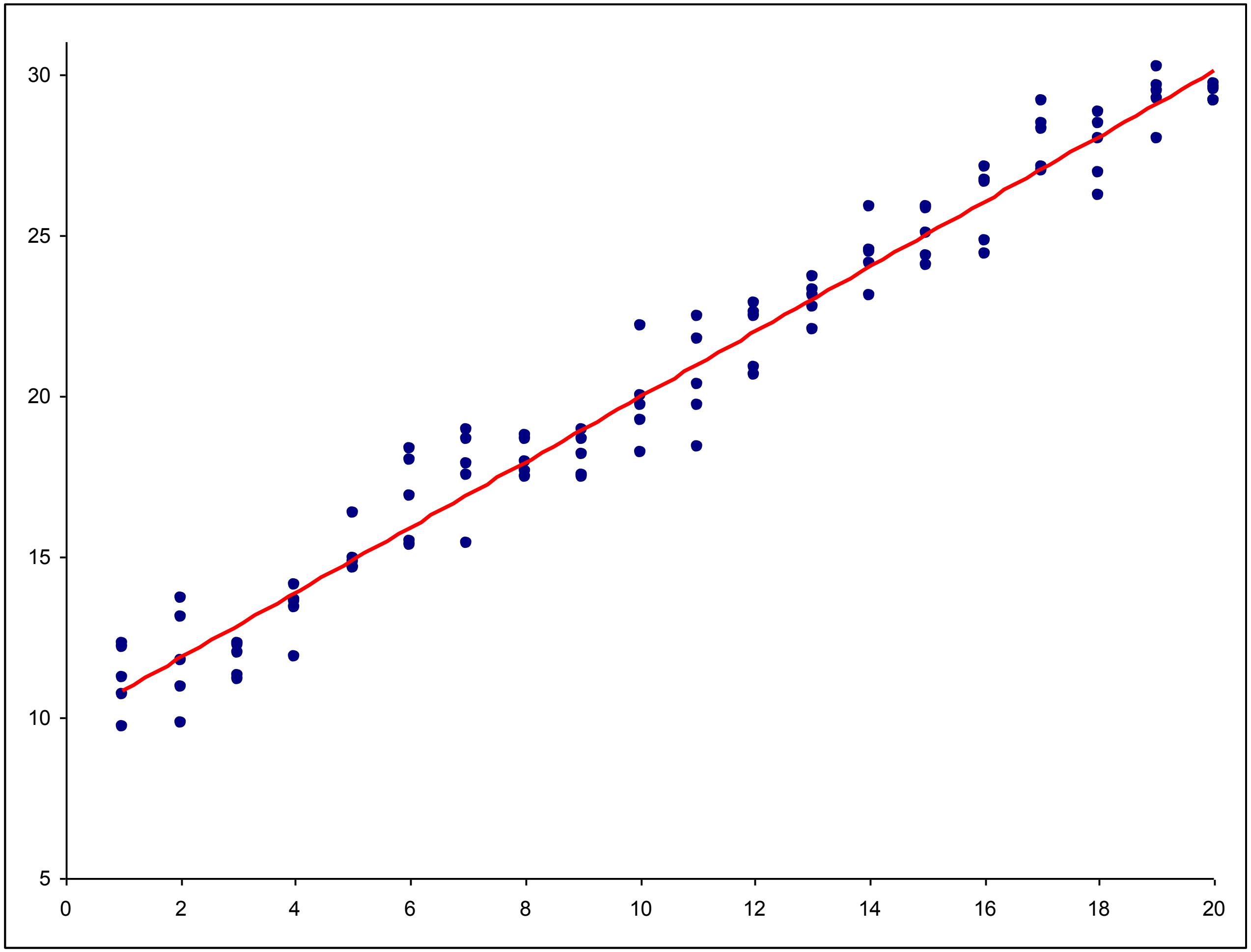
Линейная регрессия

Модель парной регрессии



В статистических данных редко встречаются точные линейные соотношения:

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

Обычно они бывают приближенными:

$$y_i \approx \beta_1 + \beta_2 x_i$$

<= Как на этом графике

Модель парной регрессии

Приблизительные взаимосвязи вида

$$y_i \approx \beta_1 + \beta_2 x_i \ ,$$

эконометристы обычно описывают следующим образом:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \ .$$

y_i — значения зависимой переменной

x_i — значения независимой переменной (регрессора)

ε_i — случайные ошибки

Модель парной регрессии

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

Откуда берутся случайные ошибки ε_i ?

1. Существуют другие, неучтенные в нашей упрощенной модели факторы. Эти факторы также оказывают влияние на зависимую переменную y
2. Присутствуют ошибки измерений зависимой переменной

Модель парной регрессии

В чем разница между β_i и $\hat{\beta}_i$?

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

β_1 и β_2 — истинные значения параметров модели, которые на практике **никогда не известны** исследователю

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

$\hat{\beta}_1$ и $\hat{\beta}_2$ — их оценки, полученные при помощи МНК, на основе случайной выборки

Следовательно, $\hat{\beta}_1$ и $\hat{\beta}_2$ — случайные величины

Модель парной регрессии

Разумеется, хочется, чтобы полученные оценки $\hat{\beta}_1$ и $\hat{\beta}_2$ были близки к истинным значениям.

При каких условиях можно на это надеяться?

Эти условия называют **предпосылками классической линейной модели парной регрессии (КЛМПР)**

Предпосылки КЛМПР

(1) Модель линейна по параметрам и правильно специфицирована

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

Как мы увидим в дальнейшем, правильная спецификация подразумевает в первую очередь отсутствие других переменных (кроме x), которые **влияют на $\boldsymbol{y}$** и одновременно **коррелируют с $\boldsymbol{x}$**.
Нарушение этого требования приводит к серьезным проблемам.

Предпосылки КЛМПР

(2) $x_1, \ldots, x_n$ — детерминированные (неслучайные) величины (не все одинаковые)

Предпосылки КЛМПР

(3) Математическое ожидание случайных ошибок равно нулю: $E(\varepsilon_i) = 0$

Предпосылки КЛМПР

(4) случайные ошибки имеют постоянную дисперсию: $V(\varepsilon_i) = \sigma^2 = const$

Предпосылки КЛМПР

(5) Случайные ошибки, соответствующие разным наблюдениям не зависят друг от друга (не коррелированны)

$Cov(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$

Теорема Гаусса — Маркова

Если выполнены условия (1)–(5),

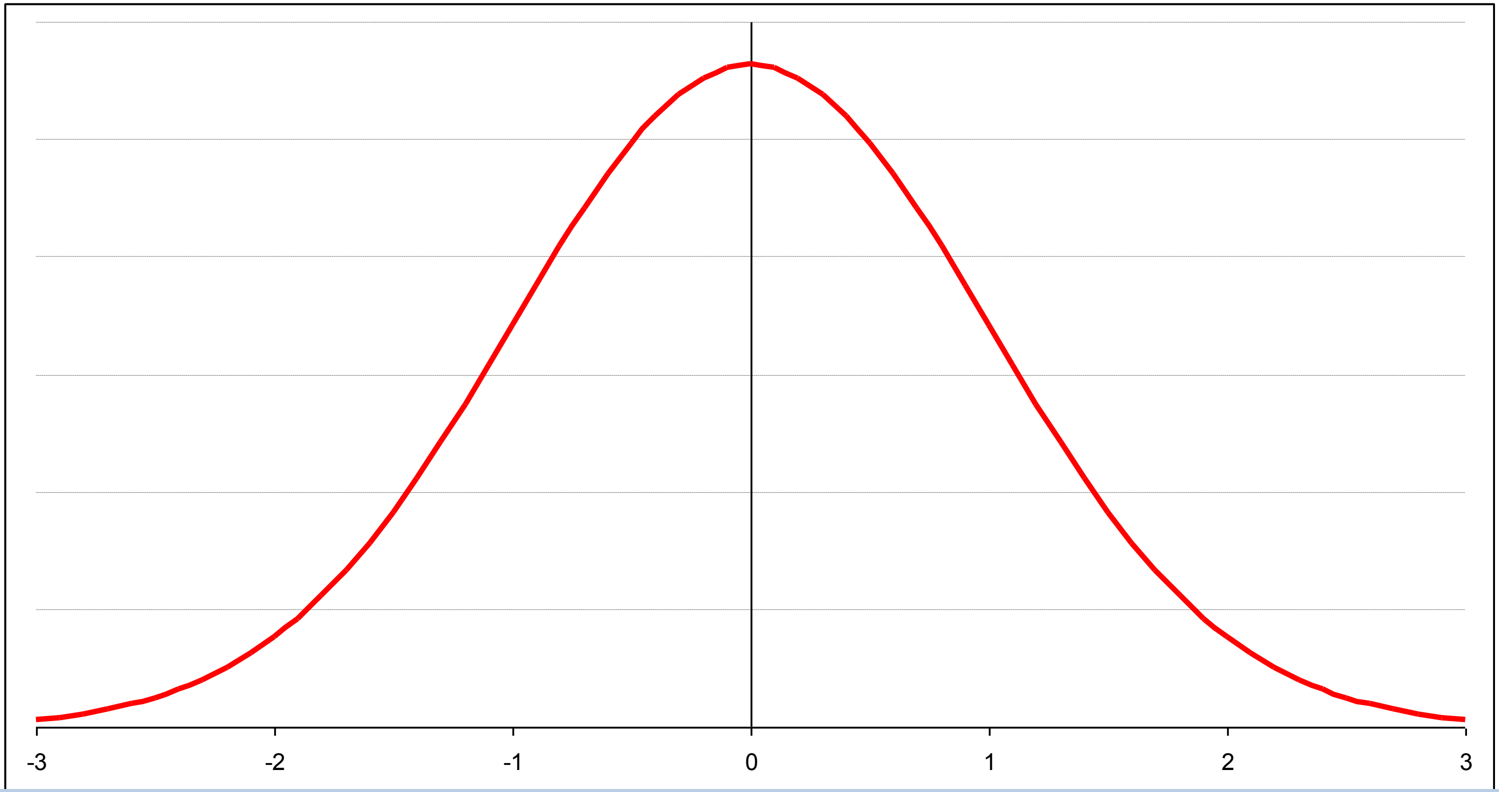
то оценки $\hat{\beta}_1$ и $\hat{\beta}_2$, полученные по методу наименьших квадратов (МНК), являются

