Множественная регрессия

$$\hat{y} = f\left(x_1, x_2, ..., x_p\right) \quad (1)$$

$$y = \alpha' + \beta_1' x_1 + \beta_2' x_2 + ... + \beta_p' x_p + \varepsilon \quad (2)$$

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p + e \quad (3)$$

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p \quad (4)$$

Матричный метод

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

$$Y = [y_1, y_2, \ldots, y_n]' \qquad B = (X'X)^{-1} X'Y \quad (10)$$

$$B = [a, b_1, b_2, \ldots, b_p]'$$

$$e = [e_1, e_2, \ldots, e_n]'$$

Семья	Накопления, ***S***	Доход, ***Y***	Имущество, ***W***
1	3	40	60
2	6	55	36
3	5	45	36
4	3,5	30	15
5	1,5	30	90

S=[3;6;5;3,5;1,5]’

B=[$a;b_1;b_2$]’

$$X=\begin{bmatrix}1 & 40 & 60\\ 1 & 55 & 36\\ 1 & 45 & 36\\ 1 & 30 & 15\\ 1 & 30 & 90\end{bmatrix}\quad X'X=\begin{bmatrix}5 & 200 & 237\\ 200 & 8450 & 9150\\ 237 & 9150 & 14517\end{bmatrix};\quad (X'X)^{-1}=\begin{bmatrix}5,6916 & -0,1074 & -0,0252\\ -0,1074 & 0,0024 & 0,00024\\ -0,0252 & 0,00024 & 0,00033\end{bmatrix}$$

$$B=(X'X)^{-1}X'Y=\begin{pmatrix}0,2787 & 0,1229 & -0,0294\end{pmatrix}$$

$$\hat{S}=0,2787+0,1229Y-0,0294W$$

Скалярный метод

$$\begin{cases} an & +b_1\sum x_1 & +b_2\sum x_2 & +\ldots+b_p\sum x_p & =\sum y \\ a\sum x_1 & +b_1\sum x_1^2 & +b_2\sum x_2x_1 & +\ldots+b_p\sum x_px_1 & =\sum yx_1 \\ \square & \square & \square & \square & \square \\ a\sum x_p & +b_1\sum x_1x_p & +b_2\sum x_2x_p & +\ldots+b_p\sum x_p^2 & =\sum yx_p \end{cases} \quad (11)$$

$$\begin{cases} an & +b_1\sum Y & +b_2\sum W & =\sum S \\ a\sum Y & +b_1\sum Y^2 & +b_2\sum WY & =\sum SY \\ a\sum W & +b_1\sum YW & +b_2\sum W^2 & =\sum SW \end{cases}$$

$$\begin{cases} 5a & +200b_1 & +237b_2 & =19 \\ 200a & +8450b_1 & +9150b_2 & =825 \\ 237a & +9150b_1 & +14517b_2 & =863{,}5 \end{cases}$$

$$\Delta = 6842700; \quad \Delta_a = 1903325; \quad \Delta_{b_1} = 840825; \quad \Delta_{b_2} = -201225.$$

$$a = \Delta_a / \Delta = 1903325 / 6842700 = 0{,}2787;$$

$$b_1 = \Delta_{b_1} / \Delta = 840825 / 6842700 = 0{,}1229;$$

$$b_2 = \Delta_{b_2} / \Delta = -201205 / 6842700 = -0{,}0294.$$

Регрессионная модель в стандартизованном масштабе

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \square + \beta_p t_{x_p} + \varepsilon \quad (12)$$

$$t_y = \frac{y - \bar{y}}{\sigma_y}; \quad t_{x_j} = \frac{x_j - \bar{x}_j}{\sigma_{x_j}}, \quad j = \overline{1, d} \quad (13)$$

