Эконометрика

с Монте-Карло и эконометрессами

в задачах и упражнениях

Дмитрий Борзых, Борис Демешев

18 ноября 2013 г.

Содержание

Todo list

1 МНК без матриц и вероятностей

1. Верно ли, что для любых векторов $a = (a_1, \ldots, a_n)$ и $b = (b_1, \ldots, b_n)$ справедливы следующие равенства?
   (a) $\sum_{i=1}^{n} (a_i - \bar{a}) = 0$
   (b) $\sum_{i=1}^{n} (a_i - \bar{a})^2 = \sum_{i=1}^{n} (a_i - \bar{a})a_i$
   (c) $\sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b}) = \sum_{i=1}^{n} (a_i - \bar{a})b_i$
   (d) $\sum_{i=1}^{n} (a_i - \bar{a})(b_i - \bar{b}) = \sum_{i=1}^{n} a_i b_i$

   да, да, да, нет
2. При помощи метода наименьших квадратов найдите оценку неизвестного параметра θ в следующих моделях:
   (a) $y_i = \theta + \theta x_i + \varepsilon_i$
   (b) $y_i = \theta - \theta x_i + \varepsilon_i$
   (c) $\ln y_i = \theta + \ln x_i + \varepsilon_i$
   (d) $y_i = \theta + x_i + \varepsilon_i$
   (e) $y_i = 1 + \theta x_i + \varepsilon_i$
   (f) $y_i = \theta / x_i + \varepsilon_i$
   (g) $y_i = \theta x_{i1} + (1 - \theta) x_{i2} + \varepsilon_i$
3. Покажите, что для моделей $y_i = \alpha + \beta x_i + \varepsilon_i$, $z_i = \gamma + \delta x_i + \upsilon_i$ и $y_i + z_i = \mu + \lambda x_i + \xi_i$ МНК-оценки связаны соотношениями $\hat{\mu} = \hat{\alpha} + \hat{\gamma}$ и $\hat{\lambda} = \hat{\beta} + \hat{\delta}$.
4. Найдите МНК-оценки параметров α и β в модели $y_i = \alpha + \beta y_i + \varepsilon_i$.
5. Рассмотрите модели $y_i = \alpha + \beta(y_i + z_i) + \varepsilon_i$, $z_i = \gamma + \delta(y_i + z_i) + \varepsilon_i$.
   (a) Как связаны между собой $al\hat{p}ha$ и $ga\hat{m}ma$?
   (b) Как связаны между собой $b\hat{e}ta$ и $de\hat{l}ta$?

   $\hat{\alpha} + \hat{\gamma} = 0$ и $\hat{\beta} + \hat{\delta} = 1$
6. Как связаны МНК-оценки параметров α, β и γ, δ в моделях $y_i = \alpha + \beta x_i + \varepsilon_i$ и $z_i = \gamma + \delta x_i + \upsilon_i$, если $z_i = 2y_i$.
7. Для модели $y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ решите условную задачу о наименьших квадратах: $Q(\beta_1, \beta_2) := \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2 \to \min\limits_{\beta_1 + \beta_2 = 1}$
8. Даны n пар чисел: (x_1, y_1), ..., (x_n, y_n). Мы прогнозируем y_i по формуле $\hat{y}_i = \hat{\beta} x_i$. Найдите $\hat{\beta}$ методом наименьших квадратов. $\hat{\beta} = \sum x_i y_i / \sum x_i^2$
9. Даны n чисел: y_1, ..., y_n. Мы прогнозируем y_i по формуле $\hat{y}_i = \hat{\beta}$. Найдите $\hat{\beta}$ методом наименьших квадратов. $\hat{\beta} = \bar{y}$
10. Даны n пар чисел: (x_1, y_1), ..., (x_n, y_n). Мы прогнозируем y_i по формуле $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$. Найдите $\hat{\beta}_1$ и $\hat{\beta}_2$ методом наименьших квадратов. $\hat{\beta}_2 = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (x_i - \bar{x})^2$, $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$
11. Даны n пар чисел: (x_1, y_1), ..., (x_n, y_n). Мы прогнозируем y_i по формуле $\hat{y}_i = 1 + \hat{\beta} x_i$. Найдите $\hat{\beta}$ методом наименьших квадратов. $\hat{\beta} = \sum x_i (y_i - 1) / \sum x_i^2$
12. Перед нами два золотых слитка и весы, производящие взвешивания с ошибками. Взвесив первый слиток, мы получили результат 300 грамм, взвесив второй слиток — 200 грамм, взвесив оба слитка — 400 грамм. Оцените вес каждого слитка методом наименьших квадратов. $(300 - \hat{\beta}_1)^2 + (200 - \hat{\beta}_2)^2 + (400 - \hat{\beta}_1 - \hat{\beta}_2)^2 \to \min$
13. Аня и Настя утверждают, что лектор опоздал на 10 минут. Таня считает, что лектор опоздал на 3 минуты. С помощью мнк оцените на сколько опоздал лектор. $2 \cdot (10 - \hat{\beta})^2 + (3 - \hat{\beta})^2 \to \min$