(а) несмещенными

(б) эффективными, то есть имеют наименьшую дисперсию в классе всех линейных по ***y*** несмещенных оценок

Предпосылки КЛМПР

(6)* Случайные ошибки ε_i имеют нормальное распределение



Это свойство не требуется для теоремы Гаусса — Маркова, но полезно для проверки гипотез и построения доверительных интервалов

Теорема Гаусса — Маркова: доказательство несмещенности (1)

Полезное замечание

$$\sum(x_i-\bar{x})=\sum x_i-\sum\bar{x}=$$

$$=\sum x_i-n\bar{x}=\sum x_i-n\frac{\sum x_i}{n}=0$$

$$\hat{\beta}_2=\frac{\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum(x_i-\bar{x})^2}=$$

$$=\frac{\sum(x_i-\bar{x})(\beta_1+\beta_2x_i+\varepsilon_i-\beta_1-\beta_2\bar{x}-\bar{\varepsilon})}{\sum(x_i-\bar{x})^2}=$$

Теорема Гаусса — Маркова: доказательство несмещенности (2)

$$= \frac{\sum (x_i - \bar{x})(\beta_2 (x_i - \bar{x}) + \varepsilon_i - \bar{\varepsilon})}{\sum (x_i - \bar{x})^2} =$$

$$= \frac{\beta_2 \sum (x_i - \bar{x})^2 + \sum (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum (x_i - \bar{x})^2} =$$

$$= \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i - \bar{\varepsilon} \sum (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} =$$

$$= \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i - \bar{\varepsilon} 0}{\sum (x_i - \bar{x})^2}$$

Теорема Гаусса — Маркова: доказательство несмещенности (3)

$$\hat{\beta}_2 = \beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2}$$

$$E(\hat{\beta}_2) = E\left(\beta_2 + \frac{\sum (x_i - \bar{x})\varepsilon_i}{\sum (x_i - \bar{x})^2} \right) =$$

$$= \beta_2 + \frac{\sum (x_i - \bar{x})E(\varepsilon_i)}{\sum (x_i - \bar{x})^2} = \beta_2$$

Вычисление дисперсии оценки коэффициента (1)

$$V(\widehat{\beta_2}) = V\left(\beta_2 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right) = V\left(\frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right) =$$

$$= \frac{1}{(\sum(x_i - \bar{x})^2)^2} * V\left(\sum(x_i - \bar{x})\varepsilon_i\right) =$$

$$= \{\text{в силу предпосылки (5) о независимости}\} =$$

$$= \frac{1}{(\sum(x_i - \bar{x})^2)^2} * \sum V\big((x_i - \bar{x})\varepsilon_i\big) =$$

Вычисление дисперсии оценки коэффициента (2)

$$= \frac{1}{\left(\sum(x_i - \bar{x})^2\right)^2} * \sum V\big((x_i - \bar{x})\varepsilon_i\big) =$$

$$= \frac{1}{\left(\sum(x_i - \bar{x})^2\right)^2} * \sum (x_i - \bar{x})^2 * V(\varepsilon_i) =$$

$$= \frac{1}{\left(\sum(x_i - \bar{x})^2\right)^2} * \sum (x_i - \bar{x})^2 * \sigma^2 =$$

$$= \frac{\sum(x_i - \bar{x})^2}{\left(\sum(x_i - \bar{x})^2\right)^2} * \sigma^2 = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Оценка дисперсии случайной ошибки

В реальной ситуации величина σ^2 нам не известна. Вместо нее самой можно вычислить ее оценку.

Если выполнены предпосылки (1)-(5), то несмещенная оценка будет иметь вид

$$\widehat{\sigma^2} = S^2 = \frac{1}{n-2} * \sum_{i=1}^{n} e_i^2$$

Чтобы показать, что эта оценка является несмещенной, нужно аккуратно вычислить ее математическое ожидание

(см. Магнус, глава2)

Стандартные ошибки коэффициентов

Как мы показали выше, дисперсия оценки коэффициента β_2 имеет вид

$$V\left(\widehat{\beta_2}\right) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Следовательно, **оценка** этой дисперсии:

$$\hat{V}\left(\widehat{\beta_2}\right) = \frac{S^2}{\sum(x_i - \bar{x})^2}$$

Корень из оценки дисперсии оценки коэффициента называется **стандартной ошибкой оценки коэффициента** (standard error, s.e.):

$$se\left(\widehat{\beta_2}\right) = \sqrt{\hat{V}\left(\widehat{\beta_2}\right)} = \sqrt{\frac{S^2}{\sum(x_i - \bar{x})^2}}$$

Стандартные ошибки коэффициентов

$$se(\widehat{\beta_2}) = \sqrt{\hat{V}(\widehat{\beta_2})} = \sqrt{\frac{S^2}{\sum(x_i - \bar{x})^2}}$$

Стандартная ошибка оценки коэффициента характеризует ее точность: чем меньше стандартная ошибка, тем точнее оценен коэффициент. Стандартные ошибки нужны для проверки гипотез и построения доверительных интервалов.

Мы подробно показали, как получается стандартная ошибка для $\widehat{\beta_2}$. Аналогично можно получить стандартную ошибку для $\widehat{\beta_1}$. Она имеет вид:

$$se(\widehat{\beta_1}) = \sqrt{\hat{V}(\widehat{\beta_1})} = \sqrt{\frac{S^2}{n} * \frac{\sum x_i^2}{\sum(x_i - \bar{x})^2}}$$

Доверительные интервалы

$$\widehat{\beta_2} = \beta_2 + \frac{\sum(x_i - \bar{x}) * \varepsilon_i}{\sum(x_i - \bar{x})^2}$$

Если выполнена предпосылка (2) о том, что x_i — детерминированные величины, и предпосылка (6) о том, что случайные ошибки распределены нормально, то $\widehat{\beta_2}$ представляет собой линейную комбинацию нормальных случайных величин.

Следовательно $\widehat{\beta_2}$ также является нормальной случайной величиной.

$$\widehat{\beta_2} \sim N\left(\beta_2, \quad \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

Доверительные интервалы

$$\widehat{\beta_2} \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right)$$

В этом случае случайная величина $\frac{\widehat{\beta_2} - \beta_2}{se(\widehat{\beta_2})}$ имеет t-распределение Стьюдента с $(n-2)$ степенями свободы.

$$\frac{\widehat{\beta_2} - \beta_2}{se(\widehat{\beta_2})} \sim t_{n-2}$$

Доказательство этого факта мы обсудим в конце темы (если останется время)

Этот факт можно использовать для построения доверительных интервалов.