$$\bar{t}_y = \bar{t}_{x_1} = \bar{t}_{x_2} = \ldots = \bar{t}_{x_p} = 0$$

$$\sigma_{t_y} = \sigma_{t_{x_j}} = 1, \; j = \overline{1, p}$$

$$\begin{cases} \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + \ldots + \beta_p r_{x_px_1} = r_{yx_1} \\ \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_3x_2} + \ldots + \beta_p r_{x_px_2} = r_{yx_2} \\ \square \quad \square \quad \square \quad \square \quad \square \\ \beta_1 r_{x_1x_p} + \beta_2 r_{x_2x_p} + \beta_3 r_{x_3x_p} + \ldots + \beta_p = r_{yx_p} \end{cases} \quad (14)$$

$$b_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}} \quad (15)$$

$$\hat{t}_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \square + \beta_p t_{x_p} \quad (16)$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p \quad (17)$$

$$R = \begin{bmatrix} 1 & -0,27149 & 0,873684 \\ -0,27149 & 1 & -0,68224 \end{bmatrix}$$

$$\begin{cases} \beta_1 + 0,27149\beta_2 = 0,873684, \\ -0,27149\beta_1 + \beta_2 = -0,68224 \end{cases}$$

Δ=0,926291; Δ_1=0,688461; Δ_2=-0,44504;

β_1=0,688461/0,926291=0,743245;

β_2=-0,44504/0,926291=-0,48045;

$$\hat{t}_S = 0,743245t_Y - 0,48045t_W$$

$$\sigma_S = 1{,}75357; \quad \sigma_Y = 10{,}6066; \quad \sigma_W = 28{,}6496,$$

$$\overline{S} = 3{,}8; \quad \overline{Y} = 40; \quad \overline{W} = 47{,}4.$$

$$b_1 = \beta_1 \frac{\sigma_y}{\sigma_{x_1}} = 0{,}743245 \cdot \frac{1{,}75357}{10{,}6066} = 0{,}1229;$$

$$b_2 = \beta_2 \frac{\sigma_y}{\sigma_{x_2}} = -0{,}48045 \cdot \frac{1{,}75357}{28{,}6496} = -0{,}0294;$$

$$a = \overline{s} - b_1\overline{Y} - b_2\overline{W} = 3{,}8 - 0{,}1229 \cdot 40 + 0{,}0294 \cdot 47{,}4 = 0{,}2787.$$

Частные уравнения регрессии

$$\begin{cases} y_{x_1 \cdot x_2, x_3, \ldots, x_p} = f(x_1), \\ y_{x_2 \cdot x_1, x_3, \ldots, x_p} = f(x_2), \\ \boxtimes \\ y_{x_p \cdot x_1, x_2, \ldots, x_{p-1}} = f(x_p). \end{cases}$$

$$y_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p} = a + b_1\bar{x}_1 + \ldots$$

$$+ b_{j-1}\bar{x}_{j-1} + b_j x_j + b_{j+1}\bar{x}_{j+1} + \ldots$$

$$+ b_p\bar{x}_p + e, \quad j = \overline{1,p}$$

$$\hat{y}_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p} = A_j + b_j x_j, \quad j = \overline{1,p}$$

$$A_j = a + b_1\bar{x}_1 + \ldots + b_{j-1}\bar{x}_{j-1} +$$

$$+ b_{j+1}\bar{x}_{j+1} + \ldots + b_p\bar{x}_p, \quad j = \overline{1,p}$$

$$Э_j = b_j \frac{x_j}{\hat{y}_{x_j \cdot x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p}} \qquad (18)$$

$$\overline{Э}_j = b_j \frac{\bar{x}_j}{\bar{y}} \qquad (19)$$

Анализ качества эмпирического уравнения множественной линейной регрессии

$$t_{b_j} = \frac{b_j}{m_{b_j}} \quad \left(t_a = \frac{a}{m_a} \right) \quad (20)$$

$$H_0 : \beta'_j = 0; \quad H_1 : \beta'_j \neq 0$$

$$\left|t_{b_j}\right| < t_{табл}(\alpha; n-p-1)$$

H_0 не отклоняется, параметр не значим

$$\left|t_{b_j}\right| \geq t_{табл}(\alpha; n-p-1)$$

H_0 отклоняется, параметр значим

$$b_j - t(\alpha; n-p-1)\cdot m_{b_j} < \beta_j' < b_j + t(\alpha; n-p-1)\cdot m_{b_j} \quad (24)$$

$$a - t(\alpha; n-p-1)\cdot m_a < \alpha' < a + t(\alpha; n-p-1)\cdot m_a \quad (24')$$

