Проверка качества уравнения регрессии

Коэффициент множественной корреляции:

$$R = \sqrt{1 - \frac{D_{\text{îñò}}}{D_y}} = \sqrt{1 - \frac{\sum_{i=1}^{n} (\hat{y} - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

принимает значения в диапазоне $0 \le R \le 1$. Чем ближе он к 1, тем теснее связь результативного признака со всем набором исследуемых факторов.

Линейный коэффициент множественной корреляции (совокупный):

$$R_{yx_1x_2,\ldots,x_p} = \sqrt{\sum \beta_{x_i} r_{yx_i}}$$

Нелинейный квази-коэффициент детерминации:

$$\text{квази}R^2 = 1 - \frac{\sum (y - anti(\ln y))^2}{\sum (y - \bar{y})^2}$$

Скорректированный (улучшенный) коэффициент множественной детерминации

$$\overline{R}^2 = 1 - \frac{\sum (y - \hat{y})^2 : (n - m - 1)}{\sum (y - \bar{y})^2 : (n - 1)} = 1 - (1 - R^2) \cdot \frac{n - 1}{n - m - 1}$$

где n – число наблюдений,

m – число параметров при переменных x.

Чем больше величина m, тем **больше** различия между коэффициентом множественной детерминации и скорректированным коэффициентом.

Чем больше объем совокупности, по которой исчислена регрессия, тем **меньше** различия между данными коэффициентами.

Оценка значимости уравнения множественной регрессии (F-критерий):

Н0: уравнение статистически не значимо

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

где m – число независимых переменных в уравнении регрессии;

n – число единиц совокупности.

Если **Fфакт > Fтабл**, то *H0* о случайной природе связи отклоняется и признается статистическая значимость и надежность уравнения.

Если **Fфакт < Fтабл**, то *H0* не отклоняется и признается статистическая незначимость уравнения регрессии.

Частный F-критерий:

оценивает статистическую значимость присутствия каждого из факторов в уравнении

$$F_{частx_i} = \frac{R^2_{yx_1 \ldots x_i \ldots x_p} - R^2_{yx_1 \ldots x_i - 1 x_i + 1 \ldots x_p}}{1 - R^2_{yx_1 \ldots x_i \ldots x_p}} \cdot \frac{n - m - 1}{1}$$

$R^2_{yx_1 x_2 \ldots x_p}$ - коэффициент множественной детерминации для модели с полным набором факторов;

$R^2_{yx_2 \ldots x_p}$ - тот же показатель, но без включения в модель фактора *x1*;

n – число наблюдений;

m – число параметров при переменных x.

t-критерий Стьюдента:

$$t_{b_i} = \sqrt{F_{x_i}}, \quad \text{или} \quad t_{b_i} = \frac{b_i}{m_{b_i}},$$

где m_{bi} – средняя квадратическая ошибка коэффициента регрессии b_i, она может быть определена по формуле:

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_1 \ldots x_p}}}{\sigma_{x_i} \cdot \sqrt{1 - R^2_{x_i x_1 \ldots x_p}}} \cdot \frac{1}{\sqrt{n - m - 1}}$$

Частная корреляция

Частные коэффициенты (или индексы) корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в модель:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_p} = \sqrt{1 - \frac{1 - R^2_{yx_1 \cdot x_2 \ldots x_i \ldots x_p}}{1 - R^2_{yx_1 \cdot x_2 \ldots x_{i-1} x_{i+1} \ldots x_p}}}$$

При i=1 формула примет вид:

$$r_{yx_1 \cdot x_2 \ldots x_p} = \sqrt{1 - \frac{1 - R^2_{yx_1 \cdot x_2 \ldots x_p}}{1 - R^2_{yx_2 \ldots x_p}}}$$

- ***Частная корреляция первого порядка*** – когда фиксируется теснота связи двух переменных при устранении влияния одного фактора: $r_{yx_1 \cdot x_2}$

  (точка отделяет фактор, значение которого элиминируется (закрепляется на неизменном уровне)).

- ***Частная корреляция второго и т.д. порядка*** – когда фиксируется теснота связи двух переменных при устранении влияния двух и более факторов, например:

  $r_{yx_1 \cdot x_2 x_3}$ - частная корреляция второго порядка при постоянном действии факторов x2 и x3;

  $r_{yx_1 \cdot x_2 x_3 x_4 x_5}$ - частная корреляция четвертого порядка при постоянном действии факторов x2, x3, x4, x5.