14. Функция $f(x)$ дифференциируема на отрезке $[0;1]$. Найдите аналог МНК-оценок для регрессии без свободного члена в непрерывном случае. Более подробно: найдите минимум по $\hat{\beta}$ для функции

$$Q(\hat{\beta}) = \int_0^1 (f(x) - \hat{\beta}x)^2\, dx \tag{1}$$

15. Есть двести наблюдений. Вовочка оценил модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ по первой сотне наблюдений. Петечка оценил модель $\hat{y} = \hat{\gamma}_1 + \hat{\gamma}_2 x$ по второй сотне наблюдений. Машенька оценила модель $\hat{y} = \hat{m}_1 + \hat{m}_2 x$ по всем наблюдениям.

    (a) Возможно ли, что $\hat{\beta}_2 > 0$, $\hat{\gamma}_2 > 0$, но $\hat{m}_2 < 0$?

    (b) Возможно ли, что $\hat{\beta}_1 > 0$, $\hat{\gamma}_1 > 0$, но $\hat{m}_1 < 0$?

    (c) Возможно ли одновременное выполнение всех упомянутых условий?

    да, возможно. Два вытянутых облачка точек. Первое облачко даёт первую регрессию, второе — вторую. Прямая, соединяющая центры облачков, — общую.

16. Вася оценил модель $y = \beta_1 + \beta_2 d + \beta_3 x + \varepsilon$. Дамми-переменная d обозначает пол, 1 для мужчин и 0 для женщин. Оказалось, что $\hat{\beta}_2 > 0$. Означает ли это, что для мужчин $\bar{y}$ больше, чем $\bar{y}$ для женщин? Нет. Коэффициенты можно интерпретировать только «при прочих равных», т.е. при равных x. Из-за разных x может оказаться, что у мужчин $\bar{y}$ меньше, чем $\bar{y}$ для женщин.

17. Какие из указанные моделей можно представить в линейном виде?

    (a) $y_i = \beta_1 + \frac{\beta_2}{x_i} + \varepsilon_i$

    (b) $y_i = \exp(\beta_1 + \beta_2 x_i + \varepsilon_i)$

    (c) $y_i = 1 + \frac{1}{\exp(\beta_1 + \beta_2 x_i + \varepsilon_i)}$

    (d) $y_i = \frac{1}{1 + \exp(\beta_1 + \beta_2 x_i + \varepsilon_i)}$

    (e) $y_i = x_i^{\beta_2} e^{\beta_1 + \varepsilon_i}$

18. У эконометриста Вовочки есть переменная 1_f, которая равна 1, если i-ый человек в выборке — женщина, и 0, если мужчина. Есть переменная 1_m, которая равна 1, если i-ый человек в выборке — мужчина, и 0, если женщина. Какие $\hat{y}$ получатся, если Вовочка попытается построить регрессии:

    (a) y на константу и 1_f

    (b) y на константу и 1_m

    (c) y на 1_f и 1_m без константы

    (d) y на константу, 1_f и 1_m

19. У эконометриста Вовочки есть три переменных: r_i — доход i-го человека в выборке, m_i — пол (1 — мальчик, 0 — девочка) и f_i — пол (1 — девочка, 0 — мальчик). Вовочка оценил две модели

Модель A $m_i = \beta_1 + \beta_2 r_i + \varepsilon_i$

Модель B $f_i = \gamma_1 + \gamma_2 r_i + u_i$

    (a) Как связаны между собой оценки $\hat{\beta}_1$ и $\hat{\gamma}_1$?