Доверительные интервалы

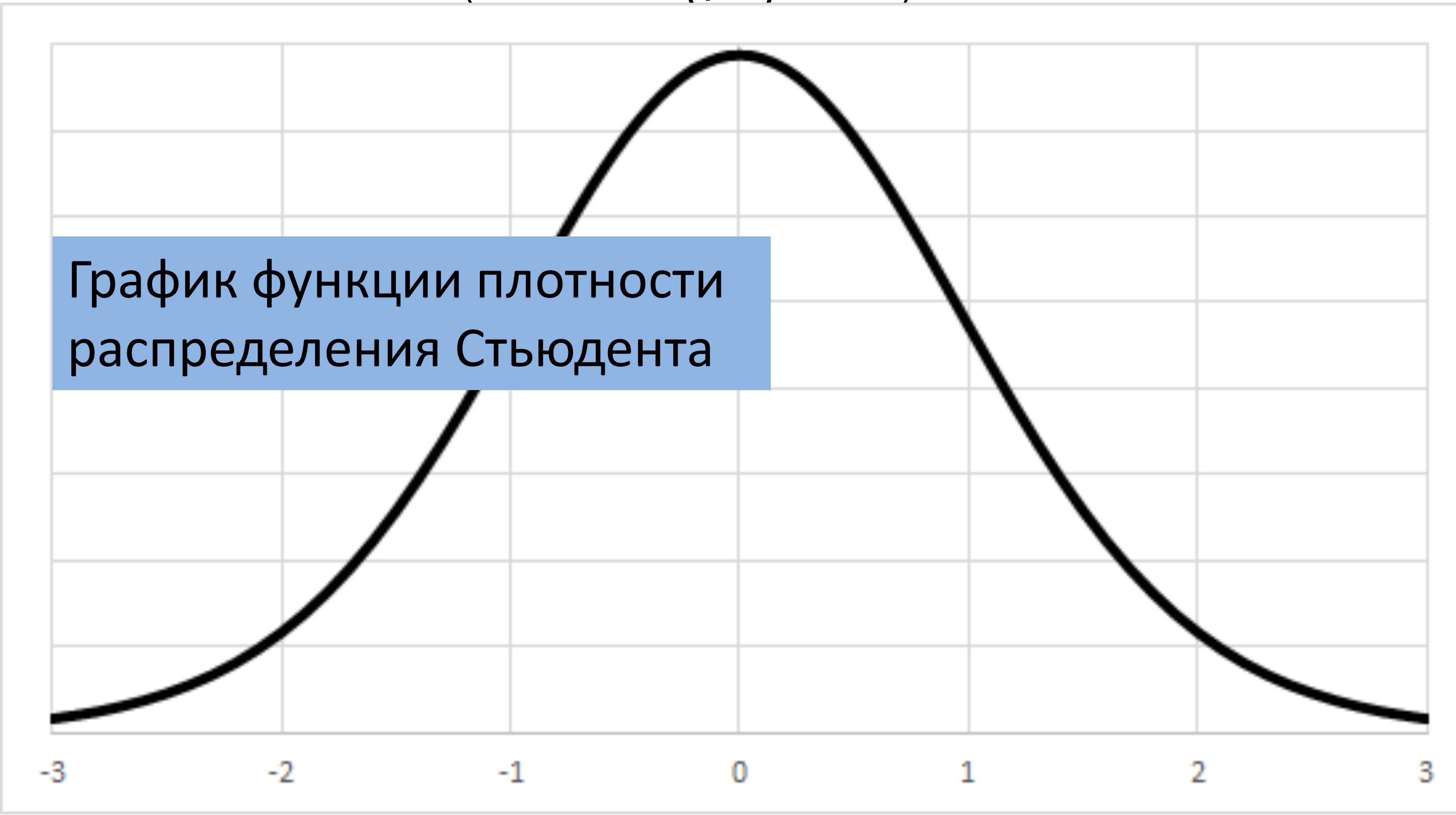
$$\frac{\widehat{\beta_2}-\beta_2}{se\left(\widehat{\beta_2}\right)}\sim t_{n-2}$$

Построим 95-процентный доверительный интервал. Назовем критическим значением $t_{\text{кр}}$ такое значение, что

$$P\left(-t_{\text{кр}}<\frac{\widehat{\beta_2}-\beta_2}{se\left(\widehat{\beta_2}\right)}< t_{\text{кр}}\right)=0{,}95$$

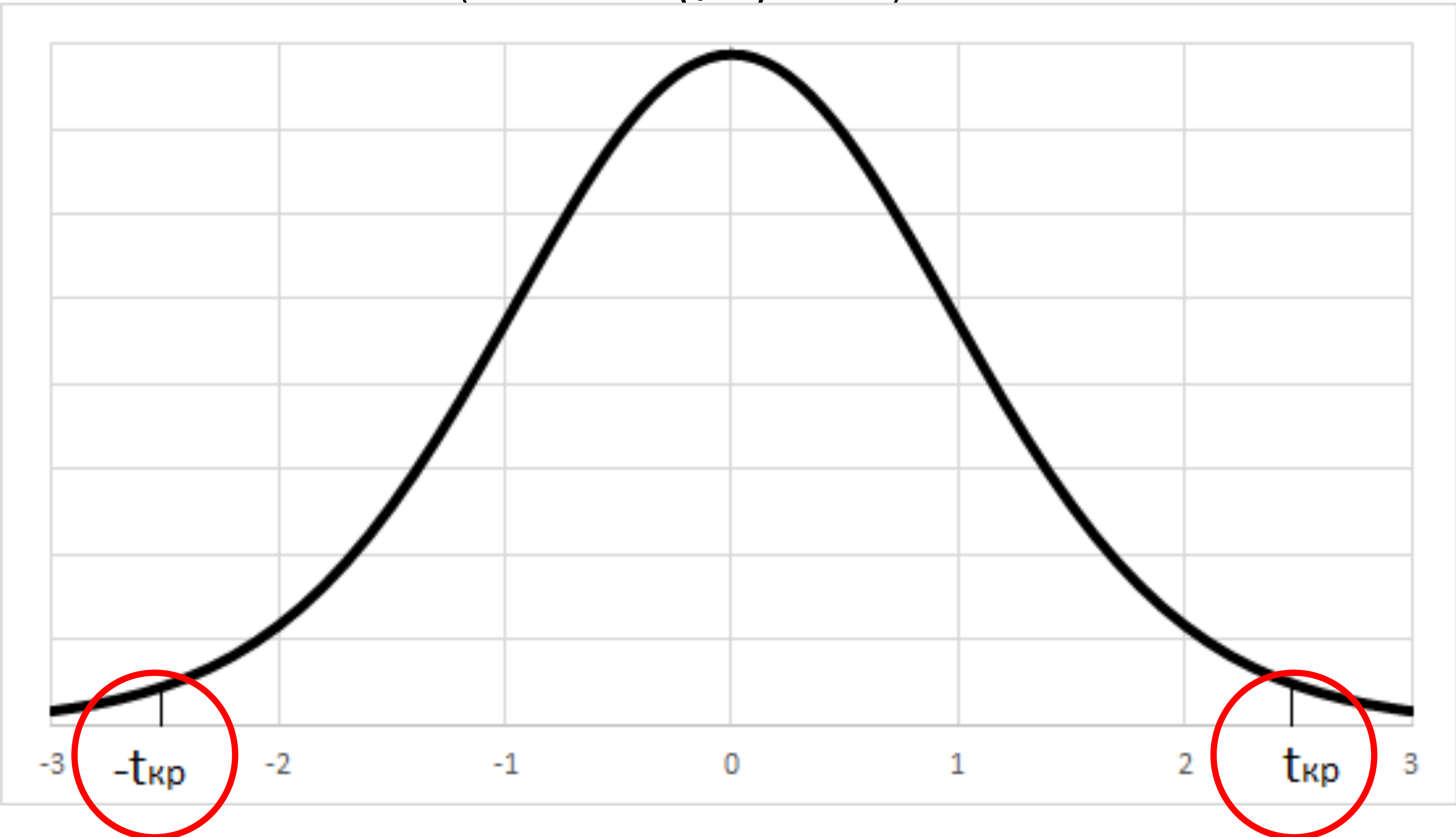
Доверительные интервалы

$$P\left(-t_{\text{кр}} < \frac{\widehat{\beta_2} - \beta_2}{se\left(\widehat{\beta_2}\right)} < t_{\text{кр}}\right) = 0{,}95$$



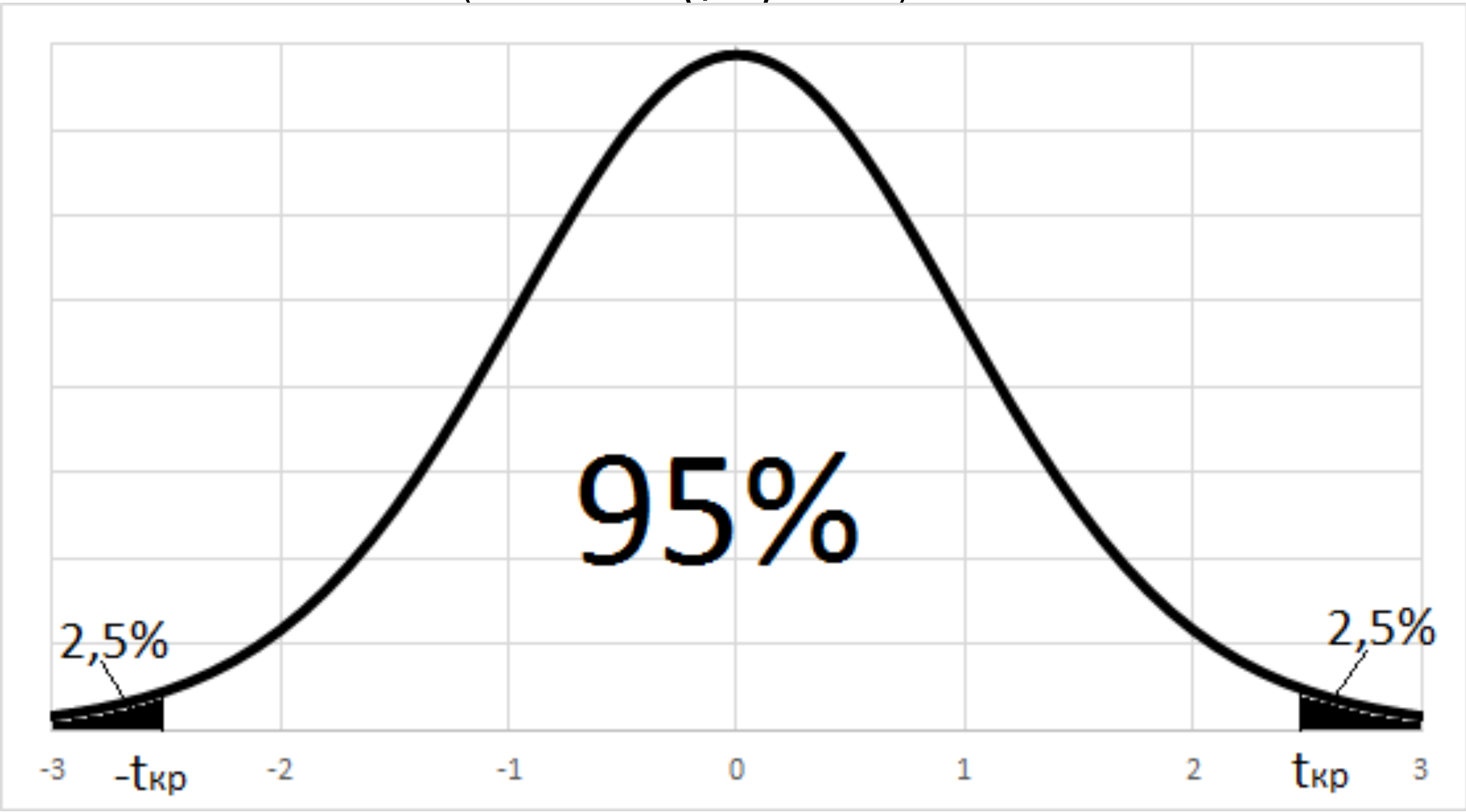
Доверительные интервалы

$$P\left(-t_{\text{кр}} < \frac{\widehat{\beta_2} - \beta_2}{se\left(\widehat{\beta_2}\right)} < t_{\text{кр}}\right) = 0{,}95$$



