК или знать ρ. Подход Дарбина—Уотсона ненамеренно даёт нам оценку ρ. Это коэффициент корреляции между ε_t и ε_{t-1}, что есть выборочный коэффициент корреляции:

$$\tilde\rho = \frac{\sum\limits_{t=2}^{T} e_t e_{t-1}}{\sum\limits_{t=1}^{T} e_t^2}, \quad d = 2-2\tilde\rho \Rightarrow \tilde\rho = 1-\frac{d}{2}$$

Нам нужно, чтобы результат, который мы получим, был хотя бы состоятельным. Всё зависит от α, T, k'.

Итак, схема: считаем OLS, остатки, проверяем статистику Дарбина—Уотсона и проверяем статистику против гипотезы. Нет проблемы: гипотеза односторонняя. Если мы не отвергаем нулевой гипотезы, то оценка — это то, что мы искали. Если нет, то берём ρ и делаем поправочки: $Y_t^* = Y_t - \tilde\rho Y_{t-1}$, $Y_1^* = \sqrt{1-\tilde\rho^2}Y_1$. Если $\tilde\rho$ — состоятельная оценка оценка истинного ρ, то в итоге мы получим состоятельные оценки наших коэффициентов, асимптотически эффективные.

На этом месте у лектора зазвонил телефон, и лекция кончилась.

21 Лекция 21

Нас ждут скоро половые праздники! — Не половые, а гендерные!

Итак, процесс вида $\vec Y = X\vec\beta + \vec\varepsilon$, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, $v_t \sim$ WN. Итак, $AR(1)$, $|\rho|<1$, $\sigma_\varepsilon^2 = \frac{\sigma_v^2}{1-\rho^2}$, находится $\mathrm{Cov}(\varepsilon)$. $Y_t^* = Y_t - \rho Y_{t-1}$, $Y_1^* = \sqrt{1-\rho^2}\cdot Y_1$. Для определения наличия автокорреляции строится статистика Дарбина—Уотсона: $\mathrm{dw} = \frac{\sum(e_t-e_{t-1})^2}{\sum e_t^2}$, сосредоточена она на интервале $[0;4]$, и Дарбин—Уотсон делят всю область на критическую область, неопределённую и некритическую. $\mathcal{H}_0\colon \rho=0$, $\mathcal{H}_1\colon \rho>0$ для левого хвоста, $\mathcal{H}_0\colon \rho=0$, $\mathcal{H}_1\colon \rho<0$ для правого хвоста. Есть пакет «SHAZAM», он рассчитывает под конкретный X. Если мы оптимист и мы попали в зону неопределённости, любим риск, говорим, что проще принять автокорреляцию за некритическую и отвергнуть. А пессимисты зону неопределённости трактуют как зону с автокорреляцией. Поправка Прейса—Уинстона существенна; она улучшает качество нашей оценки. Одной из оценок ρ является $\tilde\rho = 1-\frac{d}{2}$. Если $\rho < 0{,}3$, то численные изменения результатов несущественны.

Когда не работает статистика Дарбина—Уотсона? Когда в модели нет свободного члена, то есть нет столбца единиц. Некто Fairbrother вывел статистику dw для моделей без свободного члена. Вторая существенная особенность: в матрице X нет стохастических регрессоров, особенно Y в предыдущий момент времени. Регрессоры детерминированы. Если среди X встречаются Y_{t-i} (предыдущие наблюдения), то статистика dw неприменима. Если в EViews мы посчитаем регрессию, он автоматически посчитает dw, не защищая нас от соблазнения наделать ошибок. Кроме того, dw годится только для авторегрессии первого порядка. И в малых выборках всё осложнено.

Есть **поправка Тейла—Нагара** (Theil-Nagar): $\tilde{\tilde\rho} = \frac{n^2(1-0{,}5d)+k^2}{n^2-k^2} = \frac{n^2(1-0{,}5d)+(k'+1)^2}{n^2-(k'+1)^2}$, $k'=k-1$. Когда n^2 значительно больше, чем k^2, то формула эквивалентна предыдущей. k — это количество регрессоров с учётом свободного члена. При маленьких n это существенно.

Если есть уравнение регрессии, как в начале лекции, то $Y_t - \rho Y_{t-1} = \alpha(1-\rho) + \beta(X_t - \rho X_{t-1}) + v_t$, неизвестны α, β, ρ, σ_v^2, $t = 2, \ldots, T$. Перенесём: $Y_t = \alpha(1-\rho) + \rho Y_{t-1} + \beta X_t - \rho\beta X_{t-1} + v_t$. Мы исключили зависимость от предыстории у ε, но выскочил стохастический регрессор Y_{t-1}, что есть автокорреляция. По-прежнему мы хотим оценить эту регрессию. Это более сложное уравнение, чем квадратичная. Система нелинейная. Входят α, ρ, которые как-то завязаны. Система неприятная: у неё гарантированно не одно решение (а до пяти). И то, куда мы численным методом свалимся, существенно. А нельзя ли применить метод максимального правдоподобия? Предположим нормальное v, запишем функцию распределения, но технически Y будут выглядеть нелегко.

В 60-е годы придумана **процедура Кокрейна—Оркутта** (Кохрана—Оркатта и прочие вариации на тему). Пишутся они нестандартно и тот, и другой: Cochrane—Orcutt. Если есть оценка ρ, то к оставшемуся уравнению МНК применим. Процедура: на первом шаге проводится обычный МНК. OLS даёт остатки, по ним мы оцениваем $\tilde{\rho}_1$. Это первая итерация. На втором шаге мы рассчитываем α, β и все остальные, игнорируя первую поправку или учитывая её. Получим $\alpha_1, \beta_1, \ldots$ по FGLS. После этого у нас получаются новые остатки. Считаем новые $e_t^1 = Y_t - \tilde{\alpha}_1 - \ldots$. Мы получили новые остатки, мы можем найти новую оценку $\tilde{\rho}_2 \to FGLS$, получаем новые остатки… И пока не сойдётся. Вот это и есть итерационная схема.

Если схема $AR(1)$, то у нас присутствует коэффициент корреляции между ε_t и ε_{t-1}. Предлагается посчитать $\frac{\sum e_t e_{t-1}}{\sqrt{\sum e_t^2}\sqrt{\sum e_{t-1}^2}} = \tilde{\rho}$. Имея эту оценку, мы можем в обратную сторону всё пересчитать на основе выборочного коэффициента корреляции.

Если есть нормированная и центрированная регрессия, то тогда коэффициент регрессии равен выборочному коэффициенту корреляции. $e_t = ae_{t-1} + w_t$, $\hat{a} = \tilde{\rho}$. Почему бы не проверить простую гипотезу, что $\mathcal{H}_0 \colon a = 0$, $\mathcal{H}_1 \colon a > 0$. Но так делать нельзя!!! Что нарушается в этом предложении? Оказалось, что ситуация с решением этой итерационной задачи — это встроенная схема решения чего-то вот такого численного, поэтому у неё могут быть такие трудности, как неединственность решения (сойдётся, но неизвестно к чему), а может быть вообще несходимость. Может быть колебательность или расходимость. Люди задают ρ, генерируют данные, а потом пытаются определить ρ по методу Монте-Карло. И иногда парадоксально: две итерации лучше, чем бесконечное количество. То есть ρ_1, ε_1, ρ_2, ε_2. В пакетах не выписывается, как строится Кокрейн—Оркутт.

На заре вычислительной техники был предложен метод Хилдрейт—Лю (Hildreth—Lu). Это поиск ~~наседки~~ на сетке, или на решётке. Понятно, что $-1 \leqslant \rho \leqslant 1$, поэтому идея такая: давайте пробежимся по этой единице с каким-то шагом. Считать было не так-то просто, поэтому начинали с одной десятой. $\rho_i = -1 + 0{,}1i$. Найдём $\tilde{\tilde{\beta}}_i$, и для каждого $\tilde{\tilde{\beta}}_i$ найдём RSS. Потом возьмём более мелкий интервал. Grid search procedure. Для одномерной переменной каждая решётка — это набор точек. Находим окрестность, проводим дробное деление, находим оценки $\tilde{\rho}$ и радуемся.

Рис. 25. Поиск наседки на решётке

Самая главная неприятность, которая ждёт нас с подобными процедурами, — ограничения. Давайте с ними бороться. Самое важное: а вдруг это не марковская схема, а $AR(p)$? В инфляционных рядах это общий случай. В модель приходилось включать до 20 лагов. Была предложена двухшаговая процедура Дарбина: если мы предполагаем, что есть автокорреляция, то мы на первом шаге строим не МНК, а модель типа (ссылка на формулу, где регрессор Yt-1), и её оцениваем МНК. ρ берётся в качестве оценки неизвестного параметра автокорреляции, а на втором шаге используется GLS. Сегодня основной аппарат, с которым работают, — метод множителей Лагранжа Бройша—Годфри. Сделав ММП оценку $\ln L$, мы проверяем оценку $\theta = \theta_0$. Ограничение: $\rho = 0$, и метод Лагранжа приравнивает loglikelihood к нулю. Этот метод применим для нахождения автокорреляции любого порядка.

Общая постановка теста: $Y_t = \sum \beta_i Z_{t,i} + \varepsilon_t$. Z_i — регрессоры, $\varepsilon_t = \rho_1\varepsilon_{t-1} + \rho_2\varepsilon_{t-2} + \ldots + \rho_p\varepsilon_{t-p} + v_t$, $v_t \sim WN$ — схема $AR(p)$. В числе $Z_{t,i}$ могут быть $X_{t,i}$, Y_{t-l}, X_{t-l}, поэтому появляются переменные с лагом, лаговые переменные (backward lagged variables). Это куда более общая схема. Если мы оценим коэффициенты, то выскажем гипотезу: $\mathcal{H}_0 \colon \rho_1 = \ldots = \rho_p = 0$, $\mathcal{H}_1 \colon \sum\limits_l \rho_p^2 > 0$. Гипотеза с ограничениями. Мы рассматривали Вальда, Лагранжа, правдоподобия. Лагранж строит только регрессию с ограничениями. А это регрессия, когда нет автокорреляции. Строим OLS, и если гипотеза не отвергается, то это окончательный итог. Предлагается построить регрессию e_t на все $Z_{t,i} + r_1 e_{t-1} + \ldots + r_p e_{t-p} + u_t$. Понятно, что для первых p точек мы наблюдения потеряем. Это асимптотический тест, поэтому точек должно быть много (финансовые ряды). $\mathcal{H}_0 \colon r_1 = \ldots = r_p = 0$. Вторую модель мы не оценим, она не пройдёт. Далее строится простая статистика: $nR^2 \overset{\text{as}}{\sim} \chi_p^2$. Это основной способ проверки автокорреляции. При $p = 1$ он работает так же, как Дарбин—Уотсон, и он справедлив асимптотически. Хорошо, если мы не отвергаем нулевой гипотезы, то на первом шаге модель у нас нужная. А если мы не отвергаем гипотезы, то тогда что-то не равно нулю. Но что именно? Можно попытаться построить автокорреляционную матрицу: $Y_t - r_1 Y_{t-1} - \ldots$, и это сложно и никому не нужно.

Оказалось, что мы смогли переписать свою модель без автокоррелированных ε, записав модель с таким преобразованием, чтобы убежали ε: $Y_t - \rho_1 Y_{t-1} - \ldots - \rho_p Y_{t-p} = \ldots + v_t$. Это будет регрессия на предыдущие значения. Она должны быть без автокорреляции, и это по-прежнему будет похоже на регрессию. Добавле-

ние в правую часть переменных с лагом убивает автокорреляционную схему. На этом построена борьба с автокорреляцией. Вспомним МИП (см. стр. 15): надо избавиться от коррелированных переменных, добавив лаговые.

Ещё одно изобретение Дарбина — альтернативная статистика Дарбина. Её почти вытеснил метод множителей Лагранжа. h-статистика Дарбина предложена тогда, когда среди объясняющих регрессоров есть Y_{t-1}/ Первый шаг: строим OLS для всей регрессии, а потом берём статистику: $h = \tilde{\rho}\sqrt{\frac{n}{1-n\operatorname{Var}(\hat{\gamma})}}$, где $\hat{\gamma}$ — оценка коэффициента при Y_{t-1}. $\tilde{\rho}$ — любая оценка статистики dw. Дарбин показал, что если $\rho = 0$, то $h \overset{\text{as}}{\sim} \mathcal{N}(0;1)$. Даже таблицы не нужны. Бяка-то какая здесь? Под корнем стоит нечто, которое обязано быть положительным. Если дисперсия велика, то процедура не работает.

При борьбе с гетероскедастичностью мы считали самыми хорошими оценки Уайта. Дисперсии ошибок по-другому строятся. И Heteroscedasticity Consistency Estimator работает и при гетероскедастичности, и без неё. Нельзя ли придумать подобное для автокорреляции? Это было названо LM-оценка. В оценке Уайта суммируется такое слагаемое, как $\sum e_s^2 X_s^T X_s$, где X_s — строка. Для автокорреляции **Ньюи—Уэста** (Newey—West) $\sum e_s e_\tau$ считается с весами, где некоторые могут равняться нулю. HACE — это ещё с Autocorrelation. Там есть ширина окна и асимптотическая сходимость, и это не так повсеместно.

22 Лекция 22

«Ну вот, теперь у уборщицы не осталось шансов!»

Ошибки в спецификации моделей. Мы не знаем, какая модель истинная, поэтому задача исследователя — определить, какая модель является лучшей.

Первое требование — *простота и экономичность*. Это требование, которое говорит: если мы получим более-менее одинаково качественные модели с хорошим качеством подгонки и похожими наборами объясняющих переменных, то тогда имеет смысл отдать предпочтение более простой модели. Простая модель более ясна для интерпретации; простые зависимости более устойчивые. Робастная оценка — это грубое оценивание, при котором даже изменение предпосылок почти не влияет на результаты. А тонкий метод годится, только если мы угадали закон распределения. И переменных должно быть как можно меньше. Чтобы оценить зависимость зарплаты от возраста, учитывается зарплата и зарплата в квадрате. А если возраст не очень разбросан, то тогда высокая линейная зависимость между возрастом и возрастом в квадрате. Поэтому тогда используют линейную модель.

Однажды прогнозировали спрос для РАО ЕЭС. У Коссовой была группа из трёх студентов. Было 80 регионов. Одна девочка включила в модель температуру в сороковой степени. Она говорит: а R^2 такой высокий! Но если немножко не угадать…

Принцип идентифицируемости параметров. Должна быть возможность определить все параметры. Такой возможности нет, когда чистая мультиколлинеарность. Пусть $X_i^2 = aX_i^1 + b$, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Так беты определить нельзя. А можно подставить выражение: $Y = \beta_0 + \beta_1 X_i^1 + \beta_2 a X_i^2 + \beta_2 b + \varepsilon = \beta_0 + (\beta_1 + a\beta_2)X_1 + \beta_2 b + \varepsilon$. Если бы мы решали МНК, то матрица $X^T X$ была бы вырожденной, и было бы бесконечное множество решений. Мы стремимся к идентифицируемости параметров. У ММП была инвариантность: если мы оценили сумму и разность, то мы можем из восстановить. Но в МНК функция от оценки не есть оценка от функции.

Следующее требование — *качество подгонки модели*. Для нас оптимальней та, у которой более высокий goodness of fit. Это R^2 и $\bar{R}^2$. Но эти два показателя для не очень сведущих людей. Для умненьких есть F-статистика, которая вообще позволяет не смотреть на эр квадрат. Если же у нас в модели есть мусорные переменные, то тогда будет ниже F-статистика.

Теоретическая состоятельность модели. Это очень важное требование; студентов учат математике, а здравый смысл стоит далеко не на первом месте. Иногда модели очень трудно сравнить, и тогда руководствуются этим принципом. От модели мы чего-то ожидаем: если изучается спрос, то у цены коэффициент отрицательный, у дохода положительный. Если результат другой, то с моделью что-то не так. Если функция Кобба—Дугласа, то логично ожидать постоянную отдачу. Если же результат противоречивый, то или что-то не так с моделью, или существующая теория неверна. Может, мы строили модель с целью опровержения экономической теории. Недавно Коссова опровергала эмпирический факт лаборатории труда: умеренное потребление алкоголя положительно влияет на заработную плату. Коссова решила, что надо срочно заняться опровержением неправильной модели. Совсем продвинутый метод не сошёлся, а более простой метод дал незначимый, но отрицательный коэффициент.

Следующее свойство — *прогнозная сила модели*. Что это? Две цели обычно у эконометрического исследования: одна — объяснение, вторая — прогноз. Иногда эти цели противоречат друг другу. Если они совпадают, то это исключение. Для прогноза важна устойчивая тенденция (простота модели), а для качества подгонки нужна сложность. Полиномы где-то на прогнозе могут увести, а в линейной модели отклонение несущественно. Когда поступает новое наблюдение, то тогда смотрят, насколько они отличаются от прогнозируемых. Если нет новых наблюдений, разбивают выборку на две части, большую и маленькую. И смотрят, как основная часть прогнозирует остаточек.

Рассмотрим, какие ошибки над подстерегают при построении модели. Итак, что может быть с моделью не так? Первый пункт Гаусса—Маркова — это правильная спецификация. Мы можем оценивать некую функциональную

форму (предпочитаем линейную), а она может оказаться логлинейной. Спрос может быть с постоянной эластичность, и тогда надо оценивать $\ln Q$ от $\ln P$ и логарифма дохода. А форма может быть и линейной. Также бывает недоопределённость и переопределённость.

