

Лекции №2 - №4

Тема 2: Эконометрические модели. ***Модели парной регрессии***

План:

1. Функциональная, статистическая и корреляционная зависимости.
2. Парная линейная регрессия.
3. Ковариация и корреляция.
4. Проверка значимости выборочного коэффициента корреляции.

- 
5. Метод наименьших квадратов (МНК).
 6. Предпосылки МНК – условия Гаусса – Маркова.
 7. Ошибки исследования.
 8. Коэффициент детерминации линейной регрессии.
 9. Нелинейные модели регрессии и их линеаризация.

1. Функциональная, статистическая и корреляционная зависимости.

1. Функциональная зависимость - это полное соответствие между зависимой (результативной) переменной **Y** и независимыми (факторными) переменными **X_i**. Она задается в виде уравнения

$$y = f(x_i)$$

2. В экономике такая зависимость бывает редко, потому что на величины влияют случайные факторы ε_i , тогда уравнение примет вид $y = f(x_i) + \varepsilon_i$ в этом случае зависимость называется **статистической зависимостью.**

Пример.

Урожайности зерновых зависит от ряда факторов: удобрений, производительности труда, энерговооруженности сельского предприятия и т.п.

3. Корреляционная зависимость – это частный случай статистической зависимости, когда изменение средней величины результативной переменной **У** происходит в зависимости от математического ожидания изменения значения факторного признака **Х**.

Корреляция (от позднелат. correlatio – соотношение).

Для каждого значения $X=x$ определено условное математическое ожидание $M(Y/X=x)$ величины Y по X , которое равно среднему значению функции $\overline{y(x)}$ и называется регрессией величины Y по X .

$$\overline{y(x)} = M(Y | X = x)$$

Термин «корреляция» впервые применил французский палеонтолог Ж. Кювье, который вывел «закон корреляции частей и органов животных» (этот закон позволяет восстанавливать по частям тела облик всего животного).

В статистику термин ввел Френсис Гальтон (не просто связь –relation, а «как бы связь»-co-relation). Формулу для подсчета коэффициента корреляции разработал его ученик – математик и биолог – Карл Пирсон (1857-1936).

Примеры корреляционной связи

1. Закон Хика – скорость переработки информации пропорциональна логарифму от числа альтернатив.
2. Корреляция личной пластичности человека и склонности его к смене социальных установок.
3. Чем выше личностная тревожность, тем больше риск заболеть язвой желудка.
4. Чем боязливее особь, тем меньше у нее шансов занять ведущее положение в группе.

Задачи корреляционного анализа

1. Установление направления зависимости двух и более переменных (положительное – прямая связь или отрицательное – обратная связь).
2. Определение формы зависимости (линейная, нелинейная).
3. Измерение тесноты связи.
4. Проверка уровня значимости полученных коэффициентов корреляции.

2. Парная линейная регрессия

Понятие «регрессия» возникло в психоанализе (лат. regressio – движение назад, возвращение к более раннему состоянию или образу действий).

В математике это понятие впервые употребил Френсис Гальтон в 1886 г. как возврат к среднему значению. Он исследовал зависимость роста сыновей от роста их отцов (рост очень высоких и очень низких отцов ближе к среднему росту детей в регионе).

Эконометрический анализ построения модели парной регрессии

По имеющимся данным m наблюдений зависимости y от $x_1, x_2, \dots, x_m$ выбрать эконометрическую модель,

$$y = f(x_i) + \varepsilon_i$$

оценить ее параметры и статистически обосновать, что факторы существенны, и, что построенная функция наиболее точно соответствует данным наблюдений.

Задачи регрессионного анализа

1. Спецификация модели - определить вид уравнения регрессии.
2. Параметризация модели - оценить параметры уравнения.
3. Верификация модели – проверить адекватность уравнения эмпирическим данным и улучшить качество уравнения.
4. Сделать прогноз неизвестных значений зависимой переменной.

3. Ковариация и корреляция

Co-vary – совместное изменение.

Correlatio – соотношение.

