Содержание

1. Методы получения оценок

Методы получения оценок: метод максимального правдоподобия, метод моментов, метод наименьших квадратов.

2. Свойства оценок

Свойства оценок: несмещённость, состоятельность, эффективность в классе.

2.1. Практическое напоминание об условном ожидании и дисперсии

Вспомним условное ожидание и условную дисперсию в дискретном случае:

Задача 1. Совместный закон распределения пары случайных величин (x, y) задан таблицей:

	$y = 1$	$y = 3$
$x = 1$	0.1	0.3
$x = 2$	0.1	0.1
$x = 4$	0.2	0.2

а) Найдите $\mathbb{E}(y \mid x)$, $\mathbb{V}\text{ar}(y \mid x)$.

б) Найдите $\mathbb{E}(y)$, $\mathbb{E}(x)$, $\mathbb{C}\text{ov}(x, y)$, $\mathbb{V}\text{ar}(x)$.

в) Найдите наилучшее линейное приближение $\text{BestLin}(y \mid x)$.

Решение. а) Условное математическое ожидание и дисперсия:

$$\mathbb{E}(y \mid x) = \begin{cases} 2.5, & x = 1, \\ 2, & x = 2, 4, \end{cases}$$

$$\mathbb{V}\text{ar}(y \mid x) = \begin{cases} 0.75, & x = 1, \\ 1, & x = 2, 4. \end{cases}$$

б) Математические ожидания, ковариация и дисперсия:

$$\mathbb{E}(y) = 2.2, \quad \mathbb{E}(x) = 2.4, \quad \mathbb{C}\text{ov}(x, y) = -0.28, \quad \mathbb{V}\text{ar}(x) = 1.84.$$

в) Наилучшее линейное приближение:

$$\text{BestLin}(y \mid x) \approx 2.57 - 0.15 \cdot x.$$

Теперь вспомним, как считать условные характеристики случайных величин при наличии совместной плотности:

Задача 2. Пара случайных величин (x, y) имеет функцию плотности

$$f(x, y) = \begin{cases} (2x + 4y)/3, \text{ если } x \in [0, 1], y \in [0, 1], \\ 0, \text{ иначе.} \end{cases}$$

а) Найдите $\mathbb{E}(y \mid x)$, $\mathbb{V}\text{ar}(y \mid x)$.

б) Найдите $\mathbb{E}(y)$, $\mathbb{E}(x)$, $\mathbb{C}\text{ov}(x, y)$, $\mathbb{V}\text{ar}(x)$.

в) Найдите наилучшее линейное приближение $\text{BestLin}(y \mid x)$.

Решение.

Здесь мудрый ассист напишет решение

Особо обратим внимание на случай двумерного нормального распределения:

Задача 3. Пара случайных величин (x, y) имеет совместное нормальное распределение

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 3 \\ 2 \end{pmatrix}, \begin{pmatrix} 10 & -2 \\ -2 & 20 \end{pmatrix} \right)$$

а) Найдите $\mathbb{E}(y \mid x)$, $\mathbb{V}\text{ar}(y \mid x)$.

б) Найдите $\mathbb{E}(y)$, $\mathbb{E}(x)$, $\mathbb{C}\text{ov}(x, y)$, $\mathbb{V}\text{ar}(x)$.

в) Найдите наилучшее линейное приближение $\text{BestLin}(y \mid x)$.

Решение.

Здесь храбрый ассист напишет решение

Обратите внимание: для совместного нормального распределения условное ожидание $\mathbb{E}(y \mid x)$ и наилучшее линейное приближение $\text{BestLin}(y \mid x)$ идеально совпадают. Условное ожидание и линейное приближение совпадут и в том случае, если величина x принимает всего два значения. Убедимся в этом с помощью простой задачи

Задача 4. На первом шаге Илон Маск случайным образом выбирает одно из двух значений случайной величины x, $\mathbb{P}(x = 1) = 0.4$, $\mathbb{P}(x = 2) = 0.6$. На втором шаге Шивон Зилис выбирает значение y из экспоненциального распределения x с интенсивностью x.

а) Найдите $\mathbb{E}(y \mid x)$, $\mathbb{V}\text{ar}(y \mid x)$.

б) Найдите $\mathbb{E}(y)$, $\mathbb{E}(x)$, $\mathbb{C}\text{ov}(x, y)$, $\mathbb{V}\text{ar}(x)$.

в) Найдите наилучшее линейное приближение $\text{BestLin}(y \mid x)$.

Решение.

Здесь неотразимый ассист напишет решение

Теперь найдём условное ожидание и условную дисперсию для совместного нормального распределения в общем виде.

Определение 2.1 (наилучшее линейное приближение)**.** Наилучшее линейное приближение величины r с помощью величины s — это линейная функция от s,

$$\text{BestLin}(r \mid s) = \beta_1 + \beta_2 s,$$

где константы β_1 и β_2 находятся из решения задачи оптимизации $\mathbb{E}((r - \text{BestLin}(r, s)^2) \to \min_{\beta_1, \beta_2}$. При решении задачи окажется

$$\beta_1 = \mathbb{E}(r) - \beta_2\,\mathbb{E}(s), \quad \beta_2 = \frac{\mathbb{C}\text{ov}(r, s)}{\mathbb{V}\text{ar}(s)}$$

Определение 2.2 (линейно-независимые случайные величины)**.** Величины r и s называются линейно-независимы если $\text{BestLin}(r \mid s) = \mathbb{E}(r)$.

Линейная независимость является симметричным явлением, $\text{BestLin}(r \mid s) = \mathbb{E}(r)$, если и только если $\text{BestLin}(s \mid r) = \mathbb{E}(s)$.

Задача 5. Выразите константы β_1 и β_2 в формуле для наилучшего линейного приближения

$$\text{BestLin}(r \mid s) = \beta_1 + \beta_2 s,$$

исходя из характеристик случайных величин r и s.

Решение. Выпишем целевую функцию в виде суммы

$$\mathbb{E}((r - \text{BestLin}(r, s)^2) = \mathbb{V}\text{ar}(r - \beta_1 - \beta_2 s) + (\mathbb{E}(r - \beta_1 - \beta_2 s))^2$$

Заметим, что β_1 не влияет на первое слагаемое, так как дисперсия константы равна нулю. И при этом, выбрав $\beta_1 = \mathbb{E}(r - \beta_2 s) = \mathbb{E}(r) - \beta_2\,\mathbb{E}(s)$ мы добьёмся того, что второе слагаемое будет равно нулю, своему наименьшему возможному значению.

Остаётся минимизировать с помощью β_2 первое слагаемое.

$$\mathbb{V}\text{ar}(r - \beta s) = \mathbb{V}\text{ar}(r) + \beta_2^2\,\mathbb{V}\text{ar}(s) - 2\beta_2\,\mathbb{C}\text{ov}(r, s) \to \min_{\beta_2}.$$

Перед нами квадратичная функция от β_2, следовательно,

$$\beta_2 = \frac{\mathbb{C}\text{ov}(r, s)}{\mathbb{V}\text{ar}(s)}.$$

Обратите внимание, эта формула — родная «теоретическая» сестра «выборочной» формулы для парной регрессии

$$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx}}.$$

Аналогия между оценкой и истинным коэффициентом действует и для первого коэффициента,

$$\beta_1 = \mathbb{E}(r) - \beta_2\,\mathbb{E}(s), \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}.$$

И, попутно, мы замечаем, что условие $\mathbb{C}\mathrm{ov}(r,s) = 0$ равносильно условию $\mathrm{BestLin}(r \mid s) = \mathbb{E}(r)$ или условию $\mathrm{BestLin}(s \mid r) = \mathbb{E}(s)$.

3. Асимптотические методы

Центральная предельная теорема. Лемма Слуцкого. Дельта-метод. Построение асимптотических доверительных интервалов.

4. Святая троица тестов

Три классических теста: LM, LR, Wald.
Чёрный трек: тесты в матричной форме для вектора параметров?

4.1. Черный трек

4.2. Тройка тестов в матричной форме

Рассмотрим применение тестов W (тест Вальда), LR (тест отношения правдоподобия) и LM (тест множителей Лагранжа) для тестирования гипотез о параметрах модели.

Пусть требуется протестировать систему ограничений относительно вектора неизвестных параметров

$$H_0 : \begin{cases} g_1(\theta) = 0 \\ g_2(\theta) = 0 \\ \dots \\ g_r(\theta) = 0 \end{cases}$$

где $g_i(\theta)$ — функция, которая задаёт i-е ограничение на вектор параметров θ, $i = 1, \dots, r$.

Введём следующие обозначения:

$$\frac{\partial g}{\partial \theta^T} = \begin{pmatrix} \partial g_1/\partial\theta^T \\ \partial g_2/\partial\theta^T \\ \vdots \\ \partial g_r/\partial\theta^T \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial\theta_1} & \frac{\partial g_1}{\partial\theta_2} & \dots & \frac{\partial g_1}{\partial\theta_k} \\ \frac{\partial g_2}{\partial\theta_1} & \frac{\partial g_2}{\partial\theta_2} & \dots & \frac{\partial g_2}{\partial\theta_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial g_r}{\partial\theta_1} & \frac{\partial g_r}{\partial\theta_2} & \dots & \frac{\partial g_r}{\partial\theta_k} \end{pmatrix}$$

$$\frac{\partial g^T}{\partial \theta} = \begin{pmatrix} \frac{\partial g_1^T}{\partial\theta} & \frac{\partial g_2^T}{\partial\theta} & \dots & \frac{\partial g_r^T}{\partial\theta} \end{pmatrix} = \begin{pmatrix} \frac{\partial g_1}{\partial\theta_1} & \frac{\partial g_2}{\partial\theta_1} & \dots & \frac{\partial g_r}{\partial\theta_1} \\ \frac{\partial g_1}{\partial\theta_2} & \frac{\partial g_2}{\partial\theta_2} & \dots & \frac{\partial g_r}{\partial\theta_2} \\ \dots & \dots & \dots & \dots \\ \frac{\partial g_1}{\partial\theta_k} & \frac{\partial g_2}{\partial\theta_k} & \dots & \frac{\partial g_r}{\partial\theta_k} \end{pmatrix}, \frac{\partial \ell}{\partial \theta} = \begin{pmatrix} \frac{\partial \ell}{\partial\theta_1} \\ \frac{\partial \ell}{\partial\theta_2} \\ \vdots \\ \frac{\partial \ell}{\partial\theta_k} \end{pmatrix}$$

$$I(\theta) = -E\left(\frac{\partial^2 \ell}{\partial\theta\partial\theta^T}\right) = -\mathbb{E}\begin{pmatrix} \frac{\partial^2\ell}{\partial\theta_1\partial\theta_1} & \frac{\partial^2\ell}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2\ell}{\partial\theta_1\partial\theta_k} \\ \frac{\partial^2\ell}{\partial\theta_2\partial\theta_1} & \frac{\partial^2\ell}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2\ell}{\partial\theta_2\partial\theta_k} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2\ell}{\partial\theta_k\partial\theta_1} & \frac{\partial^2\ell}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2\ell}{\partial\theta_k\partial\theta_k} \end{pmatrix}$$

— информационная матрица Фишера

$\Theta_{UR} := \Theta$ — множество допустимых значений вектора неизвестных параметров без учёта ограничений

$\Theta_R := \{\theta \in \Theta : g(\theta) = 0\}$ — множество допустимых значений вектора неизвестных параметров с учётом ограничений

$\hat{\theta}_{UR} \in \Theta_{UR}$ — точка максимума функции ℓ на множестве Θ_{UR}

$\hat{\theta}_R \in \Theta_R$ — точка максимума функции ℓ на множестве Θ_R

Тогда для тестирования гипотезы H_0 можно воспользоваться одной из следующих ниже статистик:

$LR := -2(\ell(\hat{\theta}_R) - \ell(\hat{\theta}_{UR})) \overset{as.}{\sim} \chi^2_r$ — статистика отношения правдоподобия

$W := g^T(\hat{\theta}_{UR}) \cdot \left[\frac{\partial g}{\partial \theta^T}(\hat{\theta}_{UR}) \cdot I^{-1}(\hat{\theta}_{UR}) \cdot \frac{\partial g^T}{\partial \theta}(\hat{\theta}_{UR})\right]^{-1} g(\hat{\theta}_{UR}) \overset{as.}{\sim} \chi^2_r$ — статистика Вальда

$LM := \left[\frac{\partial \ell}{\partial \theta}(\hat{\theta}_R)\right]^T \cdot I^{-1}(\hat{\theta}_R) \cdot \left[\frac{\partial \ell}{\partial \theta}(\hat{\theta}_R)\right] \overset{as.}{\sim} \chi^2_r$ — статистика множителей Лагранжа

5. ШБ — МНК

В этой главе мы приступим к изучению регрессионного анализа. Представьте, что вы работаете риэлтором или собираетесь приобрести квартиру в Москве. В обоих случаях Вас может интересовать, от чего зависит стоимость жилья (к примеру, от площади квартиры, расположения в том или ином районе Москвы и т.д.). Регрессионный анализ позволяет не только прогнозировать стоимость нового жилья, но и определять, какие факторы влияют на цену и в какую сторону — увеличивают или уменьшают её.

Для того чтобы корректно построить модель регрессии, сперва нам необходимо поставить оптимизационную задачу метода наименьших квадратов (МНК), найти её решение и обнаружить нестатистические свойства оценок.

5.1. Оптимизационная задача

Моделью парной регрессии называется модель вида

$$y_i = \beta_1 + \beta_2 x_i + u_i,$$

где y_i — значение зависимой (или, иначе, объясняемой) переменной для i-го наблюдения, x_i — значение объясняющей переменной (иногда её называют фактором или регрессором, иногда признаком :)) для i-го наблюдения, β_1 — свободный коэффициент (константа), β_2 — коэффициент при факторе x, u_i — значение случайной ошибки для i-го наблюдения.

МНК-оценки коэффициентов модели парной регрессии находятся из решения задачи минимизации

$$Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) \to \min_{\hat{\beta}_1, \hat{\beta}_2}.$$

FOC (first order condition, условие первого порядка):

$$\begin{cases} \frac{\partial Q(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \\ \frac{\partial Q(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)x_i = 0. \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0, \\ \sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)x_i = 0. \end{cases}$$

Разрешим данную систему уравнений относительно неизвестных $\hat{\beta}_1$ и $\hat{\beta}_2$, получим МНК-оценки коэффициентов β_1 и β_2:

$$\begin{cases} \hat{\beta}_2^{\text{ols}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{sCov}(x,y)}{\text{sVar}(x)}, \\ \hat{\beta}_1^{\text{ols}} = \bar{y} - \hat{\beta}_2^{\text{ols}}\bar{x}. \end{cases}$$

Напомним, что $\text{sCov}(x, y)$ и $\text{sVar}(x)$ — это выборочная ковариация и выборочная дисперсия, определённые нами ранее в главе 2 (TODO: чекнуть, что в главе 2),

$$\text{sCov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1},$$

$$\text{sVar}(x) = \text{sCov}(x, x) = \frac{\sum(x_i - \bar{x})^2}{n - 1}$$

5.2. Три простых модельки

Рассмотрим три простых регрессионных модели.
В качестве тренировки мы предлагаем вам вывести формулы оценок коэффициентов в этих моделях.

1. В регрессии на константу

   $$y_i = \beta_1 + u_i$$

   МНК-оценка параметра β_1 определяется по формуле

   $$\hat{\beta}_1 = \bar{y}.$$

2. В модели регрессии без константы

   $$y_i = \beta_1 x_i + u_i$$

   МНК-оценка параметра β_1 равна

   $$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{x_i y_i}{x_i^2}.$$

   В идеальной ситуации после оценивания коэффициентов модели вы ещё можете объяснить окружающим, какие полезные выводы можно сделать из получившихся оценок. Проинтерпретируем коэффициент при x_i. Увеличим объясняющую переменную на единицу, при этом старый прогноз равен $\hat{y}_i^{\text{old}} = \hat{\beta}_1 x_i$, новый прогноз равен $\hat{y}_i^{\text{new}} = \hat{\beta}_1(x_i + 1)$ и разница прогнозов равна $\Delta y = \hat{y}_i^{\text{new}} - \hat{y}_i^{\text{old}} = \hat{\beta}_1$. Получается, что рост x_i на единицу приводит к изменению прогноза зависимой переменной на $\hat{\beta}_1$ единиц. Тот же самый результат можно получить взятием производной: $\frac{\partial \hat{y}_i}{\partial x_i} = \hat{\beta}_1$. Наша модель — линейная и изменение любого x_i на единицу окажет один и тот же эффект на предсказанный y_i.

   В качестве иллюстрации рассмотрим модель CAPM (Capital Asset Pricing Model) в её самом простом виде. В теории обычно предполагается, что премия за риск для некоторой ценной бумаги линейно зависит от рыночной премии за риск:

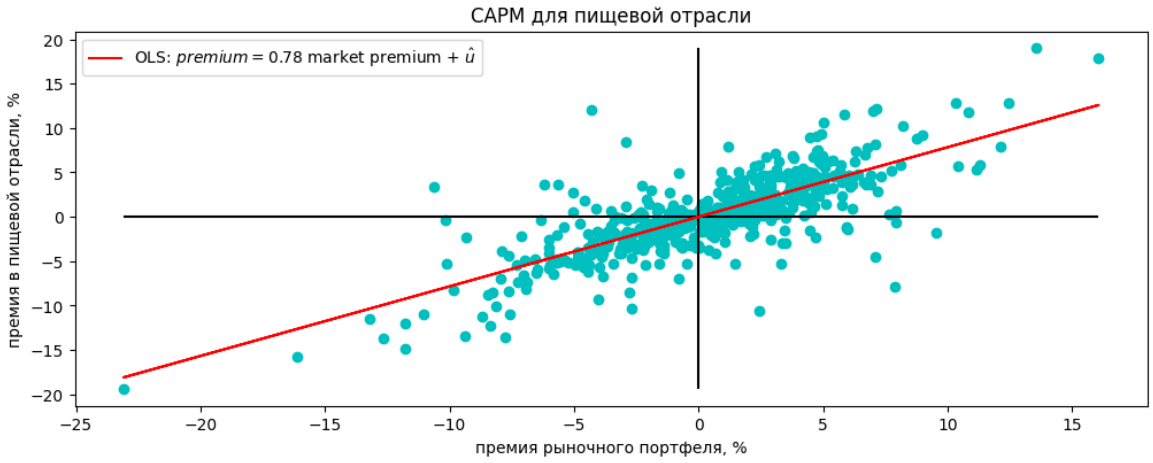


Рис. 1: CAPM модель, регрессия без константы

$$\text{premium}_i = \beta_1 \times \text{market-premium}_i + u_i$$

где premium_i — премия за риск[1] для ценной бумаги за i-й период, market-premium_i — рыночная премия за риск[2], β_1 — мера систематического (рыночного) риска бумаги (портфеля).