Начнём с недоопределённости модели: модель не включает существенных переменных (omitted variables). Правильная модель будет с бетами, неправильная — с альфами. Правильная: $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$. Неправильная модель: $Y = \alpha_0 + \alpha_1 X + u$. Тогда $\hat\alpha_1 = \frac{\sum y_i x_i}{\sum x_i^2}$. Запишем в отклонениях: $y = \beta_1 x + \beta_1 z + \tilde\varepsilon$, где $\mathbb{E}(\tilde\varepsilon) = 0$. $x_i = X_i - \bar X$. Тогда $\hat\alpha_1 = \frac{\sum \beta_1 x_i^2 + \beta_2 \sum z_i x_i + \sum \tilde\varepsilon_i x_i}{\sum x_i^2} = \beta_1 + \beta_2 \frac{\sum z_i x_1}{\sum x_i^2}$. Проматожидаем: $\mathbb{E}(\hat\alpha_1) = \beta_1 + \beta_2 \frac{\sum z_i x_i}{\sum x_i^2} + 0$. При оценке укороченной регрессии оценка будет смещённой: её матожидание должно быть β_1, а она содержит ещё добавку. На самом деле эта дробь — это коэффициент наклона в регрессии z на x. Если $z = \gamma_0 + \gamma_1 x + v$, то $\hat\gamma_1 = \frac{\sum z_i x_i}{\sum x_i^2}$. Смещение происходит на величину, пропорциональную коэффициенту наклона z на x. Когда у нас такой счастливый случай, что выборочная ковариация равна нулю или величины ортогональны, то тогда смещения нет. Остаётся надежда, что смещения нет, если мы забыли включить что-то ортогональные. В методе главных компонент у нас все переменные ортогональные, и если мы выбросим хвост, то смещения не будет. Если переменные забыть не по-умному, а из-за отсутствия данных, то смещение есть. Да и у коэффициента $\hat\alpha_0$ есть смещение:$\mathbb{E}(\hat\alpha_0) = \beta_0 + \beta_2(\bar Z - \hat\gamma_1 \bar X)$, где $\hat\gamma_1$ мы уже ранее описали: $\mathbb{E}(\hat\alpha_0) = \beta_0 + \beta_2 \left(\bar Z - \bar X \frac{\sum z_i x_i}{\sum x_i^2}\right)$. Если оценки смещены, то дисперсии можно уже не рассматривать, потому что нет состоятельности. Но всё-таки что теперь происходит с дисперсиями? Если оценки смещены и мы видим, что смещение совершенно не собирается исчезать с ростом n, то оценки не будут состоятельными. У Чебышёва вероятность отклонения оценки от матожидания стремится к нулю при $n \to \infty$. Но матожидание у нас отнюдь не β_1, состоятельности не будет. А состоятельность — это минимальное требование! Рассмотрим дисперсию $\hat\alpha_1$. $\widehat{\mathrm{Var}(\hat\alpha_1)} = \hat\sigma^2 \frac{1}{\sum x_i^2}$. $\widehat{\mathrm{Var}(\hat\beta_1)} = \hat\sigma^2 \frac{1}{\sum x_i^2 (1 - r_{zx}^2)}$. Дисперсия в неправильной модели более оптимистична, но $\hat\sigma^2$ разные! RSS в короткой модели всегда больше, чем в длинной, и возникает два противоположных эффекта. Оценка для дисперсии больше, а дробь меньше. Поэтому может быть как недооценка, так и переоценка дисперсии наклона. Оценки перестают быть несмещённые, теряют состоятельность, а об эффективности не может быть и речи.

Ходжа Насреддин рассказывает, что как-то раз поспорил с эмиром бухарским, что научит своего ишака грамоте. На это нужен кошелёк золота и двадцать лет времени. Если он не выполнит условия спора — голова с плеч. Насреддин не боится неминуемой казни: «Ведь за двадцать лет, — говорит он — кто-нибудь из троих обязательно умрёт — эмир, ишак или я!»

(Тут из-за сбоя TEXmaker кусок мог пропуститься, но немного, минут пять. У кого есть???)

Первая модель: $\hat Q_d = 89{,}97 + 0{,}107P$, статистики: 11,85, 0,118, $\hat\sigma^2 = 2{,}338$. $\hat Q_d = 92{,}05 - 0{,}142P + 0{,}263I$ с t-статистиками 5,84, 0,067, 0,037 и $\hat\sigma^2 = 1{,}952$. Получается, что настолько кардинально поменялись результаты. Изначально была переоценка модели.

В правильной модели беты, в неправильной альфы. $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \beta_{k+1} Z + \varepsilon$, $Y = \alpha_0 + \alpha_1 X_1 + \ldots + \alpha_k X_k + u$, $\mathbb{E}(\hat\alpha_j) = \beta_j + \beta_{k+1}\gamma_{zx_j}$. Таково смещение.

Если у нас нет возможности включить важные переменные (классический пример — уравнение заработной платы от интеллектуального потенциала. И часто вместо переменной, которую мы можем измерить, мы включаем инструменты, которые сильно коррелированы с той переменной, которую мы не можем пронаблюдать. В качестве инструмента включается образование. И коэффициент при этой переменной неважен. Его включают только для того, чтобы в модели было минимальным смещение. Когда оценивают зарплату, то нынешние исследователишки уже забывают о том, что это прокси, и интересуются коэффициентом. Сегодня есть модная проблема: если кто-то предлагает модель, то ему задают вопрос: а вдруг у вас есть эндогенная переменная, которая связана со случайной ошибкой ε? Только мы взяли инструмент, как он от чего-то зависит. С образованием то же самое: в уравнении дохода образование эндогенная, связанная с ошибкой. Если человек более способен, то у него зарплата и способность выше. Что тогда делают? Какой инструмент берут? Вот есть человек, есть образование. Тогда берут экзогенное образование жены или родителей. В эту игру можно долго играть, подбирая инструменты для инструментов. У женщин образование мужа — очень хороший инструмент. А опыт нехорошо коррелирует с возрастом.

Рассмотрим вторую ошибку — включение несущественной переменной. В уравнение спроса хочется включить субституты, но главное — вовремя остановиться, потому что список субститутов бесконечен. Включают много, и переменные становятся лишними. Запишем неправильное уравнение с альфами.

$Y = \beta_0 + \beta_1 X + \varepsilon$, $Y = \alpha_0 + \alpha_1 X + \alpha_2 Z + u$, $\sum yx = \hat\alpha_1 \sum x_i^2 + \hat\alpha_2 \sum xz$, $\sum yz = \hat\alpha_1 \sum xz + \hat\alpha_2 \sum z^2$. Запишем в отклонениях: $\sum(\beta_1 x + \tilde\varepsilon)x = \alpha_1(\ldots)$, $\sum(\beta_1 x + \tilde\varepsilon)z = \hat\alpha_1(\ldots)$. Тогда $\beta_1 \sum x_i^2 = \mathbb{E}(\hat\alpha_1)\sum x_i^2 + \mathbb{E}(\hat\alpha_2)\sum xz$, $\beta_1 \sum x_i z_i = \mathbb{E}(\hat\alpha_1)\sum xz + \mathbb{E}(\hat\alpha_2)\sum z_i^2$.

$\mathbb{E}(\hat\alpha_1) = \frac{\beta_1 \sum x_i^2 - \mathbb{E}(\hat\alpha_2 \sum xz)}{\sum x_i^2} = \beta_1 - \mathbb{E}(\hat\alpha_2)\frac{\sum xz}{\sum x_i^2}$. $\beta_1 \sum z_i x_i = \beta_1 \sum z_i x_i - \mathbb{E}(\hat\alpha_2)\frac{(\sum xz)^2}{\sum x_i^2} + \mathbb{E}(\hat\alpha_2 \sum z_i^2)$. У модели сумм ранг ноль. Посчитайте $\mathbb{E}(\hat\alpha_2) = 0 = \mathbb{E}(\hat\alpha_2)\left(-\frac{(\sum xz)^2}{\sum x_i^2} + \sum z_i^2\right)$.

23 Лекция 23

$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \varepsilon_i$, где на X_2^i влияет Z_i. Строится $Y_i = \alpha_0 + \alpha_1 X_i^1 + w_i$, по OLS получаются $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\sigma}^2_{\alpha_0}$, $\hat{\sigma}^2_{\alpha_1}$. Тогда нужны прокси.

Можно включить лишнюю переменную. Рассмотрим чистую схему: одна схема — мы ставим две. Оценки оказываются несмещёнными (вот так ошиблись), а $\hat{\sigma}^2_\alpha$ другие. Падает качество оценки коэффициентов. Падает точность оценок. При этом считается, что $\sigma^2_\varepsilon = \sigma^2_w$. Но остаточные суммы там $\hat{\sigma}^2_w$ и $\hat{\sigma}^2_e$. Невключение переменной представляется куда более существенной ошибкой, чем включение лишних переменных.

Пример. Целью было поставить функцию спроса на импортные товары в США. В сегодняшнем виде куда важнее спрос импортных товаров. Мерится объём импорта. А к чему его прилагать — к спросу или предложению — вопрос. Цены перевели в цены 1982 года. Построили зависимость импорта от какой-то характеристики дохода и цен. Построили зависимость от подушевого располагаемого дохода. $\hat{Y}_t = -271{,}1_{31{,}3} + 0{,}245\,2_{0{,}014\,8} X_t$. Всегда $\bar{R}^2 \leqslant R^2$. $R^2 = 0{,}938\,8$, $\bar{R}^2 = 0{,}9354$, $d.w. = 0{,}5951$. Смущает маленькое значение Дарбина—Уотсона. Значит, положительная автокорреляция. Надо бороться. Надо ещё включить цены импорта. Их трудно посчитать, так как надо взвешивать по объёму и по времени. Мы вот взяли и заменили уровень цен простой линейной зависимостью. Другое уравнение, уже с прокси: $\hat{Y}_t = -859{,}9_{31{,}3} + 0{,}647_{0{,}074} X_t - 23{,}2_{4{,}27} t$. $R^2 = 0{,}977$, $\bar{R}^2 = 0{,}975$, $dw = 1{,}36$, $\hat{\sigma}^2 = 184{,}05$. Теперь 64 цента из каждого доллара тратится на импортные товары. Не 24,5. Несуществующую переменную мы заменили прокси, и улучшилась оценка. В остатках первого уравнения сидит невключённая переменная и создаёт кажущуюся автокорреляцию. Поэтому Дарбин—Уотсон говорит как об автокорреляции, так и о пропущенной переменной. Но плоха неясность: сначала боролись Кокрейном—Оркуттом, а потом переменную добавляли. Плохая dw-статистка требует подхода и рассмотрения.

Третий тип ошибок спецификации — неправильная функциональная форма. На практике мы сужаем всё и считаем, что знаем все X_k, Но не знаем, в какой форма. $Y_t = \alpha + \beta \frac{1}{X_t} + \varepsilon_t$, $Y_t = \alpha + \beta X_t + w_t$ — модели, по-разному работающие на разных дистанциях. Та же разница: $Y_t = \alpha + \beta X_t + e_t$, $\ln Y_t = \alpha + \beta \ln X_t + w_t$. Тогда BLUE не будет. Трудно придать трактовку коэффициенту $\hat{\beta}$. Не надо писать, что $\mathbb{E}(\hat{\beta}) = \beta$, так как там это предельная производительность, а там эластичность. Мы рассматривали обычно последствия, когда мы знаем, что правильно. А в жизни мы вообще ничего не знаем.Мы стоим перед model selection. Можно написать: $\mathcal{H}_0$: короткая модель, $\mathcal{H}_1$: длинная. Мы знаем, как с этим работать. Если есть упущенные переменные, то можно рассмотреть короткую как длинную с ограничениями на параметры. Эти две гипотезы являются nested — из одного гнезда. Поэтому естественнее писать так, по-другому: $\mathcal{H}_0$: длинная модель, $\mathcal{H}_1$: короткая модель. Если есть альтернатива, что одна переменная не влияет на Y, то это одно ограничение. Нулевой гипотезе соответствует модель $Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$, при $\beta_k = 0$, альтернативной — $Y = \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon$ без ограничений.

$$t = \frac{\hat{\beta}_k}{\text{s.e.}(\hat{\beta}_k)} \overset{\text{as}}{\sim} t_{n-k}$$

Проверяется гипотеза об излишнести переменной. Если группа переменных излишняя, то тогда $\mathcal{H}_0$: $\beta_s = \beta_{s+1} = \ldots = \beta_k = 0$, $\mathcal{H}_1$: $\beta_s^2 + \ldots + \beta_k^2 > 0$. Проверяется это по F-статистике. Тогда

$$\text{F} = \frac{(\text{RSS}_{\text{restr}} - \text{RSS}_{\text{unrestr}})/(k-s+1)}{\text{RSS}_{\text{unrestr}}/(n-k)} \overset{\text{as}}{\sim} \text{F}_{k-s+1;n-k}$$

Мы знаем способы тестирования двух гипотез — как проверить, что группа переменных является излишней. Если одна, то t, если несколько, то F. Если лишние, то мы их исключаем и пересчитываем регрессию. Ещё надо помнить: если две переменные незначимые, то тогда не факт, что обе в сумме незначимы. Но это мешает нам содержательно трактовать. Если уровень осадков в Австралии незначим, то он незначим ни просто так, ни с мировой инфляции (задание: придумать, что это такое). Но в регрессии далеко не так. Поэтому есть stepwise regression — устроена просто. Вот у меня 40 переменных, не хватает степеней свободы. Тогда найдём корреляцию Y с каждым иксом, построю там, где самая большая корреляция, потом по какому-то алгоритму будет выбрана вторая переменная, потом будет выбрана третья. Современная логика: построим по всем 40, а теперь по какому-то правилу (например, по плохости t-статистики) будем выбрасывать. Но эта процедура неудовлетворительная: когда наращиваем переменные, то тогда результаты по излишне коротким регрессиям смещённые. А если выкидывать лишние, то смещения нет, но зачем же по одной?

Мы знаем: $R^2_{\text{adj}} = \bar{R}^2$, и так мы и отвечаем: та модель предпочтительнее, у которой выше R^2. Мера $\bar{R}^2$ введена Тейлом: $\bar{R}^2 \uparrow \Leftrightarrow \hat{\sigma}^2$ ниже. Если $|t| < 1$, то лучше короткая модель. Добавили переменную — потеряли df. Если $|t| > 1$, то лучше длинная модель. $|t| = 1$ — одинаковы. Но присутствует логическая сложность: на уровне 5 % критическим значением было $t \approx 2$. Мы говорили: если 1,8, то переменная незначима на уровне 5 %. Но это больше единицы! Если выбросить её, то это хуже в плане $\bar{R}^2$. Где-то надо применять априорные данные. Может, мы оцениваем производственную функцию, а там линейно-однородная функция, и там нужна такая-то политическая мера.

Если мы выбросим любую переменную, где t меньше единицы по модулю, то это улучшение. Вторую — тоже. А обе? Неизвестно. А что? Давайте- ка мы построим все модели и выберем ту, у которой R^2. А это 2^n штук, что компьютер считает легко. Надо просто смотреть, что лучше оценивает нашу модель. Результата добились

Портер и Рао (другой). Он, как Леонид Витальевич Канторович, однофамилец. Правило «корень из пэ».

Пусть построена модель с k переменными. Если $\exists p$ переменных, таких, что $|t| < \sqrt{p}$ (t-статистика для каждого коэффициента), то удаление группы этих переменных может привести к улучшению $\bar{R}^2$, что то же, что уменьшение $\hat{\sigma}^2$. Если такой группы нет, то и не ищите, не пробуйте.

Пусть у нас 5 переменных. $|t_1| = 1{,}2$, $|t_2| = 1{,}5$, $|t_3| = 1{,}6$, $|t_4| = 2{,}3$, $|t_5| = 2{,}7$. Выброс одной переменной ничего не даст: $\nexists X_i \colon |t| < 1$. Двух тоже нет. А вот для трёх есть! Выбросить первую, вторую, третью. Выброс $(1; 2; 3)$ может привести к рассмотрению модели $Y = \gamma_0 + \gamma_1 X_4 + \gamma_2 X_5 + \varepsilon$. Других групп нет. Надо, значит, рассматривать только две модели. Длинную со всеми или короткую без первых трёх. Итак, у нас есть достаточно простой механизм проверки гипотезы о том, что в модель включены лишние переменные.

Но в начале прошлой лекции было: а вдруг даже в самой большой мы включили не всё? Во-первых, новая роль dw-статистики. Если она лежит в критической зоне, то это свидетельство того, что есть невключённые переменные. Ага, если dw хороший, то опасность меньше. Он вообще вводился для временных рядов, а лишняя переменная — это cross-sections. Это такой сигнал: ага, что-то надо придумать.