Теоретической ковариацией СВ X и Y называется средняя величина отклонений этих переменных от своих средних

$$\begin{aligned}\text{Cov}(X, Y) &= M[(X - M(X)) \cdot (Y - M(Y))] = \\ &= M(XY) - M(X) \cdot M(Y)\end{aligned}$$

Теоретический коэффициент корреляции

Недостатком Cov является ее зависимость от размерности СВ X и Y. Для устранения этого вводится относительная (безразмерная) величина – теоретический коэффициент корреляции

$$\rho(X, Y) = \text{Cov}(X, Y) / \sigma(X) \cdot \sigma(Y),$$

где ρ - читается «ро»

σ (сигма) – среднее квадратическое отклонение

Выборочная ковариация

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Расчетная формула

$$\text{cov}(x, y) = \overline{xy} - \overline{x} \cdot \overline{y}$$

Выборочный коэффициент корреляции

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\left(\overline{x^2} - (\bar{x})^2\right) \cdot \left(\overline{y^2} - (\bar{y})^2\right)}}$$

Характеристика тесноты линейной связи

Если $r = 0$, то связь отсутствует.

Если $0 < r < 1$, то связь положительная,
прямая.

Если $-1 < r < 0$, то связь отрицательная,
обратная.

Если $r = +1$ или $r = -1$, то связь строгая
функциональная.

4. Проверка значимости выборочного коэффициента корреляции

Требуется проверить гипотезу H_0 о равенстве нулю истинного значения коэффициента корреляции для генеральной совокупности. Гипотезы:

$$H_0: \rho = 0$$

$H_1: \rho \neq 0$. Проверка с помощью t-критерий Стьюдента с $(n-2)$ степенями свободы и заданной доверительной вероятностью $\gamma = (0,9; 0,95; 0,98)$

$$t_{\text{наблюдаемое}} = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

Если **$|t_{\text{набл.}}| < t\text{-критерия}$** Стьюдента,
то r незначим и H_0 принимается.

Если **$|t_{\text{набл.}}| > t\text{-критерия}$** Стьюдента,
то r значим и H_0 отвергается.



Лекция № 3

Слайды 22-62

Уравнение линейной регрессии

Для генеральной совокупности
зависимость Y от X представим в виде
линейной модели первого порядка

$$\hat{Y} = \beta_0 + \beta_1 \cdot X + \varepsilon$$

Для выборочной совокупности

$$\hat{y} = b_0 + b_1 \cdot x + e$$

где $e = y - \hat{y}$ - оценка ошибки аппроксимации, отклонение, разность между выборочным и расчетным значением.

Традиционно уравнение линейной регрессии записывают в виде:

$$\hat{y} = a + b \cdot x$$

Экономический смысл параметров

$\underline{b}$ - коэффициент регрессии, показывает на сколько в среднем изменится **y** при изменении **x** на единицу.

Знак $\underline{b}$ указывает на направление связи. Если **$b > 0$** , то связь **прямая**, т.е. с увеличением **x** увеличивается **y** и наоборот.

Если **$b < 0$** , то связь **обратная**, т.е. с увеличением **x** **y** уменьшается и наоборот.

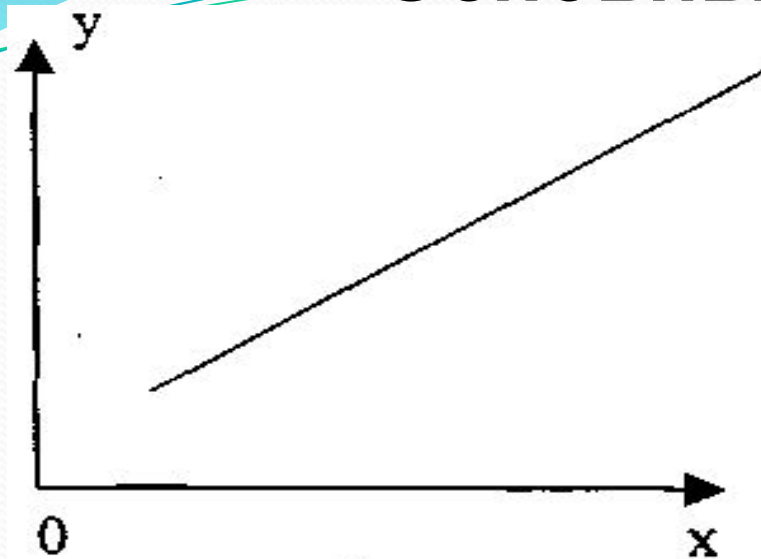
$\underline{a}$ – среднее значение **y** при **$x=0$** .

