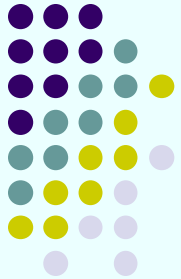
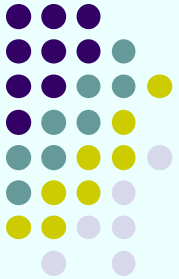
Лекция 2. Множественная регрессия



Вопросы:

1. Классическая модель множественной линейной регрессии в матричной форме
2. Оценка параметров МНК
3. Ковариационная матрица
4. Множественные коэффициенты детерминации и корреляции. Оценка значимости множественной регрессии
5. Оценка значимости параметров. Интервальная оценка параметров и прогноза
6. Понятие и проблема мультиколлинеарности факторов и способы ее преодоления
7. Коэффициент частной корреляции
8. Свойства оценок МНК
9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме.
10. Обобщенная линейная модель. ОМНК
11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК
12. Отбор наиболее существенных объясняющих переменных в модель множественной регрессии
13. Уравнение регрессии с фиктивными переменными. Критерий Чоу
14. Нелинейные модели множественной регрессии. Производственные функции
15. Вопросы для повторения и самостоятельного изучения

1. Классическая модель множественной линейной регрессии в матричной форме



$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

где i=1,2…n – номер наблюдения, число объясняющих переменных (x) равно p.

***β_i-коэффициент чистой регрессии**, показывает на сколько единиц изменится зависимая переменная, если независимая – x_i – изменится на единицу, при условии, что все остальные факторы будут зафиксированы на среднем уровне.*

1. Классическая модель множественной линейной регрессии в матричной форме

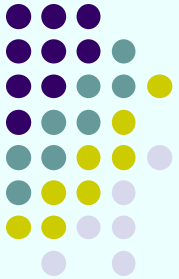
$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

матрица объясняющих переменных размера n×(p+1)

Предпосылка 6.

Векторы значений объясняющих переменных (столбцы матрицы *X*) должны быть линейно независимыми, т.е. ранг матрицы – максимальный *(X)=p+1*

1. Классическая модель множественной линейной регрессии в матричной форме



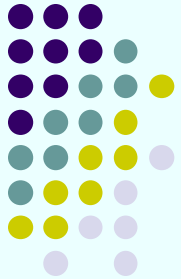
$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}; X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \hdashline 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix};$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_p \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}, \textit{тогда в матричной форме}: \; Y = X\beta + \varepsilon.$$

$$\textit{Выборочная оценка}: \; Y = Xb + e, \; \textit{где}$$

$$b = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_p \end{pmatrix}, e = \begin{pmatrix} e_0 \\ e_1 \\ \dots \\ e_n \end{pmatrix}.$$

2. Оценка параметров МНК



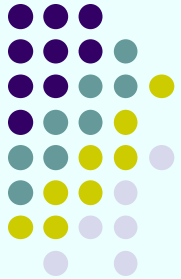
$$e'e = (e_1 e_2 \ldots e_n)\begin{pmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{pmatrix} = e_1^2 + e_2^2 + \ldots e_n^2 = \sum_{i=1}^{n} e_i^2, \text{ условие минимизации :}$$

$$\sum_{i=1}^{n} (\tilde{y}_i - y_i)^2 = \sum_{i=1}^{n} e_i^2 = e'e = (Y - Xb)'(Y - Xb) \to \min$$

$$X'Xb = X'Y$$

$$b = (X'X)^{-1} X'Y$$

2. Оценка параметров МНК



$$Пример:$$

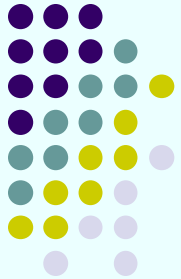
$$b_0 + b_1 x_1 = y$$

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_m \end{pmatrix}; y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix};$$

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_m \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$Тогда \;\; X'Xb = X'Y: \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} y_i x_i \end{pmatrix}$$

2. Оценка параметров МНК



$$Метод\ \ Крамера:$$

$$b_j = \frac{\Delta_j}{\Delta}$$

$$\Delta - определитель\ \ матрицы\ \ X;$$

$$\Delta_j - определитель\ \ матрицы,\ \ получающейся\ \ при\ \ замене$$

$$j - й\ независимой\ переменной\ \ матрицы\ \ X\ \ вектором\ \ Y$$

3. Ковариационная матрица

$$\Sigma_b = \begin{pmatrix} \sigma_{00} & \sigma_{01} & \dots & \sigma_{0p} \\ \sigma_{10} & \sigma_{11} & \dots & \sigma_{1p} \\ \dots & \dots & \dots & \dots \\ \sigma_{p0} & \sigma_{p1} & \dots & \sigma_{pp} \end{pmatrix}$$

$$\sigma_{ij} = M\left[(b_i - M(b_i))(b_j - M(b_J))\right], \text{поскольку } M(b_j) = \beta_j$$

$$\sigma_{ij} = M\left[(b_i - \beta_i)(b_j - \beta_j)\right] = M\left[(b-\beta)(b-\beta)'\right],$$

$$\Sigma_b = M\left[(b-\beta)(b-\beta)'\right]$$

$$\text{так как } b = \beta + (X'X)^{-1}X'\varepsilon,$$

$$\text{получим}: \Sigma_b = \left\{\left[(X'X)^{-1}X'\varepsilon\right]\left[(X'X)^{-1}X'\varepsilon\right]'\right\}$$

$$\Sigma_b = (X'X)^{-1}X'M(\varepsilon\varepsilon')X(X'X)^{-1}$$

3. Ковариационная матрица

$$\Sigma_\varepsilon = M(\varepsilon\varepsilon') = \begin{pmatrix} M(\varepsilon_1^2) & M(\varepsilon_1\varepsilon_2) & \dots & M(\varepsilon_1\varepsilon_n) \\ M(\varepsilon_2\varepsilon_1) & M(\varepsilon_2^2) & \dots & M(\varepsilon_2\varepsilon_n) \\ \dots & \dots & \dots & \dots \\ M(\varepsilon_n\varepsilon_1) & M(\varepsilon_n\varepsilon_2) & \dots & M(\varepsilon_n^2) \end{pmatrix}$$

$$M(\varepsilon_i^2) = M(\varepsilon_i - 0)^2 = \sigma^2$$

$$M(\varepsilon\varepsilon') = \sigma^2 E_n$$

$$\Sigma_b = \sigma^2 (X'X)^{-1}$$

3. Ковариационная матрица

Предпосылки множественной регрессии в матричной форме:

$$1.\ \varepsilon - случайный\ вектор, X - неслучайная\ матрица.$$

$$2. M(\varepsilon) = 0_n$$

$$3,4. \Sigma_\varepsilon = M(\varepsilon\varepsilon') = \sigma^2 E_n$$

$$5. \varepsilon - нормально\ распределенный\ вектор$$

$$6. r(X) = (p+1) < n$$

4. Множественные коэффициенты детерминации и корреляции. Оценка значимости множественной регрессии

$$R^2 = \frac{W_R}{W_y} = 1 - \frac{W_e}{W_y} = 1 - \frac{e'e}{y'y},$$

$$y = (Y - \bar{Y})$$

$$R^2_н = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

$$F = \frac{R^2(n-p-1)}{(1-R^2)p} > F_{\alpha;p;n-p-1}$$

4. Множественные коэффициенты детерминации и корреляции. Оценка значимости множественной регрессии

$$R^2 = 1 - \frac{A}{A_{11}}$$

$$R^2 = 1 - \frac{\begin{bmatrix} 1 & r_{yx_1} & r_{yx_2} \\ r_{yx_1} & 1 & r_{x_1x_2} \\ r_{yx_2} & r_{x_1x_2} & 1 \end{bmatrix}}{\begin{bmatrix} 1 & r_{x_1x_2} \\ r_{x_1x_2} & 1 \end{bmatrix}} = \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1 - r^2_{x_1x_2}}$$