Второй случай — достаточно общий тест (со всеми вытекающими недостатками) Рамсея. RESET-тест. Regression Specification Error Test. Перезагрузка: удачное название — половина дела. Первый шаг: $\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$, получаем $\vec{e}$. Второй шаг: строится регрессия: $\vec{Y} = X\vec{\beta} + \gamma_1 \hat{Y}^2 + \gamma_2 \hat{Y}^3 + \ldots + \gamma_{p-1}\hat{Y}^p + w$, обычно до четырёх степеней. Если бы были лишние переменные (вот, Госкомстат ещё добавил), то ладно. А тут ситуация наоборот: группа товарищей $\gamma + p\hat{Y}^p$ аппроксимирует пропущенные переменные. Итак, $\mathcal{H}_0 \colon \gamma_1 = \gamma_2 = \ldots = \gamma_p = 0$, $\mathcal{H}_1 \colon \gamma_1^2 + \ldots + \gamma_p^2 > 0$. Смотрится F:

$$F = \frac{(\mathrm{RSS}_{\mathrm{restr}} - \mathrm{RSS}_{\mathrm{unrestr}})/p}{\mathrm{RSS}_{\mathrm{unrestr}}/(n-p)}$$

Это 1969 год. И далёкие от математики говорят: степенной ряд аппроксимируют функцию. Тут и тест максимального правдоподобия, и тест Вальда годятся. Можно тест отношений правдоподобия, но это может оказаться из пушки по воробьям. Если мы вспомним теорему Фриша—Вауга (Во), то можно строить регрессию не на Y, а на остатки. И если статистика улетает, то есть пропущенные переменные, поищи, дружок, ещё.

Чтобы сравнивать линейную и логлинейную модель, можно действовать по Боксу—Коксу, и можно как R^2, так и $\bar{R}^2$. 5 апреля будут две метрики.

24 Лекция 24

Канторович принёс тетрадку, конспекты в которой писала ещё его родительница.

Модели с лагом. Будем всё больше затрагивать динамический аспект. Мы рассматривали автокорреляцию, и явно появился некий динамический аспект. Появился предшествующий и последующий член. $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, $Y_i = f(\vec{X}_i)$. Была разумная идея: поведение некоторой переменной в текущий момент полностью определяется значениями переменной в этот же момент времени. Покупатель в микре максимизирует функцию полезности при ценах в этот момент. Он статичен. Это не единственная теория, но классика остаётся в этой плоскости. Есть только представление о максимизации.

Есть австрийская школа: неизвестно, есть ли функция полезности. Но потребитель что-то таки покупает. Говорят: с помощью проб и ошибок. Прошлый опыт, наши рассказы, уже существующие сделки — появляется динамический момент в этом. Вот эту идею и станем развивать. Принимая решение, экономический агент учитывает не только то, что сегодня, но и то, что было во времени назад. Поднимается — это значит, вчера было ниже.

Рассмотрим условный пример автора. Студент третьего курса получает стипендию, деньги от родителей, имеет своё потребление в зависимости от располагаемого дохода. $Y_t = \alpha + \beta X_t + \varepsilon_t$. Студент собирает данные за несколько месяцев, можно посчитать marginal propensity to consumption, автономное потребление. А предположим, что студент устроился на работу. И в t_0 у него появляется заработок в 20 000. Сильно вырос доход. Конечно, он изменит своё потребление. Но как он его изменит? Рассмотрим график: время — потребление. Оно скачком поднимается в теории. А в жизни агент привыкает к новому доходу. Первый месяц он поставит пиво, потом поставит костюм с галстуком. И его расходы будут меньше, чем в соответствии с формулой. Будет казаться, что у него *mpc* меньше. Он выйдет на какой-то небольшой уровень. Потом он пообвыкнется, начнёт выходить на другой уровень потребления, и только с третьего момента он выйдет на общий уровень, соответствующий нашей функции. Выход происходит не сразу; он распределяется на несколько периодов времени:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \varepsilon_t$$

Но получается зависимость потребления от траектории располагаемого дохода в предыдущие моменты времени. Мы хотим смоделировать изменчивость Y от X, но эффект растянулся во времени. Подобные модели называются *модели с распределёнными лагами* (backward-lagged).

«Переменная со сдвигом — это намёк на ментальность».

Оператор лага: $LX_t = X_{t-1}$. И с оператором можно проводить любые арифметические действия. Тогда

$$Y_t = \alpha + (\beta_0 + \beta_1 L + \beta_2 L^2)X_t + \varepsilon_t$$

Содержательно у нас только одна переменная. Как регрессоры разные моменты истории разные, но содержательно это одно и то же.

Кстати, мы сменили индексы, β_0 стал при X_1. Коэффициент β_1 показывает, как изменится Y при изменении X_{t-1} при прочих равных. А прочие равные звучат странно: это одна и та же переменная. Но интерпретация неудобная. Более естественно трактовать β_0: до момента t нет изменения, и при скачкообразном изменении на 1 в t Y прирастёт на β_0. Это краткосрочная, мгновенная реакция. И гораздо удобнее интерпретировать $\beta_0 + \beta_1$: нарастающая сумма коэффициентов показывает суммарное изменение через период. $\beta_0 + \beta_1 + \beta_2$ — это реакция на единичное изменение, которое произошло и осталось в 2 периода времени. Сумма коэффициентов — полное изменение при изменении икса на единицу и оставании им таким же (long run). А частичная сумма — это неполная, промежуточная реакция. $\sum\limits_{i=0}^{p} \beta_i$ — долгосрочное изменение, мультипликатор, эластичность. Математически $\beta_i = \frac{\partial Y_t}{\partial X_{t-1}}$, но интерпретировать это неудобно. Очень часто ещё используют нормированные коэффициенты: $\beta_i^* = \frac{\beta_i}{\sum \beta_i}$. То есть сразу произойдёт 20 %, потом ещё 40 % — так интерпретировать.

Откуда растёт необходимость включать лаги в уравнение? Приспособление агента к резкому изменению внешних условий. Вторая очевидная причина — технологические изменения. Они никогда не происходят мгновенно, как нас и учит теория производственных функций. Capital-to-labour ratio — существенный параметр. Если оно поменялось за счёт технологии, то рациональный фирмач должен уволить народ и купить капитал. Но ему нужно время, чтобы купить и установить. Иксы изменились, и он это знает. Другой пример: есть группа потребительских товаров с быстрыми технологическими изменениями. Если мы хотим купить новый смартфон, то тогда на предыдущую модель цена падает. Мы же человек умный. Вторая модель была бы вот так вот по горло, а при выходе третьей упадёт цена на вторую. Третья группа причин — институциональныя причины. Пусть в стране резко растут розничные цены, инфляция. Профсоюз требует увеличения зарплаты. Всегда ли он резко требует? Нет, потому что никто не пойдёт платить сразу: у нормального профсоюза есть коллективный договор. Только при заключении нового договора будет учтена данная ситуация. Газовики жалуются на привязку конъюнктуры к нефти… Газ — это долгосрочные соглашения, и цену нельзя изменить сразу, даже если захочется. Экономические договоры смягчают шоки и повышают предсказуемость и устойчивость.

Из технологии управления летательными аппаратами в 40-е годы пришли распределённые лаги: эффект распределён по времени. Математически Y зависит не только от текущего значения, но и от траектории X. Ничего особо сложного у нас нет: если X_t — детерминированный регрессор, то остальные тоже детерминированные. Просто используем множественную регрессию, вот и всё! Применяем обычный МНК, и если выполнены предпосылки Гаусса—Маркова, то всё несмещённое. А сколько лагов? Никто не знает! $Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_l X_{t-l} + \varepsilon_t$. Но мы не знаем l, уравнение не идентифицировано! Нужно какое-то правило, которое позволит l определить. На заре технической революции (1942) посчитать одно уравнение регрессии было не так просто; появился метод Альта (Олта, Alt) и Тинбергена. Тинберген — один из первых нобелевских лауреатов. Наивный способ: будем наращивать. Построили модель без лагов — добавили один. Последовательность всё более усложняющихся моделей. Где остановиться? Поскольку все X — это доход, мы ожидаем, что все коэффициенты одного знака, так как они показывают реакцию величины потребления на увеличение дохода. И естественно ожидать, что коэффициенты одного знака. Здравый смысл говорит: то, что было при царе Горохе, должно оказывать меньшее влияние. Год назад — да, при Ельцине — наверно, при Иване Васильевиче — навряд ли. Если коэффициент β_1 того же знака и значим, то нормально. Останавливаемся, когда коэффициент незначим или другого знака. Но это не очень хорошо: модели, куда не включены нужные лаги, обладают невключёнными переменными, поэтому результаты смещённые. Какая-то неудовлетворённость есть, действует мистика.

В 1951 году предложили **схему Койка** (Койека, — писали ранее, но это неверно, ибо Koyck). Он учился в средней школе и предложил следующее: а давайте считать, что количество лагов бесконечно. Они просто убывают и одного знака? В школе училась геометрическая прогрессия: $\beta_i = \beta_0 \cdot \lambda^i$, $|\lambda < 1|$, $0 < \lambda < 1$. Его величина отвечает за быстрое затухание коэффициентов. Даже если у нас есть очень длинная реализация, то как оценивать такие огромные матрицы? Когда большое число лаговых переменных, то будет сильна мультиколлинеарность. Переменная через месяц не так отличается, опасная квазимультиколлинеарность. А можно вывести формулу суммы членов бесконечной геометрической ~~регрессии~~ прогрессии.

$$Y_{t-1} = \alpha + \beta_0 X_{t-1} + \ldots + \beta_l X_{t-l-1} + \ldots + \varepsilon_{t-1}$$

Домножим на λ и вычтем:

$$Y_t - \lambda Y_{t-1} = \alpha(1-\lambda) + \beta_0 X_t + \cancel{(\beta_1 X_{t-1} - \lambda\beta_0 X_{t-1})} + \varepsilon_t - \lambda e_{t-1},$$

ибо $\lambda\beta_0 = \beta_1$. Получилась квазиразность: $Y_t = \alpha(1-\lambda) + \lambda Y_{t-1} + \beta_0 X_t + v_t$. Это два представления одной и той же связи между Y и X. Вся бесконечная койковская схема ушла. Появилось запаздывающее значение версии Y, поэтому называется это чудо авторегрессионной моделью. Если же тут будут две переменные ($\gamma_i = \gamma_0 \mu^i$), то мы проведём переход дважды. Так как $\lambda \neq \mu$, то они останутся, а потом по μ вычтем. Будет ещё Y_{t-2}. Если много переменных, то появится столько лагов по игреку. Объясняющие факторы — это запаздывающие игреки. Из бесконечной схемы мы приходим к простой. Хочется применить МНК, как в авторегрессии. А 1953 году

Кейган применил схему Койка для модели гиперинфляции Кейгана. Это монетаристская модель, считающая, что инфляция зависит от динамики денежной массы и базы. Если Y — инфляция, а X — экспансия денег, то это м есть гипермодель Кейгана. Это такая инфляция, при которой ничто, кроме расширения денежной базы, на инфляции не сказывается. Если инфляция стремительно растёт в год, то надо строить модель с лагами. Идея Кейгана — применить схему Койка с бесконечным количеством лагов. В 10-х годах XXI века об оценке Кейгана мы скажем: «Нехорошо». Y — стохастический регрессор. И он даст несостоятельную оценку. Надо проверить корреляцию v_t и t. С этим надо что-то делать, хватит это терпеть — и тут макроэкономика умолкает, так как коэффициенты несостоятельные. И приходится применять метод инструментальных переменных. Давайте убедимся, что это так.

$v_t = \varepsilon_t - \lambda\varepsilon_{t-1}$. Очевидно, что $\mathbb{E}(v_t) = 0$, белый шум, $\mathrm{Var}(v_t) = (1+\lambda_2)\sigma_\varepsilon^2$. Но $\mathrm{Cov}(v_t, \varepsilon_{t-1}) = \mathbb{E}\big((\varepsilon_t - \lambda\varepsilon_{t-1})(\varepsilon_{t-1} - \lambda\varepsilon_{t-2})\big) = -\lambda\sigma_\varepsilon^2 \neq 0$. Обязательно есть корреляция. И надо применять обобщённый МНК. Y_t зависит от ε_{t-1}, можно аккуратно раскрыть: $\mathrm{Cov}(Y_t, v_t) = -\lambda\sigma_\varepsilon^2 \neq 0$. Можно применить двухшаговый МНК или МИП.

Двухшаговый МНК: у нас есть регрессоры, коррелированные со случайным членом. Строятся вспомогательные регрессии на инструменты и в конечной регрессии ставятся оценки X из инструментальных регрессий.

А почему веса убывают геометрически? Схема Койка возникает из более естественной постановки, связанной с ожиданиями. Уравнения с лагами имеют тесную связь с моделями ожиданий. Классика: агент минимизирует затраты на сегодня. А что говорят ожидания? Мы вложим X_t в проект сначала, потом и далее до X_l, а как мы будем оценивать отдачу? По ожидаемой прибыли? Мы пишем ожидаемую прибыль. Надо максимизировать ожидаемую прибыль. Это меняет ситуацию: максимизируем то, чего я не знаю.

$Y_t = \alpha + \beta X_t^* + \varepsilon_t$. Вместо икса стоит какое-то ожидание. Но X^* — ненаблюдаемая величина, у нас нет этих значений. Необходимо что-то предположить уравнения для ожидания: $X_t^* = f(X_t, \dots)$. Это как связаны ожидания с реальными величинами. Рациональные ожидания, адаптивные, наивные. Начнём с наивных. Наивные такие: завтра будет так же, как вчера (чуть изменив песенку). $X_t^* = X_{t-1}$. Половина прогнозов такие. И у мартингалов то же самое.

$X_{t+1}^* - X_t = X_t - X_{t-1}$. Иные говорят, что это от лукавого, ибо $X_{t+1}^* = 2X_t - X_{t-1}$. Наивность такая: $\frac{X_{t+1}^*}{X_t} = \frac{X_t}{X_{t-1}} \Rightarrow X_{t+1}^* = \frac{X_t^2}{X_{t-1}}$. Уже нелинейность! $Y_t = \alpha + \beta X_t + \varepsilon_t$, $Y_t = \alpha + \beta(2X_t - X_{t-1}) + \varepsilon_t$. Просто модель с ограничением. А если не от лукавого, то $Y_t = \alpha + \beta\frac{X^2}{X_{t-1}} + \varepsilon_t$. Нелинейная модель. Из наивных ожиданий возникают распределённые лаги. Этой теме уделялось большое внимание. Можно записать наивное ожидание, написать, что есть отклонения от тренда, сказать, что есть сезонность. Тогда гипотеза странная, модифицируется: $\frac{X_{t+1}^*}{X_{t-3}} = \frac{X_t}{X_{t-4}}$.

В 1950-х Фарбер сделал дополнительное обследование. В США был департамент, который каждый год разрабатывал прогноз железнодорожных тарифов. У Фарбера был ряд прогнозов и ряд реальных тарифов. Тогда он попробовал так: он наивными ожиданиями спрогнозирует вместо это департамента а посмотрит, у кого лучше. Он тогда взял $\frac{1}{n}\sum |X_{\text{ист}} - X_{\text{прогн}}|$.

25 Лекция 25

В 1969 году Хирш и Ловелл сделали аналогичную работу по расценкам в строительстве и нашли, что наивные ожидания гораздо хуже, чем департамент прогнозирования. Они показали, что специалисты и потребителя действуют лучше, чем мы о них наивно думаем. Стали думать, что агенты адаптируются к изменению: $X_{t+1}^* = \beta_0 X_t + \beta_1 X_{t-1} + \dots + \beta_l X_{t-l}$, то есть взвешенная линейная комбинация предыдущих значений. Это наивные ожидания в какой-то мере, но называется адаптивными. Первым схему эту применил Ирвинг Фишер в 1925 году. Он подставлял это в уравнение и получал модель с распределёнными лагами. Год был 1925, экономисты плохо знали матан, но Фишер что-то знал. Он считал, что коэффициенты убывают не по геометрической, а по арифметической прогрессии. То есть остаётся только начальный параметр и общее убывание. Всё сходит на ноль, а дальше их нет. Как, по Фридману, пишутся адаптивные ожидания? Были ожидания в прошлый момент (x_t^*), x_t, а x_{t+1}^* корректируется так: $x_{t+1}^* - x_t^* = (1-\lambda)(x_t - x_t^*)$. Разница есть некоторая фиксированная доля от расхождения между ожиданием в прошлым момент и фактом. Это и есть адаптивные ожидания. Раскроем скобки: $x_{t+1}^* = \cancel{x_t^*} + x_t - \cancel{x_t^*} - \lambda x_t + \lambda x_t^* = (1-\lambda)x_t + \lambda x_t^*$. Ожидания на следующий момент времени есть линейная комбинация факта и предыдущего ожидания. Если $0 < \lambda < 1$, то это называется выпуклой линейной комбинацией. Из этого выражения можно заключить: если адаптивное ожидание такое, то можно подставить x^* в выражение: $(1-\lambda)x_t + \lambda\big((1-\lambda)x_{t-1} + \lambda X_{t-1}^*\big) = (1-\lambda)x_t + \lambda(1-\lambda)x_{t-1} + \lambda^2 x_{t-1}^* = (1-\lambda)x_t + \lambda(1-\lambda)x_{t-1} + \lambda^2(1-\lambda)x_{t-2} + \lambda^3 x_{t-2}^* = (1-\lambda)x_t + \lambda(1-\lambda)x_{t-1} + \lambda^2(1-\lambda)x_{t-2} + \dots + (1-\lambda)\lambda^k x_{t-k} + \lambda^\infty x_{-\infty}^*$, $0 < \lambda < 1$. Ой, да это же просто модель с распределёнными лагами. Лямбды образуют геометрическую прогрессию. Лямбда — вес прошлого ожидания, $(1-\lambda)$ — факта. Подсчитаем сумму коэффициентов: $\sum \beta_i = \frac{1-\lambda}{1-\lambda} = 1$ — койковская схема. Ожидания на период $t+1$ есть линейная комбинация всех предыдущих значений икса со всеми положительными коэффициентами, которые равны единице.