Форма уравнения определяется на основе визуальной (зрительной) оценки.

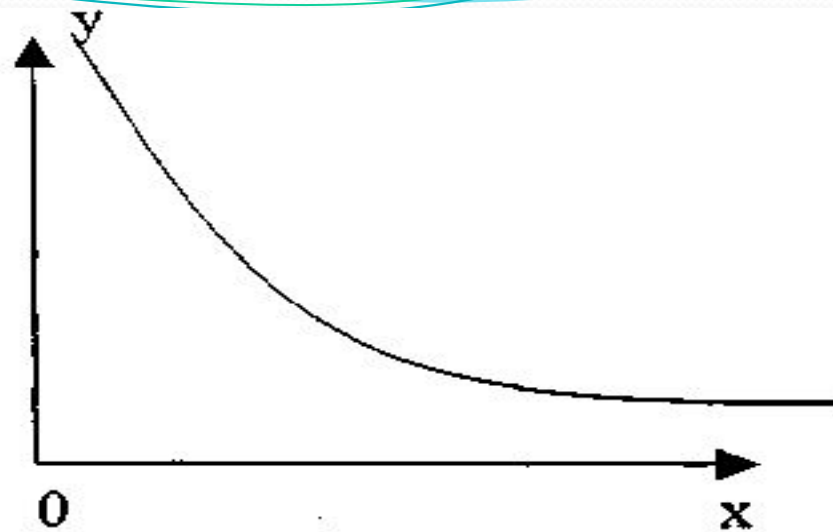
Строится график - корреляционное поле, т.е. оси абсцисс (оx) откладываются значения факторного (независимого) признака x, а по оси ординат (оy) – значения результативного признака y.

Соединяя точки графика отрезками прямой получим эмпирическую линию. По ее виду судят о наличии зависимости и о форме линии.