Доверительные интервалы

$$P\left(-t_{\text{кр}} < \frac{\widehat{\beta_2} - \beta_2}{se\left(\widehat{\beta_2}\right)} < t_{\text{кр}}\right) = 0{,}95$$



Доверительные интервалы

$$P\left(-t_{\text{кр}} < \frac{\widehat{\beta_2} - \beta_2}{se\left(\widehat{\beta_2}\right)} < t_{\text{кр}}\right) = 0{,}95$$

$$-t_{\text{кр}} < \frac{\widehat{\beta_2} - \beta_2}{se\left(\widehat{\beta_2}\right)} < t_{\text{кр}}$$

$$\widehat{\beta_2} - t_{\text{кр}} * se\left(\widehat{\beta_2}\right) < \beta_2 < \widehat{\beta_2} + t_{\text{кр}} * se\left(\widehat{\beta_2}\right)$$

$$\left(\boldsymbol{\widehat{\beta_2}} - \boldsymbol{t}_{\text{кр}} * \boldsymbol{se}\left(\boldsymbol{\widehat{\beta_2}}\right), \qquad \boldsymbol{\widehat{\beta_2}} + \boldsymbol{t}_{\text{кр}} * \boldsymbol{se}\left(\boldsymbol{\widehat{\beta_2}}\right)\right)$$

Доверительные интервалы

$$\left(\widehat{\boldsymbol{\beta}_2} - \boldsymbol{t}_{\text{кр}} * \boldsymbol{se}\left(\widehat{\boldsymbol{\beta}_2}\right), \qquad \widehat{\boldsymbol{\beta}_2} + \boldsymbol{t}_{\text{кр}} * \boldsymbol{se}\left(\widehat{\boldsymbol{\beta}_2}\right)\right)$$

Значение $t_{\text{кр}}$ вы можете получить из таблиц распределения Стьюдента или при помощи компьютера.

Например, в Excel, чтобы получить критическое значение для 95-процентного доверительного интервала для коэффициента оцененного по n=22 наблюдениям, нужно ввести:

=СТЬЮДЕНТ.ОБР(1-0,025;22-2)

Доверительные интервалы: пример

Исследуется зависимость часового заработка работника (EARNINGS) от числа законченных лет обучения (S):

$$\boldsymbol{EARNINGS = \beta_1 + \beta_2 * S_i + \varepsilon_i}$$

На основе данных о 540 работника получено следующее уравнение (в скобках — стандартные ошибки оценок коэффициентов):

$$\boldsymbol{\widehat{EARNINGS}_i = \underset{(0{,}3)}{-13{,}9} + \underset{(0{,}2)}{2{,}4} * S_i}$$

Построим 95-процентный доверительный интервал для коэффициента β_2

Доверительные интервалы: пример

$$EARNINGS = \beta_1 + \beta_2 * S_i + \varepsilon_i$$

$$\widehat{EARNINGS_i} = \underset{(0,3)}{-13{,}9} + \underset{(0,2)}{2{,}4} * S_i$$

$$n = 540, \widehat{\beta_2} = 2{,}4, se\left(\widehat{\beta_2}\right) = 0{,}2$$

$$t_{\text{кр}} = t_{n-2} = t_{538} = 1{,}96$$

$$\left(\widehat{\beta_2} - t_{\text{кр}} * se\left(\widehat{\beta_2}\right), \qquad \widehat{\beta_2} + t_{\text{кр}} * se\left(\widehat{\beta_2}\right)\right)$$

$$(2{,}4 - 1{,}96 * 0{,}2, \qquad 2{,}4 + 1{,}96 * 0{,}2)$$

$$(\mathbf{2{,}0}, \qquad \mathbf{2{,}8})$$

Доверительные интервалы: пример

Мы получили 95-процентный доверительный интервал для коэффициента β_2: $(\mathbf{2,0},\ \mathbf{2,8})$
С вероятностью 95% дополнительный год обучения увеличивает заработок работника на сумму от 2,0 до 2,8 доллара.

Все числа, которые входят в доверительный интервал, являются положительными (интервал не содержит ноль). То есть мы с высокой долей уверенности можем утверждать, что истинное значение $\beta_2 > 0$.
В этом случае говорят, что коэффициент является значимым. Можно тестировать значимость коэффициента и другим способом, без построения доверительного интервала.

Тестирование значимости коэффициента

$\hat{\beta}_1$ и $\hat{\beta}_2$ — оценки, полученные при помощи МНК, на основе случайной выборки. Следовательно, они сами являются случайными величинами.

Поэтому даже, если истинное значение коэффициента β_2 равно нулю, его оценка $\hat{\beta}_2$ может отклоняться от нуля.

Нужно уметь определять, достаточно ли сильно $\hat{\beta}_2$ отличается от нуля для того, чтобы можно было с уверенностью утверждать, что и истинное значение коэффициента также не равно нулю.

Тестирование значимости коэффициента

Нужно уметь определять, достаточно ли сильно $\hat{\beta}_2$ отличается от нуля для того, чтобы можно было с уверенностью утверждать, что и истинное значение коэффициента также не равно нулю.

На практике для решения этой задачи используется тест на значимость коэффициента.

Тестирование значимости коэффициента

Рассматриваемая модель $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

Тестируемая гипотеза

Н0: $\beta_2 = 0$ «Переменная ***x*** не оказывает значимого влияния на переменную ***y***»

Альтернативная гипотеза

Н1: $\beta_2 \neq 0$ «Переменная ***x*** оказывает значимое влияние на переменную ***y***»

Тестирование значимости коэффициента

Алгоритм проведения теста

Шаг 1
Вычисляем расчетное значение t-статистики

$$t_{расч} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)}$$

Шаг 2
Выбираем **уровень значимости**
Уровень значимости — **вероятность ошибки** первого рода, то есть вероятность отклонить гипотезу H0, если на самом деле гипотеза H0 верна.

В эконометрике обычно используется уровень значимости $\alpha = 0,01 = 1\%$ или $\alpha = 0,05 = 5\%$.

Тестирование значимости коэффициента

Шаг 3

Из таблиц t-распределения Стьюдента находим критическое значение t-статистики $t_{кр}$

Оно зависит от уровня значимости (двусторонний тест) α и от так называемого числа степеней свободы, которое в случае нашего теста равно $(n-2)$

Тестирование значимости коэффициента

Шаг 4

Сравниваем расчетное и критическое значение t-статистик

Если $\left| t_{расч} \right| < t_{кр}$,

то гипотеза H0 не отклоняется (принимается),

то есть мы делаем вывод о том, что переменная ***x*** не оказывает значимого влияния на переменную ***y***. В этом случае коэффициент при переменной ***x*** называют незначимым.

В противном случае гипотеза H0 не принимается (отклоняется).

Тестирование значимости коэффициента: пример

Исследуется зависимость часового заработка (в $) работника (EARNINGS) от числа законченных лет обучения (S):

$$EARNINGS = \beta_1 + \beta_2 S_i + \varepsilon_i$$

На основе данных о 540 работниках было получено следующее уравнение регрессии (в скобках — стандартные ошибки оценок коэффициентов):

$$EARNINGS = \underset{(3,2)}{-13,9} + \underset{(0,2)}{2,4}\, S_i$$

Можно ли утверждать, что число лет обучения значимо влияет на заработок?

Тестирование значимости коэффициента: пример

H0: $\beta_2 = 0$ «Переменная ***S*** не оказывает значимого влияния на переменную ***EARNINGS***»

$$t_{расч} = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{2,4}{0,2} = 12$$

При уровне значимости 1% и числе степеней свободы $n-2=540-2=538$

$$t_{кр} = t(538) = 2,6$$

$$\left| t_{расч} \right| > t_{кр} \, ,$$

следовательно, гипотеза H0 отвергается, и мы делаем вывод о том, что число лет обучения значимо влияет на заработок

Тестирование гипотезы $\beta_2 = A$

Аналогичным образом можно тестировать гипотезу $H_0\colon\ \beta_2 = A$

В этом случае поменяется только формула для расчетного значения тестовой статистики:

$$t_{\text{расч}} = \frac{\hat{\beta}_2 - A}{se(\hat{\beta}_2)}$$

Остальной алгоритм тестирования гипотезы сохранится без изменений

P-значение

Эконометрические пакеты при тестировании значимости обычно рассчитывают так называемое P-значение (также обозначается P-value или просто Probability).

P-значение можно определить как предельный уровень значимости, при котором тест находится на грани между отвержением и не отвержением нулевой гипотезы.