Поэтому схема Койка — это не искусственная придумка, которая возникла потому, что мы знаем, как с этим работать. Саймон Кузнец строил то, что предельная склонность к потреблению должна долгосрочно убывать. А получалось необъяснимое, не так себя она вела. Они брали агрегированное потребление и располагаемый доход. Были попытки объяснить это расхождение. Фридман придумал теорию перманентного дохода в сочетании с

адаптивными ожиданиями; он так хорошо всё сочетал, и она объяснила доход на перманентный и transitory. Там адаптивные ожидания и схема Койка.

Лаги распределены по схеме Койка и модель адаптивных ожиданий — это одно и то же. Как же на самом деле оценивать модель с адаптивными ожиданиями? Коэффициент β_0 специфический. Получаем авторегрессионную форму:

$$Y_t = a + \beta x_t^* + \varepsilon_t, \quad Y_t - \lambda Y_{t-1} = a(1-\lambda) + b(x_{t+1}^* - x_t^*) + v_t,$$
$$Y_t = a(1-\lambda) + b(1-\lambda)x_t + (\varepsilon_t - \lambda\varepsilon_{t-1}) + \lambda Y_{t-1}$$

МИП позволяет оценивать такую регрессию. Методом множителей Лагранжа можно проверить корреляцию с Y_{t-1}. Оказывается, нашу модель можно оценивать не авторегрессионно, а как будто с бесконечным лагом. Это предложил нобелиат Клейн. Было $Y_t = a + \beta x_{t+1}^* + \varepsilon_t$, подставим всю схему: $Y_t = a + b\left(\beta_0 x_t + \beta_1 x_{t-1} + \ldots + \beta_l x_{t-l} + \ldots\right) + \varepsilon_t$, $\beta_i = (1-\lambda)\lambda^i$, $\beta_0 = 1-\lambda$. $Y_t = a + b\sum\limits_{i=0}^{\infty}(1-\lambda)\lambda^i x_{t-i} + \varepsilon_t$, $t \in [1;T]$. Когда иксы уходят в отрицательную область, мы их не наблюдаем. Введём вспомогательную переменную: $z_{1t} = \sum\limits_{i=0}^{t-1}(1-\lambda)\lambda^i x_{t-i}$, $Z_{2t} = \sum\limits_{i=t}^{\infty}(1-\lambda)\lambda^i x_{t-i}$. Несколько преобразуем это выражение: $i = (i-t) + t$. Тогда $Z_{2t} = \lambda^t \sum\limits_{i=t}^{\infty}(1-\lambda)\lambda^{-(t-i)}x_{t-i}$. А что это за сумма? Линейная комбинация ненаблюдаемых иксов, то есть x_1^*, адаптивное ожидание для первого. $\lambda^t x_1^* = \mathbb{C}\cdot\lambda^t$. $Y_t = a + bZ_{1t} + b\mathbb{C}\lambda^t + \varepsilon_t$. Это регрессия на две переменные. Первая переменная обычная, вторая — искусственная. $\lambda \in (0;1)$, и Клейн предложил сделать поиск по решётке. Например, $\lambda = 0;0{,}1;\ldots;1$. Находим $Z_{1t}(\lambda)$, находим $\mathrm{RSS}(\lambda)$, находим минимизирующее это выражение λ, получаем оценки $\hat{a}(\lambda)$, $\hat{b}(\lambda)$. Методом Клейна мы смогли оценить модель прямо в такой форме, то есть в форме с бесконечным числом лагов. Всё загналось в одну переменную, правда, нелинейную, и провели численную оптимизацию. И получаются правильные оценки. Несмещённые — сказать трудно при бесконечности. И этот метод лучше ориентирован не адаптивные ожидания.

Модель Клейна пересчитали. И только бдительный Маддала нашёл одно неприятное место с точки зрения переоценивания. Что тут у нас нехорошо? Реальные остатки на счетах предприятий убывают с увеличением инфляции. $\frac{M}{P} = f(\pi)$, а зависимость от остальных переменных пропадает. Уравнения количественной теории денег пишутся в логарифмах: $p = \ln P$, $m = \ln M$. Тогда $m_t - p_t = a + b(p_{t+1}^* - p_t) + u_t$. Разница — это и есть ожидаемая инфляция, или индекс цен: $\ln\frac{P_{t+1}^*}{P_t} = \pi_t^*$. Можно сократить выражение: $y_t = a + b\pi_t^* + u_t$. Авторегрессионная форма: $y_t = a(1-\lambda) + \lambda y_{t-1} + b(1-\lambda)\pi_t + v_t$. Может, тест множителей Лагранжа скажет, что нет автокорреляции в остатках. А если так повезло, то просто эта оценка оказалась правильной. Если метод множителей Лагранжа показывает, что есть автокорреляция, то надо применять МИП.

А если мы хотим применить метод Клейна, то на мешает вот что: закон инфляции говорит, что остатки на счетах предприятий убывают от инфляции. Тогда инфляция экзогенная, остатки эндогенные. Но в моём уравнении P входит и туда, и туда. Поэтому Маддала говорит: надо переписать так, чтобы слева была чисто экзогенная переменная. Сидёж P_t слева приводит к его зависимости от u_t, то есть регрессор коррелирован с u_t. Введём вспомогательную переменную: $W_t = Z_{1t} - (1-\lambda)p_t$. Зачем? В левой части она была бы такой же. $W_t = \sum\limits_{i=0}^{t-1}(1-\lambda)\lambda^i\pi_{t-i} - (1-\lambda)p_t$. Тогда W_t не зависит от p_t. При $i=0$ имеем $p_t - p_{t-1}$, а остальные пойдут в $t-1$ и предшествующие моменты. Теперь зависимость не от p_t, а от более ранних. Тогда $m_t - p_t = a + b(1-\lambda)p_t + bW_t + c\lambda^t + u_t$. Теперь запишем это уравнение как уравнение от гиперинфляции. $p_t = \theta_0 + \theta_1 m_t + \theta_2 w_t + \theta_3\lambda^t + u_t$, где $\theta_0 = \frac{-a}{1+b(1-\lambda)}$, $\theta_1 = \frac{1}{1+b(1-\lambda)}$, $\theta_2 = \frac{-b}{1+b(1-\lambda)}$, $\theta_3 = \frac{-c}{1+b(1-\lambda)}$. Это есть регрессия логарифма цен на логарифм денежного агрегата, вспомогательную переменную и искусственную переменную. Прогоняем лямбды, берём минимальный RSS, получаем оценки $\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, $\hat{\lambda}$. Потом мы пересчитываем нужные нам коэффициенты — a и b. Оценка $\hat{b}$ — это $-\frac{\theta_2}{\theta_1}$, на c нам наплевать, $\hat{a} = -\hat{\theta}_0\left(1 + \hat{\beta}(1-\hat{\lambda})\right)$. Кстати, не всегда получается обратно обсчитать коэффициенты.

26 Лекция 26

Эконометристы победили на олимпиаде по метрике: 8 из 10 человек первые. Эконом, МИЭФ, даже кто-то с менеджмента (наверно, бывший экономист).

В прошлый раз мы рассматривали модели адаптивных ожиданий, ожиданий со схемой Койка, и у нас появилось понимание того, что модель одного и того же процесса имеет разные представления. Первое — это где среди регрессоров есть Y_{t-1}, второе — лаги. Поэтому МНК даёт несостоятельную оценку. Последнее — идея Клейна, как оценивать модель в форме бесконечного лага. И мы делали совокупность численного метода и точного МНК. Обе схемы относились к одному классу. Койк предложил возможность получить результат, и из простой схемы адаптивных ожиданий вытекала схема Койка. У Койка прогрессия, но мы собираемся обобщить это на общий класс.

ADL — autoregression distributed lag model. Мы хотим авторегрессионную схему с некоторым количеством лагов. Распределённый лаг распределён по иксу. Если регрессор один, то ADL-модель такова:

$$Y_t = \theta + \alpha_1 Y_{t-1} + \ldots + \alpha_p Y_{t-p} + \beta_0 X_t + \beta_1 X_{t-1} + \ldots + \beta_q X_{t-q} + \varepsilon_t, \quad \varepsilon_t \sim WN$$

С β_0 идёт распределённый лаг, а с α — Y в предыдущие моменты. Интерпретация: текущее значение зависит от q предыдущих значений объясняющей переменной и p предыдущих наблюдений (траектории). Политика — это не только последнее, но ми как мы её проводили. Случайность ε_t привносит сложным путём: T_{t-n} зависит от ε_{t-n}. Модель линейная, и это уже хорошо. В прошлый раз мы ввели стационарный процесс: совокупность значений переменной: Y_t, $t \in (-\infty; +\infty)$, $[0; T]$. А теперь это единый объект: случайный процесс, временной ряд. Процесс является **стационарным** (слабостационарным), если выполнены 3 свойства:

1. $\mathbb{E}(Y_t) = \text{const}$;
2. $\text{Var}(Y_t) = \sigma_Y^2 = \text{const}$ (слабо-, потому что важно только равенство матожиданий и дисперсий, но не распределений);
3. $\text{Cov}(Y_t, Y_{t+\tau}) = \gamma(\tau) = \gamma_\tau$.

Стационарность: корреляция между январём и мартом такая же, как и между июлем и сентябрём. γ — это целочисленная функция. Функция называется автоковариационной функцией: она показывает силу связи процесса, разнесённых на t шагов. Требование: она не зависит от t (covariance stationary). Далее, $\text{Corr}(Y_t, T_{t+\tau}) = \frac{\text{Cov}(Y_t, Y_{t-1})}{\sigma_Y^2} = \frac{\gamma_\tau}{\gamma_0} = \rho_\tau$. ρ_τ — это набор сдвигов, или автокорреляционная (нормированная автоковариационная) функция, или ACF.

Если нарисовать график, то он будет выходить в нуле из единицы, функция чётная (от перестановки ковариация не меняется. Часто соединяют единой линией и называют это коррелограммой.

При каких условиях Y будет стационарным? Во-первых, матожидание. Возьмём матожидание: $\mathbb{E}(Y_t) = \mu_Y$, $\mathbb{E}(X_t) = \mu_X$ (мю в роли матожидания — это святое). Берём матожидание ото всех шести: $\mu_Y = \theta + \mu_Y(\alpha_1 + \ldots + \alpha_p) + \mu_X(\beta_0 + \ldots + \beta_q)$. Перенесём всё в одну часть и вспомним, что оператор лага нужен для упрощения записи. Запишем исходное уравнение так: $\alpha_p(L) \cdot Y_t = \theta + \beta_q(L) X_t + \varepsilon_t$. Некоторый полином от оператора лага равен константе плюс оператор на икс и инновация — случайная добавка. $\mu_Y(1 - \alpha_1 - \ldots - \alpha_p) = \theta + \mu_X(\beta_0 + \beta_1 + \ldots + \beta_q)$. Можно записать это так: $\mu_Y \alpha_p(1) = \theta + \mu_X \beta_q(1)$. Это выражение, чтобы упростить жизнь. Тогда это линейное уравнение, как в пятом классе, и если $\alpha_p(1) \neq 0$, то получаем единственное решение: $\mu_Y = \frac{\theta}{\alpha_p(1)} + \frac{\beta_q(1)}{\alpha_p(1)} \mu_X$. Это связь матожиданий Y и X. Условие: $\alpha_p(1) \neq 0$, или единица не является корнем этого алгебраического полинома. И это называется проблемой такой: unit root problem. $\varepsilon_t = \rho \varepsilon_{t-1} + v_t$, или $(1 - \rho L)\varepsilon_t = v_t$, $1 - \rho \neq 1$.

Вот получили интересное выражение… Ну выражение и выражение…

Канторовича Романко учил: они одновременно пришли студентом и преподавателем.

Это похоже на характеристическое уравнение. Если Y_t подчиняется линейному уравнению, то теория линейных дифференциальных и разностных уравнений говорит: общее решение неоднородного есть частное неоднородного плюс общее однородное. Во временных рядах естественнее писать назад наблюдения. Для α_p разностное уравнение таково: $\lambda^p - \alpha_1 \lambda^{p-1} - \ldots - \alpha_p = 0$, и у этого уравнения n корней, из которых некоторые могут совпадать, а некоторые — быть комплексно-сопряжёнными. Решение — это такая сумма: $\sum_{i=1}^{p} \mathbb{C}_i \lambda_i^t +$ частное решение. Чтобы постоянное матожидание было, Y не должно убедать в бесконечность. Если хотя бы один λ по модулю больше 1, то будет бесконечность. Если $|\lambda_t| < 1$, то при $t \to \infty$ они убывают. Это условие называется условием устойчивости разностного уравнения. Тут оно называется условием слабой стационарности: $|\lambda_i < 1|$, или на комплексной плоскости они должны лежать внутри единичной окружности. Есть приёмы, как, не решая уравнения, определить устойчивость (критерий Рауса—Гурвица). Устойчивость, обратимость полинома, стационарность — это одно и то же. Можно полином по-другому записать: $1 - \alpha_1 L - \ldots - \alpha_p L^p = 0$. У этого уравнения будет корень $\frac{1}{\lambda_0}$, если λ_0 — корень исходного характеристического уравнения. Будьте внимательны в софте или в учебнике: это зависит от того, какой тип характеристического уравнения.

Матожидание мы записали, и можно посчитать дисперсию, не уходя в более громоздкие вычисления. Оказывается, если условие стационарности выполнено, то будут выполнены все три условия списка. Это не так сложно доказывается. Условие:

$$|\lambda_i| < 1 \ \forall i$$

Если оно не выполняется, то дисперсия будет бесконечной. А теперь обратим внимание и придадим ему ещё одно представление этого уравнения. Слева стоит $Y_t = Y_{t-1} + \delta Y_t$, где $\delta Y_t = Y_t - Y_{t-1} = (1 - L) Y_t$. Все хочется привести к базе из Y. Далее, $Y_{t-2} = Y_{t-1} - \Delta Y_{t-1}$, $Y_{t-3} = Y_{t-1} - \Delta Y_{t-1} - \Delta Y_{t-2}$, $Y_{t-p} = Y_{t-1} - \Delta Y_{t-1} - \ldots - \Delta Y_{t-p+1}$. Уравнение с распределёнными лагами является динамическим уравнением. Что бы назвать долгосрочным соотношением? Если они в результате долгой эволюции придут к какому-то соотношению, то долгосрочное равновесие будет таким: придя туда, они не будут изменяться. Есть тут ε_t, которая им не позволит сидеть, не меняясь, поэтому надо оговорить: в среднем матожидание должно находиться там. Если $\bar{Y}$, $\bar{X}$ — долгосрочные величины, то их соотношение должно выполняться без случайной добавки, и $\bar{Y} = \frac{\theta}{\alpha_p(1)} = \frac{\beta_q(1)}{\alpha_p(1)\bar{X}}$. А ΔY — это

краткосрочная динамика, short-run.

$$Y_{t-1}(1-\alpha_1-\ldots-\alpha_p)=\theta-\Delta Y_t-\gamma_1\Delta Y_{t-1}-\ldots-\gamma_{p-1}\Delta Y_{t-p+1}+ \\ +(\beta_0+\beta_1+\ldots+\beta_q)X_{t-1}+\delta_1\Delta X_{t-1}+\ldots+\delta_{q-1}\Delta X_{t-q+1}+\varepsilon_t,$$

где γ через α выражаются однозначно.

$$\Delta Y_t=\theta+\beta_q(1)X_{t-1}-\gamma_1\Delta Y_{t-1}-\ldots+\delta_{q-1}\Delta X_{t-q-1}+\varepsilon_t-Y_{t-1}\cdot\alpha_p(1)$$

$$\Delta Y_t=-\gamma_1\Delta Y_{t-1}-\ldots-\gamma_{p-1}\Delta Y_{t-p+1}+ \\ +\beta_0\Delta X_t+\delta_1\Delta X_{t-1}+\ldots+\delta_{q-1}\Delta X_{t-q+1}-\alpha_p(1)\left(Y_{t-1}-\frac{\theta}{\alpha_p(1)}-\frac{\beta_q(1)}{\alpha_p(1)}X_{t-1}\right)+\varepsilon_t$$

Мы получили, что краткосрочное приращение Y_t разложено на две части: авторегрессионное уравнение с распределёнными лагами другой переменной — ΔY_t — и отклонение от долгосрочного равновесия в предыдущий момент времени. Вместо долгосрочных изменений стоят отклонения. $\alpha_p(1)\neq 0$, и можно показать, что $\alpha_p(1)>1$ — это чистое поведение графика полинома. Тогда смотрим: содержательное разбиение — краткосрочное изменение и отклонение от долгосрочного равновесия в предыдущий момент. По-английски это называется error, по-русски — невязка (не вяжется). Если в предыдущем моменте мы были в равновесии, то работает только краткосрочная часть. Если не в равновесии, то скобка после $\alpha_p(1)$ больше нуля. Коэффициент $\alpha_p(1)$ больше нуля, поэтому изменение будет меньше, чем если бы оно попадало в равновесие: оно корректирует ΔY в сторону равновесия. Если же Y оказался меньше, то тогда Δ_Y — дополнительная положительная надбавка в виде доли от дизэквилибриума. Этот механизм получил название ECM, error correction model, модель коррекции ошибками, где корректируется краткосрочное поведение.