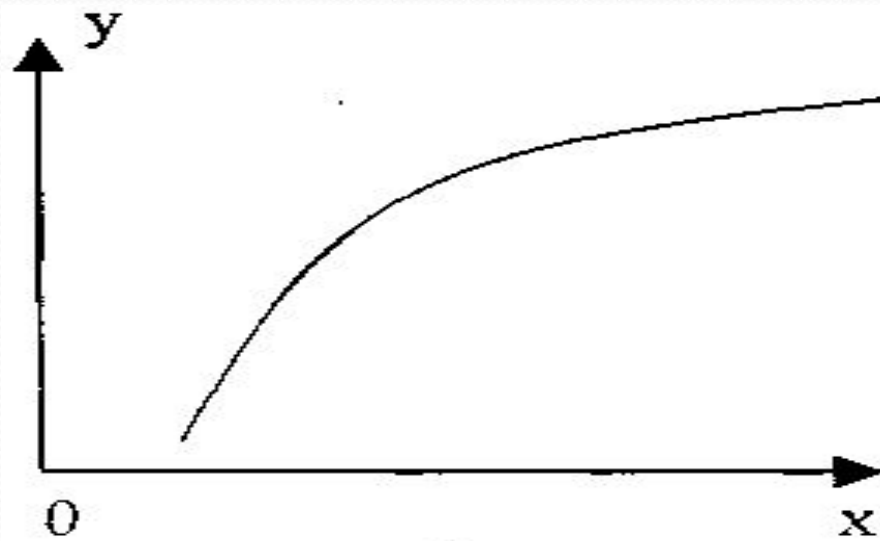
Основные типы линий



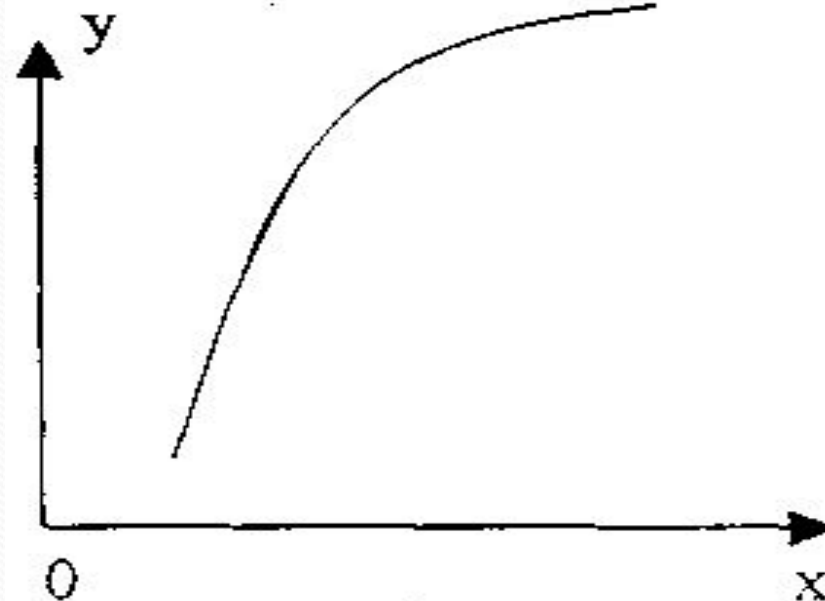
а



в



д



е

Линейная

$$y_x = a + b \cdot x;$$

Гиперболическая

$$y_x = a + b / x;$$

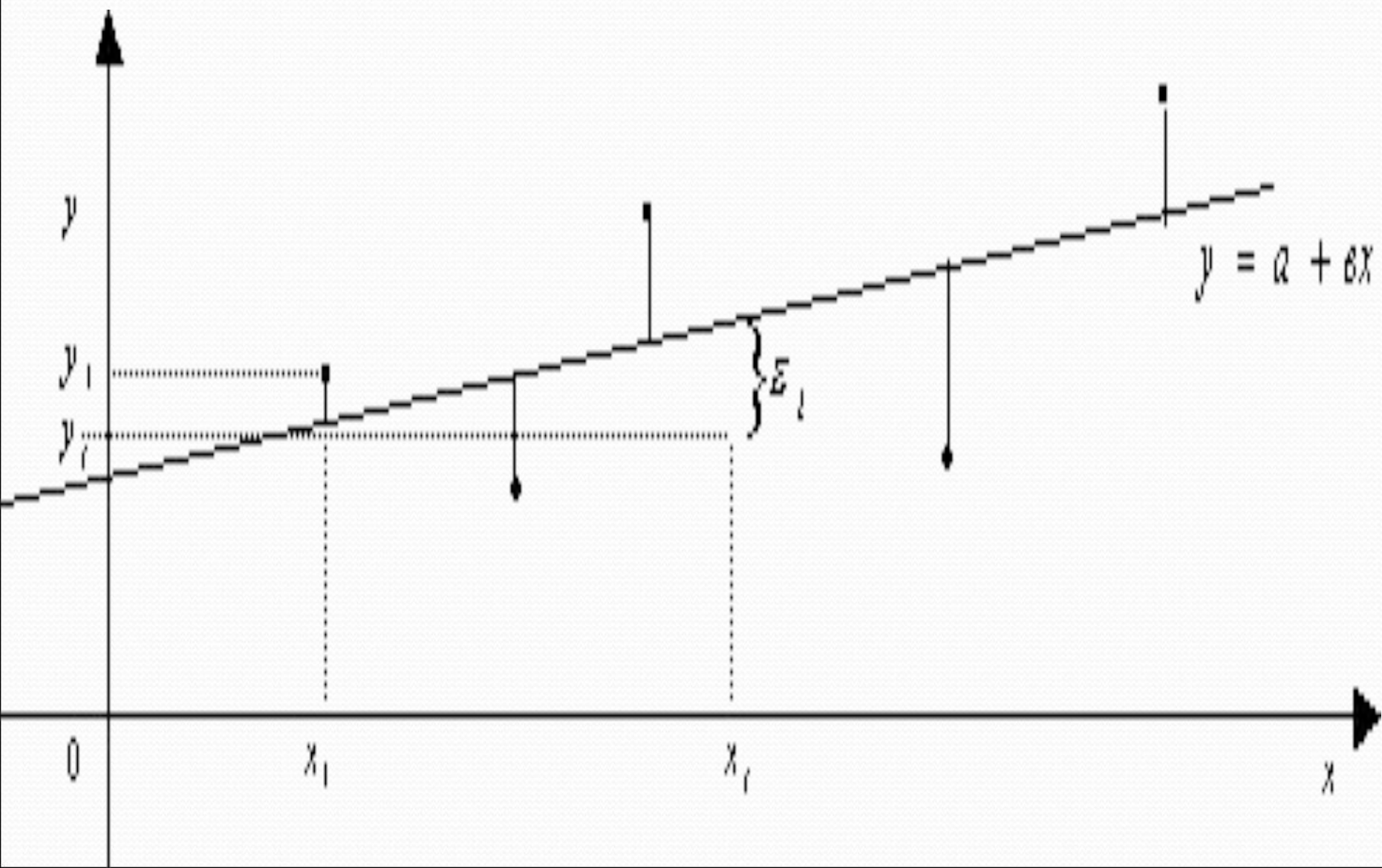
Степенная

$$y_x = a \cdot x^b;$$

Показательная

$$y_x = a \cdot b^x;$$

Корреляционное поле



Пояснения к графику

Случайная выборка значений

$(x_1, x_2, \dots, x_n)$ и $(y_1, y_2, \dots, y_n)$
Уравнение регрессии

$$\hat{y} = a + bx$$

Отклонения

$$e_i = y_i - \hat{y}_i$$

Для каждой точки (x_i, y_i) можно записать различные виды дисперсий:

Виды дисперсий

Дисперсия

y_i

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

Дисперсия

$\hat{y}_i$

$$\sigma_{\hat{y}}^2 = \frac{1}{n} \sum (\hat{y}_i - \bar{y})^2$$

Дисперсия

ε_i

$$\sigma_{\varepsilon}^2 = \frac{1}{n} \sum (\varepsilon_i - \bar{\varepsilon})^2$$

Коэффициент детерминации

$$R^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

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