Поясним это определение на примере

Р-значение

Пусть число наблюдений $n = 10$, оценка коэффициента $\widehat{\beta_2} = 8{,}0$,
а ее стандартная ошибка $se\left(\widehat{\beta_2}\right) = 4{,}0$

Тогда критическое значение тестовой статистики для проверки гипотезы $H_0\colon\ \beta_2 = 0$ при уровне значимости 5% равно:

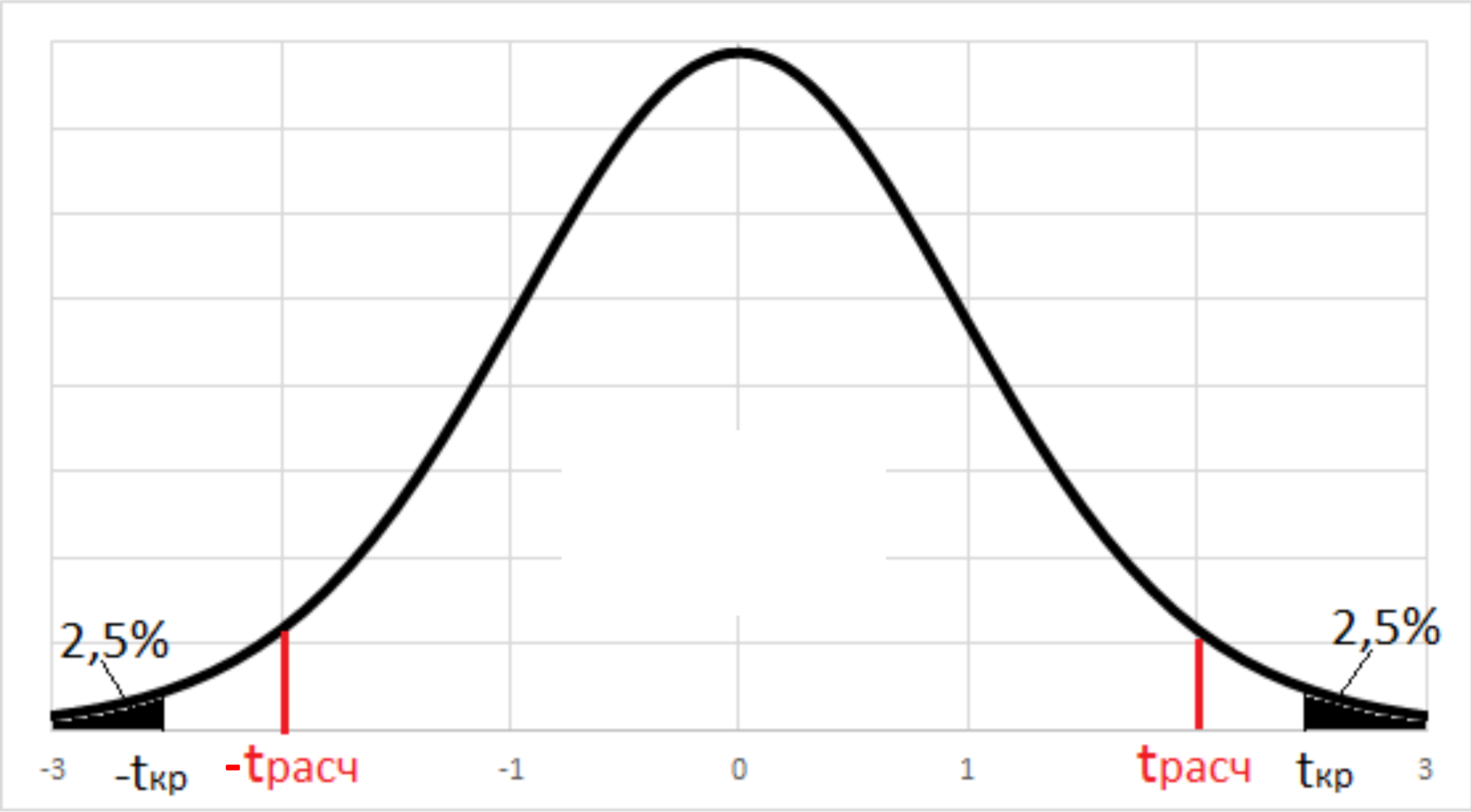
$$t_{\text{кр}} = t_8 = 2{,}3.$$

А расчетное значение статистики равно:

$$t_{\text{расч}} = \frac{\widehat{\beta_2}}{se(\widehat{\beta_2})} = \frac{8}{4} = 2$$

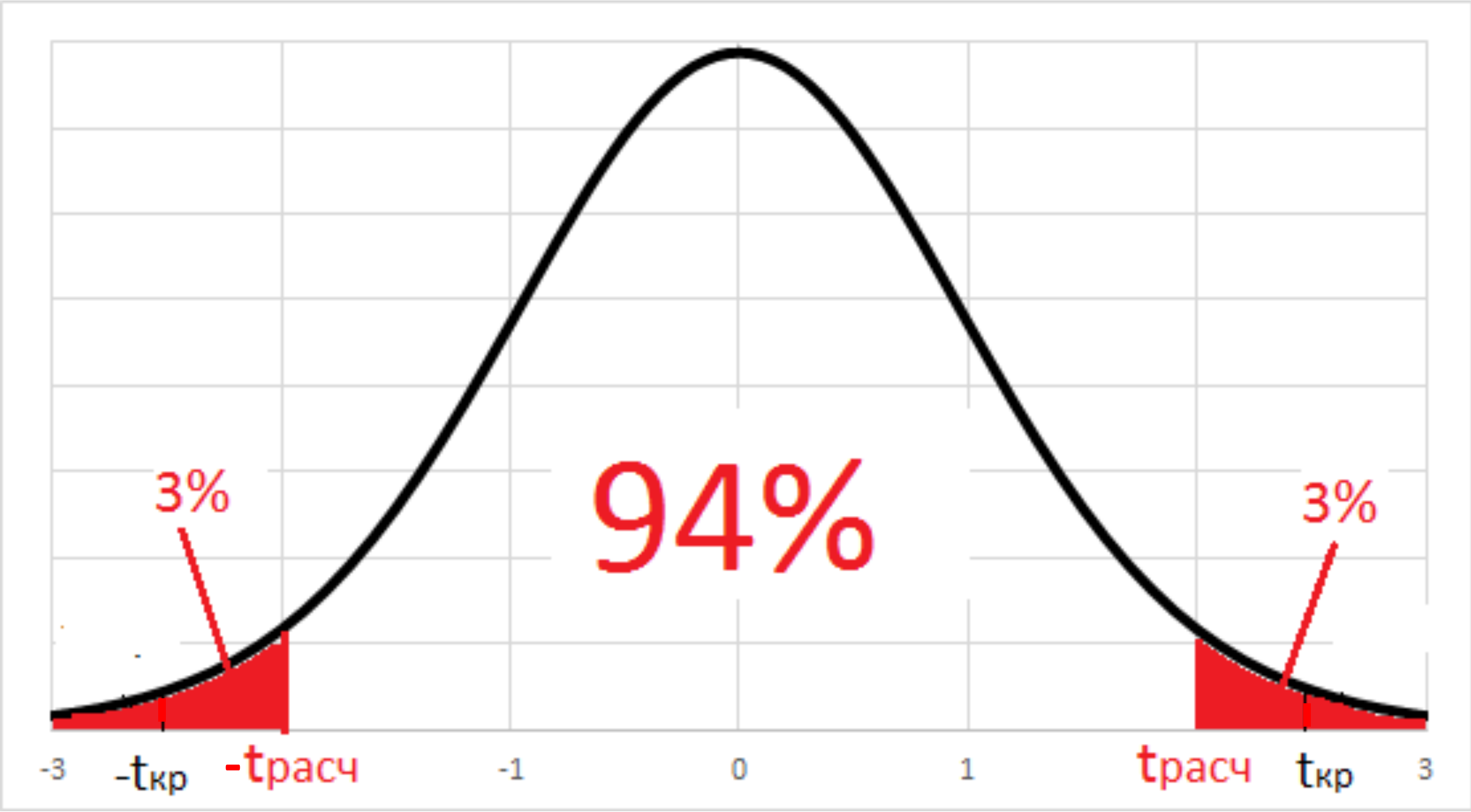
Изобразим все это на графике

P-значение



P-значение

P-значение = 3%+3% = 0,06



P-значение

Из рисунка выше следует, что P-значение больше уровня значимости тогда и только тогда, когда

$$|t_{\text{расч}}| < t_{\text{кр}}$$

Поэтому принимать решение на основе P-значения очень легко:

если P-значение больше выбранного уровня значимости, то нулевая гипотеза при данном уровне значимости не отвергается

Эконометрические пакеты рассчитывают P-значение автоматически

Тестирование значимости коэффициента: пример использования P-значения

В эконометрических программах результаты оценки уравнения представляются в виде таблицы.
Например, уравнение

$$\underset{(3,2)}{EARNINGS = -13,9} + \underset{(0,2)}{2,4}\, S_i$$

при оценке в MS Excel было представлено следующим образом:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	–13.9334	3.2198	–4.33	0.0002
S	2.4553	0.2318	10.59	0.0001

Отметим, что для коэффициента при переменной S P-value=0.0001

Так как P-value<0.01, то мы делаем вывод о том, что при уровне значимости 1% переменная S является значимой.

