Семинар 4. (Решение).

1. Пусть регрессионная модель $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, $i = 1, \ldots, n$, задана в матричном виде при помощи уравнения $y = X\beta + \varepsilon$, где $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}'$. Известно, что $E(\varepsilon) = 0$ и $Var(\varepsilon) = \sigma^2 \cdot I$. Известно также, что:

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Для удобства расчётов ниже приведены матрицы:

$$X'X = \begin{pmatrix} 5 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \text{ и } (X'X)^{-1} = \begin{pmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 4/3 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

Найдите:

(a) Рассчитайте при помощи метода наименьших квадратов оценку для вектора неизвестных коэффициентов.

Решение:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X'X)^{-1}X'y = \begin{pmatrix} 2 \\ 2 \\ 1 \end{pmatrix}.$$

(b) Рассчитайте несмещенную оценку для неизвестного параметра σ^2_ε регрессионной модели.

Решение:

$\hat{\sigma}^2_\varepsilon = \frac{RSS}{n-k} = \frac{2}{5-3} = 1.$

(c) Рассчитайте $\widehat{Var}(\hat{\beta})$, оценку для ковариационной матрицы вектора МНК-коэффициентов $\widehat{\beta}$.

Решение:

$$\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2_\varepsilon (X'X)^{-1} = 1 \cdot \begin{pmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 4/3 & -1 \\ 0 & -1 & 2 \end{pmatrix} = \begin{pmatrix} 1/3 & -1/3 & 0 \\ -1/3 & 4/3 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

(d) Сформулируйте основную и альтернативную гипотезы, которые соответствуют тесту на значимость переменной x_2 в уравнении регрессии.

Решение:

$H_0 : \beta_2 = 0$ (коэффициент при переменной x_2 незначим)

$H_1 : \beta_2 \neq 0$ (коэффициент при переменной x_2 значим)

(e) Протестируйте на значимость переменную x_2 в уравнении регрессии на уровне значимости 10%:

i. Приведите формулу для тестовой статистики.

Решение:

$t = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}}; n = 5; k = 3.$

ii. Укажите распределение тестовой статистики при верной H_0.

Решение:

$t \sim t(n-k); n = 5; k = 3.$

iii. Вычислите наблюдаемое значение тестовой статистики.

$t_{obs} = \frac{\hat{\beta}_2 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} = \frac{\hat{\beta}_2 - 0}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}} = \frac{2-0}{\sqrt{\frac{2}{5-3}1.3333}} = 1.7321.$

iv. Укажите границы области, где основная гипотеза не отвергается.

Решение:

Нижняя граница равна -2.920, верхняя граница равна 2.920.

v. Сделайте статистический вывод о значимости переменной x_2.

Решение:

Поскольку $t_{obs} = 1.7321$, что принадлежит промежутку от -2.920 до 2.920, то на основе имеющихся данных нельзя отвергнуть основную гипотезу о незначимсоти коэффициента при переменной x_2 на уровне значимости 10%.

(f) Найдите p-значение, соответствующее наблюдаемому значению тестовой статистики (t_{obs}) из предыдущего пункта. На основе полученного p-значения сделайте вывод о значимости переменной x_2.

Решение:

$p - value = P(|t| > |t_{obs}|) = 2F_t(-|t_{obs}|)$, где $F_t(-|t_{obs}|)$ — функция распределения t-распределения с $n - k = 5 - 3 = 2$ степенями свободы в точке $-|t_{obs}|$.

$p - value = P(|t| > |t_{obs}|) = 2F_t(-|t_{obs}|) = 0.2253$. Поскольку p-значение превосходит уровень значимости 10%, то основная гипотеза — $H_0 : \beta_2 = 0$ не может быть отвергнута.

(g) Постройте 90%-ый доверительный интервал для оценки коэфициента β_2.