$$R^2 = 1 - \frac{\sigma_{\varepsilon}^2}{\sigma_y^2} = 1 - \frac{\sum (e_i - \bar{e})^2}{\sum (y_i - \bar{y})^2}$$

Коэффициент детерминации оценивает **качество (точность)** уравнения регрессии, это часть дисперсии (вариации) признака у объясненная уравнением регрессии.

Например: $R^2 = 0,56$ или 56% - это доля вариации у зависящая от вариации x и $100\% - 56\% = 44\%$ вариации у зависящая от вариации других факторов, не учтенных в модели.

Для определения статистической
значимости коэффициента
детерминации используется
F– наблюдаемое Фишера по формуле

$$F = \frac{R^2 \cdot (n - 2)}{1 - R^2}$$

F- критическое определяется по
таблице распределения Фишера

$$F_{\alpha, \nu_1, \nu_2},$$

$$\text{где } \nu_1 = 1, \nu_2 = n - 2$$

Если $F > F_{\alpha, \nu_1, \nu_2}$, то R^2 значим

5. Метод наименьших квадратов (МНК)

МНК минимизирует **сумму квадратов** разностей между фактическими и расчетными значениями зависимой переменной y .

MHK

$$S = \sum (y_i - \hat{y}_i)^2 \rightarrow \min$$



Необходимым условием минимума является равенство нулю ее частных производных по параметрам регрессии.