Прогнозирование в модели парной регрессии

Прогнозирование

Для временных рядов прогнозирование — предсказание будущего значения зависимой переменной

Например, курс доллара завтра или уровень ВВП в следующем квартале

Прогнозирование

Для пространственных выборок прогнозирование — предсказание значения зависимой переменной для заданных значений объясняющих переменных

Например, рыночная стоимость квартиры с определенными жилой площадью и количеством комнат, в определенном районе

Прогнозирование: постановка задачи

Модель: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

На основе *n* наблюдений оценено уравнение регрессии: $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$

Пусть известно (n+1)-ое значение регрессора: x_{n+1}

Используя его нужно предсказать y_{n+1}

Прогнозирование: постановка задачи

Естественная идея — просто подставить это значение в уравнение регрессии:

$$\hat{y}_{n+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{n+1}$$

Это хорошая идея, так как такой прогноз

(а) несмещенный

(б) эффективный, то есть обладает наименьшей средней квадратичной ошибкой прогноза среди всех линейных несмещенных оценок

Несмещенность

Матожидание прогноза:

$$E(\hat{y}_{n+1}) = E\left(\hat{\beta}_1 + \hat{\beta}_2 x_{n+1}\right) =$$

$$= E(\hat{\beta}_1) + E(\hat{\beta}_2) x_{n+1} = \beta_1 + \beta_2 x_{n+1}$$

Матожидание истинного значения:

$$\mathrm{E}(y_{n+1}) = E(\beta_1 + \beta_2 x_{n+1} + \varepsilon_{n+1}) =$$

$$= \beta_1 + \beta_2 x_{n+1} + E(\varepsilon_{n+1}) = \beta_1 + \beta_2 x_{n+1}$$

$$\mathrm{E}(y_{n+1}) = E(\hat{y}_{n+1})$$

Точность прогноза

Дисперсия ошибки прогноза

$$E(\hat{y}_{n+1} - y_{n+1})^2 = V(e_{n+1}) =$$

$$= \sigma^2 \left(1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \right)$$

(Докажите это равенство)

Стандартная ошибка прогноза

Заменим σ^2 на ее оценку S^2 и вычислим корень из дисперсии ошибки прогноза — получим **стандартную ошибку прогноза**

$$\delta = \sqrt{S^2\left(1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)}$$

Доверительный интервал для прогноза

$$(\widehat{y_{n+1}} - \delta * t_{n-2}, \qquad \widehat{y_{n+1}} + \delta * t_{n-2})$$

Примечание:

В случае, если вы строите прогноз не для парной регрессии, а для множественной, число степеней свободы изменится, и вместо t_{n-2}, следует использовать t_{n-k}, где k — число оцениваемых параметров.

Формула для δ в случае множественной регрессии также отличается, ее можно посмотреть в Магнусе, Катышевом, Пересецком.

Точность прогноза

$$\delta = \sqrt{S^2\left(1 + \frac{1}{n} + \frac{\left(x_{n+1} - \bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)}$$

При каком x_{n+1} ошибка прогноза минимальна?

При $x_{n+1} = \bar{x}$.

=> Прогноз наиболее точен, для наблюдений «похожих» на наблюдения из исходной выборки

Следует помнить

Все сказанное выше про прогнозирование верно, при условии, что выполнены предпосылки классической линейной модели парной регрессии.

На практике эти предпосылки часто нарушаются. Как тестировать выполнение этих предпосылок, и что делать, если они не выполняются, мы обсудим в дальнейшем.

Множественная регрессия

Мотивация:
зачем нужна множественная регрессия?

Ответ на этот вопрос зависит от цели вашего исследования:

- Если ваша цель состоит в прогнозировании значения зависимой переменной, то учет большего числа факторов может позволить увеличить точность прогноза
- Если же ваша цель состоит в проверке наличия причинно-следственной связи, то в некоторых случаях множественная регрессия позволит избежать ложных выводов
    - — Пояснение — на следующих слайдах

Проблема смещения из-за пропуска существенной переменной

Пусть нас интересует, влияет ли переменная $\boldsymbol{x}$ на переменную $\boldsymbol{y}$

Пусть также в действительности на зависимую переменную влияет еще один фактор:

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i, \qquad \beta_3 > 0$$

Покажем, что если игнорировать этот факт и по-прежнему оценивать парную регрессию вместо множественной, то это может привести к получению смещенных оценок (и, следовательно, ошибочных выводов)

Проблема смещения из-за пропуска существенной переменной

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i, \qquad \beta_3 > 0$$

$$\widehat{\beta_2} = \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} = \frac{\widehat{cov}(x, \beta_1 + \beta_2 * x + \beta_3 * w + \varepsilon)}{\widehat{var}(x)} =$$

Проблема смещения из-за пропуска существенной переменной

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i, \qquad \beta_3 > 0$$

$$\widehat{\beta_2} = \frac{\widehat{cov}(x, y)}{\widehat{var}(x)} = \frac{\widehat{cov}(x, \beta_1 + \beta_2 * x + \beta_3 * w + \varepsilon)}{\widehat{var}(x)} =$$

$$= \frac{\beta_2 * \widehat{cov}(x, x) + \beta_3 * \widehat{cov}(x, w) + \widehat{cov}(x, \varepsilon)}{\widehat{var}(x)} =$$

$$= \beta_2 + \beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)} + \frac{\widehat{cov}(x, \varepsilon)}{\widehat{var}(x)}$$

Проблема смещения из-за пропуска существенной переменной

$$\widehat{\beta_2} = \beta_2 + \beta_3 \frac{\widehat{cov}(x,w)}{\widehat{var}(x)} + \frac{\widehat{cov}(x,\varepsilon)}{\widehat{var}(x)}$$

$$\widehat{cov}(x,\varepsilon) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})$$

$E\left(\widehat{cov}(x,\varepsilon)\right) = 0$, следовательно:

$$E\left(\widehat{\beta_2}\right) = \beta_2 + \beta_3 \frac{\widehat{cov}(x,w)}{\widehat{var}(x)}$$

Проблема смещения из-за пропуска существенной переменной

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i, \qquad \beta_3 > 0$$

$$E\left(\widehat{\beta_2}\right) = \beta_2 + \beta_3 \frac{\widehat{cov}(x, w)}{\widehat{var}(x)}$$

- Если $\beta_3 > 0$ и $\widehat{cov}(x, w) > 0$, то $E\left(\widehat{\beta_2}\right) > \beta_2$, то есть МНК-оценка коэффициента β_2 в парной регрессии будет **смещена** и завышена.
- **Вывод №1:** пропуск существенного фактора приводит к смещению оценок коэффициентов (***omitted variable bias***). Поэтому даже если нас в нашем исследовании интересует только эффект от переменной x, а переменная w нам не интересна, её всё равно придется включить в модель

Проблема смещения из-за пропуска существенной переменной

$$y_i = \beta_1 + \beta_2 * x_i + \beta_3 * w_i + \varepsilon_i, \qquad \beta_3 > 0$$

$$E\left(\widehat{\beta_2}\right) = \beta_2 + \beta_3 \frac{\widehat{cov}(x,w)}{\widehat{var}(x)}$$

- Если же переменная x и пропущенная переменная w **не коррелированы** $\widehat{cov}(x,w) = 0$, то МНК-оценка коэффициента в парной регрессии по-прежнему будет **несмещенной**.
- **Вывод №2:** нельзя упускать только те важные факторы, которые коррелированы с интересующей нас переменной

Классическая линейная модель множественной регрессии (КЛММР)

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \cdots + \beta_k * x_i^{(k)} + \varepsilon_i$$
$$i = 1, \dots, n$$

y_i — зависимая (объясняемая) переменная

$x_i^{(m)}$ — объясняющие переменные (регрессоры)

ε_i — случайные ошибки

k — число коэффициентов в модели

n — число наблюдений

Предпосылки КЛММР

1. Модель $y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \cdots + \beta_k * x_i^{(k)} + \varepsilon_i$ корректно специфицирована и линейна по параметрам
2. Объясняющие переменные $x_i^{(m)}$ являются детерминированными и линейно независимыми
3. Математическое ожидание случайных ошибок равно нулю $E(\varepsilon_i) = 0$
4. Случайные ошибки имеют постоянную дисперсию $var(\varepsilon_i) = \sigma^2$
5. Случайные ошибки, относящиеся к разным наблюдениям, не коррелированы $cov(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$

Теорема Гаусса — Маркова

Если выполнены предпосылки 1-5, то оценки коэффициентов модели, полученные при помощи МНК будут

(а) несмещенными,

(б) эффективными в классе всех несмещенных и линейных по y оценок

Предпосылки КЛММР

6.* Случайные ошибки модели имеют нормальное распределение

Шестая предпосылка КЛММР не требуется для теоремы Гаусса — Маркова, однако будет полезна для тестирования гипотез и построения доверительных интервалов

Предпосылки КЛММР

— *Выполняются ли все предпосылки КЛММР на практике?*

— Как правило, нет.

— *Зачем тогда изучать эту модель?*

— Это самый простой случай, на примере которого удобно обсудить некоторые важные идеи. Позже в рамках нашего курса мы откажемся от предпосылок этой модели и будем рассматривать гораздо более реалистичные наборы предпосылок.

Качество подгонки модели

1. Стандартная ошибка регрессии
2. Коэффициент детерминации R^2
3. Скорректированный (нормированный) коэффициент детерминации R^2

Стандартная ошибка регрессии

В реальной ситуации величина дисперсии случайной ошибки σ^2 нам не известна. Вместо нее самой можно вычислить ее оценку.