Коэффициенты частной корреляции более высоких порядков можно найти через коэффициенты частной корреляции более низких порядков

по рекуррентной формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_p} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{p-1}} - r_{yx_p \cdot x_1 x_2 \ldots x_{p-1}} r_{x_1 x_p \cdot x_1 x_2 \ldots x_{p-1}}}{\sqrt{(1 - r^2_{yx_p \cdot x_1 x_2 \ldots x_{p-1}})(1 - r^2_{x_i x_p \cdot x_1 x_2 \ldots x_{p-1}})}}$$

- **При i=1 и двух факторах формула примет вид:**

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{\sqrt{(1 - r^2_{yx_2})(1 - r^2_{x_1 x_2})}}$$

- **При i=2 и двух факторах:**

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{\sqrt{(1 - r^2_{yx_1})(1 - r^2_{x_1 x_2})}}$$

Предпосылки метода наименьших квадратов

$$\sum \varepsilon^2 \to \min$$

Требования, предъявляемые к ε:

1. **Несмещенность** – означает, что математическое ожидание остатков равно нулю: $M(\varepsilon_i) = 0$

т.е. при большом числе наблюдений остатки не будут накапливаться и найденный параметр регрессии b можно рассматривать как среднее значение из возможного большого количества несмещенных оценок. Если оценки обладают свойством несмещенности, то их можно сравнивать по разным выборкам.

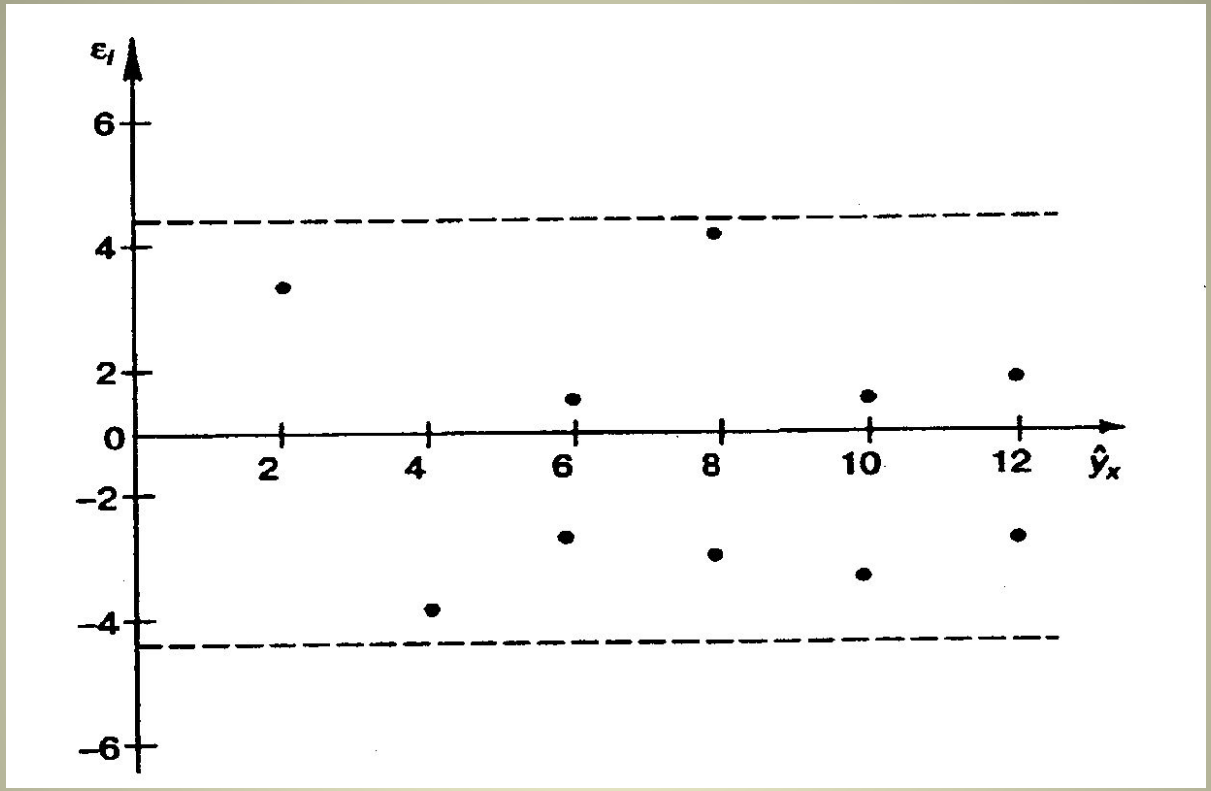
2.Эффективность – оценки считаются эффективными, если они характеризуются наименьшей дисперсией.