Для линейной регрессии получаем систему нормальных уравнений:

$$\left\{ \begin{array}{l} na + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i x_i. \end{array} \right.$$

Решая систему, получим:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$$

$$a = \bar{y} - b \cdot \bar{x}$$

Шкала Чеддока для качественной оценки тесноты связи между x и y

Теснота связи	Значение коэффициента корреляции при наличии	
	прямой связи	обратной связи
Слабая	0,1 – 0,3	(-0,1) – (-0,3)
Умеренная	0,3 – 0,5	(-0,3) – (-0,5)
Заметная	0,5 – 0,7	(-0,5) – (-0,7)
Высокая	0,7 – 0,9	(-0,7) – (-0,9)
Весьма высокая	0,9 – 0,99	(-0,9) – (-0,99)

6. Предпосылки МНК – условия Гаусса-Маркова

1. Математическое ожидание случайного отклонения ε_i равно нулю для всех наблюдений $\mathbf{M}(\varepsilon_i)=0$.
2. Дисперсия случайного отклонения постоянная для всех наблюдений
$$\mathbf{D}(\varepsilon_i) = \mathbf{D}(\varepsilon_j) = \sigma^2$$

Постоянство дисперсии отклонения называется **гомоскедастичностью**.

Непостоянство дисперсии отклонения называется **гетероскедастичностью**.

3. Случайные отклонения ε_i и ε_j должны быть независимы друг от друга. Если данное условие выполняется, то говорят об отсутствии **автокорреляции**.

4. Случайные отклонения должны быть независимы от объясняющих переменных.

Теорема Гаусса-Маркова

«Если выполняются условия 1- 4, то оценки **(a, b)**, сделанные с помощью МНК, являются наилучшими линейными несмещенными оценками параметров (β_0, β_1) , т.е. они обладают свойствами:

- 1) несмещенность;
- 2) эффективность;
- 3) состоятельность.»

- Оценка Θ_n (тэта) называется **состоятельной** , если она сходится по вероятности к значению оцениваемого параметра Θ при безграничном возрастании объема выборки.
- **Несмещенная** оценка Θ_n – это оценка параметра Θ , математическое ожидание которой равно значению оцениваемого параметра: $M(\Theta_n) = \Theta$.
- **Эффективная** оценка – это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок параметра

Лекция № 4
(слайды 49-74)

7. Ошибки измерения

Средняя ошибка аппроксимации

или среднее относительное отклонение
расчетных значений от фактических
(должно быть не более 8 – 10%)

$$\overline{A} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%$$

Коэффициент эластичности
показывает, на сколько % изменяется
функция $y=f(x)$ при изменении
независимой переменной x на 1 %.

$$\overline{\mathcal{E}_{yx}} = b \cdot \frac{\overline{x}}{y}$$

Мерой разброса переменной y служит стандартная ошибка регрессии $S^2 = S_e^2$

$$S^2 = \frac{1}{n-2} \cdot \sum e_i^2 = \frac{1}{n-2} \cdot \sum (y_i - \hat{y}_i)^2$$

$$S = \sqrt{\frac{1}{n-2} \sum e_i^2}$$

Стандартные ошибки коэффициентов
регрессии оцениваются по формулам:

$$S_a^2 = \frac{(\bar{x})^2 S^2}{n \cdot \sum (x - \bar{x})^2}, \quad S_b^2 = \frac{S^2}{n \cdot \sum (x - \bar{x})^2}$$

Оценка значимости коэффициентов
регрессии с помощью t – наблюдаемого
Стьюдента по формулам:

$$t_a = \frac{a}{S_a}, \quad t_b = \frac{b}{S_b}$$

t- критические определяются по таблицам
распределения Стьюдента

$$t_{\alpha, n-2}$$

Если $|t_a| > t_{\alpha, n-2}$ и $|t_b| > t_{\alpha, n-2}$
то коэффициенты а и b значимы

Предсказание и прогнозирование на основе линейной модели регрессии

Поиск значений Y для X , находящихся между известными значениями, называется предсказанием.

Прогнозирование – это оценка значений Y для некоторого будущего набора независимых переменных.