Найдём МНК-оценку для ценных бумаг в пищевом секторе: $\hat{\beta}_1 = 0.78$. Получается, что рост премии за риск на рынке на 1 пп (процентный пункт) приводит к тому, что доходность рассматриваемой ценной бумаги растёт на 0.78 пп. На рисунке 1 обратите внимание на то, что линия регрессии без константы проходит через начало координат. Коэффициент получился меньше единицы, значит, покупать ценные бумаги в пищевом секторе — достаточно мало рисковая операция, так как рост премии за риск в пищевом секторе слабее, чем в среднем по рынку.

3. В модели парной регрессии с константой

$$y_i = \beta_1 + \beta_2 x_i + u_i$$

МНК-оценки для параметров β_1 и β_2 выглядят следующим образом:

$$\hat{\beta}_2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2}, \quad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$$

Результаты оценивания коэффициента $\hat{\beta}_2$ можно проинтерпретировать аналогично предыдущей модели, производная прогноза по регрессору равна $\frac{\partial \hat{y}_i}{\partial x_i} = \hat{\beta}_2$. После интерпретации эффекта, который x_i оказывает на y_i, можно заняться интерпретацией константы. Оценка коэффициента $\hat{\beta}_1$ — это прогноз $\hat{y}_i$ при $x_i = 0$. Иногда оценка константы может иметь смысл. Давайте не поверим классической постановке модели и оценим модель CAPM с константой. В этом случае линия регрессии не обязана проходить через точку $(0, 0)$ на графике.

Будем использовать те же самые обозначения для премии за риск для выбранных ценных бумаг и рыночной премии за риск:

[1] разница между доходностью бумаги и безрисковой ставкой, например, по государственным облигациям
[2] разница между доходностью рыночного портфеля, например, S&P500, и той же безрисковой ставкой

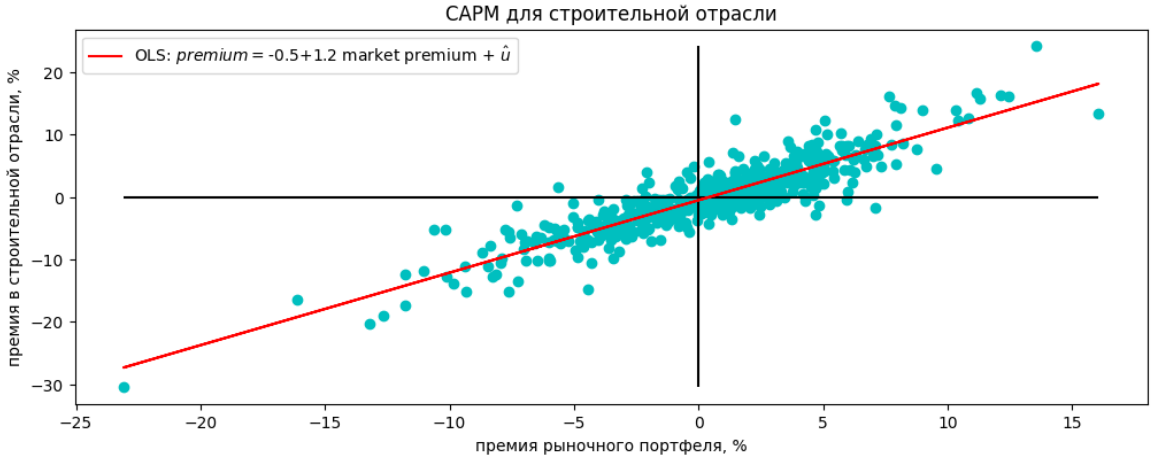


Рис. 2: CAPM модель, регрессия с константой

$$\text{premium}_i = \beta_1 + \beta_2 \times \text{market-premium}_i + u_i,$$

где β_2 – мера систематического (рыночного) риска бумаги (портфеля).

Оценённая линия регрессии для строительной отрасли получилась следующая:

$$\text{premium}_i = -0.5 + 1.2 \times \text{market-premium}_i + \hat{u}_i$$

Линия регрессии изображена на рисунке 2. При росте премии за риск для рыночного портфеля на 1 пп премия в строительной отрасли растёт на 1.2 пп, что говорит о, видимо, более высоких рисках в строительной сфере, которые хозяева фирм пытаются компенсировать. Более того, у нас оценена константа $\hat{\beta}_1 = -0.5$. Если бы доходность рыночного портфеля равнялась безрисковой ставке (рыночная премия за риск $\text{market-premium}_i = 0$), то премия за риск в строительной отрасли была бы -0.5. То есть, если бы рыночный портфель был безрисковым, то инвесторы уходили бы из строительного сектора.

На самом деле, константа не всегда осмысленна. Рассмотрим короткий пример зависимости длины остановочного пути[3] от скорости автомобиля:

$$\text{dist}_i = \beta_1 + \beta_2 \times \text{speed}_i + u_i,$$

Оценённая модель (изображена на рисунке 3) получилась следующей:

$$\text{speed}_i = -58.6 + 21.1 \times \text{dist}_i + \hat{u}_i.$$

Коэффициенты показывают, что изменение скорости на один километр в час (ceteris paribus, при прочих равных факторах) приводит к росту остановочного пути на 21 метр. Если же скорость равна нулю, то не стоит говорить, что остановочный путь должен составить -60 метров. В данном случае константа просто принимает значение параметра для наилучшего прохождения прямой через точки.

5.3. Как я перестал беспокоиться и полюбил матричное дифференцирование

Мы планируем перейти к рассмотрению регрессий с большим числом факторов, поэтому полезна будет матричная запись модели и, соответственно, матричное дифференцирование. Этому и будет посвящён данный раздел.

[3] путь, пройденный за время реакции водителя и фактического торможения

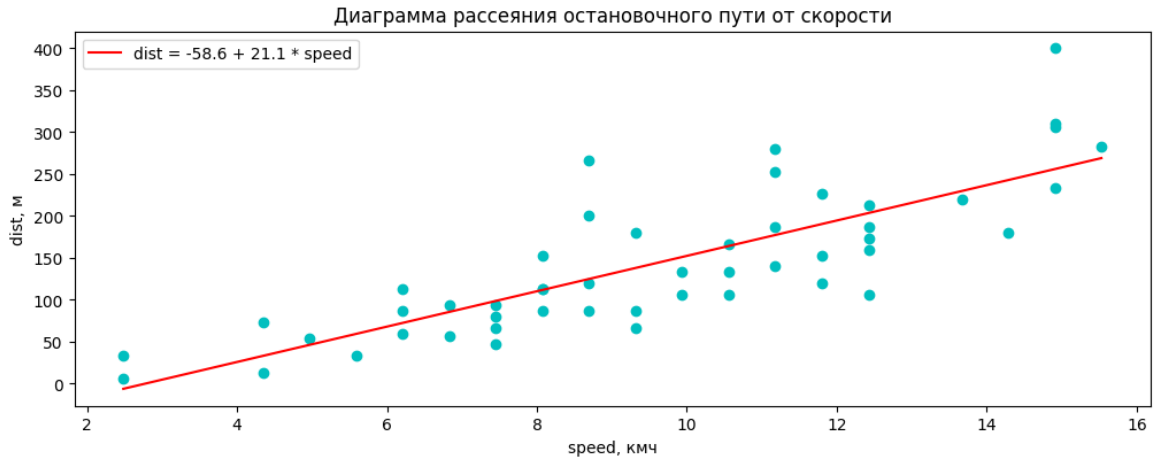


Рис. 3: Пример модели с константой, которая ничего не означает

Много полезных фактов про матрицы можно найти в шикарной книжке «Поваренная книга любителя матриц» Петерсона и Педерсона [PP12].

Нам чаще всего придётся дифференцировать скалярную функцию по векторному аргументу. По умолчанию вектор записывают столбцом и в большинстве источников производная по векторному аргументу тоже является столбцом,

$$\frac{\partial r}{\partial s} = \operatorname{grad} r = \begin{pmatrix} \partial r/\partial s_1 \\ \partial r/\partial s_2 \\ \vdots \\ \partial r/\partial s_k \end{pmatrix}$$

Это сделано для того, чтобы размер результата дифференцирования совпадал с размером вектора s, по которому дифференцируют, и равнялся $[k \times 1]$. Например, для функции $f(x) = x_1^2 + x_2^3 + x_1 \cdot x_3^4$ векторная производная равна

$$\frac{\partial f}{\partial x} = \operatorname{grad} f = \begin{pmatrix} 2x_1 + x_3^4 \\ 3x_2^2 \\ 4x_1 x_3^3 \end{pmatrix}.$$

Сформулируем основные правила дифференцирования скалярных выражений по векторному аргументу:

$$\frac{\partial a^T s}{\partial s} = \frac{\partial s^T a}{\partial s} = \frac{\partial \sum s_i a_i}{\partial s} = a = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

$$\frac{\partial s^T A s}{\partial s} = \frac{\partial \sum_{ij} A_{ij} s_i s_j}{\partial s} = (A + A^T)s$$

В частности, для суммы квадратов правило превращается в

$$\frac{\partial s^T s}{\partial s} = \frac{\partial \sum s_i^2}{\partial s} = 2s = \begin{pmatrix} 2s_1 \\ 2s_2 \\ \vdots \\ 2s_n \end{pmatrix}$$

В более сложном случае дифференцирования вектора по вектору оказывается полезной матрица Якоби. В ней строки отвечают за элементы дифференцируемой функции, а столбцы — за элементы вектора, по которому дифференцируют.

Определение 5.1 (матрица Якоби)**.** Для векторов r размера $[n \times 1]$ и s размера $[k \times 1]$ производной $\partial r/\partial s$ или матрицей Якоби называют матрицу, в которой дифференцируемые элементы записывают по строкам, а элементы, по которым берут производную, — по столбцам:

$$J = \frac{\partial r}{\partial s} = \begin{pmatrix} \partial r_1/\partial s_1 & \partial r_1/\partial s_1 & \dots & \partial r_1/\partial s_n \\ \partial r_2/\partial s_1 & \partial r_2/\partial s_1 & \dots & \partial r_2/\partial s_n \\ \vdots & & & \vdots \\ \partial r_n/\partial s_1 & \partial r_n/\partial s_1 & \dots & \partial r_n/\partial s_n \end{pmatrix}.$$

Например, для функции $f(x) = \begin{pmatrix} x_1 + 3x_2 \\ x_1 \cdot x_2^2 \end{pmatrix}$ матрица Якоби равна

$$J = \frac{\partial f}{\partial x} = \begin{pmatrix} 1 & 3 \\ x_2^2 & 2x_1x_2 \end{pmatrix}.$$

Будьте бдительны! Обозначение для матрицы Якоби часто используют ровно то же, что для производной скалярной функции по векторному аргументу, $\partial r/\partial s$. При этом для скалярной функции r матрица Якоби «положит» все производные в строчку, а производная скалярной функции по векторному аргументу положит те же производные в столбик.

$$J = (\partial r/\partial s_1, \dots, \partial r/\partial s_n) \quad \frac{\partial r}{\partial s} = \begin{pmatrix} \partial r/\partial s_1 \\ \dots \\ \partial r/\partial s_n \end{pmatrix}$$

Избежать этой путаницы с упаковкой производных то в строчку, то в столбец позволяет дифференциал. У дифференциала всегда-всегда размерность совпадает с размерностью исходного дифференцируемого объекта.

Например, для функции $f(x) = \begin{pmatrix} x_1 + 3x_2 \\ x_1 \cdot x_2^2 \end{pmatrix}$ дифференциал равен

$$df = \begin{pmatrix} dx_1 + 3dx_2 \\ x_2^2 dx_1 + 2x_1x_2dx_2 \end{pmatrix}.$$

Если брать дифференциал скалярной функции s по векторному аргументу r, то он примет вид

$$ds = (\partial s/\partial r)^T dr = \text{grad}^T\, s dr,$$

поэтому из дифференциала легко извлечь градиент.

Напишем и правила для работы с дифференциалом. Здесь A, B — постоянные матрицы; a, b — постоянные векторы; R, S — матрицы переменных; r, s — векторы переменных.

Дифференциал постоянной матрицы равен нулю:

$$dA = 0.$$

При взятии дифференциала произведения важно сохранять порядок матриц R и S:

$$d(RS) = dR \cdot S + R \cdot dS.$$

В частности,

$$d(ARB) = A \cdot dR \cdot B.$$

Для суммы квадратов правило превращается в

$$ds^T s = d(\sum s_i^2) = 2s^T ds = 2\sum s_i ds_i.$$

Линейность сохраняется для следа матрицы

$$d\,\text{trace}\,R = \text{trace}\,dR.$$

5.4. Матричное представление регрессии

Пусть теперь в модель для y включены k регрессоров $x_1, x_2, \dots x_k$. Если в модель регрессии включена константа, то мы считаем, что $x_{i1} = 1$ для всех $i = 1, \dots, n$. Модель вида

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

назовём моделью множественной регрессии.

Вектор зависимой переменной y имеет размер $[n \times 1]$, матрица признаков $X - [n \times k]$, вектор параметров модели $\beta - [k \times 1]$, вектор случайной ошибки $u - [n \times 1]$:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1k} \\ x_{21} & \dots & x_{2k} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{pmatrix}.$$

Используя введённые обозначения модель множественной регрессии

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$$

можно переписать в матричном виде

$$y = X\beta + u.$$

5.5. Решение оптимизационной задачи МНК с матрицами

Для матричную регрессионной модели

$$y = X\beta + u$$

оптимизационную задачу МНК можно переписать как

$$Q(\hat{\beta}) = (y - X\hat{\beta})^T (y - X\hat{\beta}) = \hat{u}^T \hat{u} \to \min_{\hat{\beta}}.$$

Найдём МНК-оценку вектора β, используя матричное дифференцирование.

Запишем необходимое условие для задачи минимизации:

$$\frac{\partial Q(\hat{\beta})}{\partial \hat{\beta}} = -2X^T y + 2X^T X\hat{\beta} = 0$$

$$2X^TX\hat{\beta} = 2X^Ty$$

Если матрица X^TX размера $[k \times k]$ обратима, то формула для оценок принимает вид

$$\hat{\beta}_{\text{ols}} = (X^TX)^{-1}X^Ty.$$

Для проверки достаточных условий второго порядка найдём матрицу Гессе в точке оптимума

$$\frac{\partial^2 Q(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}^T} = 2X^TX.$$

Если матрица X^TX обратима, то она положительно определена и $\hat{\beta}_{\text{ols}}$ — точка глобального минимума.

Для полноты изложения найдём условие первого порядка для оптимального $\hat{\beta}$ и матрицу Гессе альтернативным способом — через дифференциал.

Запишем дифференциал для функции $Q(\hat{\beta})$ и выделим внутри него градиент,

$$dQ(\hat{\beta}) = 2(X\hat{\beta} - y)^TXd\hat{\beta} = (\text{grad}\, Q(\hat{\beta}))^Td\hat{\beta}.$$

Приравняем градиент к нулю

$$\text{grad}\, Q(\hat{\beta}) = 2X^T(X\hat{\beta} - y) = 0,$$

и получим прежнее условие первого порядка:

$$2X^TX\hat{\beta} = 2X^Ty$$

Матрицу Гессе можно выделить внутри второго дифференциала,

$$d^2Q = d(2(X\hat{\beta} - y)^TXd\hat{\beta}) = d\hat{\beta}^T \cdot 2X^TX \cdot d\hat{\beta}.$$

Как и ранее, матрица Гессе равна $\frac{\partial^2 Q(\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}^T} = 2X^TX$.

Полученную оценку $\hat{\beta} = (X^TX)^{-1}X^Ty$ называют МНК-оценкой $\hat{\beta}_{\text{ols}}$. Далее в целях снижения занудства мы иногда будем опускать нижний индекс ols и писать $\hat{\beta}$ вместо $\hat{\beta}_{\text{ols}}$.

5.6. Геометрия МНК

Кучу интересных геометрических фактов можно найти в статье «Как встретились Гаусс, Марков и Пифагор?», [GD18].

Метод наименьших квадратов имеет шикарную геометрическую интерпретацию. Геометрия позволяет легко получить многие свойства оценок.

Для удобства рассмотрим случай двух регрессоров с константой

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 r_i + \hat{\beta}_3 s_i.$$

Обозначим вектор из сплошных единиц буквой s, $s = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$. В этом случае,

$$\hat{y}_i = \hat{\beta}_1 s_i + \hat{\beta}_2 r_i + \hat{\beta}_3 s_i$$

Или

$$\hat{y} = \hat{\beta}_1 s + \hat{\beta}_2 a + \hat{\beta}_3 b.$$

Вектор $\hat{y}$ — это линейная комбинация векторов s, a и b, $\hat{y} \in \text{Span}(s, a, b)$. Для наглядности можно представлять себе конкретный пример,

$$y = \begin{pmatrix} 2 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \quad X = \begin{pmatrix} | & | & | \\ s & a & b \\ | & | & | \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 0 & 3 \\ 1 & 0 & 4 \end{pmatrix}$$

Теперь вспомним целевую функцию метода наименьших квадратов

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \to \min_{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3}.$$

Величина $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ — это квадрат длины вектора $y - \hat{y}$, то есть $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \|y - \hat{y}\|^2$.

И теперь мы можем сформулировать метод наименьших квадратов геометрически!

Суть 5.1. Метод наименьших квадратов ищет вектор прогнозов $\hat{y}$ внутри линейной оболочки регрессоров $\text{Span}(s, a, b)$ поближе к вектору зависимой переменной y.

Конечно же, оптимальным решением оказывается проекция вектора y на линейную оболочку $\text{Span}(s, a, b)$.

Условием первого порядка будет ортогональность вектора остатков $\hat{u} = y - \hat{y}$ каждому из регрессоров,

$$\hat{u} \perp \text{Span}(s, a, b) \quad \Leftrightarrow \quad \begin{cases} \hat{u} \perp s \\ \hat{u} \perp a \\ \hat{u} \perp b \end{cases}.$$

Поскольку оптимальный вектор прогнозов $\hat{y}$ лежит в линейной оболочке $\text{Span}(s, a, b)$, то вектор остатков $\hat{u}$ перпендикулярен и ему тоже, $\hat{u} \perp \hat{y}$.

Условие ортогональности векторов означает, что скалярное произведение равно нулю, поэтому

$$\begin{pmatrix} — s^T — \\ — a^T — \\ — b^T — \end{pmatrix} \cdot \hat{u} = 0 \quad \Leftrightarrow \quad X^T \hat{u} = 0.$$

А далее из условия ортогональности регрессоров и остатков $X^T \hat{u} = 0$ можно получить и явно формулы оценок всех коэффициентов. Подставим формулу для прогнозов, $\hat{y} = X\hat{\beta}$, и решим полученное уравнение $X^T(y - X\hat{\beta}) = 0$. Раскрываем скобки,

$$X^T y - X^T X \hat{\beta} = 0 \quad \Leftrightarrow \quad X^T X \hat{\beta} = X^T y.$$

Если нам повезло, и матрица $X^T X$ обратимая, то

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Заметим, что просто сократить матрицу X^T слева и справа нельзя потому, что она точно не обратимая! Матрица X^T имеет размер $[k \times n]$.