Если Y — ВВП, X — денежный агрегат, то тогда добавочка есть краткосрочный мультипликатор. β_0 — краткосрочный, α_0 — долгосрочный. Возникает дилемма: какое соотношение использовать? Y_t или ΔY_t — ECM?

Что произойдёт с ECM для простейшей модели ADL(p,q)? Рассмотрим ADL(1,1). $Y_t=\theta+\alpha Y_{t-1}+\beta_0X_t+\beta_1X_{t-1}+\varepsilon_t$. $Y_t=Y_{t-1}+\Delta Y_t$, $X_t=X_{t-1}+\Delta X_t$. Тогда error-correction-представление будет особо простым: $\Delta Y_t=\theta-(1-\alpha)Y_{t-1}+\beta_0X_{t-1}+\beta_0\Delta X_{t-1}+\beta_1X_{t-1}+\varepsilon_t$. Тогда $\Delta Y_t=\beta_0\Delta X_t-(1-\alpha)\left(Y_{t-1}-\frac{\beta_0+\beta_1}{1-\alpha}X_{t-1}\right)+\varepsilon_t$. Вся краткосрочная схема — это β_0.

27 Лекция 27

В Америке учёба стоит 50 000 $ в год. Занятия у нас до 18 июня.

«Кто не пришёл, тот прогулял».

Мы рассмотрели модели ADL(p,q). Строим ECM, большое внимание к автокорреляциям. В знаменателе было $\alpha_p(1)\neq 0$, что значит, что единица не есть корень полинома $\alpha_p(L)$. Рассмотрим один ряд: $u_t=u_{t-1}+v_t$, где $v_t\sim WN$. Это значит, что иксов нет, записана модель AR(1), а содержательный смысл параметра ρ — корреляция между u_t и u_{t-1}, причём $|\rho|<1$, так как $\sigma_u^2=\rho^2\sigma_u^2+\sigma_v^2$. Временной ряд Y_t называется слабостационарным (weak stationary), если для этого процесса выполнено следующее:

1. $\mathbb{E}(Y_t)=\text{const}$;
2. $\text{Var}(Y_t)=\sigma_u^2=\text{const}$;
3. $\text{Cov}(Y_t,Y_{t+\tau})=\gamma(\tau)$

Есть ещё называется variance stationary. А остальные — это нестационарные. Если $\rho=1$, то $u_t=u_{t-1}+v_t$, умное слово «мартингал» выплывало на арену. Это гипотеза идеального финансового рынка. Какие свойства у этого процесса? $\mathbb{E}(u_t)=\mathbb{E}(u_{t-1})+\mathbb{E}(v_t)=\text{const}$. С дисперсией не так понятно. Этому нас учил Василий Кириллович: линейное разностное уравнение с линейной правой частью. $u_t=u_{t-2}+v_{t-1}+v_t=u_0+\sum\limits_{i=1}^{t-1}v_{t-i}$. В любой момент это есть начальное значение плюс сумма всех шоков. Но дисперсия с u_t такая: $\text{Var}(u_t)=\text{Var}(u_0)+(t-1)\sigma_v^2$, и эта величина линейно растёт при росте t, она уходит в бесконечность. Главное, что это не константа. Это пример нестационарного процесса с бесконечной дисперсией. Это что-то новое. На финансовых рынках такая ситуация, когда вся новая информация приходит только в v_t, называется EMH. Это совершенство рынка.

Рассмотрим более общий процесс: $u_t=\mu+u_{t-1}+v_t$, или случайное блуждание с дрейфом. $u_t=u_0+\mu t+\sum\limits_{i=1}^{t-1}v_{t-1}$. У него и матожидание, и дисперсия зависят от времени. Перепишем соотношение эквивалентно: $\Delta u_t=\mu+v_t=u_t-u_{t-1}=(1-L)u_t$.

В 80-х годах появились две работы. Одна — Нельсона и Канга, другая — не их. Они сделали простую вещь: сгенерировали два случайных блуждания. $Y_t=Y_{t-1}+\varepsilon_t$, $X_t=X_{t-1}+v_t$. Была построена регрессия:

$Y_t = a + bX_t + w_t$. По обычной логике они посчитали t-статистику, равную $t = \frac{\hat{b}}{\text{s.e.}(\hat{\beta})}$. Должно же быть незначимым. Значимых должно быть не больше 50 из 1 000. Ребята проводили эксперимент, и было аномально много значимых коэффициентов. Кажется, что зависимость есть, а её нет. Они это назвали кажущейся регрессией (spoolers regression). Если построить $Y_t = \alpha + \beta t + \eta_t$, то коэффициент $\hat{\beta}$ тоже часто значим. Если добавить какие-то μ_1 и μ_2 к Y_t и X_t, то эффект только усиливался. Дарбин—Уотсон показывали, что ошибки коррелированы. Окончательно стало плохо макроэкономистам, когда исследовали 13 рядов статистики США, и оказалось, что 12 из них, кроме ряда безработицы, надо отнести к такому типу. Это какие-то нестационарные ряды вроде случайного блуждания. Сарджент сказал: всё, кризис теории! Оказалось, что если наши переменные относятся к типу случайного блуждания, то рутинный МНК приводит к неправильным выводам. И в 1987 году появилась работа Дики и Фуллера, где эту ситуацию обследовали теоретически. Если это случайное блуждание, то тогда единичный корень, unit root problem.

$\hat{b} = \frac{\sum x_t y_t}{\sum x_t^2}$, а теперь мы чувствуем, где зарыта собака: дисперсия бежит в бесконечность. Мы оцениваем то, что не существует. Оказывается, тогда не работает ЦПТ: всегда последняя дисперсия мажорирует все предыдущие. Если X и Y — случайные блуждания, то рассмотрим вот что: $X_t = \rho X_{t-1} + \varepsilon_t$, $\hat{\rho} = \frac{\sum x_t x_{t-1}}{\sum x_{t-1}^2}$. $\mathcal{H}_0 \colon \rho = 1 \Leftrightarrow \exists$ unit root problem. Если $\rho > 1$, то экспоненциальный рост, это неинтересно. $\mathcal{H}_1 \colon \nexists$ unit root problem. Односторонняя гипотеза. $\frac{\hat{\rho}}{\text{s.e.}(\hat{\rho})}$ не имеет распределения Стьюдента! Получили стохастические интегралы от броуновского движения, и существует предельное распределение, отличное от распределения Стьюдента. Назвали это распределением Дики—Фуллера. Они назвали это $\tau = \frac{\hat{\rho}-1}{\text{s.e.}(\hat{\rho})} \sim \mathscr{DF}$.

t Стьюдента похоже на стандартное нормальное. А Дики—Фуллер завален влево, асимметричное распределение. Сегодня можем посчитать распределение очень просто: строим случайное блуждание, находим величину, повторяем 10 000 раз, получаем распределение. Хотите сгладить — поставьте на ночь, чтобы было гладко. Это зависит от длины выборки и некоторых параметров. Раз Дики—Фуллер скошенный, то у него хвост левее. Если наша статистика попала правее критической нормального, то ответ правильный. Если левее хвоста $\mathscr{DF}$, то тоже правильный. А если между, то тогда результаты расходятся. Чтобы не вычитать единицы, сделали $\Delta X_t = (1-\rho)X_{t-1} + \varepsilon_t$, $\Delta X_t = \gamma X_{t-1} + \varepsilon_t$, $\mathcal{H}_0 \colon \gamma = 0$, $\mathcal{H}_1 \colon \gamma < 0$. Строится регрессия приращения на уровень. Это тест Дики—Фуллера. Это обязательная проверка для макроэкономических рядов. Без этого даже в статью не возьмут.

Когда $\rho < 1$, то стационарный процесс с нулевым матожиданием. Как перейти к более общему? Добавить свободный член в регрессию. Но ситуация оказалась более сложной: если ненулевое матожидание, то надо строить $\Delta X_t = \alpha + \gamma X_{t-1} + \varepsilon_t$, смотреть статистику $\frac{\hat{\gamma}}{\text{s.e.}(\hat{\gamma})} = \tau_\mu$. Из=за добавления свободного члена возникает дрейф. Это тоже распределение Дики—Фуллера, но оно ещё сильнее завалено. Если есть матожидание, то в уравнении $X_t = \alpha + \rho X_{t-1} + \varepsilon_t$, то появляется случайное блуждание с трендом. Надо добавить детерминированный тренд и в альтернативную модель. Имеем тестовое уравнение $\Delta X_t = \alpha + \beta t + \gamma X_{t-1} + \varepsilon_t$. Дики—Фуллер назвали это «тау-тау», или тау с трендом: $\tau_\tau = \frac{\hat{\gamma}}{\text{s.e.}(\hat{\gamma})}$. Распределения аппроксимируются простыми кривыми. В STATA, gretl, R это всё есть. Чем сильнее ошибка, тем ближе мы будем трактовать это как нормальное. А в жизни-то более сложные уравнения, чем AR(1)!

Если исходный ряд подчиняется уравнению типа AR(p), то есть $X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t$, то можно этот ряд переписать в следующем виде: $\Delta X_t = \gamma X_{t-1} + \sum\limits_{i=1}^{p-1} \delta_i \Delta X_{t-i} + \varepsilon_t$. А вот эта $\gamma = \alpha_1 + \alpha_2 + \ldots + \alpha_p - 1 = -\alpha_p(1)$. Поэтому $\mathcal{H}_0 \colon \exists$ unit root $\Leftrightarrow \alpha_p(1) = 0$. Было доказано, что t-отношение, равное $\frac{\hat{\gamma}}{\text{s.e.}(\hat{\gamma})}$, имеет то же распределение, что и τ_0, τ_μ, τ_τ. Это augmented Dickey—Fuller test, расширенный тест Дики—Фуллера. Получается, что с помощью сего теста мы можем проверить стационарность ряда или единичный его корень, означающий случайное блуждание.

Рассмотрим исходный полином: $\alpha_p(L) = 1 - \alpha_1 L - \ldots - \alpha_p L^p$. Если корень равен 1, то можно полином поделить на $(1-L)$, так как это корень полинома, и получится: $\alpha_p(L) = (1-L)(\underbrace{1 - \beta_1 L - \ldots - \beta_{p-1} L^{p-1}}_{\alpha'_{p-1}(L)})$. Тогда $\alpha_p(L) X_t = \alpha'_{p-1}(L) \Delta X_t$. Если рассматривать процесс в приращениях, то он стационарный. Если ряд нестационарный, а после перехода к приращению переходит в стационарный, то тогда это называется интегрированным стационарным рядом. ΔX_t стационарный, а $X_t \sim I(1)$. Вопрос: а если два единичных корня: Ответ: то уровень интеграции 2, надо перейти к разности от разности, и это называется $I(d)$ от difference — сколько раз надо взять последовательную разность, чтобы получить стационарный. На практике больше 2 не встречается. А $I(0)$ — это уже стационарный.

Был рассмотрен более сложный ряд: $X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q}$. Эта модель достаточно общая. Первая половина — авторегрессия, вторая — взвешенный белый шум. То есть это какое-то окошко, поэтому этот хвост назван скользящим средним, moving average, и получило это чудо название ARMA(p, q). Было легко показано, что добавка на стационарность решения не влияет. Для разности Δ назвали ARMA($p-1, q-1$), а исходные модели назвали ARIMA(p, d, q), где надо d раз взять разность. Модель ARMA с X вместо ε называется ARMAX(p, q) = ADL(p, q).

Правильное определение p чрезвычайно важно. У нас есть только реализация $X_1, \ldots, X_t$. Поэтому надо надо определить, сколько p включать в Δ. Такой подход: $p = [\sqrt[4]{T}]$. Если 120 точек, то целая часть — 3. Другая

методика — строить ADF-уравнение и наращивать, пока коэффициенты значимые по t-статистике. Критерий Акаике (Akaike information criterion) определяет количество лагов сам. Тест не очень мощный, но ADF-тест всёта. KPSS — Квятковский, Филлипс, Шмидт и Шин (Kwiatkowski—Phillips—Schmidt—Shin). Нулевая гипотеза — стационарность. Если оба результата одинаковые, то тогда стационарность определить легко. Если результаты разные, то надо подумать, как строить регрессии дальше. То ли по МНК и ADL, то ли…

Бокс и Дженкинс предложили: перейдём к стационарным уровням ΔX, но тогда потеряется информация. Проблема в том, что качественно реализации ведут себя очень по-разному. А краткосрочно всё очень предсказуемо. Итак, надо переходить к уровням и строить модель. Если в макро мы говорим о долгосрочных равновесиях, то надо пользоваться тестом Дики—Фуллера. Было много литературы против этого, что есть structural breaks, а макроэкономисты бились за то, что нет столько случайных блужданий, потому что маловероятно, что нет затухания. Если при Петре I изменение ВВП на 10 рублей так же влияет на мир, то тогда все ряды нестационарные, что неправдоподобно. В 1991 году Грэнджер сделал замечательное открытие и получил Нобелевскую. Он обнаружил, что у нестационарных рядов может отсутствовать долгосрочное поведение, но существовать долгосрочная связь. Пример: если в момент t_1 мы в какой-то точке, то потом мы пойдём равновероятно туда или туда. Агент пошёл во все стороны равновероятно. Это названо по поведению пьяного. Шёл — свалился, встал, забыл, пошёл дальше. А теперь в чистом поле два пьяных. Он ходит куда хочет. Но они связаны эластичным жгутом. Они мягко связаны между собой. Это названо коинтеграцией — совместной интеграцией двух нестационарных рядом.

Поговорим об этом на первом уровне. $X_t \sim I(1)$, $Y_t \sim I(1)$. Что можно сказать о линейной комбинации этих двух процессов? У неё тоже единичный корень, так как $(1-L)$ вынесется! $\alpha X_t + \beta Y_t \sim I(1)$. А вдруг $\exists \alpha, \beta\colon \alpha X_t + \beta Y_t \sim I(0)$. Тогда они коинтегрированы, или $X_t, Y_t \sim CI$. А с чего это такой математический трюк, что два случайных блуждания будут бесконечно разбросаны и в то же время связь есть? Пример: $X \sim \mathcal{N}(0;1)$, $Y = 5 - X$, функция от случайной величины, что есть случайная величина. Тогда $Y + X$ — это всегда 5. Альфа и бета определяются с точностью до множителя. Если же $X_t \sim I(d)$, $Y_t \sim I(d)$, $\exists \alpha, \beta\colon \alpha X_t + \beta Y_t \sim I(d-b)$, $b > 0$, то тогда принято обозначение коинтегрированного процесса $CI(d,b)$. Тогда самый простой случай — это $CI(1;1)$.

Пусть $Y_t + \alpha X_t = \varepsilon_t$, $Y_t + \beta X_t = u_t$ при $\alpha \neq \beta$, $\varepsilon_t \sim WN \sim I(0)$, $u_t = u_{t-1} + v_t$, $u_t \sim I(1)$. Вычтем из одного уравнения другого и громко скажем, что мы его разрешили: $X_t = \frac{1}{\alpha-\beta}\varepsilon_t - \frac{1}{\alpha-\beta}u_t$, $Y_t = \varepsilon_t - \frac{1}{1-\alpha}\varepsilon_t + \frac{1}{\alpha-\beta}u_t = \varepsilon_t\left(-\frac{\alpha}{\alpha-\beta}\right) + \frac{1}{\alpha-\beta}u_t \sim I(1)$. Тогда каждый из них есть порождение одного и того же $I(1)$, добавлен стационарный процесс, и коинтегрированные процессы есть порождение нестационарного и стационарного. Если порождены одним и тем же, то тогда можно стационировать. Что нам это даст? Что произойдёт, если построить регрессию Y_t на X_t? $Y_t = a + bX_t + w_t$. Оба нестационарные, кажется это дело регрессией — ан нет! Тогда $\hat{b}$ — это состоятельная оценка коэффициента α! Если ряды коинтегрированные, они нестационарные, то тогда регрессия друг на друга даёт стационарную оценку, причём она сходится ещё быстрее! Это свойство называется суперсостоятельностью.

Мы хотим найти связь между двумя переменными временного ряда. Если оба стационарные, хорошо. Если оба $I(1)$, надо проверить коинтегрированность. Если есть, то хорошо, а если нет, то переходим к ΔX, ΔY.

28 Лекция 28

29 Лекция 29

Мы начали исключительно для хроники систему, которую обозначали так:

$$\boldsymbol{B}\vec{y}_t + \boldsymbol{C}\vec{x}_t = \vec{u}$$

Нельзя разделить переменные на экзогенные и эндогенные. В первом уравнении суммируются игреки и иксы. Не удаётся разделить переменные, иначе ошибка будет коррелировать с правыми частями, а оценка будет смещённой, это называется simultaneousness bias. Предполагается, что $\vec{u}_t \sim \text{i.i.d.} \sim \mathcal{N}(\vec{0}, \boldsymbol{\Sigma})$. Время существенно, порядок важен.