3.Состоятельность – характеризует увеличение точности оценок с увеличением объема выборки.

Предпосылки метода наименьших квадратов:

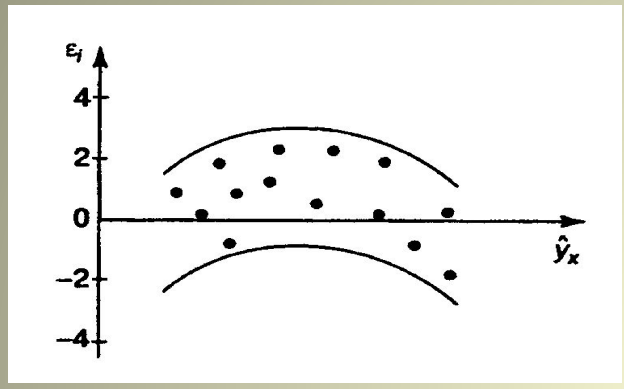
- случайный характер остатков;
- нулевая средняя величина остатков, не зависящая от x;
- гомоскедастичность;
- отсутствие автокорреляции остатков;
- нормальное распределение остатков.

1. Случайный характер остатков

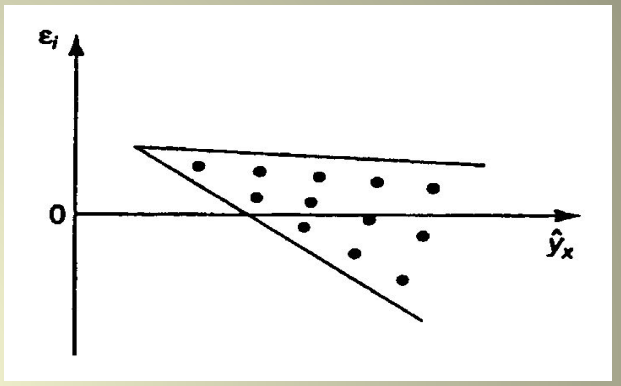


Если на графике нет направленности в расположении точек, то остатки представляют собой случайные величины и МНК оправдан.

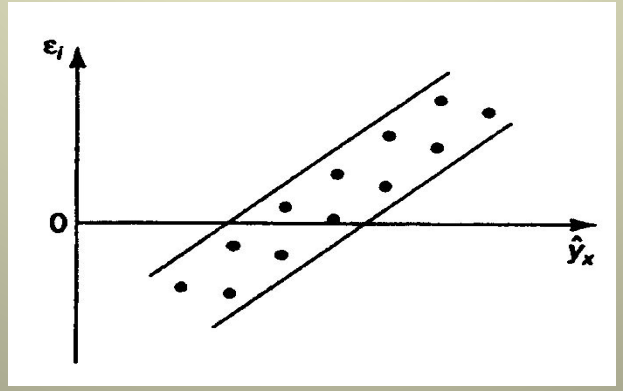
Остатки не случайны:



Остатки не имеют постоянной дисперсии:

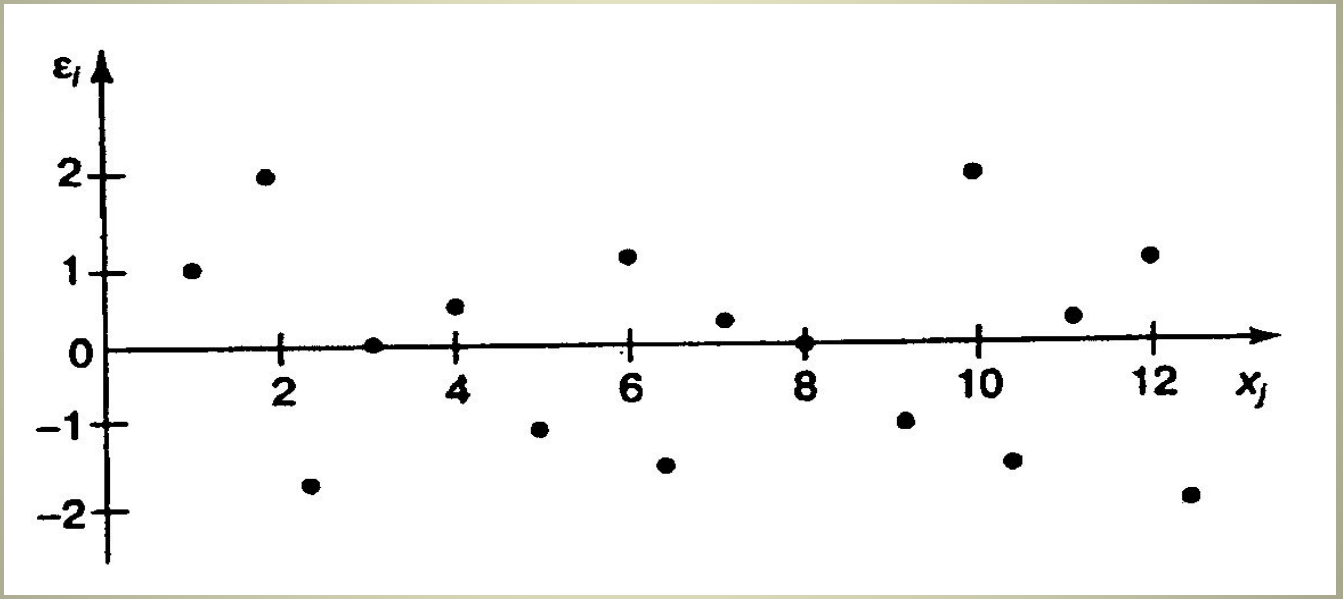


Остатки носят систематический характер:



2. Нулевая средняя величина остатков, не зависящая от x:

$$\sum\left(y-\hat{y}_x\right)=0$$

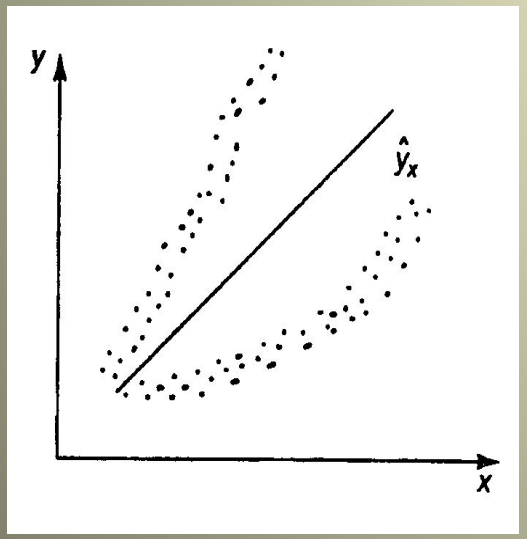


Если расположение остатков на графике не имеет направленности, то они независимы от значений x. Если же график показывает наличие зависимости, то модель неадекватна.

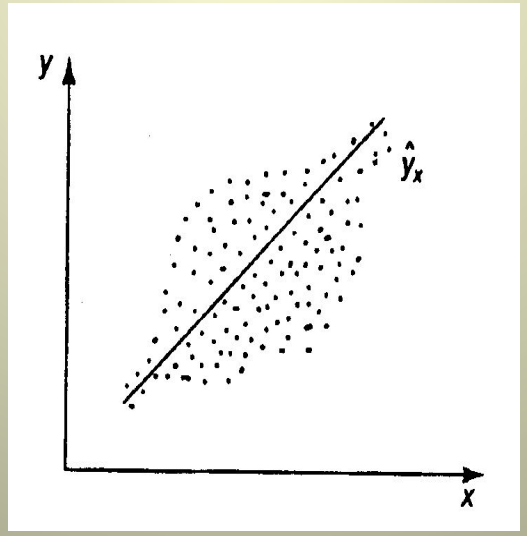
3. Гомоскедастичность

Гомоскедастичность – это однородность относительно дисперсии, т.е. дисперсия остатков одинакова для каждого значения *x*. Если это условие применения МНК не соблюдается, то имеет место ***гетероскедастичность*** (неоднородность относительно дисперсии).