Готовая формула для вектора прогнозов равна

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y.$$

Определение 5.2 (матрица-шляпница)**.** Матрица $X(X^TX)^{-1}X^T$ часто обозначается буквой H и неформально называется матрицей-шляпницей (hat matrix). Она «надевает» на y шляпку, $Hy = \hat{y}$. Формально матрица H также называется матрицей-проектором. Она проецирует любой вектор на линейную оболочку всех регрессоров $\mathrm{Span}(\mathrm{col}_1 X, \mathrm{col}_2 X, \ldots, \mathrm{col}_k X)$.

Вектор $\hat{\beta}$ состоит из k оценок $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$, поэтому его напрямую невозможно нарисовать в пространстве $\mathbb{R}^n$. Однако, оказывается, что веса, с которыми компоненты вектора зависимой переменной входят в каждую из оценок, можно изобразить! Заметим, что вектор оценок можно записать в виде

$$\hat{\beta} = (X^TX)^{-1}X^Ty = W^Ty, \text{ где } W = X(X^TX)^{-1}.$$

$$\hat{\beta} = \begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_a\\ \hat{\beta}_b\end{pmatrix} = \begin{pmatrix}— & w_1^T & —\\ — & w_2^T & —\\ — & w_3^T & —\end{pmatrix} \cdot \begin{pmatrix}|\\ y\\ |\end{pmatrix} = \begin{pmatrix}| & | & |\\ w_1 & w_2 & w_3\\ | & | & |\end{pmatrix}^T \cdot \begin{pmatrix}|\\ y\\ |\end{pmatrix} = \begin{pmatrix}1.75 & -0.5 & -0.5\\ -1.75 & 1.5 & 0.5\\ 2.25 & -1.5 & -0.5\\ -1.25 & 0.5 & 0.5\end{pmatrix}^T \cdot \begin{pmatrix}|\\ y\\ |\end{pmatrix}.$$

То есть, каждая оценка $\hat{\beta}_j$ — это взвешенные наблюдения зависимой переменной y_1, y_2, ..., y_n. Например, оценка первого коэффициента $\hat{\beta}_1$ — это скалярное произведение первого столбца W и зависимой переменной y,

$$\hat{\beta}_1 = \langle \mathrm{row}_1 W^T, y\rangle = \langle \mathrm{col}_1 W, y\rangle = \langle w_1, y\rangle = w_{11}y_1 + w_{21}y_2 + \cdots + w_{n1}y_n$$

В нашем частном случае четырёх наблюдений,

$$\hat{\beta}_1 = 1.75y_1 - 1.75y_2 + 2.25y_3 - 1.25y_4.$$

Обратим внимание на запись $W = X(X^TX)^{-1}$. Для столбца весов w_1 она означает, что

$$\begin{pmatrix}|\\ w_1\\ |\end{pmatrix} = \begin{pmatrix}1.75\\ -1.75\\ 2.25\\ -1.25\end{pmatrix} = \begin{pmatrix}| & | & |\\ s & a & b\\ | & | & |\end{pmatrix} \cdot \mathrm{col}_1(X^TX)^{-1} = \begin{pmatrix}| & | & |\\ s & a & b\\ | & | & |\end{pmatrix} \cdot \begin{pmatrix}12.75\\ -7.5\\ -3.5\end{pmatrix}$$

То есть, вектор весов $\mathrm{col}_1 W$ лежит в линейной оболочке регрессоров, $\mathrm{col}_j W \in V = \mathrm{Span}(\mathrm{col}\, X)$. Например, при проецировании каждого столбца матрицы W на линейную оболочку регрессоров ничего не происходит, $HW = W$.

В этом равенстве можно убедиться и средствами линейной алгебры,

$$HW = X(X^TX)^{-1}X^TX(X^TX)^{-1} = X(X^TX)^{-1} = W.$$

5.7. Показатели качества подгонки модели

После оценивания регрессионной модели полезно проанализировать, насколько она хороша. Для этого нужен показатель качества подгонки модели.

Назовём общей суммой квадратов (TSS) величину $\sum_{i=1}^n (y_i - \bar{y})^2$. Рассмотрим её разложение:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Слагаемое $2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$, так как остатки $\hat{u}_i = y_i - \hat{y}_i$ ортогональны регрессорам.

Таким образом, получаем

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

Введём три обозначения.

Общая сумма квадратов (Total Sum of Squares)

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

Объяснённая сумма квадратов (Explained Sum of Squares):

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

Сумма квадратов остатков (Residual sum of squares)

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

Будьте бдительны! Некоторые источники используют иные обозначения. По поводу TSS разногласий в литературе не возникает. Под RSS иногда понимают регрессионную сумму квадратов, regression sum of squares, то есть ESS в нашем курсе. Под ESS некоторые авторы подразумевают сумму квадратов остатков, error sum of squares, то есть наш RSS. Эта разница может быть критической при, например, использовании формул из интернета, которые будут говорить противоположные нашему курсу вещи.

Итоговую теорему можно записать так

Теорема 5.3. Если среди регрессоров присутствует константа, то

$$TSS = ESS + RSS.$$

Определение 5.4 (коэффициент детерминации)**.** Коэффициентом детерминации называется статистика

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

При наличии константы в модели коэффициент детерминации $R^2 \in [0, 1]$ и показывает, какая доля разброса зависимой переменной относительно её среднего объясняется регрессионной моделью.

5.8. Основные матрицы в линейной регрессии

Здесь мы приведём основные матрицы, используемые в линейной регрессии и их алгебраические свойства.

Вспомним определение матрицы-шляпницы

Определение 5.5 (матрица-шляпница)**.** Матрица $X(X^TX)^{-1}X^T$ часто обозначается буквой H и неформально называется матрицей-шляпницей (hat matrix). Она «надевает» на y шляпку, $Hy = \hat{y}$. Формально матрица H также называется матрицей-проектором. Она проецирует любой вектор на линейную оболочку всех регрессоров $\mathrm{Span}(\mathrm{col}_1 X, \mathrm{col}_2 X, \ldots, \mathrm{col}_k X)$.

Матрица-проектор (hat-matrix) H является

- симметричной, то есть $H^T = H$:

$$H^T = X(X^TX)^{-1}X^T = H$$

- идемпотентной, то есть $H^2 = H$:

$$H^2 = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

- $\text{rank}\, H = \text{trace}\, H = k$:

$$\text{trace}\, H = \text{trace}(X(X^TX)^{-1}X^T) = \text{trace}((X^TX)(X^TX)^{-1}) = \text{trace}\, I_k = k.$$

Здесь мы использовали свойство следа: $\text{trace}(ABC) = \text{trace}(CAB)$, A, B, C — матрицы.

Определим матрицу $M = I - H$. Матрица M, как и матрица H является матрицей-проектором. Она проецирует любой вектор на ортогональное дополнение к линейной оболочке всех регрессоров $\text{Span}(\text{col}_1 X, \text{col}_2 X, \ldots, \text{col}_k X)$. Несложно убедиться, что матрица M так же, как и матрица H симметричная и идемпотентная. При этом $\text{trace}\, M = \text{trace}(I_n - H) = \text{trace}\, I_n - \text{trace}\, H = n - k$. Проделайте эти упражнения самостоятельно.

Используя введённые матрицы, выразим вектор остатков в модели $y = X\beta + u$:

$$\hat{u} = y - \hat{y} = y - Hy = (I - H)y = My = M(X\beta + u) = Mu.$$

Из геометрического смысла матриц M следует, что $MX = 0$.

Пусть $s = \begin{pmatrix}1 & 1 & \cdots & 1\end{pmatrix}^T$ — вектор размерности $[n \times 1]$, состоящий из единиц. Определим матрицу $\pi = s^T(s^Ts)^{-1}s^T$. Матрица π — это матрица размерности $[n \times n]$ вида

$$\pi = \frac{1}{n}\begin{pmatrix}1 & 1 & \cdots & 1\\ 1 & 1 & \cdots & 1\\ \cdots & \cdots & \cdots & \cdots\\ 1 & 1 & \cdots & 1\end{pmatrix}.$$

В качестве домашнего упражнения покажите, что $\pi c = \bar{c}$, где c — произвольный вектор размерности $[n \times 1]$.

Тогда TSS, ESS и RSS могут быть записаны в матричном виде:

$$TSS = (y - \bar{y})^T(y - \bar{y}) = (y - \pi y)^T(y - \pi y) = (y(I - \pi))^T(y(I - \pi)) = (I - \pi)^T y^T y(I - \pi) = y^T(I - \pi)y$$

$$ESS = (\hat{y} - \bar{y})^T(\hat{y} - \bar{y}) = (Hy - \pi y)^T(Hy - \pi y) = (y(H - \pi))^T(y(H - \pi)) = (H - \pi)^T y^T y(H - \pi) = y^T(H - \pi)y$$

$$RSS = (y - \hat{y})^T(y - \hat{y}) = (y - Hy)^T(y - Hy) = (y(I - H))^T(y(I - H)) = (I - H)^T y^T y(I - H) = y^T(I - H)y$$

5.9. Теорема Фриша —Во!

Сначала рассмотрим прикольную задачку. Джеймс Бонд для конспирации строит только регрессии на одну переменную и может оценить ровно один коэффициент за одну регрессию! Как может Джеймс Бонд, сохраняя конспирацию, оценить оба коэффициента в парной регрессии $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$?

Смотрите! На первом шаге Джеймс Бонд строит регрессию y на константу $s = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$. Он получит прогнозы первого шага $\hat{y}_i = \bar{y}$. Остатками первой регрессии станут $y_i^* = y_i - \bar{y}$.

На втором шаге Джеймс Бонд аналогично строит регрессию x на константу s и, аналогично, получит остатки $x_i^* = x_i - \bar{x}$.

И, наконец, на третьем шаге Джеймс строит регрессию полученных остатков y^* на остатки x^*, $\hat{y}_i^* = \hat{\beta} x_i^*$.

По полученным ранее формулам получится оценка

$$\hat{\beta} = \frac{\sum y_i^* x_i^*}{\sum (x_i^*)^2} = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}.$$

Но именно это и есть оценка для $\hat{\beta}_2$ в парной регрессии $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$!

$$\hat{\beta}_2 = \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2}.$$

Это совпадение не случайно! Оказывается множественную регрессию с любым количеством регрессоров можно разбить на несколько шагов со вспомогательными регрессиями с меньшим числом регрессоров.

Вместо непосредственного включения переменной x в качестве регрессора в модель можно сначала «очистить» от переменной x зависимую переменную y и остальные регрессоров, а затем оценить регрессию для «очищенных» переменных.

Теорема 5.6 (Теорема Фриша–Во–Ловелла (англ. Frisch–Waugh–Lovell theorem, FWL theorem))**.** Рассмотрим два алгоритма.

Алгоритм A: оцениваем регрессию y на полный набор регрессоров X_1 и X_2 с помощью МНК:

$$\hat{y}_A = X_1 \hat{\beta}_1^A + X_2 \hat{\beta}_2^A.$$

Алгоритм B:

B1. Оцениваем регрессию y на часть регрессоров X_1 с помощью МНК:

$$\hat{y}_B = X_1 \hat{\beta}_1^B.$$

B2. Оцениваем регрессию каждого столбца из матрицы X_2 с помощью МНК:

$$\hat{X}_2^B = X_1 \hat{\beta}_1^B.$$

Уточним, что здесь $\hat{\beta}_1^B$ — это не вектор, а целая матрица, в которой содержатся оценки регрессии каждого столбца из матрицы X_2 на все регрессоры из матрицы X_1.

B3. Считаем «очищенные» переменные как остатки регрессий первых двух шагов,

$$y^* = y - \hat{y}_B, \quad X_2^* = X_2 - \hat{X}_2^B.$$

B4. Оцениваем регрессию для «очищенных переменных»

$$\hat{y}^* = X_2^* \hat{\beta}_2^B.$$

Алгоритмы A и B дают одинаковые оценки коэффициентов $\hat{\beta}_2^A = \hat{\beta}_2^B$ и финальные векторы остатков $\hat{u}_A = y - \hat{y}_A = y^* - \hat{y}^* = \hat{u}_B^*$.

Доказательство. Определим матрицу-шляпницу H_1, проецирующую на линейную оболочку столбцов блока X_1, и матрицу $M_1 = I - H_1$, проецирующую на ортогональное дополнение к линейной оболочке столбцов X_1,

$$H_1 = X_1(X_1^T X_1)^{-1} X_1^T, \quad M_1 = I - H_1.$$

По определению, $H_1 X_1 = X_1$ и $M_1 X_1 = 0$.

Возьмём результат выполнения алгоритма A

$$y = X_1\hat{\beta}_1^A + X_2\hat{\beta}_2^A + \hat{u}_A$$

и домножим его на матрицу M_1:

$$M_1 y = 0 \cdot \hat{\beta}_1^A + M_1 X_2 \hat{\beta}_2^A + M_1\hat{u}_A$$

Заметим, что остатки $\hat{u}_A$ алгоритма A ортогональны и регрессорам из блока X_1, и регрессорам из блока X_2. Сначала воспользуемся тем, что остатки $\hat{u}_A$ уже лежат в подпространстве, ортогональном регрессорам блока X_1. Дополнительное проецирование в это подпространство никак их не изменяет, $M_1\hat{u}_A = \hat{u}_A$. Следовательно,

$$M_1 y = M_1 X_2 \hat{\beta}_2^A + \hat{u}_A$$

Теперь воспользуемся тем, что остатки $\hat{u}_A$ уже лежат в подпространстве, ортогональном регрессорам блока X_2, поэтому $X_2^T\hat{u}_A = 0$.

Ортогональны ли остатки $\hat{u}_A$ и столбцы матрицы $M_1 X_2$? Проверим!

$$(M_1 X_2)^T\hat{u}_A = X_2^T M_1^T\hat{u}_A = X_2^T M_1\hat{u}_A = X_2^T\hat{u}_A = 0.$$

Остаётся лишь сказать, что умножение на матрицу M_1 очищает переменные, $M_1 y = y^*$ и $M_1 X_2 = X_2^*$,

$$y^* = X_2^*\hat{\beta}_2^A + \hat{u}_A,$$

И мы видим идеальное совпадение с разложением алгоритма B,

$$y^* = X_2^*\hat{\beta}_2^B + \hat{u}_B^*$$

В силу единственности разложения по ортогональному базису $\hat{\beta}_2^A = \hat{\beta}_2^B$ и $\hat{u}_A = \hat{u}_B^*$. □

Рассказать, что коэффициенты при стандартизации всех переменных называют частными корреляциями.

Коммент: Здесь первый раз говорим слова «строгая мультиколлинеарность».

Чёрный трэк: нелинейный мнк численно?

Задачи для доски:

МНК и R2 руками на доске

Задачи для колаба:

МНК и R2

Рост R2 с ростом числа регрессоров

Рост RSSc ростом числа наблюдений

5.10. Кросс-валидация с выкидывание отдельных наблюдений

TODO: как-то переобозначить $\hat{\beta}^-$, тк минус как индекс путается с обычным.

Определение 5.7 (LOOCV)**.** Кросс-валидация с поочередным выкидыванием отдельных наблюдений. На английском языке она часто сокращается LOOCV (leave one out cross validation).

Рассмотрим модель $y = X\beta + u$.

Оценим модель без первого наблюдения. Получим МНК-оценки $\hat{\beta}^{(-1)}$. С помощью этих оценок спрогнозируем первое наблюдение, получим прогноз $\hat{y}_1^{CV}$ и ошибку прогноза $\hat{u}_1^{CV}$.

Вернём первое наблюдение в выборку и удалим второе наблюдение. Получим МНК-оценки $\hat{\beta}^{(-2)}$. С помощью этих оценок спрогнозируем второе наблюдение, получим прогноз $\hat{y}_2^{CV}$ и ошибку прогноза $\hat{u}_2^{CV}$.

Поступим так с каждым наблюдением. На выходе получим вектор кросс-валидационных прогнозов $\hat{y}^{CV}$ и вектор кросс-валидационных ошибок прогнозов $\hat{u}^{CV}$.

Теорема 5.8 (связь обычных и кросс-валидационных остатков)**.** Если модель $y = X\beta + u$ оценивается с помощью МНК и проводится кросс-валидации с поочередным выкидыванием отдельных наблюдений, то:

$$\hat{u}_i = (1 - H_{ii}) \cdot \hat{u}_i^{CV},$$

где H — матрица-шляпница $H = X(X^TX)^{-1}X^T$, $\hat{u}$ - остатки регрессии, а $\hat{u}^{CV}$ — кросс-валидационные ошибки прогнозов.

Заметим, что сомножитель $(1 - H_{ii}) \in (0; 1)$. Другими словами, теорема численно формализует интуитивно ожидаемый результат: кросс-валидационные остатки по знаку совпадают с обычными остатками, а по абсолютной величине — больше, так как соответствующее наблюдение не используется при оценивании коэффициента.

Доказательство. Оценим модель без последнего наблюдения, $\hat{y}^- = X^-\hat{\beta}^-$.

Создадим вектор y^*, который будет отличаться от y только последним, n-м элементом: вместо настоящего y_n там будет стоять прогноз по модели без последнего наблюдения $\hat{y}_n^-$.

Раз уж мы добавили новую точку лежащую ровно на выборочной регрессии, то при оценки модели $\hat{y}^* = X\hat{\beta}^*$ мы получим в точности старые оценки $\hat{\beta}^* = \hat{\beta}^-$. Следовательно, и прогнозы эти две модели дают одинаковые, $\hat{y}_i^* = \hat{y}_i^-$.

А теперь посмотрим на последний элемент вектора $v = H(y^* - y)$.

С одной стороны, он равен последней строке матрицы H умножить на вектор $(y^* - y)$. В векторе $(y^* - y)$ только последний элемент ненулевой, поэтому $v_n = H_{nn}(\hat{y}_n^- - y_n)$.

С другой стороны, мы можем раскрыть скобки, и заметить, что $v = Hy^* - Hy$. И окажется, что $v_n = \hat{y}_n^* - \hat{y}_n = \hat{y}_n^- - \hat{y}_n$.

Отсюда

$$\hat{y}_n^- - \hat{y}_n = H_{nn}(\hat{y}_n^- - y_n)$$

Приводим подобные слагаемые и добавляем слева и справа y_n, получаем как раз то, что нужно:

$$y_n - \hat{y}_n = (1 - H_{ii})(y_n - \hat{y}_n^-)$$

□

5.11. Задачи

Задача 6. Для модели регрессии на константу $y_i = \beta_1 + u_i$ найдите

а) $\hat{\beta}_1$;

б) $\hat{y}_i$;

в) ESS;

г) R^2.