5. Оценка значимости параметров. Интервальная оценка параметров и прогноза

$$m_{b_j}^2 = S_{b_j}^2 = S_e^2 \left[(X'X)^{-1} \right]_{jj}$$

$$m_{b_j} = S_{b_j} = \sqrt{S_e^2 \left[(X'X)^{-1} \right]_{jj}}$$

$$S_e = \sqrt{\frac{e'e}{n-p-1}} = \sqrt{\frac{W_e}{n-p-1}}$$

$$X'_n = (1 x_{1n} x_{2n} \ldots x_{pn}), Y_n = X'_n b$$

$$m_{\tilde{y}_n} = \sqrt{S_e^2 X'_n (X'X)^{-1} X_n}$$

$$m_{y_n} = \sqrt{S_e^2 (1 + X'_n (X'X)^{-1} X_n)}$$

6. Понятие мультиколлинеарности и способы ее преодоления

$$T = 1 - R^2_{x_i . x_{1\ldots} x_{i-1}, x_{i+1\ldots} x_p}$$

$$-1 \cdot A^{-1}$$

$$b = \left[X'X + \lambda E_{p+1}\right]^{-1} X'Y$$

7.Коэффициент частной корреляции

Выборочным частным коэффициентом корреляции (частным коэффициентом корреляции) между переменными x_i и x_j при фиксированных значениях остальных $(p-2)$ переменных называется выражение:

$$r_{ij.1,2\ldots p} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}},$$

где через q обозначены алгебраические дополнения, например:

$q_{ij} = (-1)^{(i+j)} M_{ij}$ - алгебраическое дополнение, а M_{ij} - минор (определитель матрицы парных коэффициентов корреляции, получаемый при вычеркивании i-той строки и j-го столбца).

$$r_{yx_1.x_2} = \frac{-q_{yx_1}}{\sqrt{q_{yy}q_{x_1x_1}}} \qquad r_{yx_2.x_1} = \frac{-q_{yx_2}}{\sqrt{q_{yy}q_{x_2x_2}}} \qquad r_{x_1x_2.y} = \frac{-q_{x_1x_2}}{\sqrt{q_{x_1x_1}q_{x_2x_2}}}$$

7.Коэффициент частной корреляции

$$r_{ij.k} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}$$

Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель. В общем виде:

$$r_{ij.1,2...i-1,i+1...j-1,j+1...p} = \sqrt{\frac{W_{i1,2...i-1,i+1...j-1,j+1...p} - W_{i1,2...j...p}}{W_{i1,2...i-1,i+1...j-1,j+1...p}}},$$

где $W_{i1,2...j...p}$ - остаточный объем вариации при построении модели регрессии зависимой переменной i от всего набора (p-1) переменных;

$W_{i1,2...i-1,i+1...j-1,j+1...p}$ - остаточный объем вариации модели регрессии зависимой переменной i от (p-2) набора переменных (за исключением j).

В случае трех переменных: $r_{ij.k} = \sqrt{\frac{W_{ik} - W_{ijk}}{W_{ik}}}$.

7.Коэффициент частной корреляции

$$r_{ij.1,2\ldots i-1,i+1\ldots j-1,j+1\ldots p} = \sqrt{1 - \frac{1 - R^2_{i1,2\ldots p}}{1 - R^2_{i1,2\ldots j-1,j+1\ldots p}}}.$$

В случае трех переменных:

$$r_{ij.k} = \sqrt{1 - \frac{1 - R^2_{ijk}}{1 - r^2_{ik}}}$$

Коэффициент частной корреляции может быть найден как обычный коэффициент парной корреляции между остатками моделей регрессии x_i по x_k $(e_{x_i.x_k})$ и x_j по x_k $(e_{x_j.x_k})$.

8. Свойства оценок МНК

1. Оценки b являются несмещенными, т.е.

$$M(b_i) = \beta,$$

b_i – *оценки по всем возможным выборкам.*

$Y = X\beta + \varepsilon$

$$b = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + \varepsilon) =$$

$$= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'\varepsilon = E\beta + (X'X)^{-1}X'\varepsilon =$$

$= \beta + (X'X)^{-1}X'\varepsilon$, *т.е. оценки параметров, найденные по выборке, будут содеражать случайные ошибки*

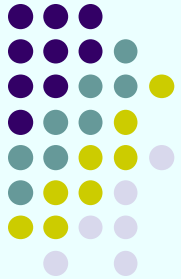
Поскольку $M(\varepsilon) = 0$, *то*

$$M(b) = M(\beta + (X'X)^{-1}X'\varepsilon) = M(\beta) + M(X'X)^{-1}X'\varepsilon) =$$

$$= M(\beta) + (X'X)^{-1}M(X'\varepsilon) = \beta$$

Требование несмещенности гарантирует отсутствует систематических ошибок при оценивании

8. Свойства оценок МНК



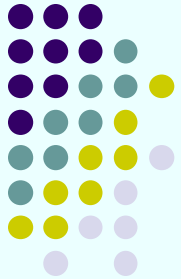
2. По *теореме Гаусса-Маркова* при выполнении предпосылок 1-4, 6 *несмещенная оценка МНК*

$$b = (X'X)^{-1} X'Y$$

является *наиболее эффективной*, т.е. обладает наименьшей дисперсией в классе линейных несмещенных оценок.

Эффективность является решающим свойством, определяющим качество оценок.

8. Свойства оценок МНК



3. Оценки b являются состоятельными, т.е. при увеличении численности выборки сходятся по вероятности к оцениваемым параметрам:

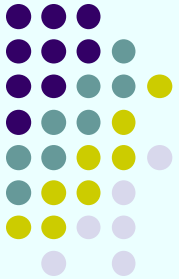
$$\lim_{n\to\infty} P(|\beta - b|) \le \varepsilon) = 1$$

или

$$b \xrightarrow[n\to\infty]{\text{P}} \beta$$

В случае использования состоятельных оценок оправдывается увеличение объема выборки, так как при этом становятся маловероятными значительные ошибки при оценивании. При построении множественной модели регрессии на каждый фактор должно приходиться по 6-10 наблюдений.

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме



Стандартизованные коэффициенты регрессии $\beta - коэффициенты$:

$$\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}$$

Величина бета-коэффициента показывает, на сколько средних квадратических отклонений изменится y, если x_j изменится на одно среднее квадратическое отклонение.

