Лекция №1 (1

2016г.

«Эконометрика»

Эконометрика как наука расположена где-то между экономикой, статистикой и математикой.

это наука, связанная с эм- пирическим выводом экономических законов.

эконометрист

формулирует экономические модели, основываясь на экономической теории или на эмпирических данных,

оценивает неиз вестные величины (параметры) в этих моделях,

делает прогнозы (и оценивает их точность,

даст рекомендации по экономической политике.

Типы моделей

Модели временных рядов

Регрессионные модели с одним уравнением

Системы одновременных уравнений

Типы данных

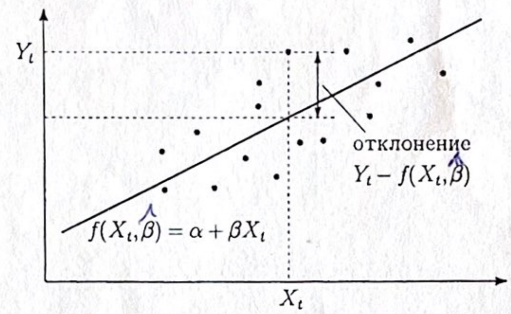
При моделировании экономических процессов используют два типа данных:

пространственные данные (*cross-sectional data*)

временные ряды (*time-series data*).

Модель парной регрессии

Пусть у нас есть набор значений двух переменных X_t, Y_t, $t = 1, \dots, n$; можно отобразить пары (X_t, Y_t) точками на плоскости X–Y (рис. 2.1).



$\hat{\beta} = (\alpha; \beta)$

Рис. 2.1

Предположим, что нашей задачей является подобрать («подогнать») функцию $Y = f(X)$ из параметрического семейства функций $f(X, \beta)$, «наилучшим» способом описывающую зависимость Y от X. т.е выбрать „наилучшее" значение $\hat{\beta}$.

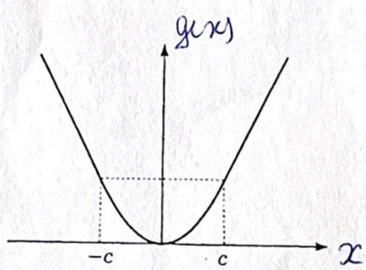
Мера отклонений.

В качестве меры отклонения функции $f(X, \beta)$ от набора наблюдений можно взять:

1) сумму квадратов отклонений $F = \sum_{t=1}^{n} (Y_t - f(X_t, \hat{\beta}))^2$,
2) сумму модулей отклонений $F = \sum_{t=1}^{n} |Y_t - f(X_t, \hat{\beta})|$, или, в общем случае,
3) $F = \sum_{t=1}^{n} g(Y_t - f(X_t, \beta))$, где g — «мера», с которой отклонение $Y_t - f(X_t, \beta)$ входит в функционал F.

Например, функция Хубера при малых отклонениях квадратична, а при больших линейна.

$$g(x) = \begin{cases} x^2, & |x| < c, \\ 2cx - c^2, & x \geqslant c, \\ -2cx - c^2, & x \leqslant -c. \end{cases}$$



Сумма квадратов отклонений

Плюсы метода:

- легкость вычислительной процедуры;
- хорошие статистические свойства, простота математических выводов делают возможным построить развитую теорию, позволяющую провести тщательную проверку различных статистических гипотез;

минусы метода:

- чувствительность к «выбросам» (*outliers*).