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \quad (25)$$

$$\bar{R}^2 = 1 - \frac{\sum e_i^2 / (n - p - 1)}{\sum (y_i - \bar{y})^2 / (n - 1)} \quad (26)$$

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

$$= \frac{n-1}{n-p-1} R^2 - \frac{p}{n-p-1} = R^2 - \frac{p}{n-p-1}(1 - R^2) \quad (27)$$

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p} \qquad (28)$$

$$H_0 : R^2 = 0 \qquad H_1 : R^2 > 0$$

$$F < F_{табл}(\alpha; p; n-p-1)$$

- H_0 не отклоняется, на уровне α уравнение не значимо

$$F \geq F_{òàáë}(\alpha; p; n-p-1)$$

- H_0 отклоняется, на уровне α уравнение значимо

Исключение k факторов:

$$\hat{y} = c + d_1 x_1 + d_2 x_2 + \ldots + d_{p-k} x_{p-k} \qquad (30)$$

$H_0 : R_1^2 - R_2^2 = 0$ исключение оправдано

$H_1 : R_1^2 > R_2^2$ исключение не оправдано

$$F = \frac{R_1^2 - R_2^2}{1 - R_1^2} \cdot \frac{n - p - 1}{k} \qquad (31) \qquad F_{кр} = F(\alpha; k; n - p - 1)$$

$F < F_{кр}$ H_0 не отклоняется, исключение оправдано

$F > F_{кр}$ H_0 отклоняется, исключение не оправдано

Включение k факторов:

$H_0 : R_1^2 - R_2^2 = 0$ включение не оправдано

$H_1 : R_2^2 > R_1^2$ включение оправдано

$$F = \frac{R_2^2 - R_1^2}{1 - R_2^2} \cdot \frac{n - p - 1}{k} \quad (32) \quad F_{кр} = F(\alpha; k; n - p - 1)$$

$F < F_{кр}$ H_0 не отклоняется, включение не оправдано

$F > F_{кр}$ H_0 отклоняется, включение оправдано

Частный случай: добавление одного фактора

$$F_{x_j} = \frac{R^2_{yx_1x_2\ldots x_p} - R^2_{yx_1x_2\ldots x_{j-1}x_{j+1}\ldots x_p}}{1 - R^2_{yx_1x_2\ldots x_p}} \frac{n-p-1}{1} \quad (50)$$

- частный *F* – критерий

$$F_{kp} = F_{tabl}(\alpha;1;n-p-1)$$

$$t_{b_j} = \sqrt{F_{x_j}} \quad (51)$$

Показатель множественной корреляции

$$R = \sqrt{1 - \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}} \qquad (33)$$

$$R = \sqrt{\sum_{j=1}^{p} \beta_j r_{yx_j}} \qquad (34)$$

Показатель частной корреляции

$$r_{yx_j \cdot x_1 x_2 \ldots x_{j-1} \ldots x_{j+1} \ldots x_p}$$

$$r_{x_i x_j \cdot y x_1 x_2 \ldots x_{i-1} \ldots x_{i+1} \ldots x_{j-1} \ldots x_{j+1} \ldots x_p}$$

$$r_{yx_j \cdot x_1 x_2 \ldots x_p} = \frac{r_{yx_j \cdot x_1 x_2 \ldots x_{p-1}} - r_{yx_p \cdot x_1 x_2 \ldots x_{p-1}} \cdot r_{x_j x_p \cdot x_1 x_2 \ldots x_{p-1}}}{\sqrt{\left(1 - r^2_{yx_p \cdot x_1 x_2 \ldots x_{p-1}}\right)\left(1 - r^2_{x_j x_p \cdot x_1 x_2 \ldots x_{p-1}}\right)}} \qquad (46)$$