Решение. а) Задача минимизации для модели регрессии на константу:

$$Q(\hat{\beta}_1) = \sum_{i=1}^{n}(y_i\hat{\beta}_1) \to \min_{\hat{\beta}_1}.$$

Условие первого порядка:

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1) = 0.$$

Решаем уравнение:

$$\sum_{i=1}^{n} y_i - n\beta_1 = 0 \implies \hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

б) Поскольку модель содержит только константу:

$$\hat{y}_i = \hat{\beta}_1 = \bar{y} \quad \text{для всех } i = 1, \ldots, n.$$

в) Вычислим ESS:

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\bar{y} - \bar{y})^2 = 0.$$

г) Вычислим компоненты:

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$ESS = 0$$
$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = TSS$$

Коэффициент детерминации:

$$R^2 = \frac{ESS}{TSS} = 0.$$

Задача 7. Для модели парной линейной регрессии $y_i = \beta_1 + \beta_2 x_i + u_i$ получите оценки коэффициентов β_1, β_2 двумя путями:

а) преобразуя формулу МНК-оценки в матричном виде $\hat{\beta} = (X^T X)^{-1} X^T y$ для скалярного случая;

б) решая задачу оптимизации методом наименьших квадратов $\sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \to \min\limits_{\hat{\beta}_1, \hat{\beta}_2}$.

Решение. а) Запишем модель в матричном виде:

$$y = X\beta + u$$

где

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Вычислим необходимые матрицы:

$$X^T X = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}, \quad X^T y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}$$

Обратная матрица:

$$(X^T X)^{-1} = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

МНК-оценка:

$$\hat{\beta} = (X^T X)^{-1} X^T y = \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i \\ n \sum x_i y_i - \sum x_i \sum y_i \end{pmatrix}$$

Таким образом:

$$\hat{\beta}_1 = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\beta}_2 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

б) Минимизируем сумму квадратов остатков:

$$Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 \to \min_{\hat{\beta}_1, \hat{\beta}_2}$$

Условия первого порядка:

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0,$$

$$\frac{\partial Q}{\partial \hat{\beta}_2} = -2 \sum x_i (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0.$$

Получаем систему нормальных уравнений:

$$\begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum x_i = \sum y_i, \\ \hat{\beta}_1 \sum x_i + \hat{\beta}_2 \sum x_i^2 = \sum x_i y_i, \end{cases}$$

Решение системы:

$$\hat{\beta}_2 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2},$$

$$\hat{\beta}_1 = \frac{\sum y_i - \hat{\beta}_2 \sum x_i}{n} = \bar{y} - \hat{\beta}_2 \bar{x}.$$

Сравнение результатов

Оба метода дают одинаковые выражения для оценок коэффициентов:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$$

$$\hat{\beta}_2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Задача 8. Покажите, что для модели парной линейной регрессии $y_i = \beta_1 + \beta_2 x_i + u_i$ выполняются следующие свойства:

а) $\sum_{i=1}^n \hat{u}_i = 0$;

б) $\sum_{i=1}^n \hat{y}_i = \bar{y}$;

в) $\sum_{i=1}^n \hat{u}_i x_i = 0$;

г) $\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}_i$

Решение. а) Оценки МНК получаются минимизацией суммы квадратов остатков:

$$Q(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2.$$

Условие первого порядка для $\hat{\beta}_1$:

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^n (y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) = 0.$$

Отсюда:

$$\sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0.$$

б) Из пункта а):

$$\sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0 \implies \sum_{i=1}^n y_i = \sum_{i=1}^n \hat{y}_i.$$

Разделив на n, получим:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \bar{\hat{y}}.$$

в) Условие первого порядка для $\hat{\beta}_2$:

$$\frac{\partial Q}{\partial \hat{\beta}_2} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)x_i = 0.$$

Отсюда:

$$\sum_{i=1}^{n}\hat{u}_i x_i = \sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = 0.$$

г) Из пункта б):

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \hat{\beta}_1 + \hat{\beta}_2\bar{x}.$$

Задача 9. Для модели парной линейной регрессии $y_i = \beta_1 + \beta_2 x_i + u_i$

а) в скалярном виде выпишите TSS, RSS, ESS,

б) покажите, что $TSS = RSS + ESS$.

Решение. а) Для модели парной линейной регрессии $y_i = \beta_1 + \beta_2 x_i + u_i$ определим следующие суммы квадратов:

- **Общая сумма квадратов (Total Sum of Squares, TSS)**:

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

где $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ — выборочное среднее зависимой переменной.

- **Объяснённая сумма квадратов (Explained Sum of Squares, ESS)**:

$$\text{ESS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

где $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ — предсказанное значение y_i по оценённой модели.

- **Остаточная сумма квадратов (Residual Sum of Squares, RSS)**:

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

где $y_i - \hat{y}_i = \hat{u}_i$ — остаток модели.

б) Докажем равенство $TSS = ESS + RSS$.

Начнём с определения TSS и разложим отклонение $y_i - \bar{y}$ на две составляющие:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

Возведём обе части в квадрат и просуммируем по всем наблюдениям:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}\left[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})\right]^2.$$

Раскроем квадрат в правой части:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y}).$$

Заметим, что:

- $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=$ RSS,
- $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=$ ESS,
- перекрёстное произведение $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$ (доказательство ниже).

Докажем, что $\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=0$. Заметим, что:

- $\hat{y}_i=\hat{\beta}_1+\hat{\beta}_2x_i$, где $\hat{\beta}_1=\bar{y}-\hat{\beta}_2\bar{x}$ (из условий первого порядка МНК).
- $\sum_{i=1}^{n}\hat{u}_i=\sum_{i=1}^{n}(y_i-\hat{y}_i)=0$ (свойство остатков МНК).
- $\sum_{i=1}^{n}\hat{u}_ix_i=0$ (условие ортогональности в МНК).

Теперь раскроем перекрёстное произведение:

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=\sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_1+\hat{\beta}_2x_i-\bar{y})=$$

(подставим $\hat{\beta}_1=\bar{y}-\hat{\beta}_2\bar{x}$):

$$=\sum_{i=1}^{n}\hat{u}_i(\bar{y}-\hat{\beta}_2\bar{x}+\hat{\beta}_2x_i-\bar{y})=\sum_{i=1}^{n}\hat{u}_i\hat{\beta}_2(x_i-\bar{x})=\hat{\beta}_2\sum_{i=1}^{n}\hat{u}_i(x_i-\bar{x})=$$

$$=\hat{\beta}_2\left(\sum_{i=1}^{n}\hat{u}_ix_i-\bar{x}\sum_{i=1}^{n}\hat{u}_i\right)=\hat{\beta}_2(0-\bar{x}\cdot 0)=0.$$

Таким образом, перекрёстное произведение равно нулю, и равенство $TSS=ESS+RSS$ доказано.

Задача 10. Для модели парной линейной регрессии без константы $y_i=\beta_1x_i+u_i$ покажите, что в общем случае

а) $\sum_{i=1}^{n}\hat{u}_i\neq 0$

б) $\bar{y}\neq\hat{\beta}_2\bar{x}$

в) $TSS\neq RSS+ESS$

г) $R^2\notin[0,1]$

Решение. а) В модели **с константой** выполняется $\sum \hat{u}_i = 0$, но в модели **без константы**:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i.$$

Эта разность равна нулю только если $\sum y_i = \hat{\beta} \sum x_i$, что выполняется не всегда.

Пример: Пусть $x = [1, 2]$, $y = [3, 5]$. Тогда:

$$\hat{\beta} = \frac{1 \cdot 3 + 2 \cdot 5}{1^2 + 2^2} = \frac{13}{5} = 2.6$$

Остатки:

$$\hat{u}_1 = 3 - 2.6 \cdot 1 = 0.4$$

$$\hat{u}_2 = 5 - 2.6 \cdot 2 = -0.2$$

Сумма остатков: $0.4 - 0.2 = 0.2 \neq 0$.

б) В модели **без константы**:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

$$\hat{\beta}_1 \bar{x} = \hat{\beta}_1 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Из пункта 1 следует, что в общем случае $\sum y_i \neq \hat{\beta}_1 \sum x_i$, значит:

$$\bar{y} \neq \hat{\beta}_1 \bar{x}.$$

в) В модели **с константой** выполняется разложение дисперсии:

$$\text{TSS} = \text{ESS} + \text{RSS},$$

где:

- $\text{TSS} = \sum (y_i - \bar{y})^2$ (общая сумма квадратов),
- $\text{ESS} = \sum (\hat{y}_i - \bar{y})^2$ (объясненная сумма квадратов),
- $\text{RSS} = \sum \hat{u}_i^2$ (остаточная сумма квадратов).

В модели **без константы** это разложение **не выполняется**, так как:

$$\sum \hat{u}_i \neq 0 \implies \text{ковариация между } \hat{y}_i \text{ и } \hat{u}_i \text{ не равна нулю.}$$

Следовательно:

$$\text{TSS} \neq \text{ESS} + \text{RSS}.$$

г) Коэффициент детерминации R^2 вычисляется по формуле:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}.$$

Поскольку для модели без константы $\text{TSS} \neq \text{ESS} + \text{RSS}$, то R^2 может выходить за пределы $[0, 1]$:

- если RSS $>$ TSS, то $R^2 < 0$;
- если ESS < 0, то $R^2 > 1$.

Задача 11. Покажите, что для регрессии вида $y_i = \beta_1 + \beta_2 x_i + u_i$ R^2 обладает следующими свойствами:

а) $R^2 = \hat{\rho}_{xy}$, где $\hat{\rho}_{xy}$ – выборочный коэффициент корреляции.

б) R^2 для заданной регрессии и R^2 для регрессии $x_i = \alpha_1 + \alpha_2 y_i + v_i$ равны.

в) Если $\hat{\beta}_1 = 0$, то $R^2 = 0$.

Задача 12. Пусть мы оценили модель $y_i = \beta_1 + \beta_2 x_i + u_i$ с помощью МНК.

а) Если данные оказались центрированными, что вы можете сказать о $\hat{\beta}_1$?

б) Ко всем наблюдениям x_i прибавили 15. Как изменятся $\hat{\beta}_1, \hat{\beta}_2$?

в) Все наблюдения x_i увеличили в 5 раз, что произойдёт с $\hat{\beta}_1, \hat{\beta}_2$ и $\hat{y}_i$?

Решение. а) Если данные центрированы ($\bar{x} = 0$, $\bar{y} = 0$), то:

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 0 - \hat{\beta}_2 \cdot 0 = 0$$

б) Ко всем x_i прибавили 15. Обозначим новый регрессор как $x'_i = x_i + 15$. Тогда новое среднее равно $\bar{x}' = \bar{x} + 15$.

- Оценка наклона не изменится:

$$\hat{\beta}'_2 = \frac{\sum(x'_i - \bar{x}')(y_i - \bar{y})}{\sum(x'_i - \bar{x}')^2} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \hat{\beta}_2.$$

- Оценка константы изменится:

$$\hat{\beta}'_1 = \bar{y} - \hat{\beta}'_2 \bar{x}' = \bar{y} - \hat{\beta}_2(\bar{x} + 15) = \hat{\beta}_1 - 15\hat{\beta}_2.$$

в) Все x_i увеличили в 5 раз. Обозначим новый регрессор как $x''_i = 5x_i$. Тогда новое среднее равно $\bar{x}'' = 5\bar{x}$.

- Оценка наклона изменится:

$$\hat{\beta}''_2 = \frac{\sum(x''_i - \bar{x}'')(y_i - \bar{y})}{\sum(x''_i - \bar{x}'')^2} = \frac{5\sum(x_i - \bar{x})(y_i - \bar{y})}{25\sum(x_i - \bar{x})^2} = \frac{\hat{\beta}_2}{5}.$$

- Оценка константы не изменится:

$$\hat{\beta}''_1 = \bar{y} - \hat{\beta}''_2 \bar{x}'' = \bar{y} - \frac{\hat{\beta}_2}{5} \cdot 5\bar{x} = \bar{y} - \hat{\beta}_2 \bar{x} = \hat{\beta}_1.$$

- Прогнозные значения:

$$\hat{y}''_i = \hat{\beta}''_1 + \hat{\beta}''_2 x''_i = \hat{\beta}_1 + \frac{\hat{\beta}_2}{5} \cdot 5x_i = \hat{\beta}_1 + \hat{\beta}_2 x_i = \hat{y}_i.$$

Задача 13. Рассмотрим регрессию $\hat{y}_i = \hat{\beta}_1 z_i + \hat{\beta}_2 x_i$. Все исходные данные поместим в матрицу X и вектор y:

$$X = \begin{pmatrix} z_1 & x_1 \\ \vdots & \vdots \\ z_n & x_n \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

а) Выпишите явно матрицы X^T, $X^T y$, $X^T X$, $y^T X$, $y^T z$ и укажите их размер.

б) Выпишите условия первого порядка для оценок $\hat{\beta}_1$ и $\hat{\beta}_2$ по методу наименьших квадратов.

в) Запишите эти же условия в виде линейной системы

$$\begin{cases} \hat{\beta}_1 \cdot \ldots + \hat{\beta}_2 \cdot \ldots = \ldots \\ \hat{\beta}_1 \cdot \ldots + \hat{\beta}_2 \cdot \ldots = \ldots \end{cases}$$

г) Как упростится данная система для регрессии $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$?

д) Запишите систему условий первого порядка с помощью матрицы X и вектора y;

Задача 14. Рассмотрим модель $y_i = \beta_1 + \beta_2 x_i + u_i$, где

$$x = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 1 \\ 4 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

а) Укажите число наблюдений

б) Найдите $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

в) С помощью МНК найдите оценку для вектора неизвестных коэффициентов без использования матриц.

г) Перепишите модель в матричном виде и получите оценку коэффициентов через матричные формулы для МНК (убедитесь, что оценки совпали :)).

д) Найдите $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

е) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оценённого уравнения регрессии.

Решение. а) Количество строк в векторе y равно 3. Следовательно, $n = 3$.

б) Вычислим среднее значение $\bar{y}$:

$$\bar{y} = \frac{1 + 1 + 4}{3} = 2.$$

Тогда общая сумма квадратов:

$$\text{TSS} = (1 - 2)^2 + (1 - 2)^2 + (4 - 2)^2 = 1 + 1 + 4 = 6.$$

в) Формулы для МНК-оценок:

$$\hat{\beta}_2 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2},$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}.$$

Вычислим необходимые величины:

$$\bar{x} = \frac{1+2+3}{3} = 2,$$

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = (1-2)(1-2) + (2-2)(1-2) + (3-2)(4-2) = 1 + 0 + 2 = 3,$$

$$\sum(x_i - \bar{x})^2 = (1-2)^2 + (2-2)^2 + (3-2)^2 = 1 + 0 + 1 = 2.$$

Тогда оценки:

$$\hat{\beta}_2 = \frac{3}{2} = 1.5,$$

$$\hat{\beta}_1 = 2 - 1.5 \times 2 = -1.$$

г) Матрица регрессоров и вектор зависимой переменной:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad Y = \begin{pmatrix} 1 \\ 1 \\ 4 \end{pmatrix}.$$

МНК-оценка коэффициентов:

$$\hat{\beta} = (X^T X)^{-1} X^T Y.$$

Вычислим:

$$X^T X = \begin{pmatrix} 3 & 6 \\ 6 & 14 \end{pmatrix}, \quad X^T Y = \begin{pmatrix} 6 \\ 15 \end{pmatrix},$$

$$(X^T X)^{-1} = \frac{1}{6} \begin{pmatrix} 14 & -6 \\ -6 & 3 \end{pmatrix},$$

$$\hat{\beta} = \frac{1}{6} \begin{pmatrix} 14 \times 6 - 6 \times 15 \\ -6 \times 6 + 3 \times 15 \end{pmatrix} = \begin{pmatrix} -1 \\ 1.5 \end{pmatrix}.$$

Результаты совпали с предыдущим пунктом.

д) Прогнозные значения:

$$\hat{y}_1 = -1 + 1.5 \times 1 = 0.5,$$

$$\hat{y}_2 = -1 + 1.5 \times 2 = 2,$$

$$\hat{y}_3 = -1 + 1.5 \times 3 = 3.5.$$

Остаточная сумма квадратов:

$$\text{RSS} = (1 - 0.5)^2 + (1 - 2)^2 + (4 - 3.5)^2 = 0.25 + 1 + 0.25 = 1.5.$$

е) Формула для R^2:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{1.5}{6} = 0.75.$$

Интерпретация: модель объясняет 75% вариации зависимой переменной, что говорит о хорошем качестве подгонки модели.

Задача 15. Рассмотрим модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + u_i$, где

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}$$

Для удобства расчётов даны матрицы:

$$X^T X = \begin{pmatrix} 5 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}, (X^T X)^{-1} = \begin{pmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 4/3 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

а) Укажите число наблюдений

б) Укажите число регрессоров в модели, учитывая свободный член.

в) Найдите $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

г) С помощью МНК найдите оценку для вектора неизвестных коэффициентов.

д) Найдите вектор прогнозов $\hat{y}$.

е) Найдите $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

ж) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оценённого уравнения регрессии.

Решение. а) Количество строк в матрице X равно 5, следовательно, $n = 5$.

б) В модель включена свободный член и две объясняющие переменные (x и z), то есть $k = 3$.

в) Вычислим среднее значение $\bar{y}$:

$$\bar{y} = \frac{1 + 2 + 3 + 4 + 5}{5} = 3.$$

Тогда общая сумма квадратов:

$$\text{TSS} = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

г) Используем матричную формулу МНК:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

Умножим обратную матрицу на $X^T y$:

$$X^T y = \begin{pmatrix} 1+2+3+4+5 \\ 0+0+0+4+5 \\ 0+0+0+0+5 \end{pmatrix} = \begin{pmatrix} 15 \\ 9 \\ 5 \end{pmatrix},$$

$$(X^T X)^{-1} X^T y = \begin{pmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 4/3 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} 15 \\ 9 \\ 5 \end{pmatrix} = \begin{pmatrix} 5-3+0 \\ -5+12-5 \\ 0-9+10 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}.$$

д) Вычислим вектор прогнозов $\hat{y} = X\hat{\beta}$:

$$\hat{y} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \\ 2 \\ 4 \\ 5 \end{pmatrix}.$$

е) Остаточная сумма квадратов:

$$\text{RSS} = (1-2)^2 + (2-2)^2 + (3-2)^2 + (4-4)^2 + (5-5)^2 = 1+0+1+0+0 = 2.$$

ж) Формула для R^2:

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{2}{10} = 0.8.$$

Задача 16. Константин оценивает влияние продаж на заработную плату менеджера: $salary_i = \beta_1 + \beta_2 sales_i + u_i$. Он оценил коэффициенты с помощью МНК, нашёл R_1^2 и ему не понравился результат.

а) Тогда Константин выкинул одно из наблюдений, переоценил модель и получил R_2^2. Может ли он сравнить модели по R^2 и выбрать наилучшую?