    (b) Как связаны между собой оценки $\hat{\beta}_2$ и $\hat{\gamma}_2$?

    Оценки МНК линейны по объясняемой переменной. Если сложить объясняемые переменные в этих двух моделях, то получится вектор из единичек. Если строить регрессию вектора из единичек на константу и r, то получатся оценки коэффициентов 1 и 0. Значит, $\hat{\beta}_1 + \hat{\gamma}_1 = 1$, $\hat{\beta}_2 + \hat{\gamma}_2 = 0$

20. Эконометрист Вовочка оценил линейную регрессионную модель, где y измерялся в тугриках. Затем он оценил ту же модель, но измерял y в мунгу (1 тугрик = 100 мунгу). Как изменятся оценки коэффициентов? Увеличатся в 100 раз
21. Возможно ли, что при оценке парной регрессии $y = \beta_1 + \beta_2 x + \varepsilon$ оказывается, что $\hat{\beta}_2 > 0$, а при оценке регрессии без константы, $y = \gamma x + \varepsilon$, оказывается, что $\hat{\gamma} < 0$? да
22. Эконометрист Вовочка оценил регрессию y только на константу. Какой коэффициент R^2 он получит? $R^2 = 0$
23. Эконометрист Вовочка оценил методом наименьших квадратов модель 1, $y = \beta_1 + \beta_2 x + \beta_3 z + \varepsilon$, а затем модель 2, $y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 w + \varepsilon$. Сравните полученные ESS, RSS, TSS и R^2. $TSS_1 = TSS_2$, $R^2_2 \geqslant R^1_2$, $ESS_2 \geqslant ESS_1$, $RSS_2 \leqslant RSS_1$
24. Создайте набор данных с тремя переменными y, x и z со следующими свойствами. При оценке модели $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ получается $\hat{\beta}_2 > 0$. При оценке модели $\hat{y} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z$ получается $\hat{\gamma}_2 < 0$. Объясните принцип, руководствуясь которым легко создать такой набор данных.
25. У меня есть набор данных с выборочным средним $\bar{y}$ и выборочной дисперсией s^2_y. Как нужно преобразовать данные, чтобы выборочное среднее равнялось 7, а выборочная дисперсия — 9? $y^*_i = 7 + 3(y_i - \bar{y})/s_y$

2 Парный МНК без матриц

1. Рассмотрим модель $y_t = \beta_1 + \beta_2 \cdot t + \varepsilon_t$, где ошибки ε_t независимы и равномерны на $[-1; 1]$. С помощью симуляций на компьютере оцените и постройте график функции плотности для $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{s}^2$, $\widehat{\mathrm{Var}}(\hat{\beta}_1)$, $\widehat{\mathrm{Var}}(\hat{\beta}_2)$ и $\widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$.
2. Пусть $y_i = \mu + \varepsilon_i$, где $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$. Найдите:
   (a) $\mathbb{E}(\overline{y})$
   (b) $\mathrm{Var}(\overline{y})$
   (c) $\mathbb{E}(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2)$
   (d) $\mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2)$, если дополнительно известно, что ε_i нормально распределены
3. Рассматривается модель $y_i = \beta x_i + \varepsilon_i$, $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$. При каких значениях параметров c_i несмещённая оценка $\hat{\beta} = \frac{\sum_{i=1}^{n} c_i y_i}{\sum_{i=1}^{n} c_i x_i}$ имеет наименьшую дисперсию?

   $c_i = c \cdot x_i$, где $c \neq 0$
4. Пусть $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ и $i = 1, \ldots, 5$ — классическая регрессионная модель. Также имеются следующие данные: $\sum_{i=1}^{5} y_i^2 = 55, \sum_{i=1}^{5} x_i^2 = 3, \sum_{i=1}^{5} x_i y_i = 12, \sum_{i=1}^{5} y_i = 15, \sum_{i=1}^{5} x_i = 3$. Используя их, найдите:
   (a) $\hat{\beta}_1$ и $\hat{\beta}_2$
   (b) $\mathrm{Corr}(\hat{\beta}_1, \hat{\beta}_2)$
   (c) TSS
   (d) ESS
   (e) RSS
   (f) R^2
   (g) $\hat{\sigma}^2$