Метод наименьших квадратов (МНК)

Рассмотрим задачу «наилучшей» аппроксимации набора наблюдений X_t, Y_t, $t = 1, \ldots, n$, линейной функцией $f(X) = a + bX$ в смысле минимизации функционала

$$F = \sum_{t=1}^{n} (Y_t - (a + bX_t))^2. \to \min_{a,b}$$

Запишем необходимые условия экстремума (*First Order Conditition, FOC*):

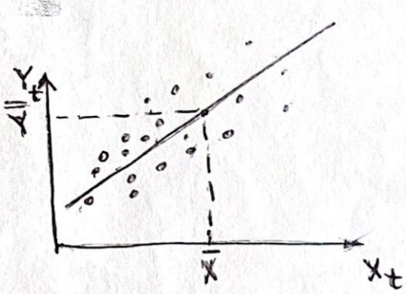
$$\frac{\partial F}{\partial a} = -2\sum_{t=1}^{n}(Y_t - a - bX_t) = 0, \qquad \frac{\partial F}{\partial b} = -2\sum_{t=1}^{n} X_t(Y_t - a - bX_t) = 0,$$

или

$$\sum_{t=1}^{n}(Y_t - a - bX_t) = 0, \qquad \sum_{t=1}^{n} X_t(Y_t - a - bX_t) = 0.$$

Раскроем скобки и получим *стандартную форму нормальных уравнений* (для краткости опустим индексы суммирования у знака суммы $\sum$):

$$an + b\sum X_t = \sum Y_t, \qquad a\sum X_t + b\sum X_t^2 = \sum X_t Y_t.$$



Решения $\widehat{a}$, $\widehat{b}$ системы (2.3) можно легко найти:

$$\widehat{b} = \frac{n\sum X_t Y_t - \left(\sum X_t\right)\left(\sum Y_t\right)}{n\sum X_t^2 - \left(\sum X_t\right)^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)},$$

$$\widehat{a} = \frac{1}{n}\sum Y_t - \frac{1}{n}\sum X_t \widehat{b} = \overline{Y} - \overline{X}\,\widehat{b}.$$

Замечание 1. Из первого уравнения системы (2.3) следует

$$\overline{Y} = \widehat{a} + \widehat{b}\,\overline{X},$$

т. е. уравнение прямой линии $Y = \widehat{a} + \widehat{b}X$, полученное в результате минимизации функционала (2.1), проходит через точку $(\overline{X}, \overline{Y})$. Здесь через $\overline{X}$ и $\overline{Y}$ обозначены выборочные средние значения переменных X_t и Y_t: $\overline{X} = (1/n)\sum X_t$, $\overline{Y} = (1/n)\sum Y_t$.

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})$$

$$Var(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Уравнения в отклонениях

Обозначим через $x_t = X_t - \overline{X}$, $y_t = Y_t - \overline{Y}$ отклонения от средних по выборке значений X_t и Y_t, $\overline{X} = (1/n)\sum X_t$, $\overline{Y} = (1/n)\sum Y_t$. (*Проверьте, что* $\overline{x} = \overline{y} = 0$.) – Д/з

Решим теперь ту же задачу: подобрать линейную функцию $f(x) = a + bx$, минимизирующую функционал

$$F = \sum_{t=1}^{n}(y_t - (a + bx_t))^2.$$

Из геометрических соображений ясно, что решением задачи будет та же прямая на плоскости (x, y), что и для исходных данных X_t, Y_t. В самом деле, в силу (2.5) переход от X, Y к отклонениям x, y означает лишь перенос начала координат в точку $(\overline{X}, \overline{Y})$. Вычисления, которые необходимо проделать для решения задачи, вполне аналогичны предыдущим (с заменой X, Y на x, y). Заменив в (2.4а), (2.4б) X_t, Y_t на x_t, y_t и учитывая, что $\overline{x} = \overline{y} = (1/n)\sum x_t = (1/n)\sum y_t = 0$, получим

$$\widehat{a} = 0, \quad \widehat{b} = \frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum (X_t - \overline{X})(Y_t - \overline{Y})}{\sum (X_t - \overline{X})^2}. \quad \text{– Д/з}$$

Таким образом, мы получили другое выражение для углового коэффициента прямой $\widehat{b}$

