Семинар 5.

1. Рассмотрим модель $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i$. При оценке модели по 24 наблюдениям оказалось, что $RSS = 15$, $\sum(y_i - \bar{y} - x_{i2} + \bar{x_2})^2 = 20$. 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 зависимые ограничения, оставим только независимые:

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

Ограниченная модель имеет вид:

$$y_i = \beta_1 + w_i + \varepsilon_i$$

Переносим w_i в левую часть, и получим оценку коэффициента β_1:

$$\hat{\beta}_1 = \bar{y} - \bar{w}$$

Теперь можно найти RSS_R:

$$RSS_R = \sum_{i=1}^{24}(y_i - \hat{\tilde{y}}_i)^2 = \sum_{i=1}^{24}(y_i - \bar{y} + \bar{w} - w_i)^2 = 20$$

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.

2. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, где $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, $y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$, $E(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$.

   На уровне значимости 5% проверьте гипотезу $H_0 : \beta_1 + \beta_2 = 2$ против альтернативной $H_a : \beta_1 + \beta_2 \neq 2$:

   (a) Приведите формулу для тестовой статистики.

   (b) Укажите распределение тестовой статистики при верной H_0.

   (c) Вычислите наблюдаемое значение тестовой статистики.

   (d) Укажите границы области, где основная гипотеза не отвергается.

   (e) Сделайте статистический вывод.

   Решение:

   (a) $T = \frac{\hat{\beta}_1 + \hat{\beta}_2 - (\beta_1 + \beta_2)}{\sqrt{\widehat{Var}(\hat{\beta}_1 + \hat{\beta}_2)}}$, где $\widehat{Var}(\hat{\beta}_1 + \hat{\beta}_2) = \widehat{Var}(\hat{\beta}_1) + \widehat{Var}(\hat{\beta}_2) + 2\widehat{Cov}(\hat{\beta}_1; \hat{\beta}_2) = \hat{\sigma}^2[(X'X)^{-1}]_{11} + 2\hat{\sigma}^2[(X'X)^{-1}]_{12} + \hat{\sigma}^2[(X'X)^{-1}]_{22} = \frac{RSS}{n-k}([(X'X)^{-1}]_{11} + 2[(X'X)^{-1}]_{12} + [(X'X)^{-1}]_{22}), \beta_1 + \beta_2 = 2$

   (b) $T \sim t_{n-k}; n = 5; k = 3$

   (c) $\hat{\beta} = (X'X)^{-1}X'y = (1.5\ 2.0\ 1.5)'$. $\widehat{Var}(\hat{\beta}_1 + \hat{\beta}_2) = \frac{1}{2}(0.5 + 1 + 2 \cdot (-0.5)) = \frac{1}{4}$. $T = \frac{1.5+2-2}{\sqrt{\frac{1}{4}}} = 3$.

   (d) Так как проверяется знак «равно» в гипотезе, то нижняя граница $-\infty$, а верхняя граница $+\infty$.

   (e) $t_{0.95,2} = 2.92 < T$, значит, гипотеза отвергается на уровне значимости 5%.

3. На основе опроса 25 человек была оценена следующая модель зависимости логарифма зарплаты (lnW) от уровня образования (Edu, в годах) и возраста (Age).

$$\widehat{lnW} = 1.7 + 0.5Edu + 0.06Age - 0.0004Age^2,$$
$$ESS = 90.3, RSS = 60.4.$$

Когда в модель были введены переменные $Fedu$ и $Medu$, учитывающие уровень образования родителей, величина ESS увеличилась до 110.3.

(a) Напишите спецификацию уравнения регрессии с учетом образования родителей.

(b) Сформулируйте и проверьте гипотезу о значимом влиянии уровня образования родителей на зарплату (уровень значимости 5%).

Решение:

Ограниченная модель (Restricted model):

$$\ln W_i = \beta_1 + \beta_2 Edu_i + \beta_3 Age_i + \beta_4 Age_i^2 + \varepsilon_i$$

Неограниченная модель (Unrestricted model):

$$\ln W_i = \beta_1 + \beta_2 Edu_i + \beta_3 Age_i + \beta_4 Age_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i$$

По условию $ESS_R = 90.3$, $RSS_R = 60.4$, $TSS = ESS_R + RSS_R = 90.3 + 60.4 = 150.7$.
Также сказано, что $ESS_{UR} = 110.3$.
Значит, $RSS_{UR} = TSS - ESS_{UR} = 150.7 - 110.3 = 40.4$

(a) Спецификация:

$$\ln W_i = \beta_1 + \beta_2 Edu_i + \beta_3 Age_i + \beta_4 Age_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i$$

(b) Проверка гипотезы:

- $H_0 : \begin{cases} \beta_5 = 0 \\ \beta_6 = 0 \end{cases}$
  $H_1 : \beta_5^2 + \beta_6^2 > 0$
- $T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k_{UR})}$, где $q = 2$ — число линейно независимых уравнений в основной гипотезе H_0, $n = 25$ — число наблюдений, $k = 6$ — число коэффициентов в модели без ограничений.
- $T \sim F(q; n - k_{UR})$
- $T_{obs} = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k_{UR})} = \frac{(60.4-40.4)/2}{40.4/(25-6)} = 4.70$
- $F_{crit} = F_{2,19;0.95} = 3.52$.
- Поскольку $T_{obs} = 4.70 < 3.52$, то на основе имеющихся данных можно отвергнуть основную гипотезу на уровне значимости 5%. Таким образом, образование родителей существенно влияет на заработную плату.