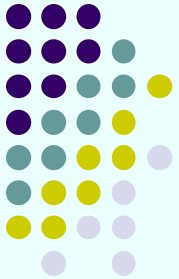
Для парной модели регрессии β-коэффициент равен коэффициенту корреляции:

$$r = \frac{\operatorname{cov}(x;y)}{\sigma_x \sigma_y} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n\sigma_x \sigma_y} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x \sigma_y}$$

$$b = \frac{\operatorname{cov}(x;y)}{\sigma_x^2}$$

$$r = \beta = b\frac{\sigma_x}{\sigma_y}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме



Уравнение прямой, проходящей через точку $M(\bar{x};\bar{y})$

$$(y-\bar{y})=b(x-\bar{x})$$

$$(y-\bar{y})=r\frac{\sigma_y}{\sigma_x}(x-\bar{x})$$

$$\frac{(y-\bar{y})}{\sigma_y}=r\frac{(x-\bar{x})}{\sigma_x}$$

$$t_y=rt_x \Rightarrow r=\frac{t_y}{t_x}$$

$$t_y=\beta t_x, r=\beta$$

$$r=\frac{1}{n}\frac{\sum(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y}=\frac{1}{n}t_xt_y=\overline{t_xt_y}$$

$$\overline{t_xt_x}=\overline{t_yt_y}=1$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме

Уравнение двухфакторной модели регрессии в стандартизованной форме:

$$\beta_1 t_{x_1} + \beta_2 t_{x_2} = t_y$$

Параметры (бета-коэффициенты могут быть найдены методом наименьших квадратов):

$$МНК: \sum (\tilde{t}_y - t_y)^2 \to \min$$

$$\sum (\beta_1 t_{x_1} + \beta_2 t_{x_2} - t_y)^2 \to \min$$

$$\begin{cases} \dfrac{\partial t_y}{\partial \beta_1} = 2t_{x_1}(\beta_1 t_{x_1} + \beta_2 t_{x_2} - t_y) = 0 \\ \dfrac{\partial t_y}{\partial \beta_2} = 2t_{x_2}(\beta_1 t_{x_1} + \beta_2 t_{x_2} - t_y) = 0 \end{cases}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме

$$\begin{cases} \beta_1 \overline{t_{x_1} t_{x_1}} + \beta_2 \overline{t_{x_1} t_{x_2}} = \overline{t_{x_1} t_y} \\ \beta_1 \overline{t_{x_2} t_{x_1}} + \beta_2 \overline{t_{x_2} t_{x_2}} = \overline{t_{x_2} t_y} \end{cases}$$

$$\overline{t_x t_x} = \overline{t_y t_y} = 1$$

$$\overline{t_y t_x} = r_{yx}$$

$$\begin{cases} \beta_1 + \beta_2 r_{x_1 x_2} = r_{x_1 y} \Rightarrow \beta_1 = r_{x_1 y} - \beta_2 r_{x_1 x_2} \\ \beta_1 r_{x_1 x_2} + \beta_2 = r_{x_2 y} \end{cases}$$

$$(r_{x_1 y} - \beta_2 r_{x_1 x_2}) r_{x_1 x_2} + \beta_2 = r_{x_2 y}$$

$$r_{x_1 y} r_{x_1 x_2} - \beta_2 r_{x_1 x_2} r_{x_1 x_2} + \beta_2 = r_{x_2 y}$$

$$\beta_2 = \frac{r_{x_2 y} - r_{x_1 y} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме

$$\beta_1 = \frac{r_{x_1 y} - r_{x_2 y} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}$$

$$\beta_2 = \frac{r_{x_2 y} - r_{x_1 y} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}$$

$$r_{x_1 y.x_2} = \frac{r_{x_1 y} - r_{x_2 y} r_{x_1 x_2}}{\sqrt{1 - r^2_{x_2 y}}\sqrt{1 - r^2_{x_1 x_2}}}$$

$$\beta_1 = r_{x_1 y.x_2} \frac{\sqrt{1 - r^2_{x_2 y}}}{\sqrt{1 - r^2_{x_1 x_2}}}$$

$$\beta_2 = r_{x_2 y.x_1} \frac{\sqrt{1 - r^2_{x_1 y}}}{\sqrt{1 - r^2_{x_1 x_2}}}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме

Коэффициенты эластичности для линейной связи определяются по формуле:

$$\overline{Э}_j = b_j \cdot \frac{\overline{x}_j}{\overline{y}}$$

Они показывают, на сколько процентов изменится признак-результат, если признак-фактор изменится на один процент.

$$\lim_{\Delta x \to 0} \left(\frac{\Delta y}{y} : \frac{\Delta x}{x} \right) = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{\Delta x} \cdot \frac{x}{y} \right) = f'(x) \frac{x}{y}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме

На основе множественного линейного уравнения регрессии

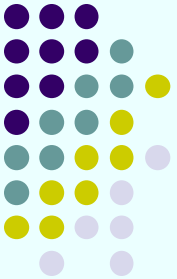
$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$$

Могут быть найдены частные уравнения регрессии:

$$\begin{cases} y_{x_1 . x_{2\ldots} x_p} = b_0 + b_1 x_1 + b_2 \bar{x}_2 + \ldots + b_p \bar{x}_p + \varepsilon \\ y_{x_2 . x_1 x_3 \ldots x_p} = b_0 + b_1 \bar{x}_1 + b_2 x_2 + \ldots + b_p \bar{x}_p + \varepsilon \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_p . x_1 \ldots x_{p-1}} = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + \ldots + b_p x_p + \varepsilon \end{cases}$$

$$\begin{cases} y_{x_1 . x_{2\ldots} x_p} = a_1 + b_1 x_1 \\ y_{x_2 . x_1 x_3 \ldots x_p} = a_2 + b_2 x_2 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_p . x_1 \ldots x_{p-1}} = a_p + b_p x_p \end{cases} \qquad \begin{cases} a_1 = b_0 + b_2 \bar{x}_2 + \ldots + b_p \bar{x}_p \\ a_2 = b_0 + b_1 \bar{x}_1 + b_3 \bar{x}_3 \ldots + b_p \bar{x}_p \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_p = b_0 + b_1 \bar{x}_1 + \ldots + b_{p-1} \bar{x}_{p-1} \end{cases}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме



В отличие от парной регрессии частные уравнения регрессии характеризуют изолированное влияние фактора на результат, поскольку все остальные факторы закреплены на среднем уровне, эффекты их влияния добавлены к свободному члену. На основе частных уравнений регрессии могут быть найдены частные коэффициенты эластичности (для каждого x_j):

$$Э_{y x_{ij}} = b_j \frac{x_{ij}}{\tilde{y}_{x_{ij}.x_1 \ldots x_{i-1}, x_{i+1} \ldots x_p}},$$

где x_{ij} –значение j-го фактора по i-му наблюдению,

$\tilde{y}_{x_{ij}.x_1 \ldots x_{i-1}, x_{i+1} \ldots x_p}$ - выравненное по частному уравнению регрессии (для x_j) значение зависимой переменной для i-наблюдения