Для предсказания достаточно поставить в уравнение регрессии нужное значение x .
Для прогноза используется понятие

доверительной вероятности

$$\gamma = (0,9; \quad 0,95; \quad 0,99)$$

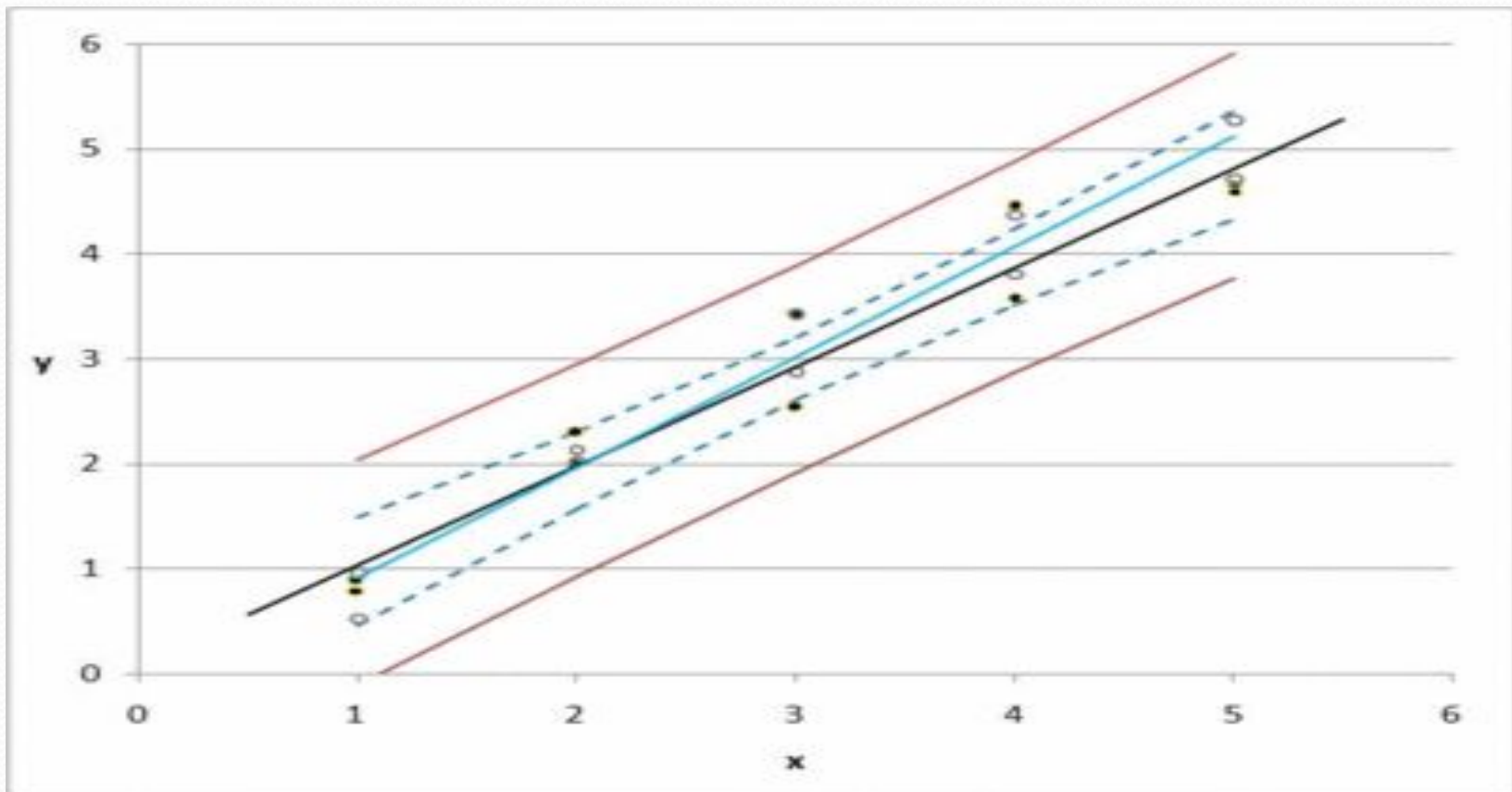
и уровня значимости

$$\alpha = 1 - \gamma = (0,1; \quad 0,05; \quad 0,01)$$



Доверительные интервалы

**Линия регрессии и 95%-е
доверительные области для линии
регрессии (пунктиром)
и для значений границы (сплошные)**



Доверительный интервал для y

$$\hat{y} - t_{\frac{1-\alpha}{2}, n-2} \cdot S_{\hat{y}} < y < \hat{y} + t_{\frac{1-\alpha}{2}, n-2} \cdot S_{\hat{y}}$$