4. Рассмотрим следующую модель зависимости цены дома $Price$ (в тысячах долларов), от его площади $Hsize$ (в м^2), площади участка $Lsize$ (в м^2), числа ванных комнат $Bath$ и числа спален BDR:

$$\widehat{Price} = \hat{\beta}_1 + \hat{\beta}_2 Hsize + \hat{\beta}_3 Lsize + \hat{\beta}_4 Bath + \hat{\beta}_5 BDR, R^2 = 0.218, n = 23.$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы: $H_0 : \beta_4 = 20\beta_5$. Дайте интерпретацию проверяемой гипотезе. Для регрессии с ограничениями был вычислен коэффициент $R^2_R = 0.136$. Протестируйте нулевую гипотезу на уровне значимости 5%.

Решение:

$$Price_i = \beta_1 + \beta_2 Hsize_i + \beta_3 Lsize_i + 20\beta_5 Bath_i + \beta_5 BDR_i + \varepsilon_i$$

Число ванных комнат в 20 раз сильнее влияет на цену дома, чем число спален.

$$\begin{cases} R^2 = \frac{ESS}{TSS} \\ TSS = ESS + RSS \\ TSS_R = TSS_{UR} = TSS \end{cases}$$

$$\begin{cases} RSS_R = TSS(1 - R^2_R) \\ RSS_{UR} = TSS(1 - R^2_{UR}) \end{cases} \rightarrow$$

$$\begin{cases} F_{obs} = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(n-k_{UR})} = \frac{(0.218-0.136)/1}{(1-0.218)/18} = 1.887 \\ F_{crit} = F_{1,18;0.95} = 4.41 \end{cases}$$

$F_{obs} < F_{crit}$ и, следовательно, H_0 не отвергается на уровне значимости 5%. Вывод: гипотеза H_0 о том, что число ванных комнат в 20 раз сильнее влияет на цену дома, чем число спален, не отвергается на уровне значимости 5%.

5. В файле dataflats.xlsx хранятся данные о стоимости квартир в Москве (тыс.долл.).

   (a) Оцените следующие модели регрессии для стоимости одного квадратного метра жилья:

$$price\_sq_i = \beta_1 + \beta_2 livesp_i + \beta_3 dist_i + \varepsilon_i,$$
$$price\_sq_i = \beta_1 + \beta_2 livesp_i + \beta_3 dist_i + \beta_4 metrdist_i + \varepsilon$$

   (b) Для построенных моделей проверьте гипотезу о незначимости модели в целом.

   (c) Используя $p - value$ коэффициентов, укажите для каждой из моделей, какие из переменных являются значимыми, а какие – незначимыми?

(d) Проинтерпретируйте оценки коэффициентов при значимых переменных. Согласуются ли знаки данных оценок с интуицией и здравым смыслом?

(e) Постройте 90%-ые доверительные интервалы для коэффициентов обеих моделей.

(f) Для каждой из моеделей роверьте гипотезу $H_0 : \beta_3 = -0.1$. Содержательно проинтерпретируйте результаты тестирования.

(g) Для второй модели проверьте гипотезу $H_0 : \beta_2 + \beta_4 = 0$. Содержательно проинтерпретируйте результаты тестирования.

(h) Переоцените модели регрессии на шакалированных данных. Какой из факторов оказывает наибольшее влияние на стоимость квартиры?

(i) Какие из трех оцененных моделе могут быть сравнены по значению коэффициента детерминации R^2? Выполните сравнение.

6. Домашнее задание. [Борзых Д.А., Вакуленко Е.С., Фурманов К.К. Эконометрика: РАБОТА С ДАННЫМИ НА КОМПЬЮТЕРЕ. ПРАКТИКУМ: Элементы теории. Практические задания. Ответы и решения].

Оценивается зависимость количества продаваемых чебуреков (в штуках) – qch от цены на чебуреки (в рублях) – pch, цены на шаурму (в рублях) – psh и цены на мороженое (в рублях) – $pmor$ в виде линейной регрессии:

$$qch_i = \beta_1 + \beta_2 pch_i + \beta_3 psh_i + \beta_4 pmor_i + \varepsilon_i, i = 1, ..., n.$$

В файле "$Regression\_9.xlsx$" приведены данные. Выполните следующие задания. Используйте 5%-ый уровень значимости.

(a) Оцените данное уравнение регрессии и выпишите оцененное значение.

(b) Является ли полученное уравнение регрессии значимым?

(c) Используя $p-value$ коэффициентов, укажите, какие из переменных являются значимыми, а какие – незначимыми?

(d) Проинтерпретируйте оценки коэффициентов при значимых переменных. Согласуются ли знаки данных оценок с интуицией и здравым смыслом?

(e) Протестируйте гипотезу $H_0 : \beta_2 = -5$ против альтернативной гипотезы $H_0 : \beta_2 - 5$.

(f) Протестируйте гипотезу $H_0 : \beta_2 = -5$ против альтернативной гипотезы $H_0 : \beta_2 < -5$.

(g) Протестируйте гипотезу $H_0 : \beta_2 = -5$ против альтернативной гипотезы $H_0 : \beta_2 > -5$.

(h) Проверьте гипотезу: "чем выше цена на шаурму, тем больше в среднем продается чебуреков". Подразумевается, что все прочие условия являются неизменными.