Так, для x_1 для 10 наблюдения частный коэффициент эластичности :

$$Э_{y x_{10.1}} = b_1 \frac{x_{10.1}}{\tilde{y}_{x_{10.1}.x_1 \ldots x_{i-1}, x_{i+1} \ldots x_p}}$$

9. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации. Модель регрессии в стандартизованной форме

Коэффициент частной детерминации определяется по формуле:

$$d_j^2 = \beta_j \cdot r_{x_j y}$$

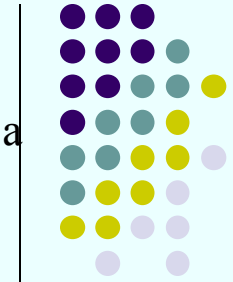
$$d_j^2 = b_j \frac{\sigma_x}{\sigma_y} \frac{\mathrm{cov}(x_j; y)}{\sigma_x \sigma_y} = b_j \frac{\mathrm{cov}(x_j; y)}{\sigma_y^2}$$

Коэффициенты частной детерминации показывают вклад каждого фактора в формирование коэффициента множественной детерминации:

$$R^2 = \sum d_j^2$$

В коэффициенте частной детерминации смешивается чистый эффект от влияния фактора, который выражается бета-коэффициентом, и смешанный (коэффициент парной корреляции), поэтому существует альтернативная форма разложения коэффициента множественной детерминации с учетом системного эффекта (η):

$$R^2 = \sum \beta_j^2 + \eta$$



- Обобщенная линейная модель. Предпосылки 1,2,6 остаются неизменными, а заменяется на

$$3,4.\Sigma_\varepsilon = M(\varepsilon\varepsilon') = \sigma^2 E_n$$
$$3,4.\Sigma_\varepsilon = M(\varepsilon\varepsilon') = \Omega$$

- Ковариационная матрица оценок параметров оказывается неприемлемой в условиях ОЛММР:

$$\Sigma_b = (X'X)^{-1} X'\Omega X (X'X)^{-1}$$

- Теорема Айткена

В классе линейных несмещенных оценок вектора β для обобщенной регрессионной модели оценка

$$b^* = (X'\Omega^{-1}X)^{-1} X'\Omega^{-1}Y$$

Имеет наименьшую ковариационную матрицу Σ_{b^*}

В случае классической модели оценка b^* ОМНК совпадает с оценкой b МНК, поскольку

$$\Sigma_\varepsilon = \Omega = \sigma^2 \mathrm{E}_n$$

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

Для проверки случайного характера остатков строится график зависимости остатков от теоретических значений результативного признака:

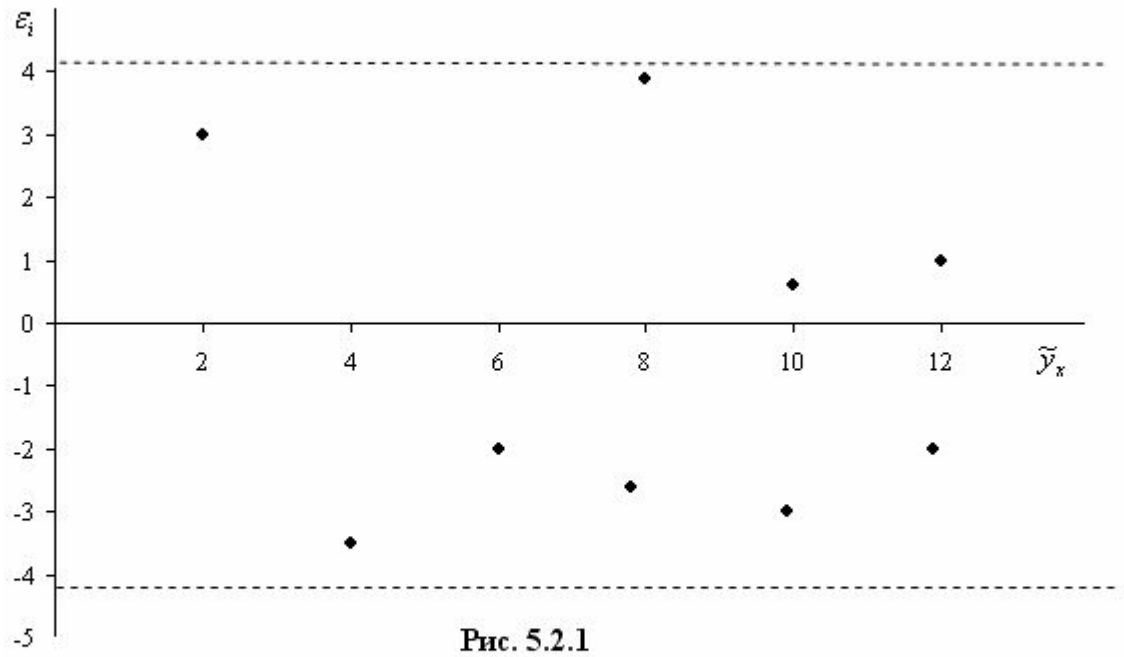


Рис. 5.2.1

Если на графике получена горизонтальная полоса, то остатки представляют собой случайные величины и МНК оправдан.

Возможны иные случаи:

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

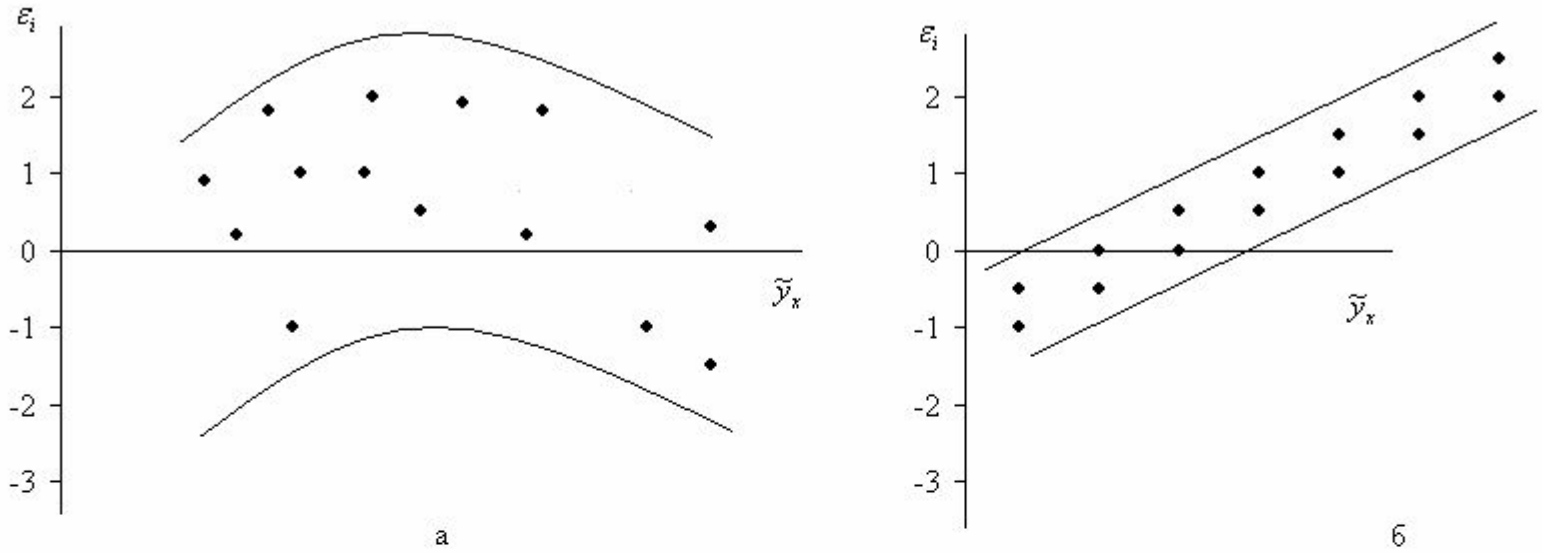
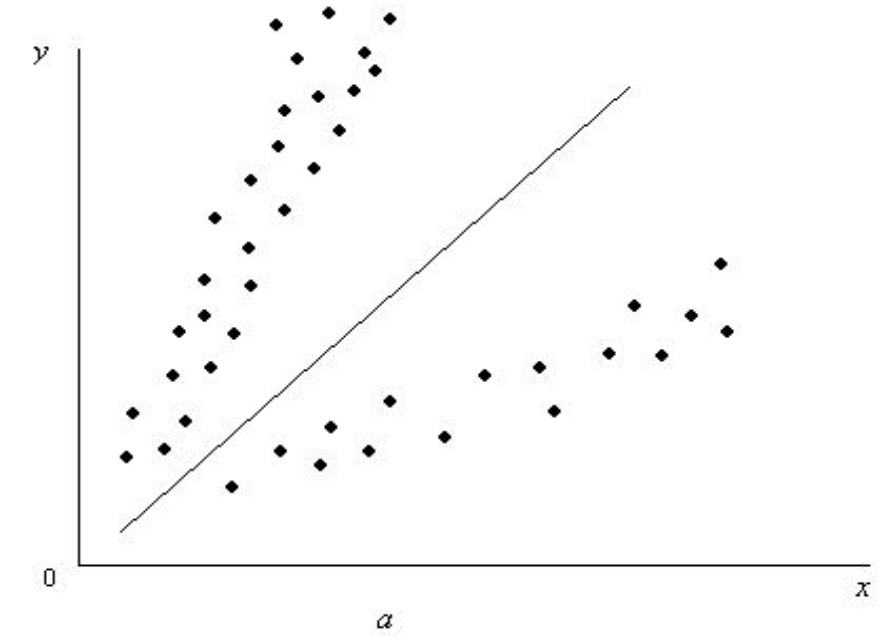


Рис. 5.2.2

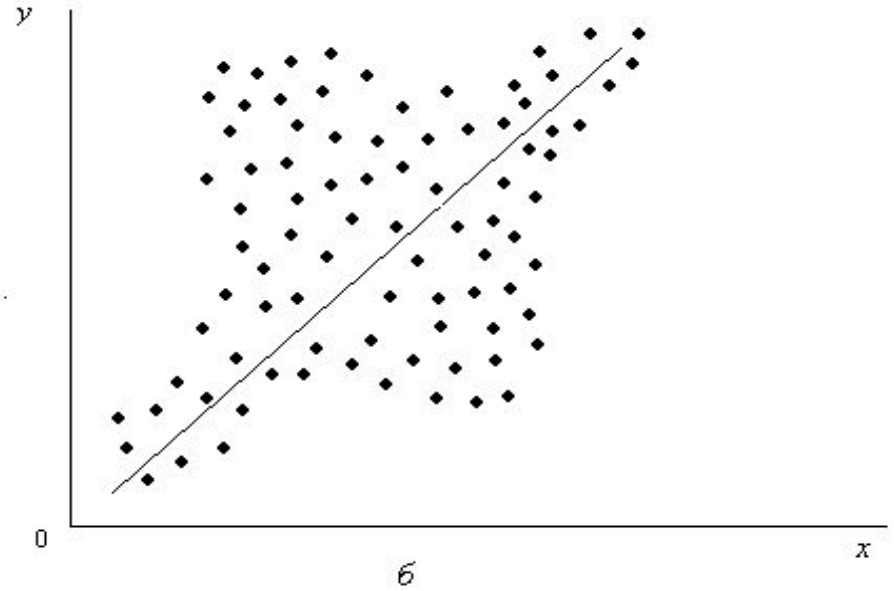
а) – остатки носят систематический характер, то есть отрицательные значения соответствуют низким значениям расчетных «у», а положительные – высоким;

б) – преобладание положительных остатков над отрицательными. В этих случаях необходимо применять либо другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии до тех пор, пока остатки не будут случайными величинами.

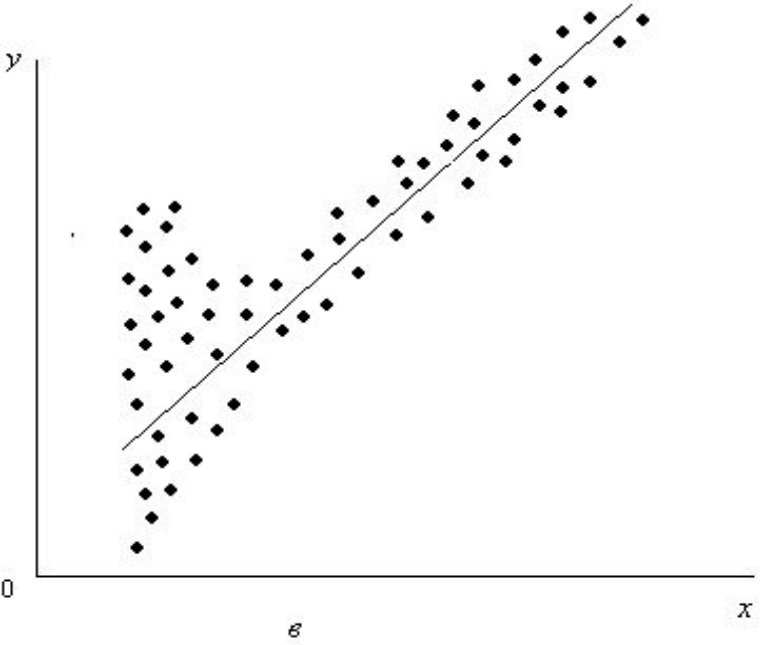
Вторая предпосылка МНК требует, чтобы дисперсия остатков была *гомоскедастичной*. Это значит, что для каждого значения фактора остатки имеют одинаковую дисперсию. Если это условие не соблюдается, то имеет место *гетероскедастичность*. Наличие гомо- или гетероскедастичности можно видеть по графику зависимости остатков от теоретических значений результативного признака (рис.5.2.3).