Матричная форма записи

Обозначим теперь через $\boldsymbol{X}$ матрицу размерности $n \times 2$

$$x = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_1 \\ \cdots & \cdots \\ 1 & X_n \end{bmatrix}, \quad y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \beta = \begin{bmatrix} a \\ b \end{bmatrix}, \quad i = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}, \quad e = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

$\hat{y} = ai + bx$, $e = y - \hat{y}$

— 2×1 матрица (вектор) коэффициентов, $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, условие (2.7) ортогональности вектора $\boldsymbol{e}$ плоскости π теперь записывается как $\boldsymbol{X}'\boldsymbol{e} = 0$, или $\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = 0$. Отсюда получаем $\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$, или

необходимых из условий экстремума

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

в предположении, конечно, что векторы $\boldsymbol{\imath}$, $\boldsymbol{x}$ линейно независимы и, следовательно, матрица $\boldsymbol{X}'\boldsymbol{X}$ обратима.

Нетрудно проверить, что (2.8) совпадает с (2.4а), (2.4б): Док-ть.

$$\widehat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} n & \sum X_t \\ \sum X_t & \sum X_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_t \\ \sum X_t Y_t \end{bmatrix} = \cdots = \begin{bmatrix} \widehat{a} \\ \widehat{b} \end{bmatrix}.$$

Отметим, что матрица $\boldsymbol{X}'\boldsymbol{X}$ невырождена, так как матрица $\boldsymbol{X}$ имеет максимальный ранг 2 $\operatorname{rank}(X^TX) = 2 \Rightarrow \exists (X^TX)^{-1}$

$x_t = X_t - \bar{X}$

$y_t = Y_t - \bar{Y}$

$\bar{x} = \bar{y} = 0$.

Док-во:

$$\bar{x} = \frac{1}{n}\sum_t (X_t - \bar{X}) = \frac{1}{n}\sum_t X_t - \frac{1}{n}\cdot n\bar{X} = 0.$$

$$\bar{y} = \frac{1}{n}\sum_t (Y_t - \bar{Y}) = \frac{1}{n}\sum_t Y_t - \frac{1}{n} n\bar{Y} = 0.$$

$$F = \sum_{t=1}^{n} (y_t - (a + b x_t))^2 \to \min_{a,b}$$

из мнк следует, что

$$\hat{a} = \bar{y} - \bar{x} b = 0.$$

$$\hat{b} = \frac{n\sum x_t y_t - (\sum x_t)(\sum y_t)}{n\sum x_t^2 - (\sum x_t)^2} = \frac{n\sum x_t y_t - (n\bar{x})(n\bar{y})}{n\sum x_t^2 - (n\bar{x})^2} =$$

$$= \frac{\sum x_t y_t}{\sum x_t^2}.$$

Линейная регрессионная модель с двумя переменными

добавим к постановке задачи некоторые статистические свойства данных.

На самом деле, для одного X мы можем наблюдать разные значения Y.

Пример 1. X — возраст индивидуума, Y — его зарплата.

Пример 2. X — доход семьи, Y — расходы на питание.

Запишем уравнение зависимости Y_t от X_t в виде

$$Y_t = a + bX_t + \varepsilon_t, \quad t = 1, \dots, n,$$

где X_t — неслучайная (детерминированная) величина, а Y_t, ε_t — случайные величины. Y_t называется объясняемой (зависимой) переменной, а X_t — объясняющей (независимой) переменной или *регрессором*. Уравнение, приведенное выше, также называется *регрессионным уравнением*.

Какова природа ошибки ε_t?

Есть две основные возможные причины случайности:

1. Наша модель является упрощением действительности и на самом деле есть еще другие параметры (пропущенные переменные, *omitted variables*), от которых зависит Y. Зарплата, например, может зависеть от уровня образования, стажа работы, пола, типа фирмы (государственная, частная) и т. п.

2. Трудности в измерении данных (присутствуют ошибки измерений). Например, данные по расходам семьи на питание составляются на основании записей участников опросов, которые, как предполагается, тщательно фиксируют свои ежедневные расходы. Разумеется, при этом возможны ошибки.