б) Константин добавил в модель регрессор числа созвонов с начальством $sinks_i$. Покажите, что R^2 у данной регрессии вырастет в любом случае.

в) Константин слышал, что некоторые исследователи логарифмируют зарплаты перед включением в модель. Объясните, можно ли сравнивать по R^2 регрессии $salary_i$ и $log(salary_i)$ на один и тот же набор регрессоров?

Задача 17. Я очень хочу тут реальный датасет с точными датами рождения людей :)

Задача 18. Рассмотрим модель множественной регрессии

$$y = X\beta + u,$$

где X — матрица признаков размерности $[n \times k]$. Определим матрицу $M = I - X(X^T X)^{-1} X^T$. Покажите, что

а) матрица M симметричная;

б) матрица M идемпотентная;

в) $\operatorname{trace} M = n - k$.

Решение. а) Вычислим транспонированную матрицу M:

$$M^T = \left(I_n - X(X^TX)^{-1}X^T\right)^T = I_n^T - \left(X(X^TX)^{-1}X^T\right)^T = I_n - X\left((X^TX)^{-1}\right)^T X^T.$$

Поскольку X^TX симметрична, то $(X^TX)^{-1}$ также симметрична:

$$\left((X^TX)^{-1}\right)^T = (X^TX)^{-1}.$$

Следовательно,

$$M^T = I_n - X(X^TX)^{-1}X^T = M.$$

Таким образом, M симметрична.

б) Матрица M называется идемпотентной, если $M^2 = M$.

Вычислим M^2:

$$M^2 = \left(I_n - X(X^TX)^{-1}X^T\right)\left(I_n - X(X^TX)^{-1}X^T\right) \tag{1}$$
$$= I_n - 2X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T. \tag{2}$$

Упростим последнее слагаемое:

$$X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T.$$

Подставим обратно:

$$M^2 = I_n - 2X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^T = I_n - X(X^TX)^{-1}X^T = M.$$

Таким образом, M идемпотентна.

в) Используем свойства следа матрицы:

$$\operatorname{trace}(M) = \operatorname{trace}\left(I_n - X(X^TX)^{-1}X^T\right) \tag{3}$$
$$= \operatorname{trace}(I_n) - \operatorname{trace}\left(X(X^TX)^{-1}X^T\right). \tag{4}$$

След единичной матрицы равен её размеру, $\operatorname{trace}(I_n) = n$. Для второго слагаемого воспользуемся свойством следа $\operatorname{trace}(AB) = \operatorname{trace}(BA)$:

$$\operatorname{trace}\left(X(X^TX)^{-1}X^T\right) = \operatorname{trace}\left((X^TX)^{-1}X^TX\right) = \operatorname{trace}(I_k) = k.$$

Таким образом,

$$\operatorname{trace}(M) = n - k.$$

Задача 19. Пусть $s = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T$ — вектор размерности $[n \times 1]$, состоящий из единиц. Определим матрицу $\pi = s^T(s^T s)^{-1}s^T$. Матрица π — это матрица размерности $[n \times n]$ вида

$$\pi = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}.$$

Покажите, что $\pi c = \bar{c}$, где c — произвольный вектор размерности $[n \times 1]$.

Решение. Вычислим произведение πc:

$$\pi c = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$$

Умножение матрицы на вектор дает:

$$\pi c = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n c_i \\ \sum_{i=1}^n c_i \\ \vdots \\ \sum_{i=1}^n c_i \end{pmatrix} = \frac{1}{n}\sum_{i=1}^n c_i \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} \bar{c} \\ \bar{c} \\ \vdots \\ \bar{c} \end{pmatrix}.$$

По определению среднего значения:

$$\pi c = \bar{c}.$$

Таким образом, мы показали, что умножение матрицы π на произвольный вектор c дает вектор, все элементы которого равны среднему значению элементов вектора c.

6. Предпосылки о математическом ожидании и дисперсии

В этой главе мы познакомимся с понятиями независимости и линейной независимости; расчётом математических ожиданий, ковариаций и дисперсий в матричном виде.

Добавим в метод наименьших квадратов ряд статистических предпосылок на ожидание и дисперсию.

Сформулируем и докажем теорему Гаусса - Маркова (которая пообещает, что МНК-оценки будут обладать свойствами несмещённости и эффективности).

6.1. Иерархия зависимостей случайных величин

Напомним определение наилучшей линейной аппроксимации

Определение 6.1 (наилучшая линейная аппроксимация)**.** Наилучшее линейное приближение величины r с помощью величины s — это линейная функция от s,

$$\text{BestLin}(r \mid s) = \beta_1 + \beta_2 s,$$

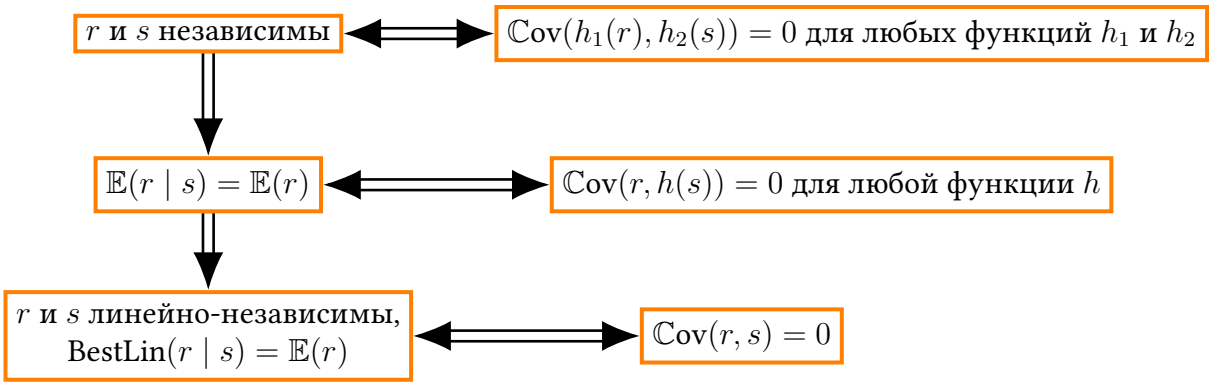
где константы β_1 и β_2 находятся из решения задачи оптимизации $\mathbb{E}((r - \text{BestLin}(r,s)^2) \to \min_{\beta_1,\beta_2}$. При решении задачи оказывается, что

$$\beta_1 = \mathbb{E}(r) - \frac{\mathbb{C}\text{ov}(r,s)}{\mathbb{V}\text{ar}(s)}\,\mathbb{E}(s), \quad \beta_2 = \frac{\mathbb{C}\text{ov}(r,s)}{\mathbb{V}\text{ar}(s)}.$$

Определение 6.2 (линейная независимость)**.** Величины r и s называются линейно-независимыми, если $\text{BestLin}(r \mid s) = \mathbb{E}(r)$

Некоторые авторы считают условие $\mathbb{C}\text{ov}(r, s) = 0$ определением линейной независимости.

Можно выделить три степени независимости случайных величин. Рассмотрим их на примере пары произвольных величин r и s.



Напомним, что определение независимых случайных величин,

Определение 6.3 (независимость случайных величин)**.** Случайные величины r и s называются независимыми если для любых[4] числовых множеств A и B независимы события $\{r \in A\}$ и $\{s \in B\}$:

$$\mathbb{P}(r \in A, s \in B) = \mathbb{P}(r \in A) \cdot \mathbb{P}(s \in B)$$

Из независимости величин r и s следует, что информация, известная об s, никак не помогает угадывать значение r. Поэтому условное математическое ожидание для r равно безусловному. Точно также из независимости r и s следует $\mathbb{E}(s \mid r) = \mathbb{E}(s)$. Обратное утверждение неверно, что показывает контрпример ниже.

Задача 20. Покажем, что из равенства условного и безусловного математических ожиданий не следует независимость случайных величин. Пусть дискретные случайные величины r характеризуют погоду (-1 снег, 1 солнце, 0 дождь), s – наличие зонта (0 нет или 1 есть) и ниже приведена таблица их совместного распределения.

	1/3	1/3	1/3
r	-1	1	0
s	0	0	1

Решение. Уже по формулировке подозреваем, что величины зависимые :).

Найдём условное ожидание зонта при условии, что мы видим погоду на улице: $\mathbb{E}(s \mid r) = \begin{cases} 0 \text{ если } r \in \{-1, \\ 1, \text{ если } r = 0. \end{cases}$

Получается, что информация о погоде помогает предсказать наличие зонтика, события не являются независимыми.

Найдём ожидания о погоде за окном, если вы можете наблюдать наличие или отсутствие зонта у человека: $\mathbb{E}(r \mid s) = \begin{cases} (-1) \times 1/6 + 1 \times 1/6 = 0, \text{ если } s = 0, \\ 0, \text{ если } s = 1. \end{cases}$.

Обычное безусловное ожидание погоды на улице: $\mathbb{E}(r) = (-1) \times 1/3 + 1 \times 1/3 + 0 \times 1/3 = 0$

Получается, что $\mathbb{E}(r \mid s) = \mathbb{E}(r) = 0$, но события зависимы.

[4] Не совсем любых, требуется измеримость множеств. В рамках нашего курса мы не будем обращать внимания на данный нюанс.

Вернемся к тому факту, что из равенства условного и безусловного математических ожиданий следует нулевая ковариация. Используя закон повторных математических ожиданий $\mathbb{C}\mathrm{ov}(r,s) = \mathbb{E}(rs) - \mathbb{E}(r)\,\mathbb{E}(s) = \mathbb{E}(\mathbb{E}(rs \mid s)) - \mathbb{E}(r)\,\mathbb{E}(s) = \mathbb{E}(s\,\mathbb{E}(r \mid s)) - \mathbb{E}(r)\,\mathbb{E}(s) = \mathbb{E}(r)\,\mathbb{E}(s) - \mathbb{E}(r)\,\mathbb{E}(s) = 0$.

Задача 21. Из нулевой ковариации не следует равенство условного и безусловного математических ожиданий (и тем более не следует независимость). Пусть случайная величина s имеет равномерное распределение на отрезке $[-1;1]$, а $r = s^2$.

Решение. Напоминаем, что для равномерно распределённой случайной величины $\mathbb{E}(s) = \frac{-1+1}{2} = 0, pdf(s) = \frac{1}{1-(-1)} = \frac{1}{2}$.

$\mathbb{E}(r \mid s) = \mathbb{E}(s^2 \mid s) = s^2 \neq 0$ в общем случае.

При этом $\mathbb{C}\mathrm{ov}(r,s) = \mathbb{E}(rs) - \mathbb{E}(r)\,\mathbb{E}(s) = \mathbb{E}(s^3) - \mathbb{E}(s^2) \times 0 = \mathbb{E}(s^3)$.

Математическое ожидание сложной функции $\mathbb{E}(g(x)) = \int_b^r g(x) pdf(x) dx$, если $x \in [a,b]$.

Найдём $\mathbb{E}(s^3) = \int_{-1}^{1} s^3 pdf(s) ds = \int_{-1}^{1} s^3 \frac{1}{2} ds = \frac{1}{8} s^4 |_{-1}^{1} = 0$. Значит, мы получили нулевую ковариацию у зависимых случайных величин.

Вывод

Существуют независимые случайные величины, но на ???

6.2. Ожидание и ковариационная матрица

Любопытный читатель снова может заглянуть в «Поваренную книга любителя матриц» Петерсона и Педерсона, [PP12].

Пусть r — случайный вектор размерности $[n \times 1]$, s — случайный вектор размерности $[k \times 1]$, A и b — неслучайные матрица и вектор соответственно, имеющие подходящие размерности.

Математическим ожиданием случайного вектора r называется вектор

$$\mathbb{E}(r) = \begin{pmatrix} \mathbb{E}(r_1) \\ \mathbb{E}(r_2) \\ \dots \\ \mathbb{E}(r_n) \end{pmatrix}.$$

Ковариационная матрица вектора r определяется следующим образом:

$$\mathbb{V}\mathrm{ar}(r) = \begin{pmatrix} \mathbb{C}\mathrm{ov}(r_1, r_1) & \mathbb{C}\mathrm{ov}(r_1, r_2) & \dots & \mathbb{C}\mathrm{ov}(r_1, r_n) \\ \mathbb{C}\mathrm{ov}(r_2, r_1) & \mathbb{C}\mathrm{ov}(r_2, r_2) & \dots & \mathbb{C}\mathrm{ov}(r_2, r_n) \\ \dots & & \dots & \dots \\ \mathbb{C}\mathrm{ov}(r_n, r_1) & \mathbb{C}\mathrm{ov}(r_n, r_2) & \dots & \mathbb{C}\mathrm{ov}(r_n, r_n) \end{pmatrix}.$$

Ковариационная матрица векторов r и s определяется следующим образом:

$$\mathbb{C}\mathrm{ov}(r,s) = \begin{pmatrix} \mathbb{C}\mathrm{ov}(r_1, s_1) & \mathbb{C}\mathrm{ov}(r_1, s_2) & \dots & \mathbb{C}\mathrm{ov}(r_1, s_k) \\ \mathbb{C}\mathrm{ov}(r_2, s_1) & \mathbb{C}\mathrm{ov}(r_2, s_2) & \dots & \mathbb{C}\mathrm{ov}(r_2, s_k) \\ \dots & \dots & \dots & \dots \\ \mathbb{C}\mathrm{ov}(r_n, s_1) & \mathbb{C}\mathrm{ov}(r_n, s_2) & \dots & \mathbb{C}\mathrm{ov}(r_n, s_k) \end{pmatrix}.$$

Свойства вектора математических ожиданий и ковариационной матрицы:

а) $\mathbb{E}(Ar + b) = A\,\mathbb{E}(r) + b$

б) $\mathbb{C}\text{ov}(r, s) = \mathbb{E}(rs^T) - \mathbb{E}(r)\,\mathbb{E}(s^T)$

в) $\mathbb{C}\text{ov}(Ar + b, s) = A\,\mathbb{C}\text{ov}(r, s)$

г) $\mathbb{C}\text{ov}(r, As + b) = \mathbb{C}\text{ov}(r, s)A^T$

д) $\mathbb{V}\text{ar}(r) = \mathbb{C}\text{ov}(r, r) = \mathbb{E}(rr^T) - \mathbb{E}(r)\,\mathbb{E}(r^T)$

е) $\mathbb{V}\text{ar}(Ar + b) = A\,\mathbb{V}\text{ar}(r)A^T$

ж) $\mathbb{E}(r^T Ar) = \text{trace}(A\,\mathbb{V}\text{ar}(r)) + \mathbb{E}(r^T)A\,\mathbb{E}(r)$

з) Если вектора r и s имеют одинаковый размер, то $\mathbb{V}\text{ar}(r+s) = \mathbb{V}\text{ar}(r)+\mathbb{V}\text{ar}(s)+\mathbb{C}\text{ov}(r, s)+\mathbb{C}\text{ov}(s, r)$

Условные ожидание и дисперсия определяются аналогично и обладают аналогичными свойствами. Главное — не забывать ставить вертикальную палочку!

Например,

написать пример

6.3. Теорема Гаусса — Маркова

$$y = X\beta + u$$

Чтобы исследовать свойства полученной точечной оценки $\hat{\beta}$ нам потребуются предпосылки о математическом ожидании и ковариационной матрице вектора u.

Мы предположим, что случайные ошибки в среднем равны нулю, а именно,

$$\mathbb{E}(u \mid X) = 0.$$

Предпосылку о математическом ожидании можно записать и в скалярном виде,

$$\mathbb{E}(u_i \mid X) = 0, \qquad \text{при } \forall i \in \{1, \ldots, n\}.$$

Важно пояснить смысл введённой предпосылки. При оценивании связи между регрессорами X и переменной y мы не предполагаем, что величины u и X независимы. В ошибки модели попадают все те факторы, которые мы забыли включить в регрессию. Эти факторы могут быть взаимосвязаны с тем, что в регрессию всё же попало. Мы делаем более слабое предположение лишь о бесполезности всей собранной в X информации для угадывания u (и следующей из неё линейной независимости между ошибками и регрессорами, в том числе о нулевой ковариации).

Теорема 6.4 (Гаусс — Марков)**.** Если

1. Модель линейна по параметрам: $y = X\beta + u$;
2. Матрица X размера $[n \times k]$ имеет полный ранг k.
3. Условное ожидание ошибок равно нулю, $\mathbb{E}(u \mid X) = 0$;
4. Условная ковариационная матрица ошибок пропорциональна единичной, $\mathbb{V}\text{ar}(u \mid X) = \sigma^2 I$;
5. Оценка $\hat{\beta}$ получена методом наименьших квадратов, $\hat{\beta} = (X^T X)^{-1} X^T y$;

то

(a) Оценка $\hat{\beta}$ является линейной по y;

(b) Оценка $\hat{\beta}$ является условно несмещённой, $\mathbb{E}(\hat{\beta} \mid X) = \beta$ и несмещённой, $\mathbb{E}(\hat{\beta}) = \beta$;

(c) Оценка любого коэффициента $\hat{\beta}_j$ является наиболее эффективной в классе линейных несмещённых оценок.

Что означает «эффективная в классе линейных несмещённых оценок»? Это означает, что у любой другой линейной по y несмещённой оценки $\hat{\beta}_j^{\text{alt}}$ дисперсия не меньше, чем у МНК-оценки.

$$\mathbb{V}\text{ar}(\hat{\beta}_j \mid X) \leq \mathbb{V}\text{ar}(\hat{\beta}_j^{\text{alt}} \mid X).$$

В иностранной литературе для простоты запоминания используется аббревиатура BLUE, best linear unbiased estimator. То есть при выполнении условий теоремы Гаусса-Маркова мы получаем несмещённые и наилучшие (в терминах эффективности) оценки в классе всех линейных оценок.

Хорошие оценки подобны хорошему подвенечному платью,

> Something Olde, Something New, Something Borrowed, Something Blue, A Sixpence in your Shoe.

Вывод теоремы можно усилить, для любой линейной комбинации коэффициентов $w^T\beta$ МНК-оценка $w^T\hat{\beta}$ эффективнее альтернативной оценки $w^T\hat{\beta}^{\text{alt}}$:

$$\mathbb{V}\text{ar}(w^T\hat{\beta}_j \mid X) \leq \mathbb{V}\text{ar}(w^T\hat{\beta}_j^{\text{alt}} \mid X).$$

Доказательство. Линейность оценки по y видна прямо из её формулы, $\hat{\beta} = (X^TX)^{-1}X^Ty$.