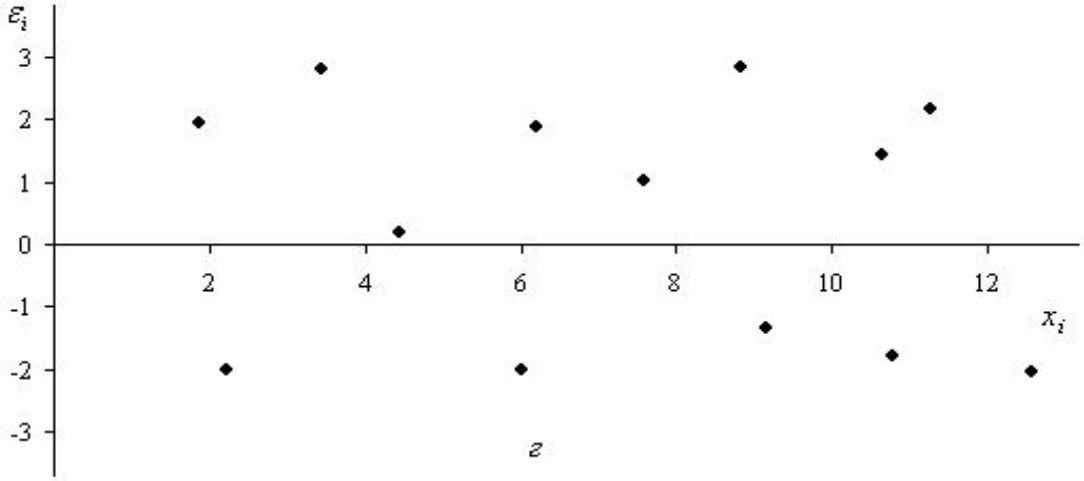
а



б



в



г

Рис. 5.2.3.

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

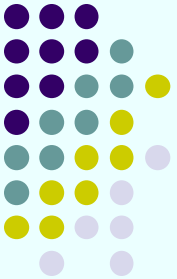
Отсутствие гетероскедастичность остатков (гомоскедастичность остатков, т.е. постоянство дисперсий остатков , для любого *i, i=1,…,n*) – важное условие (3 предпосылка), которое должно выполняться при проведении регрессионного анализа. Чтобы выявить гетероскедастичность остатков выборочной регрессии используют метод проверки статистических гипотез.
В качестве нулевой гипотезы предполагают отсутствие гетероскедастичности в генеральной совокупности, т.е.:

H_0: $\sigma_{\varepsilon_i}^2 = \sigma^2$

H_A: $\sigma_{\varepsilon_i}^2 \neq \sigma^2$

Для проверки данной гипотезы можно использовать разные тесты: Уайта, Глейзера, Спирмена, Голдфелда-Квандта и др.

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК



Тест ранговой корреляции Спирмена

предполагает, что остаточная дисперсия в генеральной совокупности – это некоторая функция от независимой переменной:

$$\sigma_\varepsilon^2 = f(x)$$

e_i – оценки σ_i, поэтому в случае гетероскедастичности абсолютные величины остатков (e_i) и значения регрессоров x_i будут коррелированы.

Для нахождения коэффициента ранговой корреляции $\rho_{x,e}$ следует ранжировать наблюдения по значениям переменной x_i и остатков e_i и вычислить коэффициент корреляции:

$$\rho_{x,e} = 1 - \frac{6\sum d_i^2}{n^3 - n},$$

где d_i – разность между рангами x_i и остатков e_i.

Коэффициент ранговой корреляции значим на уровне α при n>10, если статистика

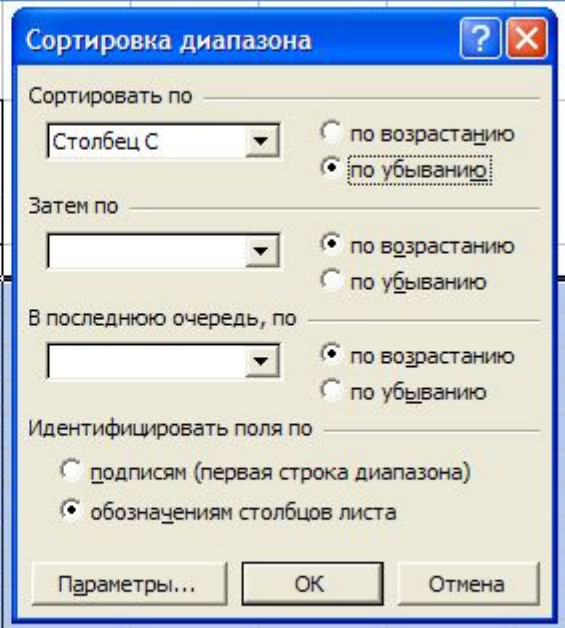
$$t = \frac{\rho_{x,e}\sqrt{n-2}}{\sqrt{1-\rho_{x,e}^2}} > t_{\alpha;n-2}$$

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

Тест Голдфелда-Квандта

1. Исходные данные сортируются по величине независимой переменной (нужно выделить весь диапазон значений зависимой и независимой переменной и произвести сортировку по убыванию x):

	A	B	C
1	1. Инвестиции в основной капитал и валовой региональный продукт на душу населения по отдельным регионам РФ		
2	№ региона	Валовой региональный продукт, тыс. руб., (y)	Инвестиции в основной капитал, руб., (x)
3		2002	2001
4	1	30,8	5514
5	2	38,3	5619
6	3	21,7	1873
7	4	42,6	4587
8	5	29,0	3199
9	6	78,8	11165
10	7	55,2	5599
11	8	49,3	6921
12	9	47,9	5195
13	10	43,0	4035
14	11	76,4	14709
15	12	36,5	7058



11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

2) Далее следует построить уравнение парной линейной регрессии y по x с использованием инструмента «Регрессия», при этом нужно предусмотреть вывод остатков и построение графика зависимости остатков от величины независимой переменной:

ВЫВОД ИТОГОВ

рессионная статистика	
Множеств	0,874486
R-квадрат	0,764726
Нормиров	0,741198
Стандартн	8,916321
Наблюден	12

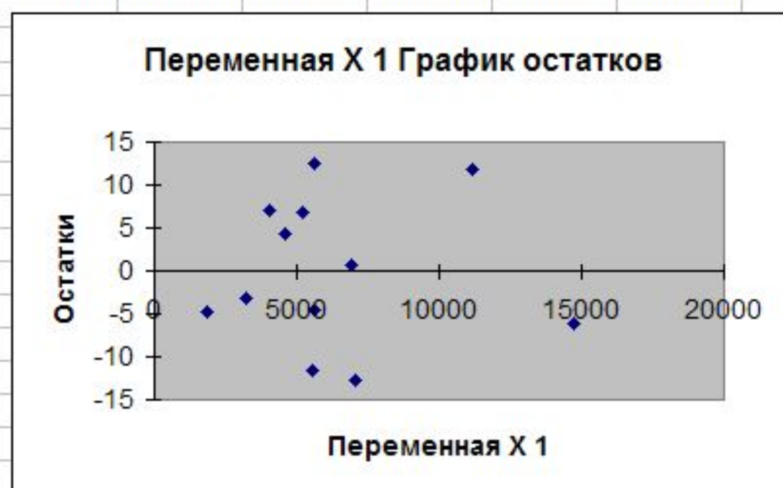
Дисперсионный анализ

	df	SS	MS	F	ачимость F
Регрессия	1	2584,057	2584,057	32,50354	0,000198
Остаток	10	795,0078	79,50078		
Итого	11	3379,065			