Таким образом, можно считать, что ε_t — случайная величина с некоторой функцией распределения, которой соответствует функция распределения случайной величины Y_t.

Основные гипотезы:

1. $Y_t = a + bX_t + \varepsilon_t$, $t = 1, \dots, n$, — спецификация модели.

2. X_t — детерминированная величина; вектор $(X_1, \dots, X_n)'$ не коллинеарен вектору $\imath = (1, \dots, 1)'$.

3a. $\mathrm{E}\varepsilon_t = 0$, $\mathrm{E}(\varepsilon_t^2) = \mathrm{V}(\varepsilon_t) = \sigma^2$ — не зависит от t. $(\mathrm{E}\varepsilon = 0, \mathrm{V}(\varepsilon) = \sigma^2 I_n)$

3b. $\mathrm{E}(\varepsilon_t \varepsilon_s) = 0$ при $t \neq s$, некоррелированность ошибок для разных наблюдений.

Часто добавляется условие:

3c. Ошибки ε_t, $t = 1, \dots, n$, имеют совместное нормальное распределение: $\varepsilon_t \sim N(0, \sigma^2)$.

В этом случае модель называется нормальной линейной регрессионной (*Classical Normal Linear Regression model*).

Условие $\mathrm{E}\varepsilon = 0$ означает, что $\mathrm{E}Y_t = a + bX_t$, т. е. при фиксированном X_t среднее ожидаемое значение Y_t равно $a + bX_t$.

Условие независимости дисперсии ошибки от номера наблюдения (от регрессора X_t): $\mathrm{E}(\varepsilon_t^2) = \mathrm{V}(\varepsilon_t) = \sigma^2$, $t = 1, \ldots, n$, называется *гомоскедастичностью* (*homoscedasticity*); случай, когда условие гомоскедастичности не выполняется, называется *гетероскедастичностью* (*heteroscedasticity*). На рис. 2.2а приведен пример типичной картинки для случая гомоскедастичности ошибок; на рис. 2.2б — пример данных с гетероскедастичными ошибками (возможно, что в этом примере $\mathrm{V}(\varepsilon_t) \sim X_t^2$).

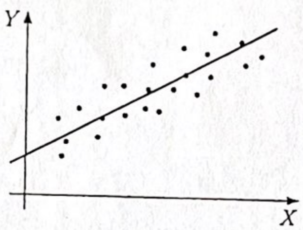


Рис. 2.2а

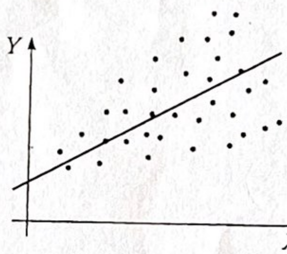
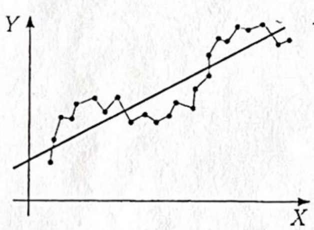
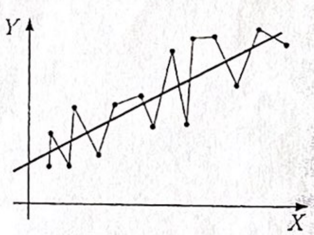


Рис. 2.2б

Условие $\mathrm{E}(\varepsilon_t \varepsilon_s) = 0$, $t \neq s$ указывает на некоррелированность ошибок для разных наблюдений. Это условие часто нарушается в случае, когда наши данные являются временными рядами. В случае, когда это условие не выполняется, говорят об *автокорреляции ошибок* (*serial correlation*).

Для простейшего случая автокорреляции ошибок, когда $\mathrm{E}(\varepsilon_t \varepsilon_{t+1}) = \rho \neq 0$, типичный вид данных представлен на рис. 2.3а ($\rho > 0$) и рис. 2.3б ($\rho < 0$).