Для двухфакторной модели

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}}$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right)\left(1 - r_{x_1x_2}^2\right)}}$$

Процедуры пошагового отбора переменных

- процедура последовательного присоединения
- процедура последовательного присоединения – удаления
- процедура последовательного удаления

Процедура «всех возможных регрессий»:

Для заданного значения k (k=1,2,…,p-1) проводится полный перебор всех возможных комбинаций из k факторов (которые отобраны из исходного набора x_1, x_2,…,x_p). Определяются такие переменные

$$x_{i_1}, x_{i_2}, \boxtimes, x_{i_k}$$

для которых коэффициент детерминации с результатом был бы максимальным.

Таким образом:

- на первом шаге (k=1) находим **один** наиболее информативный фактор (при условии, что в модель можно включать только **один** фактор из первоначального набора)
- на втором шаге (k=2) определяется уже наиболее информативная **пара** факторов (из первоначального набора), имеющая наиболее тесную связь с результатом
- на третьем шаге (k=3) будет отобрана наиболее информативная **тройка** факторов
- и т.д.

Критерий останова (завершения) процедуры:

- выбирается такое оптимальное число k_0 факторов, на котором нижняя доверительная граница коэффициента детерминации достигает своего **максимума**. Эта граница вычисляется по формуле:

$$R^2_{\min} = \overline{R}^2(k) - 2\sqrt{\frac{2k(n-k-1)}{(n-1)(n^2-1)}} \cdot \left(1 - R^2(k)\right),$$

Гетероскедастичность

-тест ранговой корреляции Спирмена

$$r_{x,e} = 1 - 6 \cdot \frac{\sum d_i^2}{n\left(n^2 - 1\right)} \quad (53)$$

$$H_0 : r_{x,e} = 0 \qquad H_1 : r_{x,e} \neq 0 \qquad t = \frac{r_{x,e}\sqrt{n-2}}{\sqrt{1 - r_{x,e}^2}} \quad (54)$$

$$|t| > t_{\text{òàáë}}\left(\alpha; n-2\right)$$

H_0 отклоняется, гетероскедастичность имеет место

-тест Голдфелда-Квандта

$$\sigma_i^2 = \sigma^2 x_{ji}^2, \quad i = \overline{1,n}$$

$$s_1 = \sum_{i=1}^{k} e_i^2 \qquad s_3 = \sum_{i=n-k+1}^{n} e_i^2$$

$$F_{íŕáë} = \frac{s_3 / (k-p-1)}{s_1 / (k-p-1)} = \frac{s_3}{s_1} \quad (53)$$

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2,$$

$$F_{набл} > F_{кр} = F_{табл}(\alpha; k-p-1; k-p-1)$$

-метод «взвешенных наименьших квадратов»

1.Дисперсии известны

$$y_i = a + bx_i + \varepsilon_i \quad (57)$$

$$\frac{y_i}{\sigma_i} = a\frac{1}{\sigma_i} + b\frac{x_i}{\sigma_i} + \frac{\varepsilon_i}{\sigma_i} \quad (58)$$

$$y_i^* = \frac{y_i}{\sigma_i}; \quad x_i^* = \frac{x_i}{\sigma_i}; \quad z_i = \frac{1}{\sigma_i}; \quad v_i = \frac{\varepsilon_i}{\sigma_i}; \quad (59)$$

$$y_i^* = az_i + bx_i^* + v_i \quad (60)$$

2.Дисперсии неизвестны

$$\sigma_i^2 = \sigma^2 x_i \quad (61)$$

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

$$\frac{y_i}{\sqrt{x_i}} = a\frac{1}{\sqrt{x_i}} + b\sqrt{x_i} + v_i \quad (62)$$