	Коэффициент	дартная ош	статисти	P-Значение	ижние 95%	ерхние 95%	ижние 95,0	рхние 95,0%
Y-пересеч	18,31898	5,463484	3,352984	0,007328	6,145573	30,49238	6,145573	30,49238
Переменн	0,004368	0,000766	5,701188	0,000198	0,002661	0,006076	0,002661	0,006076

ВЫВОД ОСТАТКА

аблюдение	дсказанно	Остатки
1	82,5739	-6,1739
2	67,09226	11,70774
3	49,15121	-12,6512
4	48,55273	0,747266
5	42,86507	-4,56507
6	42,7777	12,4223
7	42,40638	-11,6064
8	41,01286	6,88714
9	38,35687	4,272133
10	35,94551	7,054494
11	32,29352	-3,29352
12	26,50101	-4,80101



11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

3) Раздели совокупность на три равные части и по первым m наблюдениям и последним m наблюдениям определим суммы квадратов остатков: m=n/3=12/3=4

	A	B	C	D	E	F
19	ВЫВОД ОСТАТКА					
20						
21	*Наблюдение*	*Предсказанное Y*	*Остатки*		e_i^2	
22	1	82,57390192	-6,173901916		38,1	
23	2	67,09226084	11,70773916		137,1	
24	3	49,15120556	-12,65120556		160,1	
25	4	48,55273354	0,747266458		0,6	
26	5	42,86506518	-4,565065178	Итого	335,8	
27	6	42,777697	12,422303			
28	7	42,40638225	-11,60638225			
29	8	41,01285981	6,887140188			
30	9	38,35686721	4,272132788		18,3	
31	10	35,94550551	7,054494491		49,8	
32	11	32,29351568	-3,293515683		10,8	
33	12	26,50100551	-4,801005505		23,0	
34				Итого	101,9	
35						

$\sum_{i=1}^{m} e_i^2$

$\sum_{n-m-1}^{n} e_i^2$

4) Рассчитаем фактическое значение критерия Фишера:

$$F = \frac{\sum_{i=1}^{m} e_i^2}{\sum_{n-m-1}^{n} e_i^2} = \frac{335.8}{101.9} = 3.29$$

Определим его критическое значение , где p число параметров уравнения регрессии (для парной линейной регрессии p=2). Найдем критическое значение с помощь встроенной функции «FРАСПОБР()», в наем случае выполнение «FРАСПОБР(0,05;2;2)» дало значение 19,00.

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

5) Альтернативная гипотеза о наличии гетероскедастичности будет принята, если:

$$F = \frac{\sum_{i=1}^{m} e_i^2}{\sum_{n-m-1}^{n} e_i^2} > F_{\alpha;\nu_{n-p};\nu_{n-p}}$$

В нашем случае фактическое значение критерия Фишера (3,29) не превысило его критическое значение (19,00), таким образом, принимаем нулевую гипотезу о гомоскедастичности остатков уравнения парной линейной регрессии в генеральной совокупности. Следовательно, выполняется третья предпосылка регрессионного анализа и параметры уравнения могут быть оценены с помощь обычного метода наименьших квадратов.

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

Тест Глейзера

Тест Глейзера оценивает зависимость абсолютных значений остатков от значений фактора *x* в виде функции:

$$|e| = a + bx^c$$

, где *c* задается определенным числом степени. Обычно используются значения c, равные 1; 0,5; -1; -0,5.

Гипотеза о присутствии гетероскедастичности принимается в случае значимых значений b. Для аппроксимации гетероскедастичности выбирается функция с максимальным значением t_b

11. Гетероскедастичность. Тесты на гетероскедастичность. ВМНК

Тест Уайта

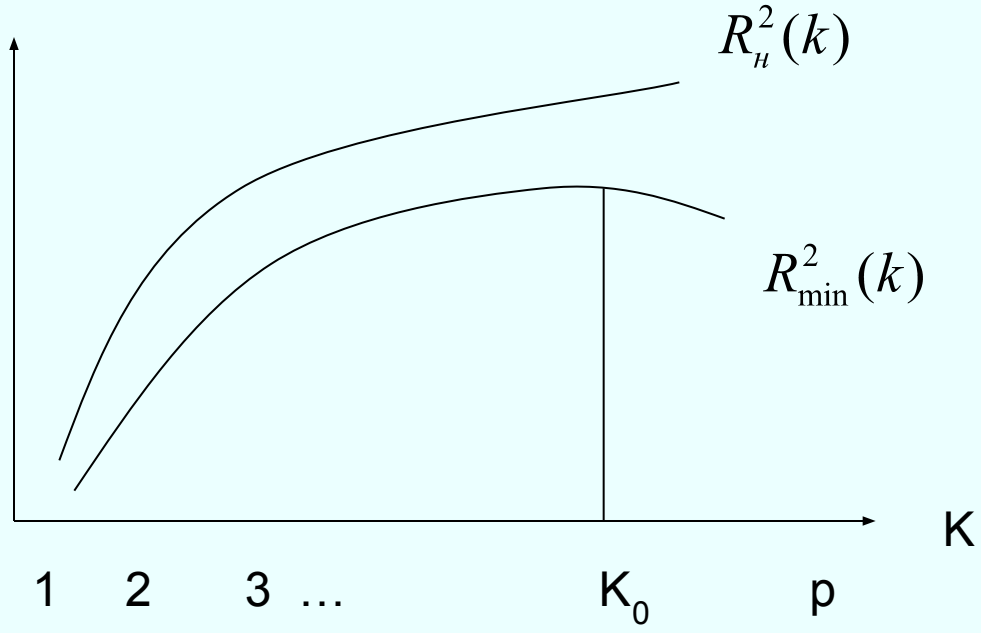
Тест Уайта предполагает, что дисперсия ошибок регрессии представляет собой квадратичную функцию от значений факторов. Тест Уайта для уравнения с двумя объясняющими переменными предполагает нахождение функции:

$$e^2 = a + b_1 x_1 + b_2 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + b_{12} x_1 x_2 + u$$

Гипотеза о присутствии гетероскедастичности принимается в случае значимости уравнения по критерию Фишера.

12. Отбор наиболее существенных объясняющих переменных в модель множественной регрессии

$$R^2_{\min} = R^2_{н(k)} - 2\sqrt{\frac{2k(n-k-1)}{(n-1)(n^2-1)}}(1-R^2_{(k)})$$



13. Уравнение регрессии с фиктивными переменными. Критерий Чоу

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \alpha_1 z_{i1} + \ldots \alpha_i z_{if-1} + \varepsilon_i,$$

где i=1…n, p – число факторных количественных переменных, f-число факторных качественных или категориальных переменных (число фиктивных переменных должно быть на единицу, чем число факторов)

$$y_i = \beta_0 + \beta_1 x_{i1} + \alpha_1 z_{i1} + \varepsilon_i,$$

где z_{i1}=

- **1, если домохозяйство расположено в городской местности**
- **0, если домохозяйство расположено в сельской местности**

Модель регрессии с фиктивными переменными для совокупности предприятий, по которой проведена типизация (выделены три группы):

$$y_i = \beta_0 + \beta_1 x_{i1} + \alpha_1 z_{i1} + \alpha_2 z_{i2} + \varepsilon_i,$$