Теорема Гаусса–Маркова. Оценка дисперсии ошибок σ^2

Итак, мы имеем набор данных (наблюдений) (X_t, Y_t), $t = 1, \ldots, n$, и модель 1–3ab. Наша задача — оценить все три параметра модели: a, b, σ^2.

Мы хотим оценить параметры a и b «наилучшим» способом. Что значит «наилучшим»? Например, найти в классе линейных (по Y_t) несмещенных оценок наилучшую в смысле минимальной дисперсии (*Best Linear Unbiased Estimator, BLUE*).

Заметим, что когда такая оценка найдена, это вовсе не означает, что не существует нелинейной несмещенной оценки с меньшей дисперсией. Кроме того, например, можно отбросить требование несмещенности оценки и минимизировать среднеквадратичное отклонение оценки от истинного значения: $\mathrm{E}(\widehat{b} - b)^2$.

Теорема Гаусса–Маркова. *В предположениях модели* 1–3ab:

1. $Y_t = a + bX_t + \varepsilon_t$, $t = 1, \ldots, n$;

2. X_t — *детерминированная величина*;

3a. $\mathrm{E}\varepsilon_t = 0$, $\mathrm{E}(\varepsilon_t^2) = \mathrm{V}(\varepsilon_t) = \sigma^2$;

3b. $\mathrm{E}(\varepsilon_t \varepsilon_s) = 0$, *при* $t \neq s$;

оценки $\widehat{a}$, $\widehat{b}$ (2.4а), (2.4б), *полученные по методу наименьших квадратов (МНК),*

$$\widehat{b} = \frac{n\sum X_t Y_t - \left(\sum X_t\right)\left(\sum Y_t\right)}{n\sum X_t^2 - \left(\sum X_t\right)^2}$$

$$\widehat{a} = \frac{1}{n}\sum Y_t - \frac{1}{n}\sum X_t \widehat{b} = \overline{Y} - \overline{X}\,\widehat{b}.$$

имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Теорема Гаусса-Маркова

Док-во: 1) несмещенность $\hat{a}$ и $\hat{b}$
2) знач. дисп. $V(\hat{a}), V(\hat{b})$ вычисл.
3) проверка, что ∃ оценок с меньшей дисп

1) Проверим несмещенность МНК-оценок $\hat{a}$ и $\hat{b}$.

$$E\hat{b} = E\,\frac{\sum x_t y_t}{\sum x_t^2} = \frac{\sum x_t E y_t}{\sum x_t^2} \overset{(*)}{=} \frac{b\sum x_t^2}{\sum x_t^2} = b.$$

$$(*)\ E y_t = E(Y_t - \bar{Y}) = E(a + bX_t + \varepsilon_t - (a + b\bar{X} + \bar{\varepsilon})) = E(b(X_t - \bar{X})) = b x_t$$

$$E\hat{a} = E(\bar{Y} - \hat{b}\bar{X}) = a + b\bar{X} - \bar{X}E(\hat{b}) = a + b\bar{X} - b\bar{X} = a$$

2) Вычислим дисперсии оценок $\hat{a}$ и $\hat{b}$.

$$\hat{b} = \frac{\sum x_t y_t}{\sum x_t^2} = \sum w_t y_t, \text{ где } w_t = \frac{x_t}{\sum_s x_s^2}$$

w_t удовлетворяют следующим условиям:

1) $\sum w_t = 0$
2) $\sum w_t x_t = \sum w_t X_t = 1$
3) $\sum w_t^2 = 1/\sum x_t^2$
4) $\sum w_t y_t = \sum w_t Y_t$

Проверим эти условия.