Проверим условную несмещённость,

$$\mathbb{E}(\hat{\beta} \mid X) = \mathbb{E}((X^TX)^{-1}X^Ty \mid X) = (X^TX)^{-1}X^T\,\mathbb{E}(y \mid X).$$

Для удобства посчитаем $\mathbb{E}(y \mid X)$ отдельно,

$$\mathbb{E}(y \mid X) = \mathbb{E}(X\beta + u \mid X) = X\beta + \mathbb{E}(u \mid X) = X\beta.$$

И теперь завершаем вычисление $\mathbb{E}(\hat{\beta} \mid X)$:

$$\mathbb{E}(\hat{\beta} \mid X) = (X^TX)^{-1}X^T\,\mathbb{E}(y \mid X) = (X^TX)^{-1}X^TX\beta = \beta.$$

Мы доказали условную несмещённость оценки, $\mathbb{E}(\hat{\beta} \mid X) = \beta$. Безусловная несмещённость следует из свойства условного ожидания,

$$\mathbb{E}(\hat{\beta}) = \mathbb{E}(\mathbb{E}(\hat{\beta} \mid X)) = \mathbb{E}(\beta) = \beta.$$

Эффективность МНК-оценок — это реинкарнация теоремы Пифагора. Мы увидим, что дисперсия МНК-оценки — это квадрат длины катета, дисперсия альтернативной несмещённой оценки — квадрат длины гипотенузы.

Для примера рассмотрим оценку первого коэффициента бета, $\hat{\beta}_1$. Доказательство не меняется ни капли, если рассмотреть оценку другого коэффициента, скажем, $\hat{\beta}_7$ или даже оценку произвольной линейной комбинацию коэффициентов бета, например, $\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$.

Итак, у нас есть две оценки, $\hat{\beta}_1$ и $\hat{\beta}_1^{\text{alt}}$. Обе они линейны по y, следовательно, $\hat{\beta}_1 = a^Ty$ и $\hat{\beta}_1^{\text{alt}} = a_{\text{alt}}^Ty$.

Замечаем, что $\mathbb{V}\text{ar}(\hat{\beta}_1 \mid X) = \sigma^2 a^T a$, и $\mathbb{V}\text{ar}(\hat{\beta}_1^{\text{alt}}) = \sigma^2 r_{\text{alt}}^T r_{\text{alt}}$. То есть дисперсии пропорциональны квадратам длин векторов a и a^{alt}. Осталось доказать, что вектор a не длиннее вектора a^{alt} :)

Для этого мы докажем, что вектор a^{alt} — это гипотенуза, а вектор a — катет. Нам нужно доказать, что вектор $a - a^{\text{alt}}$ перпендикулярен вектору a.

Разобъём доказательство перпендикулярности a и $a - a^{\text{alt}}$ на два шага:

Шаг 1. Вектор $a - a^{\text{alt}}$ перпендикулярен любому столбцу матрицы X.

Шаг 2. Вектор a является линейной комбинацией столбцов матрицы X.

здесь простая картинка с теоремой Пифагора!

Приступаем к шагу 1. Обе оценки несмещённые, поэтому для любых β должно выполняться:

$$\mathbb{E}(\hat{\beta}_1 \mid X) = \mathbb{E}(\hat{\beta}_1^{\text{alt}} \mid X)$$

Переносим всё в левую сторону:

$$\mathbb{E}((a^T - a_{\text{alt}}^T)(X\beta + u) \mid X) = 0$$

Получаем, что для любых β должно быть выполнено условие:

$$(a - r_{\text{alt}})^T X\beta = 0$$

Это возможно только, если вектор $(a - r_{\text{alt}})^T X$ равен нулю. Следовательно, вектор $(a - r_{\text{alt}})$ перпендикулярен любому столбцу X.

Приступаем к шагу 2.

Вспоминаем, что $\hat{\beta} = (X^T X)^{-1} X^T y$. Следовательно, нужная строка весов a^T — это первая строка в матрице $(X^T X)^{-1} X^T$. Замечаем, что выражение имеет вид $A \cdot X^T$.

Вспоминаем из линейной алгебры, что при умножении матриц AB получается матрица C, на которую можно взглянуть несколькими способами! Можно считать, что C — это разные линейные комбинации столбцов левой матрицы A. Можно считать, что C — это разные линейные комбинаций строк правой матрицы B.

Применим второй взгляд :) Получаем, что строка a^T — линейная комбинация строк матрицы X^T. Или, другими словами, столбец a — линейная комбинация столбцов матрицы X. □

Классическое доказательство эффективности, которое можно найти во многих учебниках, не замечает связи с теоремой Пифагора и исследует разницу ковариационных матриц. Приведём его здесь для демонстрации альтернативной техники!

Доказательство. У нас есть две линейных по y оценки: МНК-оценка и оценка-конкурент,

$$\hat{\beta} = (X^T X)^{-1} X^T y \text{ и } \hat{\beta}_{\text{alt}} = A_{\text{alt}}^T y.$$

Оценки ковариационных матриц этих оценок равны

$$\mathbb{V}\text{ar}(\hat{\beta} \mid X) = (X^T X)^{-1}\sigma^2 \text{ и } \mathbb{V}\text{ar}(\hat{\beta}_{\text{alt}} \mid X) = A^T A\sigma^2.$$

□

Условие несмещённости альтернативной оценки имеет вид

$$\mathbb{E}(\hat{\beta}_{\text{alt}} \mid X) = \mathbb{E}(A^T y \mid X) = A^T X\beta = \beta.$$

То есть для несмещённости альтернативной оценки должно выполняться условие $A^TX = I$. Для простоты рассмотрим случай $\sigma^2 = 1$. Мы докажем, что разница этих матриц $D = A^TA - (X^TX)^{-1}$ является положительно полуопределённой матрицы.

Вспомним из линейной алгебры определение и свойства положительно полуопределённой матрицы.

Определение 6.5 (положительно полуопределённая форма)**.** Матрица D или квадратичная форма $q(v) = v^TDv$ называется положительно полуопределённой, если $q(v) \geq 0$ для любого вектора v.

Теорема 6.6 (свойства положительно полуопределённой матрицы)**.** Матрица D является положительно полуопределённой, если и только если её можно записать в виде произведения $D = B^TB$.

У положительно полуопределённой матрицы D на диагонали находятся неотрицательные числа.

Если $D = A^TA - (X^TX)^{-1}$ — положительно полуопределена, то $d_{ii} \geq 0$ и, следовательно, $[A^TA]_{ii} \geq [(X^TX)^{-1}]_{ii}$, то есть, дисперсии альтернативных оценок не меньше дисперсий МНК-оценок.

Перейдём к доказательству положительной полуопределённости D:

Доказательство. Возьмём $B = A - X(X^TX)^{-1}$ и найдём B^TB:

$$B^TB = (A-X(X^TX)^{-1})^T(A-X(X^TX)^{-1}) = A^TA-A^TX(X^TX)^{-1}-(X^TX)^{-1}X^TA+(X^TX)^{-1}X^TX(X^TX)^{-1}$$

В силу несмещённости $A^TX = I$ или $X^TA = I$, поэтому

$$B^TB = A^TA - (X^TX)^{-1} - (X^TX)^{-1} + (X^TX)^{-1} = A^TA - (X^TX)^{-1}.$$

Мы видим, что матрица $D = A^TA - (X^TX)^{-1}$ оказалась разложенной в произведение $D = B^TB$ и, следовательно, матрица D положительно полуопределена. □

6.4. Статистические свойства остатков

Используя матричное представление для остатков $\hat{u} = My = Mu$, вычислим вектор математических ожиданий и ковариационной матрицы:

$$\mathbb{E}(\hat{u} \mid X) = \mathbb{E}(My \mid X) = M\,\mathbb{E}(y \mid X) = MX\beta = 0, \text{ так как } MX = 0.$$

Ожидаемое значение остатков равно нулю, также как и ожидаемое значение ошибок, $\mathbb{E}(\hat{u} \mid X) = \mathbb{E}(u \mid X) = 0$.

$$\mathbb{V}\text{ar}(\hat{u} \mid X) = \mathbb{V}\text{ar}(My \mid X) = M\,\mathbb{V}\text{ar}(y \mid X)M^T = M\sigma^2I_nM^T = \sigma^2MM^T = \sigma^2M.$$

Вспомним, что у ковариационной матрицы ошибок $\mathbb{V}\text{ar}(u \mid X) = \sigma^2I$ на диагонали стоят одинаковые элементы, а вне диагонали стоят нули. А у ковариационной матрицы остатков $\mathbb{V}\text{ar}(\hat{u} \mid X) = \sigma^2M$ на диагоналях находятся разные элементы и вне диагонали элементы в общем случае не равны нулю.

Другими словами, остатки $\hat{u}_i$ зависимы между собой и имеют разную дисперсию $\mathbb{V}\text{ar}(\hat{u}_i)$. Например, при наличии константы в регрессии остатки обязательно удовлетворяют соотношению $\sum \hat{u}_i = 0$.

Посчитаем ковариационную матрицу вектора остатков и вектора прогнозов:

$$\mathbb{C}\text{ov}(\hat{u}, \hat{y} \mid X) = \mathbb{C}\text{ov}(Mu, Py \mid X) = \mathbb{C}\text{ov}(Mu, P(X\beta + u) \mid X) = \mathbb{C}\text{ov}(Mu, X\beta + Pu \mid X) =$$

$$= \mathbb{C}\text{ov}(Mu, Pu \mid X) = M\,\mathbb{C}\text{ov}(u, u \mid X)P = M\sigma^2I_nP^T = \sigma^2MP = 0, \text{так как } P^T = P \text{ и } MP = 0.$$

Следовательно, вектор остатков и вектор прогнозов линейно независимы. Метод наименьших квадратов даёт наилучший линейный прогноз, то есть даже зная прогнозные значения $\hat{y}$ нет возможности уменьшить остатки модели.

Посчитаем ковариационную матрицу вектора остатков и МНК-оценки вектора параметров β:

$$\mathbb{C}\text{ov}(\hat{u}, \hat{\beta} \mid X) = \mathbb{C}\text{ov}(Mu, \beta + (X^TX)^{-1}X^Tu \mid X) = \mathbb{C}\text{ov}(Mu, (X^TX)^{-1}X^Tu \mid X) =$$

$$= M\,\mathbb{C}\text{ov}(u, u \mid X)X(X^TX)^{-1} = M\,\mathbb{C}\text{ov}(u, u \mid X)X(X^TX)^{-1} = \sigma^2 MX(X^TX)^{-1}, \text{ так как } MX = 0.$$

Следовательно, вектор остатков и вектор МНК-оценок параметров модели.

6.5. Оценивание дисперсии

Метод наименьших квадратов позволяет оценить вектор параметров β, однако ~~власти скрывают настоящую дисперсию~~ никак не оценивает неизвестный параметр σ^2. Интуиция говорит, что высокая дисперсия ошибок u_i должна проявляться в высоком разбросе $\hat{u}_i$, поэтому разумно попробовать построить оценку $\hat{\sigma}^2$ на базе $RSS = \sum \hat{u}_i^2$.

Для построения оценки $\hat{\sigma}^2$ найдём ожидание $\mathbb{E}(RSS \mid X)$:

Теорема 6.7 (ожидание суммы квадратов остатков)**.** Если выполнены предпосылки теоремы Гаусса — Маркова,

1. Модель линейна по параметрам: $y = X\beta + u$;
2. Матрица X размера $[n \times k]$ имеет полный ранг k.
3. Условное ожидание ошибок равно нулю, $\mathbb{E}(u \mid X) = 0$;
4. Условная ковариационная матрица ошибок пропорциональна единичной, $\mathbb{V}\text{ar}(u \mid X) = \sigma^2 I$;
5. Оценка $\hat{\beta}$ получена методом наименьших квадратов, $\hat{\beta} = (X^TX)^{-1}X^Ty$;

то $\mathbb{E}(RSS \mid X) = \mathbb{E}(\sum \hat{u}_i^2 \mid X) = (n-k)\sigma^2$.

Из этой теоремы следует, что оценка $\hat{\sigma}^2 = RSS/(n-k)$ — несмещённая оценка для неизвестной дисперсии σ^2.

Доказательство. На помощь нам придёт след матрицы! След матрицы прекрасен двумя свойствами. Во-первых, его можно менять местами с математическим ожиданием, $\mathbb{E}(\text{trace}\, W)) = \text{trace}\, \mathbb{E}(W)$. Во-вторых, внутри следа можно переставлять местами перемножаемые матрицы, $\text{trace}(AB) = \text{trace}(BA)$. Кроме того, на скалярную величину след можно навесить совершенно бесплатно! Если величина R — не вектор, а скаляр, то $\text{trace}\, R = R$.

Продолжаем,

$$\mathbb{E}(\hat{u}^T\hat{u} \mid X) = \mathbb{E}(\text{trace}(\hat{u}^T\hat{u}) \mid X) = \mathbb{E}(\text{trace}(\hat{u}\hat{u}^T) \mid X) = \text{trace}\, \mathbb{E}(\hat{u}\hat{u}^T \mid X).$$

Подумаем о середине,

$$\mathbb{E}(\hat{u}\hat{u}^T \mid X) = \mathbb{E}(Mu(Mu)^T \mid X) = \mathbb{E}(Muu^TM^T \mid X) = M\,\mathbb{E}(uu^T \mid X)M^T.$$

Вспомним, что матрица M — проектор, поэтому $M^T = M$, $M^2 = M$. У матрицы uu^T на диагонали стоят u_i^2, вне диагонали — u_iu_j. Поэтому $\mathbb{E}(uu^T \mid X) = \sigma^2 I$. Завершаем вычисления,

$$\mathbb{E}(\hat{u}\hat{u}^T \mid X) = M\,\mathbb{E}(uu^T \mid X)M^T = M \cdot \sigma^2 I \cdot M^T = \sigma^2 M^2 = \sigma^2 M$$

След проектора равен размерности пространства, на которое он проецирует, поэтому trace $M = n - k$ и

$$\mathbb{E}(RSS \mid X) = \text{trace}(\sigma^2 M) = (n-k)\sigma^2$$

И мы легко строим несмещённую оценку, $\hat{\sigma}^2 = RSS/(n-k)$,

$$\mathbb{E}(\hat{\sigma}^2 \bar{X}) = \mathbb{E}\left(\frac{RSS}{n-k} \mid X\right) = \frac{(n-k)\sigma^2}{n-k} = \sigma^2$$

□

Выборочная дисперсия при случайной выборке

Заметим, что данная теорема обобщает старый факт про выборочную дисперсию! Вспомним, что для выборки из независимых y_i с ожиданием $\mathbb{E}(y_i) = \mu$ и дисперсией $\mathbb{V}\text{ar}(y_i) = \sigma^2$ несмещённая оценка дисперсии имеет вид

$$\hat{\sigma}^2 = \frac{\sum(y_i - \bar{y})^2}{n-1}.$$

В данном случае величины y_i можно представить в виде $y_i = \mu + u_i$. Тогда предпосылки теоремы Гаусса — Маркова выполнены, матрица регрессоров X — это просто единственный столбец-регрессор из единиц, $k = 1$, $\beta = \mu$. В этом случае $\hat{\beta} = \bar{y}$, все прогнозы равны $\hat{u}_i = \bar{y}$ и $RSS = \sum(y_i - \hat{y}_i)^2 = \sum(y_i - \bar{y})^2$. И мы видим, что новая оценка совпадает в этом случае со старой:

$$\hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\sum(y_i - \hat{y}_i)^2}{n-1} = \frac{\sum(y_i - \bar{y})^2}{n-1}$$

Оценка дисперсии оценок коэффициентов

Для построения доверительных интервалов для коэффициентов β_j нам понадобятся оценки дисперсий $\mathbb{V}\text{ar}(\hat{\beta}_j \mid X)$. К счастью, у нас есть несмещённая оценка $\hat{\sigma}^2$ для σ^2. Из неё мы легко построим оценку и для неизвестной ковариационной матрицы $\mathbb{V}\text{ar}(\hat{\beta} \mid X) = \sigma^2(X^TX)^{-1}$. А именно, мы просто подставим оценку дисперсии вместо неизвестной дисперсии:

$$\widehat{\mathbb{V}\text{ar}}(\hat{\beta} \mid X) = \hat{\sigma}^2(X^TX)^{-1} = \frac{RSS}{n-k}(X^TX)^{-1}.$$

Уточним, что эту оценку мы вывели из предпосылок теоремы Гаусса — Маркова. Если использовать другие предпосылки, то ковариационная матрица $\mathbb{V}\text{ar}(\hat{\beta} \mid X)$ перестанет быть равной $\sigma^2(X^TX)^{-1}$ и нам потребуется другой способ оценивания.

Оценки корней из дисперсий оценок называются стандартными ошибками оценок (standard errors).

$$\text{se}(\hat{\beta}_j) = \sqrt{\widehat{\mathbb{V}\text{ar}}(\hat{\beta}_j \mid X)} = \sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}$$

Число $[(X^TX)^{-1}]_{jj}$ мы берём в матрице $(X^TX)^{-1}$ из j-й строки j-го столбца.

6.6. Неправильная спецификация модели

Одной из предпосылок теоремы Гаусса — Маркова является правильный выбор спецификации, при котором мы регрессируем y в точности на набор истинных регрессоров. В реальности такое условие вряд ли выполнимо, так как не до всех регрессоров мы способны догадаться. А если догадаемся, то не все сможем измерить или собрать. Можно ли допустить неполную спецификацию модели, но получить BLUE-оценки (несмещённые и эффективные в классе линейных) для собранных регрессоров?

Рассмотрим для начала случай, когда при оценивании модели мы пропускаем часть важных регрессоров. Истинная модель имеет вид

$$y = W\beta + V\gamma + u,$$

где W — матрица регрессоров размерности $[n \times k_1]$, V — матрица регрессоров размерности $[n \times k_2]$. Обозначим через $X = [W \quad V]$ $[n \times k]$ матрицу всех регрессоров, где $k = k_1 + k_2$.

Вместо истинной модели оценивается следующая модель:

$$y = W\beta + \nu,$$

где ν — вектор случайных ошибок в оцениваемой модели.

Утверждение 6.1. тут должно быть утверждение про смещённость

Доказательство. Пусть X — истинный набор регрессоров, а W — собранный датасет. При этом $X = [W \quad V]$. Тогда новая МНК-оценка получается из изменившейся предпосылки о правильности спецификации $\tilde{\beta} = (W^T W)^{-1} W^T y$. Мы бы всё равно хотели получать несмещённую оценку.

$$\mathbb{E}(\tilde{\beta} \mid W) =$$

□

Утверждение 6.2. Пусть $\mathbb{V}\text{ar}(\hat{\beta} \mid W, V)$ — ковариационная матрица вектора оценок β, полученного по полному набору регрессоров $X = [W \quad V]$, а $\mathbb{V}\text{ar}(\tilde{\beta} \mid W)$ — ковариационная матрица вектора оценок β, полученного по регрессорам из матрицы W. Тогда матрица $\mathbb{V}\text{ar}(\hat{\beta} \mid W, V) - \mathbb{V}\text{ar}(\tilde{\beta} \mid W)$ является положительно полуопределённой матрицей.