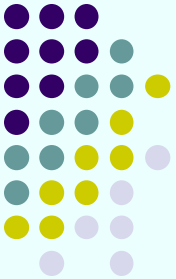
где z_{i1}=

- **1, если хозяйство принадлежит первой типической группе**
- **0, если хозяйство входит в другие группы**

где z_{i1}=

- **если хозяйство принадлежит второй типической группе**
- **0, если хозяйство входит в другие группы**

13. Уравнение регрессии с фиктивными переменными. Критерий Чоу



Data from the 2000 Census, US

	доход	возраст	пол
1	21393,00	22,50	м
2	32085,00	29,00	м
3	40741,00	39,50	м
4	44836,00	49,50	м
5	42313,00	59,50	м
6	32718,00	69,50	м
7	19108,00	22,50	ж
8	26788,00	29,00	ж
9	29520,00	39,50	ж
10	30960,00	49,50	ж
11	28360,00	59,50	ж
12	23261,00	69,50	ж

	доход	возраст	пол
1	21393,00	22,50	1
2	32085,00	29,00	1
3	40741,00	39,50	1
4	44836,00	49,50	1
5	42313,00	59,50	1
6	32718,00	69,50	1
7	19108,00	22,50	0
8	26788,00	29,00	0
9	29520,00	39,50	0
10	30960,00	49,50	0
11	28360,00	59,50	0
12	23261,00	69,50	0

13. Уравнение регрессии с фиктивными переменными. Критерий Чоу

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.	95,0%% доверительный интервал для B	
		B	Стд. Ошибка	Бета			Нижняя граница	Верхняя граница
1	(Константа)	19018,459	5849,140		3,251	,010	5786,786	32250,132
	пол	9348,167	3803,946	,596	2,457	,036	743,044	17953,289
	возраст	162,843	115,636	,341	1,408	,193	-98,743	424,430
2	(Константа)	26332,833	2818,919		9,341	,000	20051,891	32613,776
	пол	9348,167	3986,553	,596	2,345	,041	465,572	18230,761

a. Зависимая переменная: доход

Модель		Корреляции			Статистики коллинеарности	
		Нулевой порядок	Частные	Часть	Толерантность	КРД
1	(Константа)					
	пол	,596	,634	,596	1,000	1,000
	возраст	,341	,425	,341	1,000	1,000
2	(Константа)					
	пол	,596	,596	,596	1,000	1,000

13. Уравнение регрессии с фиктивными переменными. Критерий Чоу

$$H_0 : \beta' = \beta''; D(\varepsilon') = D(\varepsilon'') = \sigma^2$$

$$F = \frac{\left(\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n_1} e_i^2 - \sum_{i=n_1+1}^{n} e_i^2 \right) \cdot (n - 2p - 2)}{\left(\sum_{i=1}^{n_1} e_i^2 + \sum_{i=n_1+1}^{n} e_i^2 \right) \cdot (p + 1)} > F_{\alpha; p+1; n-2p-2}$$

.

где p – число параметров без свободного члена,

$\sum_{i=1}^{n} e_i^2$ **- остаточная сумма квадратов при построении модели по всей совокупности,**

$\sum_{i=1}^{n_1} e_i^2, \sum_{i=n_1+1}^{n} e_i^2$ **- остаточные суммы квадратов для первой и второй группы.**

13. Уравнение регрессии с фиктивными переменными. Критерий Чоу

$$y = b_0 + b_1 x + b_2 z + e$$

.

где y - средний балл по итогам контрольной недели, x - удельный вес пропущенных занятий

где z=

- **1, если студент обучается по специальности «Финансы и кредит»**
- **0, если студент обучается по специальности «Прикладная информатика»**

$$\underset{}{\tilde{Y}} = \underset{(0,00)}{4,385} - \underset{(0,00)}{0,042}\, x + \underset{(0,01)}{0,317}\, z$$

Тест Чоу:

y=4,573-0,0438x – общая модель
y=4,78-0,0476x – «Финансы и кредит»
y=4,339-0,04x – «Прикладная информатика»

$F_{факт}$= 4,15

$F_{крит}$= 3,18

14. Нелинейные модели множественной регрессии. Производственные функции

$$Q = AK^{\alpha} L^{1-\alpha}$$

Если объем производства Q будет постоянным, то дифференциал этой функции будет равен нулю:

$$dQ = 0 \quad \text{или}$$

$$\frac{\delta Q}{\delta K}\Delta K + \frac{\delta Q}{\delta L}\Delta L = 0, тогда$$

$$\Delta K = -\Delta L \frac{\delta Q}{\delta K} : \frac{\delta Q}{\delta L} или$$

$$\Delta K = -\frac{1-\alpha}{\alpha}\frac{K}{L}\Delta L$$

14. Нелинейные модели множественной регрессии. Производственные функции

$$Q = AK^{\alpha} L^{1-\alpha}$$

В относительных величинах мы имеем отношение соответствующих эластичностей:

$$\frac{\Delta K}{K} = -\frac{1-\alpha}{\alpha}\frac{\Delta L}{L}$$

- для компенсации измененияресурса труда на 1% следует изменить ресурс капитала на –(1-α)/α процентов.

Предельная норма замены трудовых ресурсов капиталом равна:

$$\frac{dK}{dL} = -\frac{1-\alpha}{\alpha}\frac{K}{L}$$

15. Вопросы для повторения и самостоятельного изучения

1. Классическая линейная модель множественной регрессии
2. Представление и отыскание параметров модели множественной регрессии в матричной форме
3. Ковариационная матрица дисперсий вектора оценок коэффициентов регрессии $\sum_b$, ее использование
4. Свойства оценок выборочных коэффициентов регрессии, полученных методом наименьших квадратов. Теорема Гаусса-Маркова
5. Обратная матрица $(X'X)^{-1}$ и ее использование во множественном регрессионном анализе
6. Оценка значимости множественной регрессии
7. Ошибки коэффициентов регрессии и прогноза в матричной форме
8. Ковариационная матрица вектора возмущений. Шестая предпосылка множественного регрессионного анализа в матричной форме
9. Понятие мультиколлинеарности факторов. Диагностика и способы устранения
10. Ридж-регрессия
11. Факторный анализ. Построение модели регрессии на главных компонентах
12. Коэффициент частной корреляции: понятие и способы расчета
13. Стандартизованные коэффициенты регрессии, коэффициенты раздельной детерминации
14. Понятие о гомо- и гетероскедастичности остатков. Последствия и подходы к выявлению гетероскедастичнсоти остатков
15. Тест Гольдфельда-Квандта
16. Тест Спирмена
17. Тест Бреуша-Пагана
18. Тест Уайта
19. Тест Глейзера
20. Тест Парка
21. Обобщенная линейная модель множественной линейной регрессии
22. Обобщенный метод наименьших квадратов
23. Взвешенный метод наименьших квадратов
24. Отбор факторов в модель регрессии. Пошаговые процедуры отбора
25. Частные уравнения регрессии, частные коэффициенты эластичности
26. Нелинейные модели множественной регрессии. Производственная функция Кобба-Дугласа, замена факторов