1) $\sum w_t = \dfrac{\sum x_t}{\sum_s x_s^2} = \dfrac{n\bar{x}}{\sum_s x_s^2} = 0 \ (\bar{x} = 0) \Rightarrow \sum w_t = 0.$

2) $\sum w_t x_t = \dfrac{\sum x_t^2}{\sum_s x_s^2} = 1 \Rightarrow \sum w_t x_t = 1.$

$$\sum w_t x_t = \sum w_t (X_t - \bar{X}) = \sum w_t X_t - \bar{X}\underbrace{\sum w_t}_{0} = \sum w_t X_t$$

$\Rightarrow \sum w_t x_t = \sum w_t X_t.$

3) $\sum w_t^2 = \sum \left[\frac{x_t}{\sum_s x_s^2} \right]^2 = \frac{\sum_t x_t^2}{\left(\sum_s x_s^2\right)^2} = \frac{1}{\sum_s x_s^2}$

4) $\sum w_t y_t = \sum w_t (Y_t - \bar{Y}) = \sum w_t Y_t - \bar{Y} \underbrace{\sum w_t}_{0} =$

$= \sum w_t Y_t \Rightarrow \sum w_t y_t = \sum w_t Y_t.$

Дисперсия оценки параметра в:

$$\boxed{V(\hat{b})} = V(\sum w_t y_t) = V(\sum w_t Y_t) = \sum w_t^2 \sigma^2 = \boxed{\frac{\sigma^2}{\sum x_t^2}}$$

Дисперсия параметра а:

$\hat{a} = \bar{Y} - \bar{X}\hat{b} = \bar{Y} - \bar{X} \sum w_t y_t = \bar{Y} - \bar{X} \sum w_t Y_t =$

$= \frac{1}{n} \sum_t Y_t - \bar{X} \sum w_t Y_t = \sum_t \left(\frac{1}{n} - \bar{X} w_t \right) Y_t$

Понадобится ещё одно тождество:

$\sum x_t^2 = \sum_t X_t^2 - n\bar{X}^2.$ $\quad (*)$

Докажем его:

$\sum x_t^2 = \sum (X_t - \bar{X})^2 = \sum X_t^2 - 2\bar{X} \sum X_t + \bar{X}^2 \cdot n =$

$= \sum X_t^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum X_t^2 - n\bar{X}^2.$

Тогда:

$\boxed{V(\hat{a})} = V\left(\sum_t \left(\frac{1}{n} - \bar{X} w_t \right) Y_t \right) = \sigma^2 \sum \left(\frac{1}{n} - \bar{X} w_t \right)^2 =$

$= \sigma^2 \sum \left(\frac{1}{n^2} - 2\frac{1}{n}\bar{X} w_t + \bar{X}^2 w_t^2 \right) = \sigma^2 \left(\frac{1}{n} - 2\frac{1}{n}\bar{X} \underbrace{\sum w_t}_{0} + \bar{X}^2 \sum w_t^2 \right) =$

$= \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum x_t^2} \right) \underset{(*)}{=} \sigma^2 \left(\frac{1}{n} + \frac{\sum X_t^2 - \sum x_t^2}{n \sum x_t^2} \right) = \boxed{\sigma^2 \frac{\sum X_t^2}{n \sum x_t^2}}$

3) Покажем, что МНК-оценки обладают наименьшей дисперсией в классе всех линейных несмещенных оценок.

Пусть $\tilde{b} = \sum c_t Y_t$ - любая другая несмещенная оценка.
Представим c_t в виде:

$$c_t = w_t + d_t.$$

Тогда в силу несмещенности $\hat{b}$ и $\tilde{b}$:

$$E(\tilde{b} - \hat{b}) = 0 = E\left(\left(\sum w_t Y_t + \sum d_t Y_t\right) - \sum w_t y_t\right) =$$

$$= E\left(\sum w_t Y_t + \sum d_t Y_t - \sum w_t Y_t\right) = E\left(\sum d_t Y_t\right) =$$

$= \sum d_t (a + bX_t)$ для всех a и b.

