Семинар 5.
Множественная регрессия.
Стандартизированные данные. Тестирование гипотез.

1. 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:
$\hat{\beta}_1' = \bar{y}^* - \hat{\beta}_2'\bar{x}^* = 0$, так как $\bar{y}^* = \bar{x}^* = 0$.

(b) Как связаны между собой $\hat{\beta}_2$, $\hat{\beta}_2'$ и $\hat{\beta}_2''$?
Решение:
$\hat{\beta}_2' = \frac{\widehat{Cov}(y_i^*, x_i^*)}{\widehat{Var}(x_i^*)} = \frac{\widehat{Cov}((y_i-\bar{y})/\hat{\sigma}_y^2, (x_i-\bar{x})/\hat{\sigma}_x^2)}{\widehat{Var}((x_i-\bar{x})/\sigma_x)} = \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\frac{\widehat{Cov}((y_i, x_i)}{\widehat{Var}(x_i)} = \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\hat{\beta}_2$
$\hat{\beta}_2'' = \frac{\sum_{i=1}^n y_i^* x_i^*}{\sum_{i=1}^n x_i^{*2}} = \frac{\sum_{i=1}^n (x_i-\bar{x})/\hat{\sigma}_x (y_i-\bar{y})/\hat{\sigma}_y}{\sum_{i=1}^n ((x_i-\bar{x})/\hat{\sigma}_x)^2} = \frac{\hat{\sigma}_x}{\hat{\sigma}_y}\hat{\beta}_2$.

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

(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:
$RSS = \sum_{i=1}^n e_i^2 = \hat{\sigma}_y^2 \sum_{i=1}^n e_i'^2 = \hat{\sigma}_y^2 RSS'$, $RSS' = \sum_{i=1}^n e_i'^2 = RSS''$.
$X'X = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}$
$X'X_{new} = \begin{pmatrix} n & \sum_{i=1}^n (x_i - \bar{x})/\hat{\sigma}_x \\ \sum_{i=1}^n (x_i - \bar{x})/\hat{\sigma}_x & \sum_{i=1}^n (x_i - \bar{x})^2/\hat{\sigma}_x^2 \end{pmatrix} = \begin{pmatrix} n & 0 \\ 0 & \sum_{i=1}^n (x_i - \bar{x})^2/\hat{\sigma}_x^2 \end{pmatrix}$

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

(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.

2. Рассмотрим модель $y_i = \beta_1+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\varepsilon_i$. При оценке модели по 24 наблюдениям оказалось, что $RSS = 15$, $\sum(y_i-\bar y-x_{i3}+\bar{x_3})^2 = 20$. На уровне значимости 1% протестируйте гипотезу

$$H_0: \begin{cases}\beta_2+\beta_3+\beta_4 = 1\\ \beta_2 = 0\\ \beta_3 = 1\\ \beta_4 = 0\end{cases}.$$

Решение:

Заметим, что в основной гипотезе есть линейно зависимые ограничения, оставим только линейно независимые:

$$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 1 \\ \beta_4 = 0 \end{cases}$$

Ограниченная модель имеет вид:

$$y_i = \beta_1 + x_{i3} + \varepsilon_i$$

Переносим x_{3i} в левую часть, и получим оценку коэффициента β_1:

$$\hat{\beta}_1 = \bar{y} - \bar{x}_3$$

Теперь можно найти RSS_R:

$$RSS_R = \sum_{i=1}^{24}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{24}(y_i - \bar{y} + \bar{x}_3 - x_{3i})^2 = 20$$

Осталось найти значение F-статистики, которая при верной H_0 имеет распределение $F_{3,20}$:

$$F_{obs} = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n - k_{UR})} = \frac{(20 - 15)/3}{15/(24 - 4)} = 20/9$$

Так как $F_{obs} < F_{3,20;0.99} = 4.94$, оснований отвергать нулевую гипотезу нет.