Утверждение 6.2 означает, что на диагонали матрицы $\mathbb{V}\text{ar}(\hat{\beta} \mid W, V) - \mathbb{V}\text{ar}(\tilde{\beta} \mid W)$ стоят неотрицательные значения. В свою очередь, диагональный элемент с индексами jj представляет собой разницу дисперсий оценок коэффициента β_j, полученных по полному и по сокращенному набору переменных. Это означает, что $\mathbb{V}\text{ar}(\hat{\beta}_j \mid W, V) - \mathbb{V}\text{ar}(\tilde{\beta}_j \mid V) \geq 0$, то есть оценка $\tilde{\beta}_j$ имеет меньшую условную дисперсию. Из-за меньшей условной дисперсии оценка $\tilde{\beta}_j$ может получиться более эффективной по сравнению с оценкой $\hat{\beta}_j$.

Утверждение 6.3. Оценка дисперсии случайной ошибки $\tilde{\sigma}^2 = \frac{RSS}{n-k_1}$, полученная по модели с пропущенными переменными, является смещённой,

$$\mathbb{E}(\tilde{\sigma}^2 \mid W) \neq \sigma^2$$

Доказательство. Вспомним матричное представление RSS:

$$RSS = y^T M y, \text{ где } M = I_n - W(W^T W)^{-1} W^T.$$

Рассчитаем математическое ожидание RSS, учитывая, что истинной моделью является модель по набору регрессоров $X = [W \quad V]$:

$$\mathbb{E}(RSS \mid W, V) = \mathbb{E}(y^T M y \mid W, V) = \mathbb{E}((W\beta + V\gamma + u)^T M (W\beta + V\gamma + u) \mid W, V) =$$
$$= \mathbb{E}(u^T M u + 2\gamma^T V^T M u + \gamma^T V^T M V \gamma \mid W, V) = \sigma^2(n - k_1) + \gamma^T V^T M V \gamma.$$

Выше мы использовали следующие результаты:

- $MW = 0$;
- $\mathbb{E}(2\gamma^T V^T M u \mid W, V) = 2\gamma^T V^T M \, \mathbb{E}(u \mid W, V) = 0$;
- $\mathbb{E}(u^T M u \mid W, V) = \sigma^2(n - k_1)$.

Таким образом, получаем, что

$$\mathbb{E}(\tilde{\sigma}^2 \mid W, V) = \mathbb{E}\left(\frac{RSS}{n - k_1} \mid W, V\right) = \frac{1}{n - k_1}(\sigma^2(n - k_1) + \gamma^T V^T M V \gamma) = \sigma^2 + \frac{1}{n - k_1}\gamma^T V^T M V \gamma.$$

Оценка дисперсии $\tilde{\sigma}^2$ будет несмещённой только, если $\gamma = 0$. Равенство $\gamma = 0$ означает, что пропущенных переменных нет и $X = W$. Заметим также, что $V^T M V = (MV)^T MV$, что означает, что матрица $V^T MV$ является положительно полуопределённой. Следовательно, смещение оценки $\tilde{\sigma}^2$ в общем случае положительное. □

Далее проанализируем, что происходит со свойствами несмещённости и эффективности МНК-оценок в случае включения в модель лишних регрессоров.

Теперь истинной моделью является

$$y = X\beta + u.$$

Вместо истинной модели оценивается следующая модель:

$$y = X\beta + R\gamma + \nu,$$

где R — матрица лишних регрессоров.

Утверждение 6.4. При включении лишних регрессоров МНК-оценка $\tilde{\beta}$, полученная в модели с набором регрессоров $\begin{pmatrix} X & R \end{pmatrix}$, остаётся несмещенной, то есть $\mathbb{E}(\tilde{\beta} \mid X, R) = \beta$.

Утверждение 6.5. Пусть $\mathbb{V}\text{ar}(\hat{\beta} \mid X$ — ковариационная матрица вектора оценок β, полученного по истинному набору регрессоров X, а $\mathbb{V}\text{ar}(\tilde{\beta} \mid X, R)$ — ковариационная матрица вектора оценок β, полученного по регрессорам из матрицы

$$\begin{pmatrix} X & R \end{pmatrix}$$

Тогда матрица $\mathbb{V}\text{ar}(\tilde{\beta} \mid X, R) - \mathbb{V}\text{ar}(\hat{\beta} \mid X)$ является положительно полуопределённой матрицей.

Утверждение 6.5 означает, что на диагонали матрицы $\mathbb{V}\text{ar}(\tilde{\beta} \mid X, R) - \mathbb{V}\text{ar}(\hat{\beta} \mid X)$ стоят неотрицательные значения. В свою очередь, диагональный элемент с индексами jj представляет собой разницу дисперсий оценок коэффициента β_j, полученных по расширенному и по истинному наборам переменных. Это означает, что $\mathbb{V}\text{ar}(\tilde{\beta}_j \mid X, R) - \mathbb{V}\text{ar}(\hat{\beta}_j \mid X) \geq 0$, то есть оценка $\tilde{\beta}_j$ имеет большую условную дисперсию. Из-за большей условной дисперсии оценка $\tilde{\beta}_j$ может получиться менее эффективной по сравнению с оценкой $\hat{\beta}_j$.

6.7. Задачи для семинара:

Задача 22. Исследовательница Мишель собрала данные по 20 студентам. Переменная y_i — количество решённых задач по эконометрике i-м студентом, а x_i — количество просмотренных серий любимого сериала за прошедший год. Оказалось, что $\sum y_i = 10$, $\sum x_i = 0$, $\sum x_i^2 = 40$, $\sum y_i^2 = 50$, $\sum x_i y_i = 60$.

а) Найдите МНК-оценки коэффициентов парной регрессии.

б) В рамках предположения $\mathbb{E}(u_i \mid X) = 0$ найдите $\mathbb{E}(y_i \mid X)$, $\mathbb{E}(\hat\beta_j \mid X)$, $\mathbb{E}(\hat u_i \mid X)$, $\mathbb{E}(\hat y_i \mid X)$.

в) Предположим дополнительно, что $\mathbb{V}\mathrm{ar}(u_i \mid X) = \sigma^2$ и u_i при фиксированных X независимы. Найдите $\mathbb{V}\mathrm{ar}(y_i \mid X)$, $\mathbb{V}\mathrm{ar}(y_i(x_i - \bar x) \mid X)$, $\mathbb{V}\mathrm{ar}(\sum y_i(x_i - \bar x) \mid X)$, $\mathbb{V}\mathrm{ar}(\hat\beta_2 \mid X)$.

Решение.

Здесь нужны решения

Задача 23. Рассмотрим модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + u_i$, где

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{pmatrix}.$$

Случайные ошибки u_i независимы и нормально распределены с $\mathbb{E}(u \mid X) = 0$ и $\mathbb{V}\mathrm{ar}(u \mid X) = \sigma^2 I$.

Для удобства расчётов даны матрицы: $X'X$ и $(X'X)^{-1}$

$$X'X = \begin{pmatrix} 5 & 3 & 1 \\ 3 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \quad (X'X)^{-1} = \begin{pmatrix} 0.5 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}.$$

а) Найдите $\mathbb{E}(\hat\sigma^2 \mid X)$, $\hat\sigma^2$.

б) Найдите $\mathbb{V}\mathrm{ar}(u_1)$, $\mathbb{V}\mathrm{ar}(\beta_1)$, $\mathbb{V}\mathrm{ar}(\hat\beta_1 \mid X)$, $\widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_1 \mid X)$, $\mathbb{E}(\hat\beta_1^2 \mid X) - \beta_1^2$;

в) Найдите $\mathbb{C}\mathrm{ov}(\hat\beta_2, \hat\beta_3 \mid X)$, $\widehat{\mathbb{C}\mathrm{ov}}(\hat\beta_2, \hat\beta_3 \mid X)$, $\mathbb{V}\mathrm{ar}(\hat\beta_2 - \hat\beta_3 \mid X)$, $\widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_2 - \hat\beta_3 \mid X)$;

г) Найдите $\mathbb{V}\mathrm{ar}(\beta_2 - \beta_3)$, $\mathbb{C}\mathrm{orr}(\hat\beta_2, \hat\beta_3 \mid X)$, $\widehat{\mathbb{C}\mathrm{orr}}(\hat\beta_2, \hat\beta_3 \mid X)$;

Решение.

replace 4 by σ^2

check order of questions

а) $\mathbb{V}\mathrm{ar}(u_1) = \mathbb{V}\mathrm{ar}(u)_{(1,1)} = 4 \cdot I_{(1,1)} = 4$

б) $\mathbb{V}\mathrm{ar}(\beta_1) = 0$, так как β_1 — детерминированная величина.

в) $\mathbb{V}\mathrm{ar}(\hat\beta_1) = \sigma^2 (X'X)^{-1}_{(1,1)} = 0.5\sigma^2 = 0.5 \cdot 4 = 2$

г) $\widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_1) = \hat\sigma^2(X'X)^{-1}_{(1,1)} = 0.5\hat\sigma^2_{(1,1)} = 0.5\frac{RSS}{5-3} = 0.25RSS = 0.25y'(I - X(X'X)^{-1}X')y = 0.25\cdot 1 = 0.25$
$\hat\sigma^2 = \frac{RSS}{n-k} = \frac{1}{2}$.

д) Так как оценки МНК являются несмещёнными, то $\mathbb{E}(\hat\beta) = \beta$, значит:

$$\mathbb{E}(\hat\beta_1) - \beta_1^2 = \mathbb{E}(\hat\beta_1) - (\mathbb{E}(\hat\beta_1))^2 = \widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_1) = 0.25$$

е) $\mathbb{C}\mathrm{ov}(\hat\beta_2, \hat\beta_3) = \sigma^2(X'X)^{-1}_{(2,3)} = 4\cdot\left(-\frac{1}{2}\right) = -2$

ж) $\widehat{\mathbb{C}\mathrm{ov}}(\hat\beta_2, \hat\beta_3) = \widehat{\mathbb{V}\mathrm{ar}}(\hat\beta)_{(2,3)} = \hat\sigma^2(X'X)^{-1}_{(2,3)} = \frac{1}{2}\cdot\left(-\frac{1}{2}\right) = -\frac{1}{4}$

з) $\mathbb{V}\mathrm{ar}(\hat\beta_2 - \hat\beta_3) = \mathbb{V}\mathrm{ar}(\hat\beta_2) + \mathbb{V}\mathrm{ar}(\hat\beta_3) + 2\,\mathbb{C}\mathrm{ov}(\hat\beta_2, \hat\beta_3) = \sigma^2((X'X)^{-1}_{(2,2)} + (X'X)^{-1}_{(3,3)} + 2(X'X)^{-1}_{(2,3)} = 4(1 + 1.5 + 2\cdot(-0.5)) = 6$

и) $\widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_2 - \hat\beta_3) = \widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_2) + \widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_3) + 2\,\widehat{\mathbb{C}\mathrm{ov}}(\hat\beta_2, \hat\beta_3) = \hat\sigma^2((X'X)^{-1}_{(2,2)} + (X'X)^{-1}_{(3,3)} + 2(X'X)^{-1}_{(2,3)} = \frac{1}{2}\cdot 1.5 = 0.75$

к) $\mathbb{V}\mathrm{ar}(\beta_2 - \beta_3) = 0$

л) $\mathbb{C}\mathrm{orr}(\hat\beta_2, \hat\beta_3) = \frac{\mathbb{C}\mathrm{ov}(\hat\beta_2, \hat\beta_3)}{\sqrt{\mathbb{V}\mathrm{ar}(\hat\beta_2)\,\mathbb{V}\mathrm{ar}(\hat\beta_3)}} = \frac{-2}{\sqrt{4\cdot 6}} = -\frac{\sqrt{6}}{6}$

м) $\widehat{\mathbb{C}\mathrm{orr}}(\beta_2, \beta_3) = \frac{\widehat{\mathbb{C}\mathrm{ov}}(\hat\beta_2, \hat\beta_3)}{\sqrt{\widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_2)\,\widehat{\mathbb{V}\mathrm{ar}}(\hat\beta_3)}} = \frac{-\frac{1}{4}}{\sqrt{\frac{1}{2}\cdot\frac{3}{4}}} = -\frac{\sqrt{6}}{6}$

н) $(n-k)\frac{\hat\sigma^2}{\sigma^2} \sim \chi^2_{n-k}$.

$$\mathbb{E}\left((n-k)\frac{\hat\sigma^2}{\sigma^2}\right) = n - k$$

$$\mathbb{E}\left(\frac{\hat\sigma^2}{2}\right) = 1$$

$$\mathbb{E}(\hat\sigma^2) = 2$$

о) $\hat\sigma^2 = \frac{RSS}{n-k} = \frac{1}{2}$

Задача 24. Рассмотрим классическую линейную модель $y = X\beta + u$ с предпосылками Гаусса — Маркова: $\mathbb{E}(u \mid X) = 0$ и $\mathbb{V}\mathrm{ar}(u \mid X) = \sigma^2 I$. Для всех случайных векторов (y, $\hat y$, $\hat\beta$, u, $\hat u$, $\bar y$) найдите все возможные ожидания и ковариационные матрицы $\mathbb{E}(\cdot)$, $\mathbb{V}\mathrm{ar}(\cdot)$, $\mathbb{C}\mathrm{ov}(\cdot,\cdot)$.

Решение.

Здесь нужны решения

Задача 25. Рассмотрим модель $y_i = \beta x_i + u_i$ с двумя наблюдениями, $x_1 = 1$, $x_2 = 2$. Величины u_1 и u_2 независимы и равновероятно равны $+1$ или -1.

а) Найдите оценку $\hat\beta_{\text{ols}}$ для β с помощью метода наименьших квадратов.

б) Чему равна дисперсия $\mathbb{V}\mathrm{ar}(\hat\beta_{\text{ols}} \mid x)$ и ожидание $\mathbb{E}(\hat\beta_{\text{ols}} \mid x)$?

в) Постройте несмещённую оценку $\hat\beta_{\text{best}}$ с наименьшей дисперсией.

г) Чему равна дисперсия $\mathbb{V}\text{ar}(\hat{\beta}_{\text{best}} \mid x)$?

д) А как же теорема Гаусса — Маркова? Почему в данном примере удаётся построить оценку с дисперсией меньше, чем у оценки методом наименьших квадратов?

Решение. а) $\hat{\beta}_{\text{ols}} = (y_1 + 2y_2)/5$;

б) $\mathbb{V}\text{ar}(\hat{\beta}_{\text{ols}} \mid x) = 1/5$;

в) Заметим, что по величине $2y_1 - y_2$ можно однозначно восстановить величины ошибок u_1 и u_2. Например, если $2y_1 - y_2 = 3$, то $u_1 = 1$, $u_2 = -1$.

$$\hat{\beta}_{\text{best}} = \begin{cases} y_1 + 1, \text{ если } 2y_1 - y_2 < 0, \\ y_1 - 1, \text{ если } 2y_1 - y_2 > 0. \end{cases}$$

г) Шок контент, $\mathbb{V}\text{ar}(\hat{\beta}_{\text{best}} \mid x) = 0$.

д) Построенная оценка $\hat{\beta}_{\text{best}}$ является нелинейной по y, а теорема Гаусса — Маркова гарантирует только, что метод наименьших квадратов порождает несмещённую оценку с наименьшей дисперсией среди линейных по y оценок.

Задача 26. Предположим, что все предпосылки теоремы Гаусса — Маркова выполнены. Вычислите математические ожидания для TSS, ESS и RSS, используя их матричные представления.

Решение.
Здесь нужны решения

Задача 27. (Hansen 4.14)
Задана модель $y = X\beta + u$, для которой выполняются предпосылки теоремы Гаусса — Маркова. Вас интересует величина $\theta = \beta^2$. Получены МНК-оценки коэффициентов: $\hat{\beta}$, $V_{\hat{\beta}} = \mathbb{V}\text{ar}[\hat{\beta} \mid X]$. Кажется, неплохой идеей будет оценить θ как $\hat{\theta} = \hat{\beta}^2$.

а) Найдите $\mathbb{E}[\hat{\theta} \mid X]$. Является ли $\hat{\theta}$ смещённой?

б) Предложите способ коррекции смещения для получения несмещённой оценки $\hat{\theta}^*$, используя результаты предыдущего пункта.

Решение.
Здесь нужны решения

Задача 28. Рассмотрим модель регрессии $y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i$. Все предпосылки теоремы Гаусса — Маркова выполнены. Дополнительно предположим, что $u_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$. Дополнительно известно, что на самом деле $\beta_2 = \cdots = \beta_k = 0$.

а) Найдите $\mathbb{E}(R^2)$.

Решение:
Модель без ограничений:

$$y_i = \beta_1 + \beta_2 x_{i1} + \cdots + \beta_k x_{ik} + u_i.$$

Модель с ограничениями (истинная модель!):

$$y_i = \beta_1 + u_i.$$

Тогда F-статистика имеет следующий вид:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \sim F(k-1, n-k).$$

Выразим R^2:

$$R^2(n-k) = F(1-R^2)(n-k)$$

Факт дня №1: Если $X \sim F(k_1, k_2)$, то $Y = \frac{\frac{k_1}{k_2}X}{1+\frac{k_1}{k_2}X} \sim \text{Beta}\left(\frac{k_1}{2}, \frac{k_2}{2}\right)$.

Используя факт дня №1, получаем:

$$R^2 = \frac{(k-1)F}{(n-k)+(k-1)F} = \frac{\frac{k-1}{n-k}F}{1+\frac{k-1}{n-k}F} \sim \text{Beta}\left(\frac{k-1}{2}, \frac{n-k}{2}\right).$$

Тогда чтобы посчитать математическое ожидание R^2, надо вспомнить, чему равно математическое ожидание для $Beta\left(\frac{k-1}{2}, \frac{n-k}{2}\right)$:

$$E(R^2) = \frac{\frac{k-1}{2}}{\frac{k-1}{2}+\frac{n-k}{2}} = \frac{k-1}{n-1}.$$

Что нам даёт полученный результат? Математическое ожидание коэффициента детерминации линейно по k. То есть даже при включении в модель лишних факторов R^2 все равно продолжает линейно расти!

б) Найдите $\mathbb{E}(R^2_{\text{adj}})$.
Решение:
Скорректированный коэффициент детерминации имеет вид:

$$R^2_{\text{adj}} = 1 - \left(1-R^2\right)\frac{n-1}{n-k}.$$

Рассчитаем математическое ожидание:

$$E(R^2_{\text{adj}}) = E\left(1 - \left(1-R^2\right)\frac{n-1}{n-k}\right) = 1 - \frac{n-1}{n-k} + \frac{n-1}{n-k}E(R^2) =$$

$$= 1 - \frac{n-1}{n-k} + \frac{n-1}{n-k}\frac{k-1}{n-1} = 0.$$

Скорректированный R^2 помог решить проблему линейного роста по k!

Задача 29. У овечки Долли был набор данных из n наблюдений для которого были выполнены предпосылки теоремы Гаусса — Маркова. Овечка Долли клонировала каждое наблюдение по одному разу и дописала каждое наблюдение-клон сразу после исходного наблюдения.