Проблема: если матрица $\boldsymbol{B}$ не невырожденная, то мы не можем решить систему относительно всех входящих переменных. Разрешается она в хорошем случае так: $\vec{y} = -\boldsymbol{B}^{-1}\boldsymbol{C}\vec{x}_t + \boldsymbol{B}^{-1}\vec{u}_t \rightsquigarrow \vec{y}_t = \boldsymbol{\Pi}\vec{x}_t + \vec{v}_t$, что есть приведённая (reduced) форма. Если переменные x не коррелируют с v, то тогда хорошо применять МНК. В x_t включаются две группы переменных: экзогенные (госрасходы, налоги) и эндогенные, которые не связаны с моментом t (например, x_{t-1}, который уже известен и predetermined). И у нас был очевидный результат: уравнение в приведённой форме поддаётся оценке МНК. Если x детерминированные, то оценки BLUE, а если случайные, то состоятельные. v_t — это линейная комбинация ошибок u_t. Если иксы не коррелируют с u, то они и с v не коррелируют.

Если мы собираемся прогнозировать и знаем прогнозные их значения, то оценки матрицы $\hat{\boldsymbol{\Pi}}$ нам достаточно. Но иногда может потребоваться для объяснения экономических эффектов восстановить коэффициенты исходной формы. Зная оценки $\hat{\boldsymbol{\Pi}}$, можно попробовать найти оценки $\hat{\boldsymbol{B}}$ и $\hat{\boldsymbol{C}}$. Структурные коэффициенты удастся определить через приведённые. Тогда скажем, что уравнение точно идентифицируемо. Вопрос надо решать для каждого структурного уравнения в отдельности. Новое явление: коэффициенты некоторого структурного уравнения могут дать две оценки. Возникло 3 понятия: точно идентифицируемое, недоидентифицируемое

и сверхидентифицируемое. Как определить, к какому классу мы относимся? Вопрос идентифицируемости не связан с методами оценивания. Надо решить уравнение в приведённой форме (одно эндогенное на все эндогенные), а потом по коэффициентам пересчитать коэффициенты $\hat{\mathbf{\Pi}}$. Называется это ILS (Indirect Least Squares, **косвенный МНК**).

Удобно переписать это в привычном виде с блочными матрицами. Найти коэффициенты структурной формы — это и $\boldsymbol{B}$, и $\boldsymbol{C}$. Мы собрали все переменные как $\vec{z}_t = \begin{bmatrix}\vec{y}_t \\ \vec{x}_t\end{bmatrix}$. Действия с ней так же производятся: $\boldsymbol{A}\vec{z}_t = \begin{pmatrix}\boldsymbol{B} & \boldsymbol{C}\end{pmatrix}\begin{pmatrix}\vec{y}_t & \vec{x}_t\end{pmatrix} = \boldsymbol{B}\vec{y} + \boldsymbol{C}\vec{x}_t = \vec{u}_t$. Если $\dim \vec{y}_t = G$ и $\dim \vec{x}_t = K$, $B \mapsto G \times G$, $C \mapsto G \times K$.

Если никакие коэффициенты заранее не известны, то уравнения очень похожи одно на другое. Отличаются они только набором экзогенных переменных. Некоторые эндогенные могут не входить. Надо тогда задать информацию о том, что некоторые коэффициенты суть нулевые. Как это записать? Договорились записывать при помощи матрицы $\mathbf{\Phi}$. Рассматривается столбец, в котором на одном месте единица, на остальных — нули. И такая матрица есть для каждой строки.

$$\mathbf{\Phi}_i = \begin{pmatrix}\text{столбцы} & \text{переменных}\end{pmatrix},\ \mathbf{\Phi}_i \mapsto (G+K) \times R_i,$$

где R_i — количество ограничений в i-м уравнении.

$\vec{\alpha}_1$ — строка коэффициентов первой строки уравнения. Пусть $\boldsymbol{A} = \begin{pmatrix}\cdots & \vec{\alpha}^1 & \cdots \\ \cdots & \vdots & \cdots \\ \cdots & \vec{\alpha}_G & \cdots\end{pmatrix}$. Верно, что $\vec{\alpha}_1 = \vec{z}_t = u_{1t}$, $\vec{\alpha}_1\mathbf{\Phi}_1 = 0$. Введённые до сих пор матрицы структурной и приведённой формы обладают следующими связями: $\mathbf{\Pi} = -\boldsymbol{B}^{-1}\boldsymbol{C} \rightsquigarrow \boldsymbol{B}\mathbf{\Pi} + \boldsymbol{C} = 0$. Матричная форма: $\begin{pmatrix}\boldsymbol{B} & \boldsymbol{C}\end{pmatrix}\begin{pmatrix}\mathbf{\Pi} & \boldsymbol{I}_k\end{pmatrix}$. Одинаковое соотношение для всех уравнений: $\boldsymbol{A}\boldsymbol{W} = 0$. Симметричная структура, из которой следует, что $\vec{\alpha}_i\boldsymbol{W} = 0\ \forall i$. Одни и те же коэффициенты зашиты в строке и умножены на матрицу. Продолжим трюк: $\vec{\alpha}_1(\boldsymbol{W}\mathbf{\Phi}_1) = 0$. Это равно вот чему:

$$\begin{pmatrix}b_{1;1} & b_{1;2} & \cdots & b_{1G} & c_{1;1} & \cdots & c_{1K}\end{pmatrix}\begin{pmatrix}\boldsymbol{W}\mathbf{\Phi}_1\end{pmatrix} = 0$$

$G + K$ неизвестных. Каковы размерности? $(\boldsymbol{W}\mathbf{\Phi}_1) \mapsto (G+K) \times (K + R_1)$.

Далее, линейные однородные уравнения, справа ноль. Система всегда имеет решения, но она имеет либо тривиальное, либо бесконечное количество. И нас нужно не просто бесконечное, а с единственным свободным членом. Чтобы уравнение было точно идентифицируемо, необходимо и достаточно иметь только одну свободную переменную, то есть ранг транспонированной систем должен быть на единицу меньше. Rank condition: ранг матрицы, составленной из $\boldsymbol{W}$ и $\mathbf{\Phi}_1$, должен быть равен $G+K-1$. Пусть все наши ограничения — это ограничения невключения в уравнение. Сколько априорных ограничений? R. Что нам не нравится в соотношении ранга? Смущает матрица $\boldsymbol{W}$ — матрица структурных коэффициентов, которых мы не знаем. $\boldsymbol{B}^{-1}\boldsymbol{C}$. Доказано, что условие ранга может быть упрощено за счёт того, что в него не будет входить матрица $\mathbf{\Pi}$.

$$\operatorname{rank}(\boldsymbol{W}\mathbf{\Phi}_i) \geqslant G + K - 1 \Leftrightarrow \operatorname{rank}(\boldsymbol{A}\mathbf{\Phi}_i) = G - 1$$

Условие тяжёлое. Для большой системы проверка в общем виде сложна. Как проверять в общем случае в параметрической матрице миноры? В матрице 2×2 можно что-то лёгкое выписать, а если 7×7, то громоздко.

Order condition (необходимое условие порядка): число столбцов в матрице больше или равно числу неизвестных:

$$K + R_i \geqslant G + K - 1 \Leftrightarrow R_i \geqslant G - 1$$

Условие порядка легко проверяется. Если нарушено, то точно нельзя идентифицировать, а если нет, то ещё неизвестно. R_i — число априорно наложенных ограничений, которое должно быть не меньше, чем количество уравнений минус один. Если все ограничения типа исключения, то можно перефразировать: количество ограничений — это количество невключённых переменных. Пусть g_i — количество переменных, включённых в i-е уравнение. k_i — количество предопределённых переменных, включённых в данное уравнение. Тогда $R_i = (K - k_i) + (G - g_i)$. Тогда $K - K_i \geqslant g_i - 1$. $K - k_i$ — количество предопределённых переменных, исключённых из i-го уравнения. Если исключить слишком много, что тогда недоидентифицированность (нам нужно больше экзогенных переменных, что говорят эндогенные). Если исключить слишком мало, то будет переидентифицированность, и нам придётся употреблять линейные комбинации.

Завершающие слова: если уравнения точно идентифицируемые, то косвенный МНК даст результат. Лёгкий вывод: КМНК даёт точно такие же оценки, как и двухшаговый МНК, что приведёт к решению. Из всевозможных комбинаций инструментов подходит та, которая даёт самую точную ковариационную матрицу. Однако для некоторых уравнений мы ничего не получим, а для идентифицируемых и переидентифицируемых всё вроде получается. Придумал это Тейл.

Двухшаговый МНК. Первый шаг — построить OLS-регрессию на все остальные переменные. $\hat{\vec{y}}_t = \hat{\mathbf{\Pi}}\vec{x}_t$. Второй шаг — оценивать каждое структурное уравнение, то есть в правых частях заменять эндогенные переменные на их оценки с первого шага. Вместо y_2, например, возьмётся $\hat{y}_2$. Это частный случай МИП, где подбираются те, которые асимптотически обеспечивают наименьшую дисперсию.

Рассмотрим для примера некоторую макросистему. Потребление: $C = a_1 + b_1Y - c_1T + d_1R + u_1$. Real consumption ($C$), real income ($Y$), real tax ($T$), interest rate ($R$). Второе уравнение — функция инвестиций: $I = a_2 + b_2Y + c_2R + u_2$. Третье уравнение: $Y = C + I + G$. Четвёртое уравнение — liquidity preference: $M = a_3 + b_3Y + c_3R + d_3P + u_3$. Пятое соотношение — производственная функция: $Y = a_4 + b_4N + u_4$. Рынок труда, спрос: $N = a_5 + b_5W + c_5P + u_5$, где W — темп прироста денежных доходов (вариант кривой Филлипса). Предложение труда: $N = a_6 + b_6W + c_6P + u_6$.

Получается, 7 эндогенные переменных, $G = 7$: C, Y, I, R, N, P, W. $K = 3$, осталось G, T, M. Надо сконструировать $\boldsymbol{A\Phi}_1$. В первом уравнении 3 эндогенных и одна экзогенная. Выпишем таблицу присутствия переменных в уравнениях:

	C	I	N	P	R	Y	W	G	T	M
1	1	0	0	0	1	1	0	0	1	0
2	0	1	0	0	1	1	0	0	0	0
3	1	1	0	0	0	1	0	1	0	0
4	0	0	0	1	1	1	0	0	0	1
5	0	0	1	0	0	1	0	0	0	0
6	0	0	1	1	0	0	1	0	0	0
7	0	0	1	1	0	0	1	0	0	0

Выбрасываем строку уравнения, выписываем оставшиеся столбцы, соответствующие нулям в столбцах. Будет вот что:

I	N	P	W	G	M
1	0	0	0	0	0
1	0	0	0	1	0
0	0	1	0	0	1
0	1	0	0	0	0
0	1	1	1	0	0
0	1	1	1	0	0

Нас интересует равенство ранга шести. Единички — это какое-то ненулевое число. Ранг равен 5. Получается сверхидентифицированность. Проверим условие порядка: $R_i = 6$ (нулей в первой строке). $6 \geqslant 7 - 1$, условие порядка выполнено.

Третье переидентифицированное, последние два переидентифицируемые. По условию ранга последние два неидентифицируемые. Такую систему бессмысленно оценивать. Для разумных результатов надо в спрос или предложение труда добавить какие-то экзогенные переменные. Минимальная зарплата (предложение), а спрос — искать в производстве.

30 Лекция 30

Панельные данные — это то, что кормит зарубежных студентов с самого начала.

Cross-section — это рассмотрение кого-то в один момент (y_i). Если смотреть за поведением одного агента, то это time series (y_t). Третий тип данных — это то, что сочетает свойства и cross-section, и pooled data, — панельные данные. Это данные, которые обладают более сложной структурой ($y_t i$). Во временных рядах есть упорядоченность.

Началось это с того, что строили производственные функции для предприятий одной и той же отрасли. Все экономические агенты идут по следуемым траекториям. Longitudional survey — опросы населения. Бывают индивиды, бывают домохозяйства. Как правило, производят бюджетные обследования. Уровень образования, место жительства. Один индекс показывает номер индивидуума (номер его в cross-section), а второй — период времени (упорядоченность). Это последовательность cross-sections или много-много временных рядов.

Что есть несбалансированная панель? Один разорился, два объединились, объекты выбывают. Пытаются что-то заменить кем-то похожим, люди женились, развелись, дедушка с бабушкой умерли. Второй подход — заменить типичным.

Мы будем рассматривать сбалансированную панель. Договорились, что обычно, когда мы говорим о панельных данных, у нас $y_i t$ — переменная с номером наблюдения и временем. $i \in [1; n]$, $t \in [1, T]$. Панельные данные характеризуются тем, что $n \gg T$. Стандартная задача на панельных данных — отдача от образования. Какой выигрыш зарплате даёт такое и сякое образование. Она зависит от стажа, его квадрата, интеллекта и прочих факторов. И их у нас X_{it}^j — j-й фактор ($j \in [1; K]$). Здесь не зватает матриц, надо выстраивать третье измерение. Делают кронекеровские матрицы.

Введём следующие обозначения: $\vec{y}_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}$, $\boldsymbol{X}_i = \begin{pmatrix} X_{i1}^1 & X_{i1}^2 & \cdots & X_{i1}^k \\ \vdots & \vdots & \ddots & \vdots \\ X_{iT}^1 & X_{iT}^2 & \cdots & X_{iT}^k \end{pmatrix}$. Примем: $\vec{y}_{it} = \boldsymbol{X}_{it} \cdot \vec{\beta} + \varepsilon_{it}$.

Конечно, надо что-то сказать об ошибках. Введём ещё вектор: $\vec{y} = \begin{pmatrix} \vec{y}_1 \\ \vdots \\ \vec{y}_n \end{pmatrix}$ — и его размерность равна $N = nT$ — общее число наблюдений за все моменты времени. Для факторов то же самое: $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix}$, $\vec{\varepsilon} = \begin{pmatrix} \vec{\varepsilon}_1 \\ \vdots \\ \vec{\varepsilon}_n \end{pmatrix}$. Общая модель: $\vec{y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$. Нулевое матожидание ошибок, отсутствие корреляции, а мы введём более сильное условие: $\varepsilon_{it} \sim \text{i.i.d.}(0; \sigma_\varepsilon^2)$. Это слегка странно, ведь почему у нас нет корреляции самих с собой? Ладно, оставим это, ведь если это выполнено, то OLS-оценка $\hat{\vec{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$ будет BLUE. Это называется pooled regression — сквозная регрессия, которая игнорирует панельность структуры.

Однако самый интерес наблюдается тогда, когда мы проводим Russian Longitudional Monitoring Service. Российский мониторинг экономики и здоровья, если по-нашему. Он начался в конце 1990-х с участием зарубежных учёных. Это опрос, идущий волнами, большого числа домохозяйств. Говорят с главой домохозяйства. Они построили выборку по всей стране. Есть ядро одинаковых вопросов. Туда входят доходы, трансферты, трансферты, место жительства, состав семьи, статус, наличие детей, укрупнённые группы затрат (образование, здоровье). Понятно, что среди этих переменных — и это существенно — есть данные двух типов: общие (доход), а также индивидуальные (пол, несмотря на то, что современная медицина творит чудеса), которые со временем не меняются. Поэтому $\varepsilon_{it} = \alpha_i + u_{it}$, где α_i — индивидуальный эффект, инвариантный во времени. На самом деле это некоторая идеализация, но идея проста: α_i может немножко меняться, но между двумя наблюдениями за одним и тем же объектом есть больше сходства, чем между наблюдениями за двумя объектами. Выгодно различать индивидуальные и общие эффекты.

Первый подход: α_i детерминирована. Она fixed, и это названо fixed effects model (FE). Переводят в лоб: модель с фиксированными эффектами, но на самом деле с детерминированными. А чем он детерминирован? Домохозяйством, фирмой. Понятно, что в функции Кобба—Дугласа в общих стоят качество труда, капитала, менеджмента.

Второй подход: α_i индивидуален, но случаен (random effect, RE).

Рассмотрим быстро модель с одной переменной. В первом случае при наличии мультиколлинеарности надо договориться, что мы среди иксов константу не числим. Договоримся, что K на единицу меньше, чем обычно. Тогда зададим вопрос: если свободный член выделен и он разный для разных объектов, то это решается с помощью дамми-переменных. Fixed effect — Это применение дамми-переменных. После введения дамми у нас появятся простые матрицы из нулей и единичек. Если всё остальное выполнено, применяем OLS. Такая модель получила название LSDV (least squares, dummy variables). Ту же оценку можно получить более простым и обобщаемым путём.