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

$$\frac{y_i}{\sqrt{\hat{y}_i}} = a\frac{1}{\sqrt{\hat{y}_i}} + b_1\frac{x_{i1}}{\sqrt{\hat{y}_i}} + \ldots + b_p\frac{x_{ip}}{\sqrt{\hat{y}_i}} + \frac{\varepsilon_i}{\sqrt{\hat{y}_i}} \quad (63)$$

$$\sigma_i^2 = \sigma^2 x_i^2$$

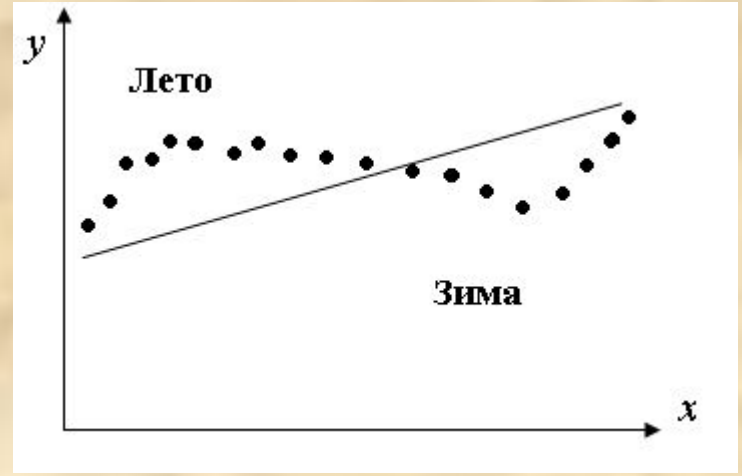
$$\frac{y_i}{x_i} = \frac{a}{x_i} + b + \frac{\varepsilon_i}{x_i}$$

$$\frac{y_i}{x_i} = a\frac{1}{x_i} + b + v_i \quad (64)$$

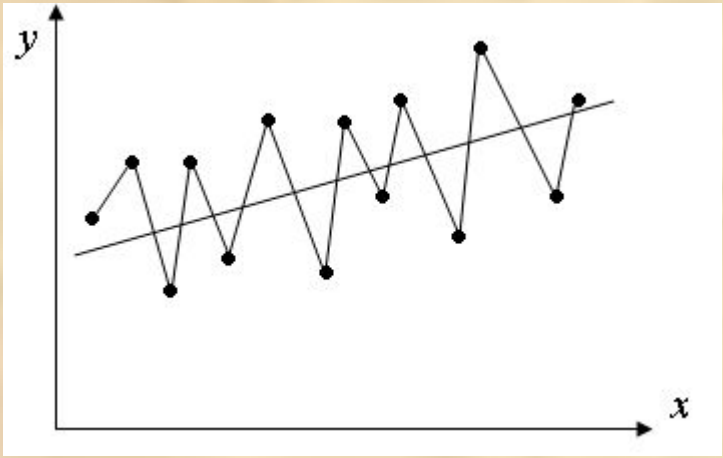
$$y_i^* = \frac{y_i}{x_i}; \quad z_i = \frac{1}{x_i}$$

Автокорреляция остатков

$\left(\mathrm{cov}\left(\varepsilon_{i-1}, \varepsilon_i\right) > 0\right)$ -положительная автокорреляция



$\left(\mathrm{cov}\left(\varepsilon_{i-1}, \varepsilon_i\right) < 0\right)$ -отрицательная автокорреляция



-метод рядов

$$M(k)=\frac{2n_1n_2}{n_1+n_2}+1;$$

$$D(k)=\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

$$M(k)-u_{\alpha/2}\cdot\sqrt{D(k)}<k<M(k)+u_{\alpha/2}\cdot\sqrt{D(k)}$$

-нет автокорреляции

$k\le M(k)-u_{\alpha/2}\cdot\sqrt{D(k)}$ -положительная автокорреляция

$k\ge M(k)+u_{\alpha/2}\cdot\sqrt{D(k)}$ -отрицательная автокорреляция

-критерий Дарбина - Уотсона

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad (65)$$

$0 \le DW \le d_l$ - положительная автокорреляция

$d_l < DW < d_u$ - зона неопределенности

$d_u \le DW \le 4 - d_u$ - автокорреляция отсутствует;