   Проверьте следующие гипотезы:

   (a) $\begin{cases} H_0 : \beta_2 = 2 \\ H_a : \beta_2 \neq 2 \end{cases}$

(b) $\begin{cases} H_0 : \beta_1 + \beta_2 = 1 \\ H_a : \beta_1 + \beta_2 \neq 1 \end{cases}$

5. Пусть $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ и $i = 1, \ldots, 5$ — классическая регрессионная модель. Также имеются следующие данные: $\sum_{i=1}^5 y_i^2 = 55, \sum_{i=1}^5 x_i^2 = 2, \sum_{i=1}^5 x_i y_i = 9, \sum_{i=1}^5 y_i = 15, \sum_{i=1}^5 x_i = 2$. Используя их, найдите:

   (a) $\hat{\beta}_1$ и $\hat{\beta}_2$

   (b) $\mathrm{Corr}(\hat{\beta}_1, \hat{\beta}_2)$

   (c) TSS

   (d) ESS

   (e) RSS

   (f) R^2

   (g) $\hat{\sigma}^2$

   Проверьте следующие гипотезы:

   (a) $\begin{cases} H_0 : \beta_2 = 2 \\ H_a : \beta_2 \neq 2 \end{cases}$

   (b) $\begin{cases} H_0 : \beta_1 + \beta_2 = 1 \\ H_a : \beta_1 + \beta_2 \neq 1 \end{cases}$

6. Рассмотрите классическую линейную регрессионную модель $y_i = \beta x_i + \varepsilon_i$. Найдите $\mathbb{E}\hat{\beta}$. Какие из следующих оценок параметра β являются несмещенными:

   (a) $\hat{\beta} = \frac{y_1}{x_1}$

   (b) $\hat{\beta} = \frac{1}{2}\frac{y_1}{x_1} + \frac{1}{2}\frac{y_n}{x_n}$

   (c) $\hat{\beta} = \frac{1}{n}\frac{y_1}{x_1} + \ldots + \frac{y_n}{x_n}$

   (d) $\hat{\beta} = \frac{\overline{y}}{\overline{x}}$

   (e) $\hat{\beta} = \frac{y_n - y_1}{x_n - x_1}$

   (f) $\hat{\beta} = \frac{1}{2}\frac{y_2 - y_1}{x_2 - x_1} + \frac{1}{2}\frac{y_n - y_{n-1}}{x_n - x_{n-1}}$

   (g) $\hat{\beta} = \frac{1}{n}\frac{y_2 - y_1}{x_2 - x_1} + \frac{1}{n}\frac{y_3 - y_2}{x_3 - x_2} + \ldots + \frac{1}{n}\frac{y_n - y_{n-1}}{x_n - x_{n-1}}$

   (h) $\hat{\beta} = \frac{1}{n-1}\frac{y_2 - y_1}{x_2 - x_1} + \frac{y_3 - y_2}{x_3 - x_2} + \ldots + \frac{y_n - y_{n-1}}{x_n - x_{n-1}}$

   (i) $\hat{\beta} = \frac{x_1 y_1 + \ldots + x_n y_n}{x_1^2 + \ldots + x_n^2}$

   (j) $\hat{\beta} = \frac{1}{2}\frac{y_n - y_1}{x_n - x_1} + \frac{1}{2n}\frac{y_1}{x_1} + \ldots + \frac{y_n}{x_n}$

   (k) $\hat{\beta} = \frac{1}{2}\frac{y_n - y_1}{x_n - x_1} + \frac{1}{2}\frac{x_1 y_1 + \ldots + x_n y_n}{x_1^2 + \ldots + x_n^2}$

   (l) $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x^2})^2}$

   (m) $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \overline{x})(\overline{y} - y_i)}{\sum_{i=1}^n (x_i - \overline{x^2})^2}$

   (n) $\hat{\beta} = \frac{y_1 + 2y_2 + \ldots + n y_n}{x_1 + 2x_2 + \ldots + n x_n}$

   (o) $\hat{\beta} = \frac{\sum_{i=1}^n i(y_i - \overline{y})}{\sum_{i=1}^n i(x_i - \overline{x})}$

   (p) $\hat{\beta} = \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}$

(q) $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\overline{y}}{x_i-\overline{x}}$

7. Рассмотрите классическую линейную регрессионную модель $y_i = \beta x_i + \varepsilon_i$. Найдите $\mathrm{Var}(\hat{\beta})$.