а) Как выглядит ковариационная матрица ошибок для нового набора данных?

б) Как изменится ответ на (а), если Долли клонирует только последнее наблюдение n раз?

Решение. а) Ковариационная матрица будет содержать блоки B на диагонали

$$\mathbb{V}\text{ar}(u) = \begin{pmatrix} B & 0 & 0 & \dots \\ 0 & B & 0 & \dots \\ 0 & 0 & B & \dots \\ \dots & & & \end{pmatrix},$$

где каждый блок равен $B = \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2 \end{pmatrix}$.

б) Ковариационная матрица будет состоять из четырех блоков: два блока нулевые, левые верхний блок пропорционален единичной матрицы, а все элементы правого нижнего блока равны σ^2:

$$\mathbb{V}\text{ar}(u) = \begin{pmatrix} \sigma^2 \cdot I & 0 \\ 0 & S \end{pmatrix},$$

где I — единичная матрица, а все $S_{ij} = \sigma^2$.

6.8. Компьютерные задачи для семинара:

Генерация R2 для вывода распределения
Генерация смещения
Генерация лишних регрессоров
Реальный пример с лишним регрессорами (тип знаки зодиака и ретроградный)
Какая-то длинная задача, которую из темы в тему и в ней находить потом нарушения предпосылок?
https://colab.research.google.com/drive/1wFrLyGcVVETx96jS93I4z8asgAQwqIdw?usp=sharing

6.9. Домашнее задание:

6.10. Чёрный трэк:

Умножение блочных матриц. Если размеры блоков допускают операцию умножения, то:

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \cdot \left[\begin{array}{c|c} E & F \\ \hline G & H \end{array}\right] = \left[\begin{array}{c|c} AE + BG & AF + BH \\ \hline CE + DG & CF + DH \end{array}\right].$$

Формула Фробениуса (блочное обращение).

$$\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]^{-1} = \left[\begin{array}{c|c} A^{-1} + A^{-1}BH^{-1}CA^{-1} & -A^{-1}BH^{-1} \\ \hline -H^{-1}CA^{-1} & H^{-1} \end{array}\right],$$

где A — невырожденная квадратная матрица размерности $[n \times n]$, D — квадратная матрица размерности $[k \times k]$, $H = D - CA^{-1}B$.

Задача 30. Пусть истинной является модель $y = X_1\beta_1 + X_2\beta_2 + u$, где X_1, X_2 — матрицы признаков размерностей $[n \times k_1]$ и $[n \times k_2]$ соответственно. Вместо истинной модели вы оцениваете модель вида $y = X_1\beta_1 + v$, где v — вектор случайной ошибки, удовлетворяющий предпосылкам теоремы Гаусса — Маркова.

а) Будет ли МНК-оценка вектора параметров β_1 несмещённой?

б) Будет ли несмещённой МНК-оценка дисперсии случайной ошибки?

в) Рассчитайте $\mathbb{V}\text{ar}(\hat{\beta}_1)$. Не противоречит ли полученной результат теореме Гаусса — Маркова?

Задача 31. Пусть истинной является модель $y = X_1\beta_1 + u$, где X_1 — матрица признаков размерности $[n \times k_1]$. Вместо истинной модели вы оцениваете модель вида $y = X_1\beta_1 + X_2\beta_2 + v$, где X_2 — матрица признаков размерности $[n \times k_2]$, v — вектор случайной ошибки, удовлетворяющий предпосылкам теоремы Гаусса — Маркова.

а) Будет ли МНК-оценка вектора параметров β_1 несмещённой?

б) Будет ли несмещённой МНК-оценка дисперсии случайной ошибки?

в) Рассчитайте $\mathbb{V}\text{ar}(\hat{\beta}_1)$. Не противоречит ли полученной результат теореме Гаусса — Маркова?

7. Доверительные интервалы для коэффициентов

Построение доверительных интервалов для МНК оценок. Проверка гипотез. Асимптотика без нормальности ошибок. Нормальность ошибок.

7.1. Случай многомерного нормального распределения

сопроводить оценкой правдоподобия и показать, что она совпадает с МНК

Напомним несколько фактов про многомерное нормальное распределение.

Начнём с классического определения:

Определение 7.1 (многомерное нормальное распределение)**.** Вектор v имеет многомерное невырожденное нормальное распределение, $v \sim \mathcal{N}(\mu, C)$, если его совместная функция плотности равна

$$f(v) = (2\pi)^{-n/2} \det(C)^{-1/2} \exp\left(-\frac{1}{2}(v-\mu)^T C^{-1}(v-\mu)\right),$$

где n — размерность вектора v.

Заметим, что совместный закон распределения нормального вектора v полностью определён его ожиданием $\mathbb{E}(v)$ и его ковариационной матрицей $\mathbb{V}\text{ar}(v)$. Никакие другие параметры в совместную функцию плотности не входят.

Для многомерного нормального распределения нет разницы между независимостью и некоррелированностью:

Теорема 7.2 (некоррелированность и независимость для нормального вектора)**.** Если нормальный вектор v состоит из двух подвекторов, $v = (x, y)$, то $\mathbb{C}\text{ov}(x, y) = 0$ если и только если подвекторы x и y независимы.

Доказательство. Докажем в одну сторону. Если подвекторы x и y независимы, то $\mathbb{C}\text{ov}(x, y) = 0$. А теперь изящно докажем в обратную сторону. Если $\mathbb{C}\text{ov}(x, y) = 0$, то вся ковариационная матрица $\mathbb{V}\text{ar}(v)$ ровно такая же как и в случае независимых x и y. Остаётся лишь вспомнить, что $\mathbb{E}(v)$ и $\mathbb{V}\text{ar}(v)$ полностью определяют закон распределения нормального вектора v, а значит компоненты обязаны быть независимы. □

Также для многомерного нормального распределения нет разницы между условным ожиданием $\mathbb{E}(y \mid x)$ и наилучшим линейным приближением $\text{BestLin}(y \mid x)$, другими словами функция $\mathbb{E}(y \mid x)$ линейна по x.

Задача 32. Рассмотрим совместное нормальное распределение

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{pmatrix} \right)$$

а) Найдите наилучшее линейное приближение $\text{BestLin}(y \mid x)$.

б) Найдите условное ожидание $\mathbb{E}(y \mid x)$.

в) Найдиту условную дисперсию $\mathbb{V}\text{ar}(y \mid x)$.

Решение.

Здесь рассказать про определение bestlin в векторном случае?

а) Пусть $\text{BestLin}(y \mid x) = a + Bx$. Мы хотим, чтобы ошибка линейной аппроксимации $r = y - \text{BestLin}(y \mid x)$ была некоррелирована с x,

$$\mathbb{C}\text{ov}(y - \text{BestLin}(y \mid x), x) = 0.$$

Другими словами,

$$\mathbb{C}\text{ov}(y, x) = \mathbb{C}\text{ov}(\text{BestLin}(y \mid x), x) = 0.$$

Подставим $\text{BestLin}(y \mid x) = a + Bx$.

$$\mathbb{C}\text{ov}(y, x) = \mathbb{C}\text{ov}(a + Bx, x) = \mathbb{C}\text{ov}(Bx, x) = B\,\mathbb{C}\text{ov}(x, x) = B\,\mathbb{V}\text{ar}(x).$$

Отсюда $C_{yx} = BC_{xx}$ и $B = C_{yx}C_{xx}^{-1}$. Кроме того, ошибка линейной аппроксимации должна иметь нулевое ожидание, следовательно,

$$\mathbb{E}(y) = \mathbb{E}(\text{BestLin}(y \mid x)) = a + B\,\mathbb{E}(x).$$

Получаем уравнение на a:

$$\mu_y = a + C_{yx}C_{xx}^{-1}\mu_x$$

И ответ,

$$\begin{cases} B = C_{yx}C_{xx}^{-1} \\ a = \mu_y - C_{yx}C_{xx}^{-1}\mu_x \end{cases}$$

б) Для нормально распределённой пары векторов нулевая ковариация равносильная независимости. Следовательно, ошибка аппроксимации $r = y - \text{BestLin}(y \mid x)$ и x независимы. Отсюда мы получаем, что для многомерного нормально распределённого вектора (x, y)

$$\mathbb{E}(y \mid x) = \text{BestLin}(y \mid x) = a + Bx$$

в) Условная дисперсия — это безусловная дисперсия ошибки прогноза,

$$\mathbb{V}\text{ar}(y \mid x) = \mathbb{V}\text{ar}(a + Bx + r \mid x) = \mathbb{V}\text{ar}(r \mid x) = \mathbb{V}\text{ar}(r).$$

Осталось вспомнить, что $y = a + Bx + r$, прогноз $a + Bx$ и ошибка r некоррелированы,

$$\mathbb{V}\text{ar}(y) = \mathbb{V}\text{ar}(a + Bx) + \mathbb{V}\text{ar}(r).$$

Значит,

$$\mathbb{V}\text{ar}(y \mid x) = \mathbb{V}\text{ar}(r) = \mathbb{V}\text{ar}(y) - B\,\mathbb{V}\text{ar}(x)B^T = C_{yy} - C_{yx}C_{xx}^{-1}C_{xx}C_{xx}^{-1}C_{xy} = C_{yy} - C_{yx}C_{xx}^{-1}C_{xy}.$$

Отметим, что для компонент x и y нормального вектора (x, y) условная дисперсия получилась постоянной и не зависящей от x.

Для ненормального распределения условное ожидание $\mathbb{E}(y \mid x)$ и условная дисперсия $\mathbb{V}\text{ar}(y \mid x)$ вполне могут быть нелинейными.

Введём дополнительную предпосылку $(u \mid X) \sim \mathcal{N}(0, \sigma^2 I)$. Учитывая, что $\hat{\beta} = (X^TX)^{-1}X^Ty = \beta + (X^TX)^{-1}X^Tu$, получаем

$$(\hat{\beta} \mid X) \sim \mathcal{N}(\beta, \sigma^2(X^TX)^{-1}).$$

7.2. Независимость оценок β и $\hat{\sigma}^2$

МНК-оценка вектора коэффициентов β имеет вид

$$\hat{\beta} = (X^TX)^{-1}X^Ty.$$

Несмещённая оценка дисперсии случайной ошибки:

$$\hat{\sigma}^2 = \frac{RSS}{n-k} = \frac{\hat{u}^T\hat{u}}{n-k}.$$

Распишем

$$\hat{\beta} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X+u) = \beta + (X^TX)^{-1}X^Tu = \beta + Au,$$

где $A = (X^TX)^{-1}X^T$.

В случае, когда случайный вектор ошибок u является нормальным, можно показать, что оценки $\hat{\beta}$ и $\hat{u}$ будут независимыми.

При $u \sim \mathcal{N}(0, \sigma^2 I_n)$ случайные векторы $\hat{\beta}$ и $\hat{u}$ имеют совместное многомерное нормальное распределение. Покажем, что $\hat{\beta}$ и $\hat{u}$ являются некоррелированными, из чего следует, что они также будет и независимыми, что справедливо для нормально распределенных векторов:

$$\mathbb{C}\text{ov}(\hat{\beta}, \hat{u}) = \mathbb{C}\text{ov}(\beta + Au, Mu) = \mathbb{C}\text{ov}(Au, Mu) = AM\,\mathbb{V}\text{ar}(u) = \sigma^2 AM = 0, \textbf{ так как } AM = 0.$$

Так как $\hat{\sigma}^2$ есть функция от случайного вектора $\hat{u}$, то оценки $\hat{\beta}$ и $\hat{\sigma}^2$ также независимы.

7.3. Проверка гипотез о параметрах

$$H_0 : \beta_j = \beta_j^0$$
$$H_1 : \beta_j \neq \beta_j^0$$

$$t = \frac{\hat{\beta}_j - \beta_j^0}{\mathrm{se}(\hat{\beta}_j)} \overset{H_0}{\sim} t(n-k)$$

Проверка гипотезы о незначимости модели в целом

$$H_0 : \beta_2 = \cdots = \beta_k = 0$$

$$H_1 : \sum_{j=2}^{k} \beta_j^2 > 0$$

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k} \overset{H_0}{\sim} F(k-1, n-k).$$

8. Бутстрэп

Бутстрэп. Классический бутстрэп до регрессии и бутстрэп в регрессии. Метод наименьших модулей.
Чёрный трэк: возможно, разные варианты бутстрэпа в регрессии? BCA-бутстрэп до регрессии?

9. Выбор функциональной формы

Дамми-переменные и их интерпретация. Функциональные формы: полиномы, логарифмы, интерпретация коэффициентов. Информационные критерии.
Чёрный трэк: Структурные сдвиги. Тест Чоу. Локально-линейная регрессия (LOESS).

10. Гетероскедастичность

Гетероскедастичность. Тестирование гетероскедастичности. Робастные оценки. Доступный обобщённый МНК.
Задачи для доски:
Хансен: во сколько раз может быть недооценена дисперсия из-за гетероскедастичности
Коммент: акцент на робастных ошибках, тестирование и обобщённый МНК — кратко.

11. Мультиколлинеарность и метод главных компонент

Мультиколлинеарность и метод главных компонент.
Чёрный трэк: несколько взглядов на метод главных компонент? LASSO?

12. Эндогенность

Эндогенность. Инструментальные переменные. Ошибка измерения регрессора. Двухшаговый МНК.

13. Эффекты воздействия

Оценка эффектов воздействия. ATE. LATE. Четкий (sharp) и нечеткий (fuzzy) разрывный регрессионный дизайн (RDD).

Чёрный трэк: Метод разность разностей (DiD). Динамический метода разность разностей (Event Study).

14. Задачи

15. Логистическая регрессия: точечные оценки

Логистическая регрессия: Бинарный и упорядоченный логит. Точечные оценки, прогнозы. Интерпретация предельных эффектов.

Чёрный трэк: Множественные логиты. Неупорядоченные, условные, смешанные логиты.

16. Логистическая регрессия: доверительные интервал

Логистическая регрессия: доверительные интервалы и проверка гипотез.

Чёрный трэк: разные хоббиты

16.1. Смещение, цензурирование и ■■■■■

Представим себе ситуацию, в которой зависимая количественная не всегда наблюдаема. Для моделирования этой ситуации мы введём скрытую латентная переменная y_i^*, которая линейно зависит от предиктора x_i, как обычно,

$$y_i^* = x_i^T\beta + u_i, \quad y^* = X^T\beta + u$$

Бинарная переменная $z_i \in \{0, 1\}$ равна 1 в случае, если мы наблюдаем y_i^*.

Возможно несколько случаев:

	наблюдаемость y^*	наблюдаемость x	наблюдаемость
Цензурирование censored model	зависит от y^*	всегда	
Усечение truncated model	зависит от y^*	если наблюдаем y^*	
Выборочное смещение sample selection	зависит от w	всегда	всегда
Переключающиеся режимы switching regimes	всегда, w переключает тип зависимости	всегда	всегда

Представим себе, что мы открыли дорогой ресторан. К нам заглядывают клиенты. Часть клиентов ужасаются от ценника и убегают, $y_i^* < 0$. Часть клиентов остаются и ужинают у нас, $y_i^* > 0$. Вместо нуля можно выбрать другой порог, но с нулём чуть-чуть удобнее.

16.2. Цензурирование

Рассмотрим самый распространённый вариант цензурирования: вместо отрицательных значений латентной переменной y_i^* мы видим нули.

Эта модель известна как тобит модель типа I, type I Tobit model.

$$\begin{cases} y_i^* = x_i^T\beta + u_i, \quad y^* = X^T\beta + u \\ (u \mid X) \sim \mathcal{N}(0, \sigma^2 I) \\ y_i = \max\{y_i^*, 0\} \\ (x_i, y_i) \text{ наблюдаемы при любых } i \end{cases}$$

Лог-функция правдоподобия равна

$$\ell(\beta, \sigma) = \sum_{y_i=0} \ln F(-x_i^T\beta/\sigma) + \sum_{y_i>0} \ln f((y_i - x_i^T\beta)/\sigma) - \sum_{y_i>0} \ln \sigma$$

16.3. Усечение

$$\begin{cases} y_i^* = x_i^T\beta + u_i, \quad y^* = X^T\beta + u \\ (u \mid X) \sim \mathcal{N}(0, \sigma^2 I) \\ y_i = \max\{y_i^*, 0\} \\ (x_i, y_i) \text{ наблюдаемы, если } y_i > 0 \end{cases}$$

Лог-функция правдоподобия равна

$$\ell(\beta, \sigma) = \sum_{y_i>0} \ln f((y_i - x_i^T\beta)/\sigma) - \sum_{y_i>0} \ln F(x_i^T\beta/\sigma) - \sum_{y_i>0} \ln \sigma$$

16.4. Три осмысленных условных ожидания

Ожидание латентной переменной показывает, сколько в среднем планирует потратить гость ресторана на ужин, ещё не видевший цен, полезность от ужина,

$$m^*(x_i) = \mathbb{E}(y_i^* \mid x_i) = x_i^T\beta$$

Предельный эффект для латентной переменной

$$\partial\, \mathbb{E}(y_i^* \mid x_{ij})/\partial x_{ij} = \beta_j$$

Ожидание цензурированной переменной, $y_i = \max\{y_i^*, 0\}$, сколько в среднем потратит человек, заглянувший в ресторан, с учётом того, что часть уйдёт испугавшись ценника

$$m(x_i) = \mathbb{E}(y_i \mid x_i) = x_i^T\beta F(x_i^T\beta/\sigma) + \sigma f(x_i^T\beta/\sigma)$$

Предельный эффект для цензурированной переменной

$$\partial\, \mathbb{E}(y_i \mid x_{ij})/\partial x_{ij} =$$

Условное ожидание усечённой переменной, $(y_i \mid y_i^* > 0)$, средний чек в ресторане

$$m^{\#}(x_i) = \mathbb{E}(y_i \mid x_i, y_i^* > 0) = x_i^T\beta + \sigma\, \text{IMR}(x_i^T\beta/\sigma),$$

где $\text{IMR}(s)$ — обратное отношение Миллса, inverse Mills ratio,

$$\text{IMR}(s) = \mathbb{E}(v \mid v + s > 0) = f(s)/F(s), \quad v \sim \mathcal{N}(0; 1)$$

Предельный эффект для ожидания усечённой переменной

Выборочное смещение

Переключающиеся режимы

List of Theorems

[PP12] K. B. Petersen и M. S. Pedersen. The Matrix Cookbook. 2012. URL: http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html. Матричные тождества, матричные распределения... Всё, что вы хотели узнать о матрицах, но боялись спросить! Обратите внимание на фамилии авторов :)

[GD18] Olya Gnilova и Boris Demeshev. How Gauss and Markov met Pythagoras: geometry in econometrics. 2018. URL: https://github.com/olyagnilova/gauss-markov-pythagoras. Как встретились Гаусс, Марков и Пифагор? Куча прикольных геометрических фактов и интерпретаций!