Тогда

$\sum d_t = 0$, $\sum d_t X_t = \sum d_t x_t = 0$ (проверить).

$$V(\tilde{b}) = V\left(\sum c_t Y_t\right) = \sigma^2 \sum c_t^2 = \sigma^2 \sum (w_t + d_t)^2 =$$

$$= \sigma^2\left(\sum w_t^2 + 2\sum w_t d_t + \sum d_t^2\right) = \sigma^2\left(\sum w_t^2 + \sum d_t^2\right) =$$

$$= V(\tilde{b}) + \sigma^2 \sum d_t^2.$$

$$\Rightarrow V(\tilde{b}) \geq V(\hat{b})$$

здесь использовали, что $\sum w_t d_t = 0$ (проверить!).

$$\sum w_t d_t = \frac{\sum x_t d_t}{\sum_s x_s^2} = \frac{0}{\sum_s x_s^2} = 0.$$

Аналогично, можно доказать, что
$V(\tilde{a}) \geq V(\hat{a})$.

Оценка дисперсии ошибок σ^2

С помощью Теоремы Гаусса - Маркова получим „наилучшие" оценки $\hat{a}$ и $\hat{b}$ коэффициентов регрессии a и b.

В регрессионном уравнении, помимо коэффициентов регрессии a, b, есть еще один параметр – дисперсия ошибок σ^2.

Пусть $\hat{Y}_t = \hat{a} + \hat{b}X_t$ - прогноз значения Y_t в точке X_t.

Остатки регрессии e_t определяются из уравнения

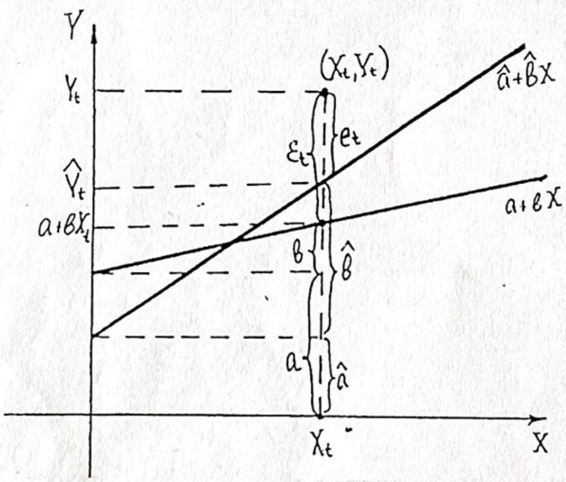
$Y_t = \hat{Y}_t + e_t = \hat{a} + \hat{b}X_t + e_t$; $e_t = Y_t - \hat{Y}_t$

Остатки регрессии e_t - случайные величины, также как и ошибки регрессии ε_t в уравнении $Y_t = a + bX_t + \varepsilon_t$.

Тем не менее, не следует их путать:

ε_t получается из уравнения истинной регрессии,

а e_t - из оценки истинной линии регрессии, т.е. e_t наблюдаемы.



Гипотеза: между оценкой σ^2 и суммой квадратов остатков регрессии $e_t = Y_t - \hat{a} - \hat{b}X_t$ существует связь.

Для её проверки надо рассмотреть $E\sum_t e_t^2$.

$E\sum_t e_t^2 = (n-2)\sigma^2$.

Таким образом,

$$\boxed{s^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum e_t^2}$$

является несмещенной оценкой дисперсии ошибок σ^2.

Дисперсии оценок $\hat{a}$, $\hat{b}$ коэффициентов регрессии, полученные при доказательстве Теоремы Гаусса-Маркова, могут быть вычислены в случае, если дисперсия ошибок σ^2 известна.