Если только одна объясняющая переменная, то $y_{it} = \alpha_i + \beta x_{it} + u_{it}$, $u_{it} \sim \text{i.i.d.}(0; \sigma_u^2)$. Опять-таки, чтобы применять ТГМ, то не надо знать i.i.d., но если i.i.d., то тогда ТГМ выполняется.

$$\min\theta = \min\sum_{i,t}(y_{it} - \hat{\alpha}_i - \hat{\beta}x_{it})^2$$

$$\frac{\partial\theta}{\partial\alpha_i} = 0 \rightsquigarrow \sum_t (y_{it} - \hat{\alpha}_i - \hat{\beta}x_{it}) = 0,\ i = 1; \ldots; n$$

$$\frac{\partial\theta}{\partial\beta} = 0 \rightsquigarrow \sum_i x_{it}(y_{it} - \hat{\alpha}_i - \hat{\beta}x_{it}) = 0$$

Тогда $\sum\limits_t y_{it} = T\hat{\alpha}_i + \hat{\beta}\sum\limits_t x_{it}$, $\frac{1}{T}\sum\limits_t y_{it} = \hat{\alpha}_i + \frac{\hat{\beta}}{T}\sum\limits_T x_{it}$. Есть внутригрупповые средние: $\bar{y}_i = \frac{1}{T}\sum\limits_t y_{it}$, $\bar{x}_i = \frac{1}{T}\sum\limits_t x_{it}$. Тогда уравнение переписывается по средним: $\bar{y}_i = \hat{\alpha}_i + \hat{\beta}\bar{x}_i$. Далее подставим выражение $\hat{\alpha}$ в уравнение для частной производной по β:

$$\frac{\partial\theta}{\partial\beta} = 0 \rightsquigarrow \sum i, t x_{it}(y_{it} - \bar{y}_i + \hat{\beta}\bar{x}_i - \hat{\beta}x_{it}) = 0$$

Введём ещё одно обозначение: $W_{xxi} = \sum\limits_t (x_{it} - \bar{x}_i)^2$, что похоже на выборочную внутри групповую дисперсию. Аналогично, $W_{yyi} = \sum\limits_t (y_{it} - \bar{y}_i)^2$, $W_{xyi} = \sum\limits_t (x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)$. Последний шаг: $W_{xx} = \sum\limits_i W_{xxi}$, $W_{yy} = \sum\limits_i W_{yyi}$, $W_{xy} = \sum\limits_i W_{xyi}$. Суммы квадратов от общего среднего, где-то TSS, а где-то попарные произведения для

ковариации. Если учесть все наши преобразования, то тогда

$$\hat{\beta} = \frac{W_{xy}}{W_{xx}}$$

Если у нас есть такие средние, что $\bar{y} = \frac{1}{nT}\sum i, t y_{it}$ и тому подобные $\bar{x}$, то $\sum\limits_{i,t}(x_{it} - \bar{x})^2 = T_{xx}$ (total), $T_{yy} = \text{TSS}$, $\hat{\beta}_{\text{OLS}} = \frac{T_{xy}}{T_{xx}}$. Мы раскладываем сумму квадратов отклонений на две части: внутригрупповую и межгрупповую. Для последний модели RSS известен $\left(\text{RSS} = T_{yy} - \frac{T_{xy}^2}{T_{xx}}\right)$, а для $\frac{W_{xy}}{W_{xx}}$ $\text{ResSS} = W_{yy} - \frac{W_{xy}^2}{X_{xx}}$.

А если рассматривать несколько объясняющих переменных, то появятся несколько β, появятся допонительные уравнения, но нам неудобно считать много выражений. Поэтому проще использовать матричные обозначения. Вместо числа W_{xx} появится число «все иксы между собой», то есть матрица W_{xx} размерности $k \times k$, W_{yy} останется числом, а W_{xy} будет вектором.

В общем случае модель множественной регрессии даёт следующие результаты: $\hat{\alpha}_i = \bar{y}_i - \hat{\vec{\beta}}^T \bar{x}_i^j$. Тогда

$$\hat{\vec{\beta}} = \boldsymbol{W}_{xx}^{-1}\boldsymbol{W}_{xy}, \quad \hat{\vec{\beta}}_w = \hat{\vec{\beta}}_{\text{LSDV}}$$

$$\alpha_1 = \alpha_2 = \ldots = \alpha_n \Rightarrow y_{it} = \alpha + \beta x_{it} + u_{it}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\vec{\beta}} = T_{xx}^{-1} T_{xy} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$$

Рассмотрим более важную альтернативу — это random effects model. Там α_i есть случайная составляющая. Эта величина определяется случайно. Придётся говорить о взаимодействии источников случайности между собой. Второе название моделей — variance components model, модель компонентов регрессии. $u_{it} \sim \text{i.i.d.}(0; \sigma_u^2)$, $\alpha_i \sim \text{i.i.d.}(0; \sigma_\alpha^2)$, α_i и u_{jt} статистически независимы. $\text{Var}(\varepsilon_{it}) = \text{Var}(\alpha_i + u_{it}) = \sigma_\alpha^2 + \sigma_u^2$. Но $\text{Cov}(\varepsilon_{it}, \varepsilon_{is}) = \sigma_\alpha^2 + \sigma_u^2$ при $t = s$ и σ_α^2 при $t \neq s$. Ещё $\text{Cov}(\varepsilon_{it}, \varepsilon_{js}) = 0$ при $t \neq s$, $\forall i, j$

$$\text{Cov}(\vec{\varepsilon_i}) = \begin{pmatrix} \sigma_\alpha^2 + \sigma_u^2 & \sigma_\alpha^2 & \\ \sigma_\alpha^2 & \ddots & \\ & & \sigma_\alpha^2 + \sigma_u^2 \end{pmatrix} = \boldsymbol{\Sigma}, \quad \text{Cov}(\vec{\varepsilon}) = \begin{pmatrix} \boldsymbol{\Sigma} & 0 \\ 0 & \boldsymbol{\Sigma} \end{pmatrix}$$

Пойдём от регрессии с одной объясняющей переменной. Если определены $T_{xx} - W_{xx} = B_{xx}$, то это назовём between. $T_{yy} - W_{yy} = B_{yy}$, $T_{xy} - W_{xy} = B_{xy}$, а также $\theta = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$. Тогда

$$\hat{\beta}_{\text{GLS}} = \frac{W_{xy} + \theta B_{xy}}{W_{xx} + \theta B_{xx}}$$

Рассмотрим два крайних случая. Если $\theta = 1$, то $\hat{\beta}_{\text{GLS}} = \frac{T_{xy}}{T_{xx}}$, то есть когда альфы неслучайны, но одинаковы. Если дисперсии альфы нет, то pooled regression — частный случай. Если $\theta = 0$, то тогда fixed effect, LSDV, $\hat{\beta}_{\text{GLS}} = \hat{\beta}_{\text{LSDV}}$. Ноль — это когда дисперсия u равна нулю, а в пределе — когда дисперсия альф значительно больше. Надо разобрать, в каких случаях использовать детерминированный эффект, в каких — случайный.

31 Лекция 31

$$y_{it} = x_{it}\vec{\beta} + \varepsilon_{it}$$

Мы смотрели, что можно игнорировать структуру данных и строить сквозную (pooled) регрессию методом OLS, что не очень хорошо. Мы можем использовать, какое наблюдение есть какое: есть cross-section (n объектов), а есть маленькие траектории (T периодов) каждого из агентов ($n \geqslant T$). Если перемножить n и T, то тогда это предполагает однородность агентов. Но, кроме общих эффектов, бывают индивидуальные эффекты.

$$y_{it} = \alpha_i + x_{it}\vec{\beta} + \varepsilon_{it}$$

α_i — это коэффициент индивидуального эффекта, и возможность его выделить для нас часто важна. Мы интересуемся и α, и β. Если для разных объектов различается α, то можно ввести дамми на каждого. α_i — детерминированная константа, или fixed effect model.

$$\hat{\vec{\beta}}_{\text{FE}} \equiv \hat{\vec{\beta}}_{\text{LSDV}} \equiv \hat{\vec{\beta}}_w$$

Оказыватся, формула очень похожа на обычный OLS, только необходимо ввести $\bar{x}_i$ — среднее по времени для одного и того же объекта (внутригрупповое среднее, если группа — это время). Если рассмотреть $x_{it} - \bar{x}_i$,

то это называется within-оценки, которые даются within-~~императором~~ оператором.

$$\tilde{\vec{\beta}}_w = \boldsymbol{W}_{xx}^{-1}\vec{\boldsymbol{W}}_{xy}$$

α_i не зависит от ошибки. В этом случае очевидно, что наблюдения внутри одной и той же группы имеют дисперсию, зависящую от двух составляющих: α и ε. Это приводит к тому, что ковариационная матрица перестаёт быть диагональной; требуется применять GLS. Оказалось, что не надо палить из пушки по воробьям. Если мы введём $\boldsymbol{B}_{xx} = \boldsymbol{T}_{xx} - \boldsymbol{W}_{xx}$ ($T_{yy} = W_{yy} + B_{yy}$, или сумма квадратов отклонений вокруг групповых средний и между группами есть общая сумма квадратов).

$$\tilde{\beta}_{\text{GLS}} = \frac{W_{xy} + \theta B_{xy}}{W_{xx} + \theta B_{xx}}, \quad \theta = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

Если $\theta = 1$, то это обычная формула для МНК. Применили сквозную регрессию, и $\theta = 1$, когда нет индивидуального эффекта. Второй крайний случай — $\theta \to 0$, и в этом случае $\tilde{\beta}_{\text{GLS}} \to \tilde{\beta}_w$. А когда это может быть? Два случая. Первый — когда нет ошибки ε_{it} (не очень правдоподобно), второй — когда $T\sigma_a^2$ велико. Либо T велико, либо произведение само большое. Оба подхода предполагают, что мы принимаем во внимание индивидуальный эффект — индивидуальные различия агентов, составляющих выборку.

Модель fixed effect может требовать большого числа параметра. Из простого уравнения определяется каждая константа. Получается, что может остаться немного степеней свободы. Разница вот в чём: когда мы берём общий разброс для расчёта моделей с фиксированными эффектами (LSDV), она использует только within-информацию, то есть о несовпадении данных внутри группы. Если $W_{yy} \gg B_{yy}$, то мы можем потерять часть информации, что есть не очень хорошо.

Второй момент на содержательном уровне: наш общий эконометрический подход говорит: случайная составляющая мимикрирует собой все невключённые переменные, то есть отражает незнание или невозможность использовать больше информации. Используемые случайные составляющие передают наши незнания об общей структуре модели. Если у нас есть индивидуальный эффект, то мы не всё о нём знаем. Логичнее считать, что случайные α_i несут информации только об индивидуальных эффектах, которые мы не знаем. Её логично трактовать случайной. Часто предполагается, что мы предпочитаем RE, а не FE (коэффициентов меньше).

Но зачем мы строим модель? Ну да, хотим найти общую зависимость между темпами роста страны и какими-то характеристиками. Для каждой страны мы можем получить нечто, отличающее её от среднего в целом. Если мы добавляем страны, то тогда коэффициент α_i интересует. Random effect говорит: это штука случайная, так как есть большая генеральная совокупность, из которой мы выбрали. Мы не всех же опросили. Мы как-то их выбрали. Нас не интересует семья Ивановых и Рабиновичей; нам нужен общий эффект, а опрашивающий решил, что мы этого спросим. Поэтому естественнее коэффициент трактовать как случайный из большой выборки, и тогда мы обобщаем модель на всю выборку. Стран же мало, и 150 стран — это почти генеральная совокупность (с других планет мы их набрать не можем). Естественнее у фирмы и домохозяйств выделять random effect и интересоваться общими параметрами. А у стран важны индивидуальные валюты: доллар к йене, а не доллар в среднем.

Мы считаем, что если есть общие параметры, а есть индивидуальный эффект, который зависит от своих каких-то параметров. Кроме общих факторов, могут быть индивидуальные (нефть, население). Тогда $y_{it} = \alpha_i + \vec{\gamma}^T z_i + \vec{\beta}^T x_{it} + \varepsilon_{it}$. То есть в fixed effect α_i и $\vec{\gamma}^T z_i$ сливаются в одну большую константу, у которой нет индекса t (time invariant variables). Государственное устройство за 3 года не меняется. А в random effect не видно α_i, и всё хорошо различается.

Nerlove предложил следующий поход: сначала строим модель с фиксированными эффектами (LSDV). Получаем $\tilde{\vec{\beta}}_w$, $\hat{\sigma}_u^2$, а для оценки $\hat{\sigma}_\alpha^2$ предложено взять оценку дисперсии $\widehat{\text{Var}(\hat{\alpha})}$ в МНК, что даст $\tilde{\theta}$. Это не есть ни состоятельное, ни несмещённое. Но получаются лучшие оценки, чем некоторые методы, предложенные другими.

$$\tilde{\vec{\beta}}_{\text{GLS}} = \tilde{\vec{\beta}}_{\text{RE}} = (W_{xx} + \theta B_{xx})^{-1}(W_{xy} + \theta B_{xy})$$

Раз это GLS, то у этого подхода должен существовать результат: преобразовать данные и применить OLS. Это связано с именами Фуллера и Баттеза (Fuller, Battese). Они честно написали матрицы и посмотрели на преобразование.

$$y_{it} - \lambda\bar{y}_i, \quad x_{it} - \lambda\bar{x}_i, \quad \lambda = 1 - \sqrt{\theta}$$

Это сводится к OLS-регрессии. В компьютере это так и реализуется. Построили LSDV, посчитали оценку λ, преобразовали, а дальше обычный OLS. Многие потом это применяли. Но почти у всех, кто это применял, оценка LSDV очень близка к Random Effect. Так что эти два подходи не так уж сильно отличаются. Этот подход легко распространяется, если появляются коэффициенты зависят от времени и коэффициенты индивидуальных эффектов становятся случайными.

Перед нами стоит вопрос: FE против RE. Некто Mundlock показал, что противопоставление не так уж и

очевидно.

$$y_{it} = \alpha_i + \beta x_{it} + \varepsilon_{it}, \quad \alpha_i = \pi \bar{x}_i + w_i$$

$$y_{it} = \pi \bar{x}_i + \beta x_{it} + w_i + \varepsilon_{it} \rightsquigarrow \text{GLS}$$

$$y_{it} - \lambda \bar{y}_i = \pi(\bar{x}_i - \lambda \bar{x}_i) + \beta(x_{it} - \lambda \bar{x}_i) + v_{it} = \beta(x_{it} - \bar{x}_i) + \delta \bar{x}_i + v_{it}$$

Отклонения от среднего ортогональны. $\delta = (\pi+\beta)(1+\lambda)$, $\operatorname{Cov}\left((x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)\right) = W_{xy}$. Тогда $\hat{\vec{\beta}} = W_{xx}^{-1} W_{xy}$. Мы пошли за Мандлоком (Random Effect), а получили результат Fixed Effect. Если нет сильной разницы у α_i при внутригрупповом среднем, то тогда FE и RE похожи. Надо смотреть, сильно ли зависит α_i от $\bar{x}_i$. Если не зависит, то RE.

Применяется **тест Хаусмана** на неправильную спецификацию модели (Hausman Specification Error Test). Французы всё равно его назовут Осман. $\mathcal{H}_0$: модель правильно специфицирована. $\mathcal{H}_1$: модель неправильно специфицирована, причём каким-то определённым образом. Скажем, $\vec{Y} = \boldsymbol{X}\vec{\beta} + \vec{\varepsilon}$. $\mathcal{H}_0$: $\boldsymbol{X}$ и $\vec{\varepsilon}$ независимы ($\mathbb{E}(\vec{\varepsilon} \mid \boldsymbol{X}) = 0$), $\mathcal{H}_1 \colon \mathbb{E}(\vec{\varepsilon} \mid \boldsymbol{X}) \neq 0$, есть корреляция. То есть мы прямо специфицируем, где мы можем ошибаться.

Сравниваются два оценивателя: $\tilde{\vec{\beta}}_0$, $\tilde{\vec{\beta}}_1$. $\tilde{\vec{\beta}}_0$ состоятелен и эффективен (асимптотически) при $\mathcal{H}_0$ и несостоятелен при $\mathcal{H}_1$. В нашем примере это МНК. Про $\tilde{\vec{\beta}}_1$ мы знаем, что он даёт состоятельные оценки и при $\mathcal{H}_0$, и при $\mathcal{H}_1$. Но при $\mathcal{H}_0$ он будет не самым хорошим, не самым эффективным. Для $\mathcal{H}_0$ это OLS, для $\mathcal{H}_0$ это МИП.

Рассмотрим $\tilde{\vec{q}} = \tilde{\vec{\beta}}_1 - \tilde{\vec{\beta}}_0$. Было показано, что $\operatorname{Cov}(\tilde{\vec{q}}) = \operatorname{Cov}(\tilde{\vec{\beta}}_1) - \operatorname{Cov}(\tilde{\vec{\beta}}_0)$. Именно, что минус. Эти оценки берутся при $\mathcal{H}_0$. И тогда тестовая статистика есть квадратичная формочка:

$$m = \tilde{\vec{q}}^T \widetilde{\mathbf{Cov}(\tilde{\vec{q}})}^{-1} \tilde{\vec{q}} \overset{\text{as}}{\sim} \chi^2_k \tag{31.1}$$

Чтобы прочувствовать это, рассмотрим одномерный случай. При $\mathcal{H}_0$ $\operatorname{plim} \tilde{\beta}_0 = \operatorname{plim} \tilde{\beta}_1 = \beta \rightsquigarrow \operatorname{plim} \tilde{q} = 0$. Теперь $\tilde{d} = \tilde{\beta}_0 + \lambda \tilde{q}\ \forall \lambda$. Тогда $\operatorname{plim} \tilde{d} = \beta$, оценка состоятельная. Мы получили целый класс состоятельных оценок, и самое интересное, что все они состоятельные — лишь дисперсия разная. Тогда $\operatorname{Var}(\tilde{d}) = \operatorname{Var}(\tilde{\beta}_0) + \lambda^2 \operatorname{Var}(\tilde{q}) + 2\lambda \operatorname{Cov}(\tilde{\beta}_0, \tilde{q}) = V_0 + \lambda^2 \operatorname{Var}(\tilde{q}) + 2\lambda \operatorname{Cov}(\tilde{\beta}, \tilde{q}) \geqslant V_0$, и из обеих частей вычтем V_0. Если $\operatorname{Var}(\tilde{q}) = a^2$ и $\operatorname{Cov}(\tilde{\beta}_0, \tilde{q}) = b$, то получается парабола рожками вверх. Вспоминаем пятый класс: $a^2x^2 + 2\beta x \geqslant 0\ \forall x$. Дискриминант меньше нуля, откуда $b^2 \leqslant 0 \rightsquigarrow \beta = 0$.