Решение:

Доверительный интервал для коэффициента $beta_j$ в общем виде имеет следующий вид:

$$\hat{\beta}_j - t_{(1-\alpha/2;n-k)}\hat{\sigma}_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + t_{1-\alpha/2;n-k)}\hat{\sigma}_{\hat{\beta}_j}.$$

Тогда для нашей задачи доверительный интервал для β_2 имеет следующий вид:

$$\hat{\beta}_2 - t_{(1-0.05;5-3)}\hat{\sigma}_{\hat{\beta}_2} \leq \beta_2 \leq \hat{\beta}_2 + t_{1-0.05;5-3)}\hat{\sigma}_{\hat{\beta}_2}.$$

$$2 - 2.92 \cdot 1.13333 \leq \beta_2 \leq 2 + 2.92 \cdot 1.13333.$$

$$-1.893 \le \beta_2 \le 5.893.$$

2. Используя матрицы $M = X(X'X)^{-1}X'$ и $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$, запишите TSS, RSS и ESS в матричной форме. (Решили полностью на семинаре.)

3. Вася оценил исходную модель:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Для надежности Вася стандартизировал переменные, т.е. перешёл к $y_i^* = (y_i - \bar{y})/\hat{\sigma}_y$ и $x_i^* = (x_i - \bar{x})/\hat{\sigma}_x$. Затем Вася оценил ещё две модели:

$$y_i^* = \beta_1' + \beta_2' x_i^* + \varepsilon_i'$$

и

$$y_i^* = \beta_2'' x_i^* + \varepsilon_i''.$$

В решении можно считать $\hat{\sigma}_x$ и $\hat{\sigma}_y$ известными.

(a) Найдите $\hat{\beta}_1'$.
Решение:
$\hat{\beta}_1' = \bar{y}^* - \hat{\beta}_2'\bar{x}^* = 0$

(b) Как связаны между собой $\hat{\beta}_2$, $\hat{\beta}_2'$ и $\hat{\beta}_2''$?
Решение:
$\hat{\beta}_2' = \frac{\widehat{Cov}(y^*,x^*)}{\widehat{Var}(x^*)} = \frac{\widehat{Cov}(y-\bar{y},x-\bar{x})\hat{\sigma}_x/\hat{\sigma}_y}{\widehat{Var}(x-\bar{x})} = \frac{\widehat{Cov}(y,x)\hat{\sigma}_x/\hat{\sigma}_y}{\widehat{Var}(x)} = \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\hat{\beta}_2$
$\hat{\beta}_2'' = \frac{\widehat{Cov}(y^*,x^*)}{\widehat{Var}(x^*)} = \hat{\beta}_2' = \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\hat{\beta}_2$

(c) Как связаны между собой e_i, e_i' и e_i''?
Решение:
$\hat{u}_i' = y_i^* - \hat{y}_i^* = y_i^* - \hat{\beta}_1' - \hat{\beta}_2' x_i^* = y_i^* - \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\hat{\beta}_2 x_i^* = \frac{y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x})}{\hat{\sigma}_y} = \frac{y_i - \hat{\beta}_2 x_i - (\bar{y} - \hat{\beta}_2\bar{x})}{\hat{\sigma}_y} = \frac{y_i - \hat{\beta}_2 x_i - \hat{\beta}_1}{\hat{\sigma}_y} = \frac{\hat{u}_i}{\hat{\sigma}_y}$
$\hat{u}_i'' = y_i^* - \hat{y}_i^* = y_i^* - \hat{\beta}_2'' x_i^* = y_i^* - \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\hat{\beta}_2' x_i^* = \frac{y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x})}{\hat{\sigma}_y} = \frac{y_i - \hat{\beta}_2 x_i - (\bar{y} - \hat{\beta}_2\bar{x})}{\hat{\sigma}_y} = \frac{y_i - \hat{\beta}_2 x_i - \hat{\beta}_1}{\hat{\sigma}_y} = \frac{u_i}{\hat{\sigma}_y} = \hat{u}_i'$