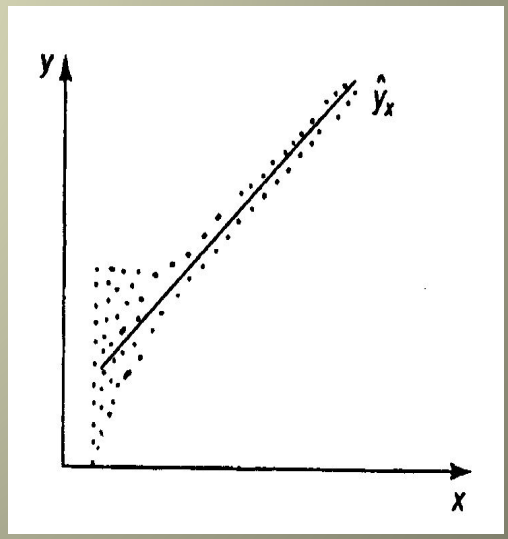
Дисперсия остатков растёет по мере увеличения x:



Дисперсия остатков достигает максимальной величины при средних значениях x



Максимальная дисперсия остатков при малых значениях x



Методы проверки предпосылки МНК о гомоскедастичности остатков:

- Тест Гольдфельда-Квандта
- Тест ранговой корреляции Спирмена
- Тест Глейзера
- и другие.

4. Отсутствие автокорреляции остатков

Под ***автокорреляцией остатков*** понимают зависимость распределения значений остатков друг от друга. Это означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений.

$$r_{\varepsilon_i \varepsilon_j} = \frac{\overline{\varepsilon_i \varepsilon_j} - \bar{\varepsilon}_i \cdot \bar{\varepsilon}_j}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_j}}$$

Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированны.

Обобщенный метод наименьших квадратов (ОМНК)

$$\sigma_{\varepsilon_i}^2 = \sigma^2 \cdot K_i$$

где $\sigma_{\varepsilon_i}^2$ – дисперсия ошибки при конкретном i-м значении фактора;

σ^2 – постоянная дисперсия ошибки при соблюдении предпосылки о гомоскедастичности остатков;

K_i – коэффициент пропорциональности, меняющийся с изменением величины фактора, что и обусловливает неоднородность дисперсии.

В общем виде для уравнения

$$y_i = a + bx_i + \varepsilon_i \text{ при } \sigma_{\varepsilon_i}^2 = \sigma^2 \cdot K_i$$

модель примет вид: $y_i = a + bx_i + \sqrt{K_i}\,\varepsilon_i$.

Поделим все переменные на $\sqrt{K_i}$. От регрессии y по x мы перейдем к регрессии на новых переменных: $y/\sqrt{K}$ и $x/\sqrt{K}$.

Уравнение регрессии примет вид:

$$\frac{y_i}{\sqrt{K_i}} = \frac{a}{\sqrt{K_i}} + b \cdot \frac{x_i}{\sqrt{K_i}} + \varepsilon_i$$

$$S(a,b)=\sum_{i=1}^{n}\frac{1}{K_i}(y_i-a-bx_i)^2\to\min$$

Соответственно получим следующую систему нормальных уравнений:

$$\begin{cases}\sum\frac{y}{K}=a\cdot\sum\frac{1}{K}+b\cdot\sum\frac{x}{K},\\ \sum\frac{y\cdot x}{K}=a\cdot\sum\frac{x}{K}+b\cdot\sum\frac{x^2}{K}.\end{cases}$$

Если преобразованные переменные x и y взять в отклонениях от средних уровней, то коэффициент регрессии b можно определить как

$$b=\frac{\sum\frac{1}{K}\cdot x\cdot y}{\sum\frac{1}{K}\cdot x^2}.$$

Пример. Пусть y – издержки производства,

x_1 – обьем продукции,

x_2 – основные производственные фонды,

x_3 – численность работников,

тогда уравнение $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$ – модель издержек производства.

Предполагая, что $\sigma_{\varepsilon_i}^2$ *пропорциональна квадрату численности работников* x_3, получим в качестве результативного признака затраты на одного работника y/x_3, а в качестве факторов следующие показатели: производительность труда x_1/x_3 и фондовооруженность труда x_2/x_3.

Соответственно трансформированная модель примет вид:

$$\frac{y}{x_3} = b_3 + b_1 \frac{x_1}{x_3} + b_2 \frac{x_2}{x_3} + \varepsilon,$$