Доверительный интервал для $\overline{y}$

$$\hat{y} - t_{\frac{1-\alpha}{2}, n-2} \cdot S_y < \overline{y} < \hat{y} + t_{\frac{1-\alpha}{2}, n-2} \cdot S_y$$

Пояснения к формулам доверительных интервалов (m – кратность измерений y)

$$S_e^2 = \sum \frac{(y_i - \hat{y})^2}{n - 2}$$

$$S_{\hat{y}} = S_e \cdot \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

$$S_y = S_e \cdot \sqrt{\frac{1}{m} + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Коэффициент детерминации **для линейной регрессии**

$$R^2 = r^2$$

показывает долю общей вариации
(дисперсии) зависимой переменной y ,
обусловленной регрессией или
изменчивостью объясняемой переменной x .

9. Нелинейные модели регрессии и их линеаризация.

Экспоненциальная регрессия

$$\tilde{y} = e^{ax+b}$$

Линеаризующие преобразования

$$x' = x$$

$$y' = \ln y$$

Параметры уравнения экспоненциальной регрессии

$$b = \frac{n \sum_{i=1}^n (x_i \ln y_i) - \sum_{i=1}^n x_i \sum_{i=1}^n \ln y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}$$

$$a = \frac{1}{n} \sum_{i=1}^n \ln y_i - \frac{1}{n} b \sum_{i=1}^n x_i$$

Логарифмическая регрессия

$$\tilde{y} = a + b \ln x$$

$$x' = \ln x$$

$$y' = y$$

Параметры уравнения логарифмической регрессии

$$b = \frac{n \sum_{i=1}^n (\ln x_i \cdot y_i) - \sum_{i=1}^n \ln x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n (\ln x_i)^2 - \left(\sum_{i=1}^n \ln x_i \right)^2}$$

$$a = \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} b \sum_{i=1}^n \ln x_i$$

Решение задач на компьютере

Microsoft Excel

Вычисление выборочной средней

$$\bar{x} = \frac{1}{n} \sum x_i \quad \bar{x} = \text{СРЗНАЧ}(\text{массив } x)$$

Выборочная дисперсия (вариация)

$$d_B = D_B = \text{var}(x) = \frac{1}{n} \sum (x_i - \bar{x})^2 = \overline{x^2} - (\bar{x})^2$$

$$d_B = \text{ДИСПР}(\text{массив } x)$$

Исправленная дисперсия

$$S_x^2 = \frac{n}{n-1} \cdot d_B = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{n}{n-1} \cdot [\overline{x^2} - (\bar{x})^2]$$

$$S_x^2 = \text{ДИСП}(\text{массив } x)$$

Стандартное отклонение

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

$$S_x = \text{СТАНДОТКЛОН}(\text{массив } x)$$

Коэффициент корреляции Пирсона

$$r_{xy} = \sqrt{\frac{\sum (\tilde{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2) \cdot (\overline{y^2} - \bar{y}^2)}}$$

$$r_{xy} = PEARSON(\text{массив } x, \text{массив } y) = \\ = KОРРЕЛ(\text{массив } x, \text{массив } y)$$

Коэффициенты уравнения регрессии

a = ОТРЕЗОК (массив x, массив y)

b = НАКЛОН (массив x, массив y)

t- критическое Стьюдента

$t = \text{СТЮДРАСПОБР} (\alpha; n-2; 2)$

F – критическое Фишера значимости
коэффициента детерминации

$$F_{\text{критич.}} = F_{\text{РАСПОБР}}(\alpha; 1; n-2)$$