Если выполнены предпосылки 1-5, то несмещенная **оценка дисперсии случайной ошибки** будет иметь вид:

$$\widehat{\sigma^2} = S^2 = \frac{1}{n - \boldsymbol{k}} * \sum_{i=1}^{n} e_i^2$$

Стандартная ошибка регрессии

Стандартная ошибка регрессии,
standard error of estimate, SEE (не путать с ESS):

$$SEE = \sqrt{S^2} = \sqrt{\frac{1}{n-k} * \sum_{i=1}^{n} e_i^2}$$

- Мера точности модели. Чем меньше стандартная ошибка регрессии, тем лучше модель соответствует данным.
- При помощи SEE можно сравнивать между собой модели с одинаковой зависимой переменной, но разным набором регрессоров

Коэффициент детерминации R^2

Все аналогично случаю парной регрессии:

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

$TSS=\sum_{i=1}^{n}(y_i-\bar{y})^2$ общая сумма квадратов (total sum of squares)

$ESS=\sum_{i=1}^{n}e_i^2$ сумма квадратов остатков (error sum of squares)

$RSS=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ объясненная сумма квадратов (regression sum of squares)

Напоминаем, что в некоторых учебниках обозначения сильно отличаются

Коэффициент детерминации R^2

Коэффициент детерминации R^2 показывает долю дисперсии зависимой переменной, «объясненной» уравнением регрессии

$$R^2 = \frac{\widehat{Var}(\hat{y})}{\widehat{Var}(y)} = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

$$0 \leq R^2 \leq 1$$

Чем лучше модель соответствует данным, тем ближе R^2 к единице

Некоторые предостережения по поводу R^2

- При добавлении в модель новых переменных R^2 не может уменьшаться, поэтому сравнивать при помощи этого показателя модели с разным числом переменных некорректно. Для этого лучше использовать скорректированный R^2 (следующий слайд)
- Равенство TSS=RSS+ESS, вообще говоря, верно только в случае, если в оцениваемой модели есть константа (свободный член). В противном случае оно может нарушаться. При этом R^2 теряет свою стандартную интерпретацию и даже не обязательно лежит на отрезке [0, 1]
- Как и в случае парной регрессии, высокий R^2 говорит о хорошей подгонке модели, но ничего не говорит о наличии или отсутствии причинно-следственной связи между переменными

Скорректированный (нормированный) R^2

R^2 с учетом штрафа за большое количество переменных

$$R_{adj}^2 = R^2 - \frac{k-1}{n-k} * (1 - R^2)$$

Если вам нужно сравнить между собой модели с одинаковой зависимой переменной, но разным числом регрессоров, то для этого лучше использовать не R^2, а R_{adj}^2

Гипотезы и доверительные интервалы

1. Тестирование незначимости коэффициента
2. Тестирование гипотезы $\beta_j = A$
3. Доверительный интервал для коэффициента
4. Тестирование незначимости уравнения
5. Сравнение «короткой» и «длинной» регрессий

Тестирование незначимости коэффициента

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \cdots + \beta_k * x_i^{(k)} + \varepsilon_i$$

Тестируемая гипотеза $H_0 : \beta_j = 0$
«Переменная $x^{(j)}$ не влияет на переменную y»

Альтернативная гипотеза $H_1 : \beta_j \neq 0$
«Переменная $x^{(j)}$ влияет на переменную y»

Расчетное значение тестовой статистки $t_{\text{расч}} = \dfrac{\widehat{\beta_j}}{se(\widehat{\beta_j})}$

Тестирование незначимости коэффициента

1. Вычисляем расчетное значение тестовой статистки:

$$t_{\text{расч}} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

2. Находим критическое значение $t_{\text{кр}}$ из таблиц распределения Стьюдента для $(n-k)$ степеней свободы и выбранного уровня значимости α (чаще всего $\alpha = 1\%$ или 5%)

3. Сравниваем и делаем вывод:

- Если $|t_{\text{расч}}| < t_{\text{кр}}$, то гипотеза H_0 не отвергается, то есть мы делаем вывод о том, что $\boldsymbol{x}^{(j)}$ не влияет на $\boldsymbol{y}$. В этом случае говорят, что переменная $\boldsymbol{x}^{(j)}$ является незначимой при уровне значимости α
- В противном случае гипотеза H_0 отвергается

Доверительный интервал для коэффициента

Тестирование гипотезы H_0: $\beta_k = A$

$$t_{\text{расч}} = \frac{\hat{\beta}_k - A}{se(\hat{\beta}_k)}$$

Доверительный интервал:

$$\left(\hat{\beta}_k - t_{n-k} * se(\hat{\beta}_k), \qquad \hat{\beta}_k + t_{n-k} * se(\hat{\beta}_k)\right)$$

Тестирование незначимости уравнения

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \beta_3 * x_i^{(3)} + \cdots + \beta_k * x_i^{(k)} + \varepsilon_i$$

Тестируемая гипотеза $H_0: \beta_{\mathbf{2}} = \beta_3 = \cdots = \beta_k = 0$
«Ни одна из объясняющих переменных не влияет на $\boldsymbol{y}$»

Альтернативная гипотеза H_1: хотя бы один из коэффициентов $\beta_{\mathbf{2}}, \beta_3, \dots, \beta_k$ не равен нулю

Тестирование незначимости уравнения

1. Вычисляем расчетное значение тестовой статистки:

$$F_{\text{расч}} = \frac{R^2}{1 - R^2} * \frac{n - k}{k - 1} = \frac{RSS}{ESS} * \frac{n - k}{k - 1}$$

2. Находим критическое значение $F_{\text{кр}}$ из таблиц распределения Фишера для $(k - 1)$ и $(n - k)$ степеней свободы и выбранного уровня значимости α

3. Сравниваем и делаем вывод:

- Если $F_{\text{расч}} < F_{\text{кр}}$, то гипотеза H_0 не отвергается. Мы делаем вывод о том, что ни один из регрессоров не влияет на $\boldsymbol{y}$.
  В этом случае говорят, что уравнение в целом является незначимым при уровне значимости α
- В противном случае гипотеза H_0 отвергается

«Короткая» или «длинная» регрессия?

«Короткая» регрессия

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \cdots + \beta_m * x_i^{(m)} + \varepsilon_i$$

«Длинная» регрессия

$$y_i = \beta_1 + \beta_2 * x_i^{(2)} + \cdots + \beta_m * x_i^{(m)} + \\ + \beta_m * x_i^{(m+1)} + \cdots + \beta_m * x_i^{(m+q)} + \varepsilon_i$$

Тестируемая гипотеза $H_0: \beta_{m+1} = \cdots = \beta_{m+q} = 0$

«Ни одна из добавленных переменных не влияет на y»

(= «короткая» модель лучше)

q — количество добавленных переменных,

$m + q = k$ — количество коэффициентов в длинной регрессии

«Короткая» или «длинная» регрессия?

Тестируемая гипотеза $H_0\colon \beta_{m+1} = \cdots = \beta_{m+q} = 0$
«Ни одна из добавленных переменных не влияет на y»

q — количество добавленных переменных,

$m + q = k$ — количество коэффициентов в длинной регрессии

$$F_{\text{расч}} = \frac{R_{UR}^2 - R_R^2}{1 - R_{UR}^2} * \frac{n-k}{q} \ \sim \ F(q, n-k)$$

R_R^2 — это R^2 в «короткой» регрессии (R^2 restricted)
R_{UR}^2 — это R^2в «длинной» регрессии (R^2 unrestricted)

Обзор проблем, возникающих при оценке регрессии

Предпосылки классической линейной модели:

1. Модель линейна по параметрам и корректно специфицирована: $y = \mathrm{X}\beta + \varepsilon$

2. Объясняющие переменные x_{ik} детерминированы и линейно независимы

3. Математическое ожидание случайных ошибок равно нулю $\mathbb{E}(\varepsilon) = 0$

4. Случайные ошибки, относящиеся к разным наблюдениям независимы и обладают равной дисперсией

$$Var(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Нелинейность

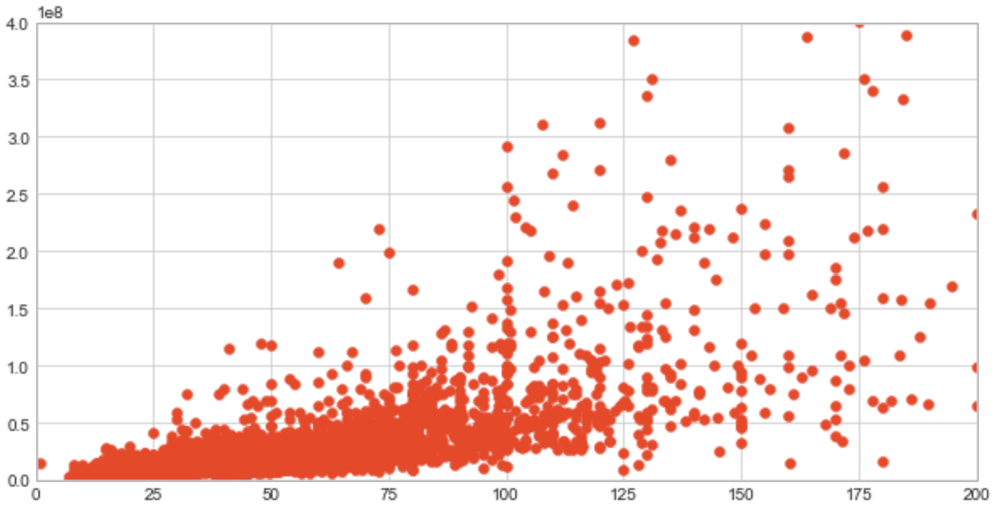
1. Модель линейна по параметрам и корректно специфицирована: $y = X\beta + \varepsilon$

В точности не выполняется никогда, все модели неверны. При сильных отклонениях от линейности оценки смещены и несостоятельны.