$4 - d_u < DW < 4 - d_l$ - зона неопределенности;

$4 - d_l \le DW \le 4$ - отрицательная автокорреляция.

$$DW \approx 2\left(1 - r_{e_{t-1}e_t}\right) \quad (67)$$

-авторегрессионная схема 1-го порядка AR(1)

$$e_t = \rho e_{t-1} + v_t \quad (68) \qquad \hat{\rho} = 1 - DW/2$$

$$y = a + bx + e \quad (71)$$

$$y_t = a + bx_t + e_t \quad (72)$$

$$y_{t-1} = a + bx_{t-1} + e_{t-1} \quad (73)$$

$$y_t - \rho y_{t-1} = a(1-\rho) + b\left(x_t - \rho x_{t-1}\right) + \left(e_t - \rho e_{t-1}\right) \quad (74)$$

$$y_t^* = y_t - \rho y_{t-1}; \quad x_t^* = x_t - \rho x_{t-1}; \quad a^* = a(1-\rho) \quad (75)$$

$$y_t^* = a^* + bx_t^* + v_t \quad (76)$$

$\rho = 1$:

$$y_t - y_{t-1} = b(x_t - x_{t-1}) + (e_t - e_{t-1})$$

$$\Delta y_t = b\Delta x_t + v_t$$

$$\Delta y_t = y_t - y_{t-1}, \quad \Delta x_t = x_t - x_{t-1}$$

$\rho = -1$:

$$y_t + y_{t-1} = 2a + b(x_t + x_{t-1}) + v_t$$

$$\frac{y_t + y_{t-1}}{2} = a + b\frac{(x_t + x_{t-1})}{2} + v_t$$

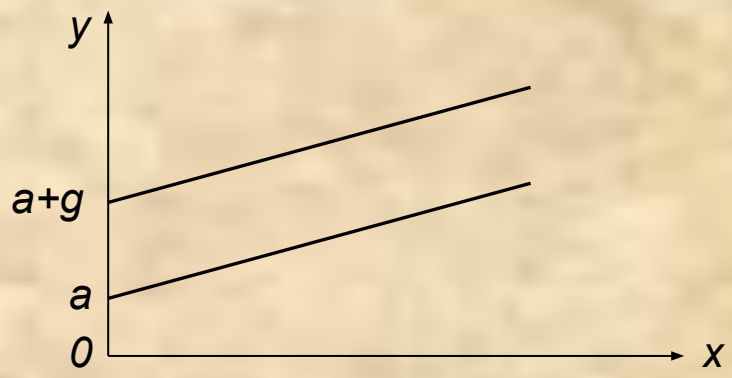
Фиктивные переменные в регрессионных моделях

$$y = a + gD + e \qquad \hat{y} = a + g \cdot 0 = a \qquad \hat{y} = a + g \cdot 1 = a + g$$

$$y = a + bx + gD + e, \quad (80)$$

$$D = \begin{cases} 0, & \text{если} \quad \textit{сотрудник} - \textit{женщина} \\ 1, & \text{если} \quad \textit{сотрудник} - \textit{мужчина}. \end{cases}$$

$$\hat{y} = a + bx, \qquad \hat{y} = a + bx + g = (a + g) + bx.$$



$$D_1 = \begin{cases} 0, & \text{страна не являетс развитой} \\ 1, & \text{страна развитая} \end{cases}$$