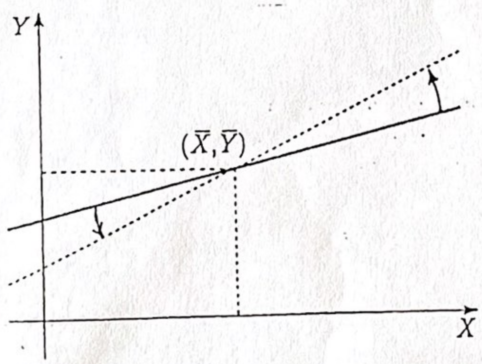
На практике дисперсия ошибок σ^2, как правило, не известна и рассчитывается одновременно с коэффициентами регрессии a, b.

В этом случае могут быть получены только оценки дисперсий $\hat{a}$, $\hat{b}$ (при замене σ^2 на s^2):

$$\boxed{\begin{aligned} \hat{V}(\hat{b}) &= s^2 \frac{1}{\sum x_i^2} = \frac{s^2}{\sum (X_i - \overline{X})^2}, \\ \hat{V}(\hat{a}) &= s^2 \frac{\sum X_i^2}{n \sum x_i^2} = \frac{s^2 \sum X_i^2}{n \cdot \sum (X_i - \overline{X})^2}, \end{aligned}}$$

В статистических пакетах используются вышеуказанные формулы для нахождения стандартных отклонений оценок коэффициентов регрессии: $s_{\hat{b}} = \sqrt{\hat{V}(\hat{b})}$, $s_{\hat{a}} = \sqrt{\hat{V}(\hat{a})}$.

Поскольку график уравнения регрессии $Y = \hat{a} + \hat{b}X$ проходит через точку $(\overline{X}, \overline{Y})$, то при увеличении $\hat{b}$ величина $\hat{a}$ уменьшается.

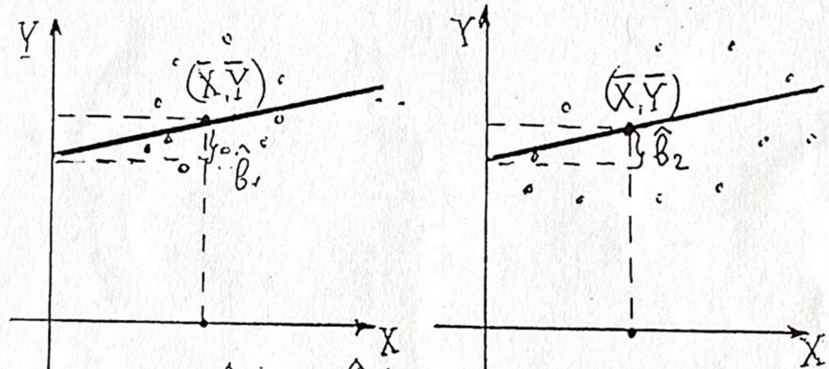


Замечание1. Предположим, что изучается зависимость Y от X и число наблюдений n задано, но можно выбирать набор значений X, т.е. $\left(X_1, X_2, \ldots, X_n\right)$.

Как выбрать X_i, чтобы точность оценки коэффициента b была наибольшей?

Из формулы для дисперсии оценки $\hat{b}$ коэффициента регрессии видим, что чем больше $\sum x_i^2 = \sum \left(X_i - \overline{X}\right)^2$, тем меньше дисперсия $V\left(\hat{b}\right)$.

Следовательно, если выбираем X_i с большим разбросом вокруг среднего значения, то получаем меньшую оценку дисперсии, т.е. большую точность оценки коэффициента регрессии b.



$$V(\hat{b}_1) > V(\hat{b}_2)$$

Семинар № 1

№ 2.1.(1) $\hat{\alpha}, \hat{\beta}$ - ?

№ 2.9.(а) построить регрессию со свободным членом.

Д/З 1) № 2.1. (1) $\hat{\sigma}^2$ - ?

2) № 2.9. (б) построить регрессию бу свободного члена

3) Доказать, что $\bar{x} = \bar{y} = 0$

$\hat{a} = 0$

$\hat{b} = \frac{\sum x_t y_t}{\sum x_t^2}$

4) Док-во теоремы Г-М.