(d) Как связаны между собой $\widehat{Var}\left(\hat{\beta}_2\right)$, $\widehat{Var}\left(\hat{\beta}_2'\right)$ и $\widehat{Var}\left(\hat{\beta}_2''\right)$?
Решение:
$RSS = \sum_{i=1}^n \hat{u}_i^2 = \hat{\sigma}_y^2 \sum_{i=1}^n \hat{u}_i'^2 = \hat{\sigma}_y^2 RSS'$, $RSS' = \sum_{i=1}^n \hat{u}_i'^2 = RSS''$.
$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$
$X'X_{new} = \begin{pmatrix} n & \sum_{i=1}^n (x_i - \bar{x})/\hat{\sigma}_x \\ \sum_{i=1}^n (x_i - \bar{x})/\hat{\sigma}_x & \sum_{i=1}^n (x_i - \bar{x})^2/\hat{\sigma}_x^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2/\hat{\sigma}_x^2 \end{pmatrix}$

$\widehat{Var}(\hat\beta_2) = \frac{RSS}{n-2}(X'X)^{-1}_{(2,2)} = \frac{RSS' s_y^2}{n-2}\frac{1}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}\begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}_{(2,2)} =$
$\frac{RSS'\hat\sigma_y^2}{n-2}\frac{n}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{RSS'\hat\sigma_y^2}{n-2}\frac{n/\hat\sigma_x^2}{(n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2)/\hat\sigma_x^2} = \widehat{Var}(\hat\beta_2')\frac{\hat\sigma_y^2}{\hat\sigma_x^2}$
$\widehat{Var}(\hat\beta_2') = \frac{RSS'}{n-2}(X'X_{new})^{-1}_{(2,2)} = \frac{RSS''}{n-1}(X'X_{new})^{-1}_{(2,2)}\frac{n-1}{n-2} = \widehat{Var}(\hat\beta_2'')\frac{n-1}{n-2}$

(e) Как выглядит матрица $\widehat{Var}\left(\hat\beta'\right)$?
Решение:
$\widehat{Var}(\hat\beta') = \frac{RSS'}{n-2}(X'X_{new})^{-1} = \frac{\sum_{i=1}^n (y_i^* - \hat\beta_1' - \hat\beta_2' x_i^*)^2}{n-2}\begin{pmatrix} (\sum_{i=1}^n x_i^2 - n\bar x^2)/\hat\sigma_x^2 & 0 \\ 0 & n \end{pmatrix}$ где
$\hat\beta' = (X'X_{new})^{-1}\begin{pmatrix} 0 \\ \sum_{i=1}^n (x_i - \bar x) y_i \end{pmatrix}$, в частности, $\hat\beta_1' = 0$.

(f) Как связаны между собой t-статистики $t_{\hat\beta_2}$, $t_{\hat\beta_2'}$ и $t_{\hat\beta_2''}$?
Решение:

$t_{\hat\beta_2} = \frac{\hat\beta}{\sqrt{\widehat{Var}(\hat\beta_2)}} = \frac{\hat\beta'\hat\sigma_y/\hat\sigma_x}{\sqrt{\widehat{Var}(\hat\beta_2')}\hat\sigma_y/\hat\sigma_x} = t_{\hat\beta_2'} = \sqrt{\frac{n-2}{n-1}}t_{\hat\beta_2''}$

(g) Как связаны между собой R^2, $R^{2\prime}$ и $R^{2\prime\prime}$?
Решение:

$TSS' = TSS'' = \frac{\sum_{i=1}^n (y_i - \bar y)^2}{\hat\sigma_y^2} = \frac{TSS}{\hat\sigma_y^2}$.
$R'^2 = R''^2$, так как соответствующие TSS и RSS равны.
$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\hat\sigma_y^2 RSS'}{TSS'\hat\sigma_y^2} = 1 - \frac{RSS'}{TSS'} = R'^2 = R''^2$.

(h) В нескольких предложениях прокомментируйте последствия перехода к стандартизированным переменным.
Решение:
При переходе к стандартизированным переменным оценка свободного члена становится равной нулю. Оценки коэффициентов наклона и остатки регрессии изменяются пропорционально стандартным отклонениям переменным. Однако при выкидывании свободного члена из регрессии слегка изменяется оценка дисперсии коэффициента наклона. Коэффициент детерминации R^2 не изменяется.