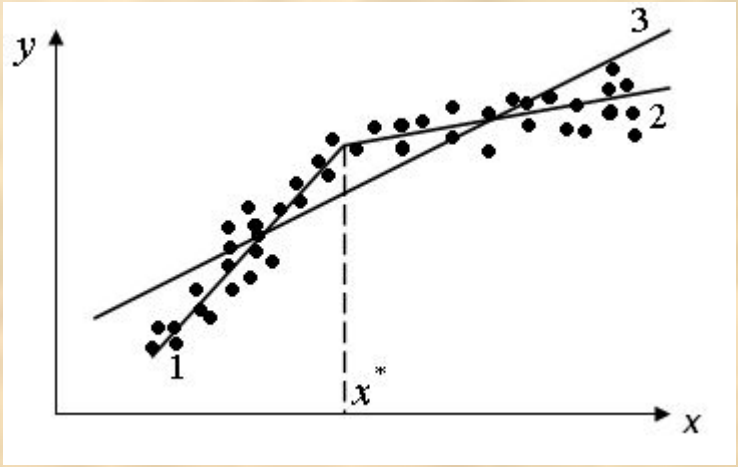
$$D_2 = \begin{cases} 0, & \text{страна не является развивающейся} \\ 1, & \text{страна развивающаяся} \end{cases}$$

$$y = a + bx + g_1 D + g_2 Dx + e \quad (81)$$

$$D = \begin{cases} 0, & \textit{до изменения условий}, \\ 1, & \textit{после изменения условий}. \end{cases}$$

$$\hat{y} = a + bx, \qquad D = 0$$

$$\hat{y} = (a + g_1) + (b + g_2)x \qquad , D = 1$$



-тест Чоу

$$F_{íŕáë} = \frac{s_3 - (s_1 + s_2)}{s_1 + s_2} \cdot \frac{n - 2p - 2}{p + 1} \quad (82)$$

$$F_{набл} < F(\alpha; p+1; n-2p-2)$$

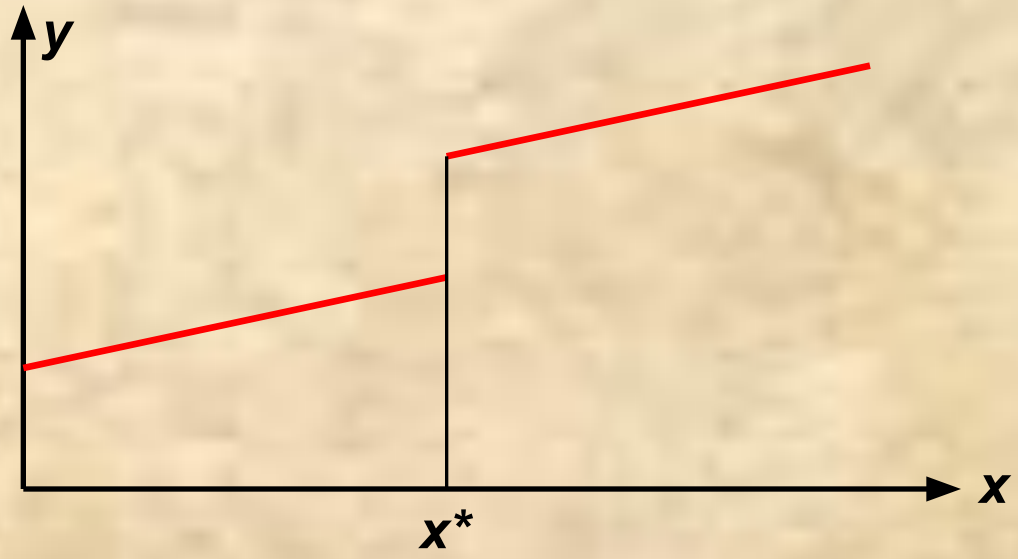
$$F_{набл} > F(\alpha; p+1; n-2p-2)$$

Данные в подвыборках описываются двумя уравнениями регрессии:

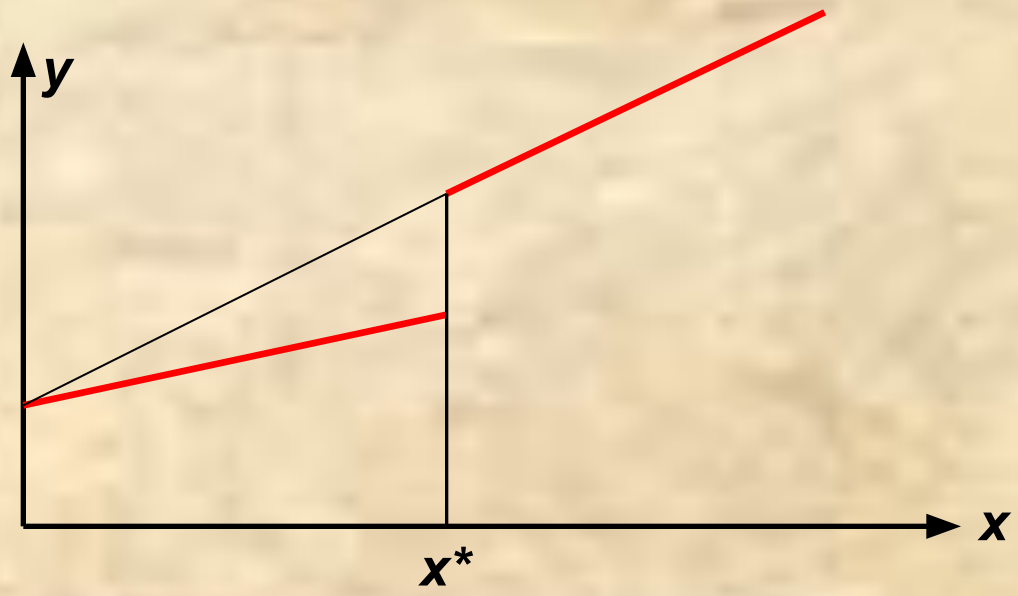
$$\hat{y} = a_1 + b_1 x,$$

$$\hat{y} = a_2 + b_2 x.$$

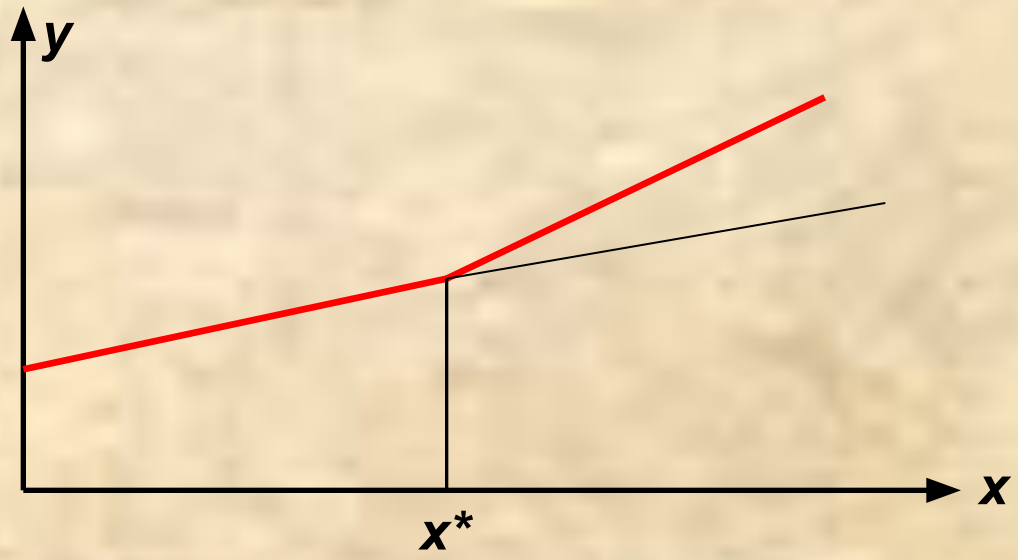
1. Различие между a_1 и a_2 значимо, между b_1 и b_2 – нет:



2. Различие между b_1 и b_2 значимо, между a_1 и a_2 – нет:



3. Различия между b_1 и b_2, а также между a_1 и a_2 значимы:



- методика Гуйарати:

$$y = a + bD + cx + dDx + e$$

$$a_1 = (a + b); \quad b_1 = (c + d) \quad (D = 1)$$

$$a_2 = a; \quad b_2 = c \quad (D = 0)$$