Решение: Графический анализ, различные тесты на спецификацию модели (тест Рамсея)

Линеаризация зависимости

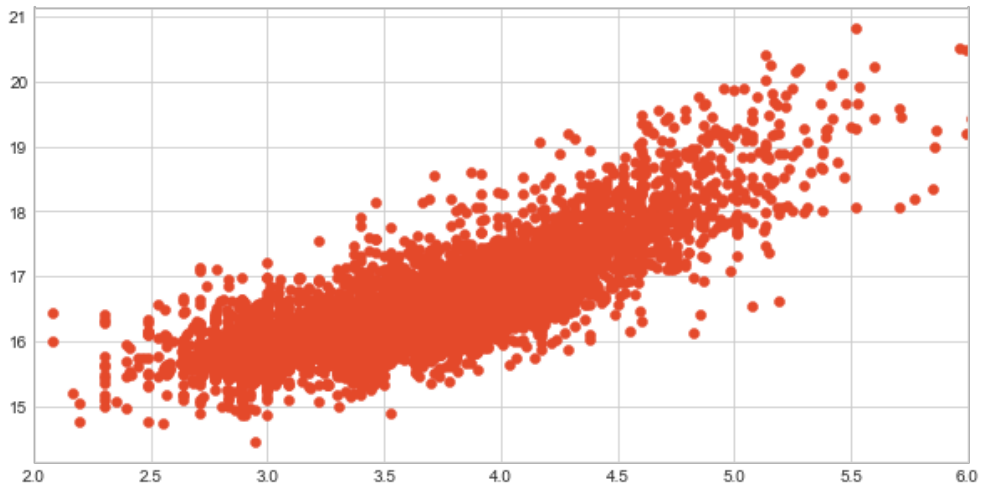
y



x

$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_t$$

$\ln(y)$



$\ln(x)$

$$\ln y_i = \beta_0 + \beta_1 \cdot \ln x_i + \varepsilon_t$$

Мультиколлинеарность

2. Объясняющие переменные x_{ik} детерминированы и линейно независимы

Если переменные зависимы, возникает проблема мультиколлинеарности, мы не можем найти МНК-оценку, так как определитель матрицы $X^T X$ оказывается близок к нулю

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

Решение: следить, чтобы среди регрессоров не было переменных, связь между которыми близка к линейной

Случайные регрессоры

2. Объясняющие переменные x_{ik} детерминированы и линейно независимы

От предпосылки, что x_{ik} детерминированы обычно отказываются и рассматривают модель со случайными регрессорами

- Доказательства теорем из-за этого становятся более сложными
- На вектор ошибок накладывается дополнительное ограничение $\mathrm{Cov}(\varepsilon_i, x_j) = 0$ либо $\mathbb{E}(\varepsilon_i \mid x_j) = 0$
- Если эта предпосылка нарушена, говорят о **проблеме эндогенности**

Эндогенность

- Если $\text{Cov}(\varepsilon_i, x_j) \neq 0$, значит среди “прочих” факторов есть такие, которые связаны с x_j

- Можно показать, что это приводит к несостоятельным и смещённым оценкам коэффициентов

- Эндогенность может возникать из-за разных причин:

  1. наблюдаемая пропущенная переменная
  2. ненаблюдаемая пропущенная переменная
  3. ошибки измерения
  4. двухсторонняя причинно-следственная связь

Эндогенность

Решение: разработка более сложных статистических процедур, которые помогут получить состоятельные несмещённые оценки (или хотя бы просто состоятельные оценки):

- Метод инструментальных переменных
- Двухшаговый МНК
- Панельные данные

Математическое ожидание ошибок

3. Математическое ожидание случайных ошибок равно нулю $\mathbb{E}(\varepsilon) = 0$

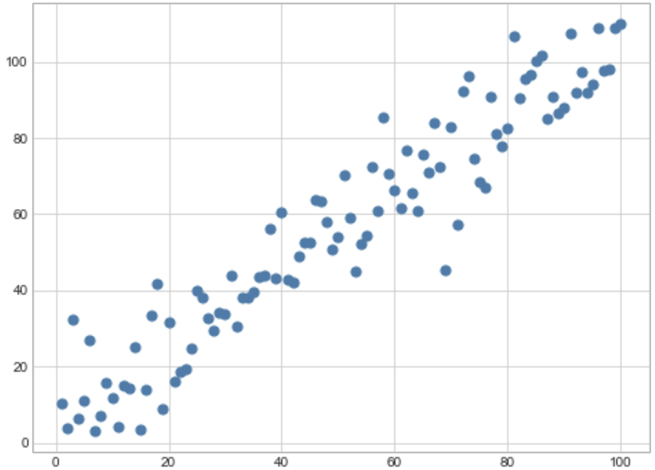
Условное математическое ожидание случайных ошибок равно нулю $\mathbb{E}\left(\varepsilon_i \mid x_j\right) = 0$

❗ Мы не включили в модель какие-то важные факторы. В условиях стохастических регрессоров мы получаем несостоятельные смещённые оценки.

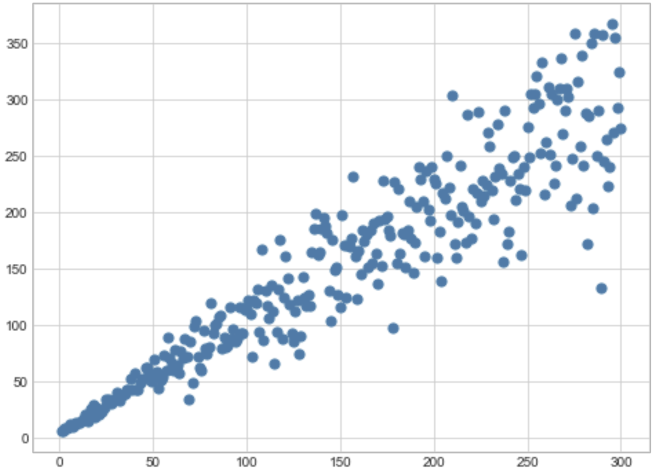
Гетероскедастичность

4. Случайные ошибки, относящиеся к разным наблюдениям независимы и обладают равной дисперсией

Гетероскедастичность



Гомоскедастичность

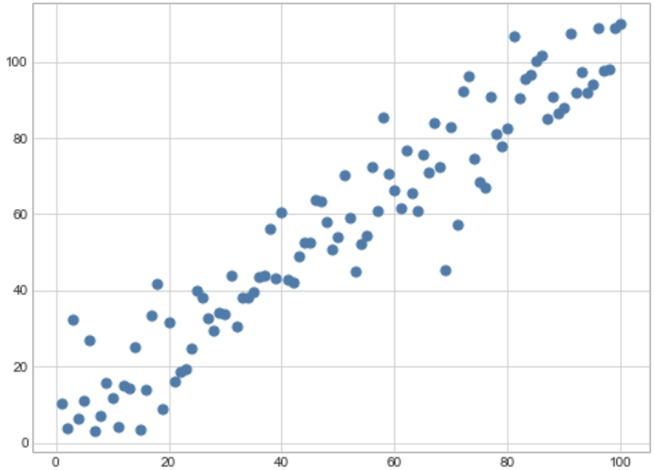


Гетероскедастичность

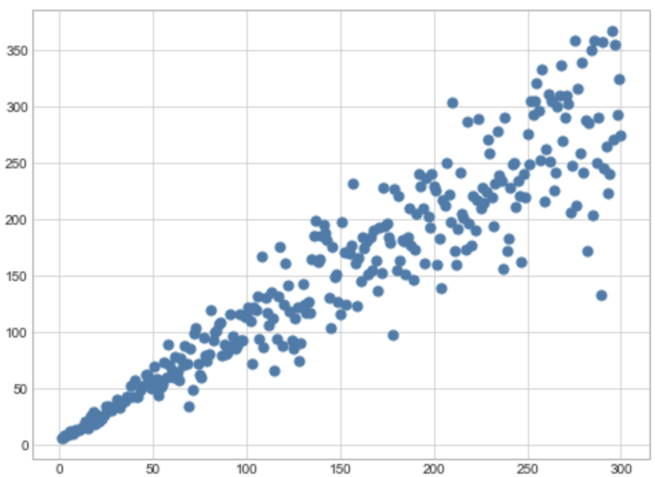
❗ Оценки коэффициентов останутся несмещёнными и состоятельными, но перестанут быть эффективными, это приведёт к искажению доверительных интервалов.

$$Var(\varepsilon) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Гетероскедастичность



Гомоскедастичность



Гетероскедастичность

Возможные решения -

*** различные процедуры коррекции оценок дисперсии**
*** обобщённый метод наименьших квадратов**

Корреляция и причинность

- Наличие значимого коэффициента в модели вовсе не означает причинно-следственной связи между переменными

- Значимый коэффициент означает, что между переменными есть корреляция

Решение: разработка более сложных статистических процедур, которые помогут выявить причинно-следственные связи, а также опора на теорию и здравый смысл