Имеем: $\tilde{\beta}_1 = \tilde{\beta}_0 + \tilde{q} \rightsquigarrow V_1 = V_0 + \operatorname{Var}(\tilde{q}) \rightsquigarrow \operatorname{Var}(\tilde{q}) = V_1 - V_0$. Так и получилась разность двух ковариационных матриц. Тест Хаусмана показал, что статистика имеет распределение χ^2.

Применение теста Хаусмана для нашей задачи. Гипотеза $\mathcal{H}_0$: α_i не коррелированы с $\bar{x}_i$. Гипотеза $\mathcal{H}_1$: α_i коррелированы с $\bar{x}_i$. кто у нас β_0? Надо применять GLS, $\tilde{\vec{\beta}}_{\text{GLS}}$ будет и состоятельной, и асимптотически эффективной оценкой при $\mathcal{H}_0$ и не будет таковой при $\mathcal{H}_1$. А $\hat{\vec{\beta}}_w$ является состоятельным и при той гипотезе, и при той. Предлагается сравнить $\tilde{\vec{q}} = \hat{\vec{\beta}}_w - \tilde{\vec{\beta}}_{\text{GLS}}$. Получится $\widetilde{\operatorname{Cov}(\tilde{\vec{q}})} = \tilde{V}_1 - \tilde{V}_0 = \widetilde{V(\tilde{\vec{q}})}$, где распределение получается χ^2.

32 Лекция 32

Модели вероятностного выбора. Мы рассматривали задачу, где в качестве зависимой переменной была какая-то непрерывная переменная, а в объясняющей части были и количественные, и качественные переменные (дамми). Но есть ряд задач, в которых зависимая переменная есть переменная качественная. Человек принимает решение о виде транспорта, чтобы добраться на работу. Автомобиль или метро. Чтобы посмотреть на переменную, надо посмотреть не на регрессионную модель. Почему не подходит обычная регрессионная модель? Потому что мы считаем, что $Y_i = x_i'\beta + \varepsilon_i$, и в этой классике мы предполагаем, что модель правильно специфицирована, что нет систематической ошибки, ковариационная матрица имеет диагональный вид, а также прочие порождения Гаусса—Маркова. Что же будет, если мы применим к такой переменной обычный МНК?

Так как $\mathbb{E}(\varepsilon) = 0$, то из этого немедленно следует, что $\mathbb{E}(Y_i) = x_i'\beta$. Индикатор события имеет матожиданием вероятность. Тогда $Y_i = \begin{cases} 1, & p_i \\ 0, & 1 - p_i \end{cases}$. Проблема в том, что при определённых $x_i'\beta$ мы можем своей вероятностью в минус уйти в модели без ограничений. Но у нас $\mathbb{E}(Y_i) = x_i'\beta = p_i$ — линейная вероятность.

Y_i	ε_i	p_i
1	$1 - x_i'\beta$	$x_i'\beta$
0	$-x_i'\beta$	$1 - x_i'\beta$

$\operatorname{Var}(\varepsilon_i) = \operatorname{Var}(Y_i) = x_i'\beta(1 - x_i'\beta)$. В разных наблюдениях разная дисперсия. Нарушается нормальность и гомоскедастичность. Если мы не ставим задачи проверять гипотезы (что странно), то с гетероскедастичностью мы можем справиться через FGLS: $Y = x_i'\beta + \varepsilon_i \rightsquigarrow \hat{\beta}$, $w_i = \sqrt{x_i'\hat{\beta}(1 - x_i'\hat{\beta})}$, а на втором шаге $\frac{Y_i}{w_i} = \frac{x_i'}{w_i}\beta + \frac{\varepsilon_i}{w_i}$, откуда получаем $\hat{\hat{\beta}}$. Но даже так прогноз вероятности может уйти за $[0; 1]$. Эти модели очень активно использовались до 70-х. Ещё 10 лет назад в банках вероятность возврата таким образом считали. Если мы уверены, что функция выпукла вверх, то в качестве начального значения очень полезно выбирать $\hat{\hat{\beta}}$. Основная проблема продвинутых

методов — они чувствительны к начальному значению. Линейную вероятностную модель — грубую — используем там, где нужны состоятельные оценки. Если результаты согласуются друг с другом, то это добавляет доверия.

Как же загнать вероятность в рамки $[0;1]$? Достаточно просто использовать любую функцию распределения от линейной комбинации объясняющих факторов с неизвестными коэффициентами. $\mathbb{P}(Y_i = 1) = F(x_i'\beta)$. А с какой стати мы можем рассмотреть такую вероятность? Мы считаем, что $Y_i \in \{1;0\}$ — пороговые значения. Он выбирает единичку, когда полезность от авто выше полезности общественного транспорта. Есть какая-то латентная переменная Y_i^* (полезность), которая есть комбинация объясняющих факторов. И тогда

$$Y_i = \begin{cases} 1, & Y_i^* = x_i'\beta + \varepsilon_i > 0; \\ 0, & Y_i^* \leqslant 0. \end{cases}$$

Рассмотрим вопрос: будете ли вы голосовать за какого-то кандидата? У него есть не полезность от выбора кандидата, а внутреннее отношение к кандидату. И эта латентная переменная определяется возрастом, образованием, семейным положением. То, что мы видим, — галочка, а не это отношение. Переменная $\{1;0\}$ выскакивает, когда латентная переменная перешагивает определённый уровень.

Что мы выигрываем с этой латентной переменной? $\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i^* > 0) = \mathbb{P}(\varepsilon_i > -x_i'\beta) = 1 - F(-x_i'\beta)$, где F — функция распределения. Ошибка суммируется из большого числа ошибок, поэтому ошибки близки к нормальным. Так как две функции распределения у нас симметричные (нормальное, логистическое), то $1 - F(-x_i'\beta) = F(x_i'\beta)$.

Почему нормальное? Потому что много факторов, роль каждого невелика. Почему логистическое? Потому что не было вычислительной техники, а хотелось распределение, более близкое к нормальному. У него чуть потяжелее хвостики. Логистическое распределение:

$$F(t) = \frac{e^t}{1 + e^t}$$

Оценки коэффициентов очень близки между собой, они почти совпадают. Рассматривается стандартное нормальное, так как матожидание равно нулю. А почему дисперсия единичная? Пусть она не единица. Мы можем сделать $\frac{\varepsilon_i}{\sigma} > -x_i'\frac{\beta}{\sigma}$. Поэтому напрямую нормальное и логистическое сравнивать нельзя, надо поделить на логистическую дисперсию.

Есть распределение, неизвестны параметры. Тогда только метод максимального правдоподобия!

$$L = \prod_{i=1}^{n} [F(x_i'\beta)]^{y_i} \cdot [1 - F(x_i'\beta)]^{1-y_i}$$

Логарифмируем и записываем FOC.

$$\ln L = \sum y_i \ln F_i + \sum (1 - y_i) \ln(1 - F_i)$$

Продифференцируем:

$$\frac{\partial \ln L}{\partial \beta} = \sum y_i \frac{f_i \cdot x_i}{F_i} + \sum (1 - y_i) \frac{-f_i \cdot x_i}{1 - F_i}\bigg|_{\beta=\hat{\beta}} = 0$$

Это должно обращаться в ноль в $\beta = \hat{\beta}$. Это уравнение векторное, поэтому в явном виде это уравнение не решается. Даже если логистическая функция, интеграл не решается. Так как функция выпукла вверх, у уравнения существует единственное решение, получаемый либо градиентным методом, либо процедурой Ньютона—Рафсона. Мы получаем $\hat{\beta}$ — оценки МП. Если у нас единственный максимум, то оценки асимптотически несмещённые, асимптотически эффективные и асимптотически нормальные, а также состоятельные. Для нас все свойства желанные, особенно нормальность, так как мы сможем проверять набор гипотез. Асимптотическая нормальность:

$$\sqrt{n}(\hat{\beta} - \beta) \overset{\text{as}}{\approx} \mathcal{N}(0; \boldsymbol{I}^{-1}(\beta))$$

Поэтому данные модели работают только при большом числе наблюдений. Если их меньше ста, то вообще смысла нет. Сто — это бедность. Нужно over nine thousand obs! Но там индивиды и микроэконометрика, домохозяйства, тысячи найдутся. Кстати, чем больше выборка, тем труднее ожидать однородность и структурную стабильность.

В эконометрике две цели: как влияет тот или иной фактор и сделать прогноз. Как рассчитать предельный эффект?

$$\text{M.E.}(s)\frac{\partial \mathbb{P}(Y_i = 1)}{\partial \mathbb{P} x_s} = f(x_i'\beta)\beta_s, \quad \delta\mathbb{P} \approx \text{M.E.}(s) \cdot \delta x_s$$

Утешает, что численно эффект будет разным, однако направление действия можно определить сразу: функция плотности неотрицательна, и положительный коэффициент увеличивает вероятность за счёт со-

ответствующего фактора. Интерпретация: как вырастет вероятность высшего образования сына при росте дохода на тысячу баксов в год. Однако для каждого наблюдения предельный эффект считать — это сколько же их! А можно выбрать среднего представителя выборки, то есть типичного. Можно взять со средними характеристиками, но проблемы, где пол. Можно взять типичного мальчика, типичного девочка. Средние или медианные? Медианные харатеристики устойчивее. Тогда будем принимать средний эффект для непрерывных переменных $f(\bar{x}'\beta)\beta_s$. Иногда правильнее рассматривать возрастную группу. Чтобы рассмотреть предельный эффект по полу, рассмотрим предельный эффект по дамми:

$$\Delta P = \mathbb{P}(Y_i = 1 \mid D = 1) - \mathbb{P}(Y_i = 1 \mid D = 0)$$

Коэффициенты — это направление, надо об этом помнить. Пробит-анализ в зависимости от типа функции распределения называется пробит- или логит-моделью. Можно рассмотреть и другие распределения, но только в том случае, если объясняющие переменные качественные.

Рассмотрим пример вымирания призывников в ВС Англии. Выборка была 30 000 призывников. Единица — если призывник дожил до конца призыва. Объясняющие переменные — это качественные. У них там < 18, 18–19, > 19 — три группы. Группа интеллекта. Уровень образования. Статус и раса. Получается 180 возможных ячеек, а реально 150. Тогда для каждой ячейки считается частота вымирается, и для неё используется регрессионная модель. И всё оценивается по МНК.

Ясно, что модели эти не для прогноза. Но нужны прогнозы, хотя прогнозная сила не впечатлеет. Имея $\hat{\beta}$, мы можем оценить $\hat{p}_i = F(x_i'\hat{b})$. Функция от оценки будет функцией от самого параметра. Иначе говоря, инвериантность ММП: $\widehat{g(\beta)} = g(\hat{\beta})$. На олимпиаде в магистратуту матожидание обладало всепроникающим свойством. А эта оценка асимптотически несмещённая, эффективная, нормальная.

Логично ожидать, что $Y_0 = \begin{cases} 1, & \hat{p}_i > 0{,}5; \\ 0, & \hat{p}_i \leqslant 0{,}5. \end{cases}$ Ошибка первого типа: мы спрогнозировали перца как единичку, а он оказался нулём. Ошибка второго типа: спрогнозировали нолик и получили единицу. Цена этих ошибок не всегда равна: потерять тело кредита или проценты на кредите. У нас из одной группы в другую переходят 4–5 %, но если вероятность мелка, то это не значит, что её не надо рассматривать. Нам главное — не пропустить ноликов, поэтому мы поднимаем планку.

В одном крупном немецком банке просчитали модель, и там владение загородным домом связано с невозвратом. Выяснили, что в базе был ряд мошеннических операций. А мошенники всегда указывали, что владеют загородным домом, хотя мы-то рассчитываем на правдивую информацию. Если событие у нас очень редкое, то мы можем почти всё предсказать единицами или нулями. Там очень мало что можно выиграть, особенно если выборке несбалансированная.

Проверка гипотез. $\mathcal{H}_0\colon \boldsymbol{H}\vec{\beta} = q$. Такая гипотеза проверяется с помощью статистики Вальда:

$$W = (\boldsymbol{H}\hat{\vec{\beta}} - q)^T \Big(\boldsymbol{H} \cdot \widehat{\mathrm{Cov}_{\mathrm{AS}}(\hat{\vec{\beta}})}\boldsymbol{H}^T\Big)(\boldsymbol{H}\hat{\vec{\beta}} - q) \overset{\text{as}}{\sim} \chi^2_{\dim(q)}$$

Вальда легче программировать. Выдача очень похожа на выдачу экселя: оценки, стандартная ошибка z-статистики и p-value.

Григорий Гельмутович говорит, что есть эр-квадрат МакФаддена: $R^2_{\mathrm{MF}} = 1 - \underbrace{\dfrac{\ln L}{\ln L_0}}$ (без объясняющей переменной). Считается, что 5 % — это хорошо. А 50 % — это ого-го, бог! Просто МНК с этой характеристикой не связано, в принципе логарифм функции правдоподобия. Функция правдоподобия — это произведения тысяч того, кто меньше единицы. Смотреть на это — только расстраиваться. Смотреть только тогда, когда выбираем разное количество объясняющих факторов. Одна сотая и пять сотых — это та ещё радость. Часто в модель включается не только возраст, но и возраст в квадрате. Зависимость образования от доходе также квадратичная.

Есть красивый пример, как МакФадден вычислял, строить дорогу или не строить. И за вклад он получил в 2000-м Нобеля.

Рассмотрим модель множественного выбора. На кого пойти учиться: (1) врач, (2) экономист, (3) инженер. Такие значения называются unordered, неупорядоченные. Для них одни модели. Это Multiple Choice Model. А есть модель множественного выбора, где модель может быть упорядочена. Задаётся вопрос: как вы оцениваете своё социальное положение ($Y \in \{1; 2; 3\}$) (бесправные, средние, власть имущие). Unordered-модели разные, приравнивается каждый выбор к полезности, та максимизируется. Для трёх выборов нужно совместное нормальное распределение. И это вызывает напряжение. А ordered-модели можем оценить: насколько вам удалось себя реализовать в жизни. Выбирают чётное число ступенек. Если выбрать нечётное число, то тогда выберут в середину. И эта латентная переменная непрерывная, зависящая от индивидуальных характеристик. Тогда

$$Y_i = \begin{cases} 1, & Y_i^* < \alpha_1; \\ 2, & \alpha_i \leqslant Y_i^* < \alpha_2; \\ 3, & Y_i^* \geqslant \alpha_2; \end{cases}$$

Как записать функцию правдоподобия? $\mathbb{P}(Y_i = 1) = \mathbb{P}(\alpha_1 \geqslant x_i'\beta + \varepsilon_i < \alpha_2) = F(\alpha_2 - x_i'\beta) - F(\alpha_1 - x_i'\beta)$, и дефолтно считается, что $\alpha_0 = -\infty$, $\alpha_3 = +\infty$

$$L = \prod_{i:\ y_i=1} F(\alpha_1 - x_i'\beta) \prod_{i:\ y_i=2} [F(\alpha_2 - x_i'\beta) - F(\alpha_1 - x_i'\beta)] \prod_{i:\ y_i=3} [1 - F(\alpha_2 - x_i'\beta)]$$

Из этого можно определить, в каком направлении переменная действует на латентную переменную, но не на вероятность! С ростом дохода отличники переходили в хорошистов, а хорошисты переходили в троечников. И средний эффект надо считать для каждого наблюдения. И даже направление неизвестно!

33 Лекция 33

Говорят, что средний класс очень важен с точки зрения макро, так как они и производят, и потребляют. Рассмотрим вероятность перехода в средний класс или из среднего класса. Это РЛМС, панельные данные Вышки, где есть данные по домохозяйствам и индивидам. Была рассмотрена куча концепций среднего класса. Можно отсечь хвосты доходов. Но другие говорят, что о доходах они молчат, а по расходам можно оценить. Третьи говорят, что надо рассматривать профессиональные группы (белые воротнички).

Образование влияет хорошо. Предположили, что должна быть разница по регионам. Если человек работает в сфере ЖКХ, то туда не войдёт. Рассмотрели семейное хозяйство; если глава в браке, то там будет стабильно и хорошо. Надо было поделить доход домохозяйства на средний доход по региону. Надо было ещё осреднять по членам домохозяйств. Первый — просто поделить. Второй — с учётом экономии от масштаба при проживании количества совместных людей (поделили на корень из числа). Третий — главе приписать единичку, взрослым — 0,5, детям — 0,1. Сложить и поделить. Ещё логично предположить, что среднее по выборке равно единице: отношение среднего дохода в семте и среднего в регионе. Но это не так. Просто средний доход — это 0,6, что плохо. Второй и третий способ близки к единице. Нижняя группа — это ниже третьего дециля и регионального прожиточного минимума. Верхняя группа — это три верхних дециля. Далее определили переход: 2 года домохозяйство должно находиться в одной группе. Один год оно может быть где угодно. А потом два года подряд в новой доходной группе. Если смотреть средний доход, то слишком много из грязи в князи. Потом определили домохозяйства в группы по расходу. В выборке РЛМС нет богатых людей. Однако по расходам попадаются богатые индивиды, которые тратят куда больше, чем говорят что зарабатывают.

Список иллюстраций

Предметный указатель