   (a) $\hat{\beta} = \frac{y_1}{x_1}$

   (b) $\hat{\beta} = \frac{1}{2}\frac{y_1}{x_1} + \frac{1}{2}\frac{y_n}{x_n}$

   (c) $\hat{\beta} = \frac{1}{n}\frac{y_1}{x_1} + \ldots + \frac{y_n}{x_n}$

   (d) $\hat{\beta} = \frac{\overline{y}}{\overline{x}}$

   (e) $\hat{\beta} = \frac{y_n-y_1}{x_n-x_1}$

   (f) $\hat{\beta} = \frac{1}{2}\frac{y_2-y_1}{x_2-x_1} + \frac{1}{2}\frac{y_n-y_{n-1}}{x_n-x_{n-1}}$

   (g) $\hat{\beta} = \frac{x_1y_1+\ldots+x_ny_n}{x_1^2+\ldots+x_n^2}$

   (h) $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x^2})^2}$

   (i) $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(\overline{y}-y_i)}{\sum_{i=1}^{n}(x_i-\overline{x^2})^2}$

   (j) $\hat{\beta} = \frac{y_1+2y_2+\ldots+ny_n}{x_1+2x_2+\ldots+nx_n}$

   (k) $\hat{\beta} = \frac{\sum_{i=1}^{n} i(y_i-\overline{y})}{\sum_{i=1}^{n} i(x_i-\overline{x})}$

   (l) $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}$

   (m) $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i-\overline{y}}{x_i-\overline{x}}$

8. Рассмотрите классическую линейную регрессионную модель $y_i = \beta\cdot i + \varepsilon_i$, $i = 1,\ldots,n$. Какая из оценок $\hat{\beta}$ и $\tilde{\beta}$ является более эффективной?

   (a) $\hat{\beta} = y_1$ и $\tilde{\beta} = y_2/2$

   (b) $\hat{\beta} = y_1$ и $\tilde{\beta} = \frac{1}{2}y_1 + \frac{1}{2}\frac{y_2}{2}$

   (c) $\hat{\beta} = \frac{1}{n}\frac{y_1}{1} + \ldots + \frac{y_n}{n}$ и $\tilde{\beta} = \frac{1\cdot y_1+\ldots+n\cdot y_n}{1^2+\ldots+n^2}$

9. На основе 100 наблюдений была оценена функция спроса:

$$\underset{(s.e.)}{\widehat{\ln Q}} = \underset{(0.04)}{0.87} - \underset{(0.02)}{1.23}\ln P$$

   Значимо ли коэффициент эластичности спроса по цене отличается от -1? Рассмотрите уровень значимости 5%.

10. На основе 100 наблюдений была оценена функция спроса:

$$\underset{(s.e.)}{\widehat{\ln Q}} = \underset{(0.04)}{2.87} - \underset{(0.02)}{1.12}\ln P$$

   На уровне значимости 5% проверьте гипотезу $H_0 : \beta_{\ln P} = -1$ против альтернативной $H_a : \beta_{\ln P} < -1$. Дайте экономическую интерпретацию проверяемой гипотезе и альтернативе.

11. Используя годовые данные с 1960 по 2005 г., была построена кривая Филлипса, связывающая уровень инфляции Inf и уровень безработицы $Unem$:

$$\widehat{Inf} = 2.34 - 0.23Unem$$

$$\sqrt{\widehat{Var}(\hat{\beta}_{Unem})} = 0.04, R^2 = 0.12$$

   На уровне значимости 1% проверьте гипотезу $H_0 : \beta_{Unem} = 0$ против альтернативной $H_a : \beta_{Unem} \neq 0$.

12. Пусть $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ и $i = 1, \ldots, 18$ — классическая регрессионная модель, где $\mathbb{E}(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$. Также имеются следующие данные: $\sum_{i=1}^{18} y_i^2 = 4256, \sum_{i=1}^{18} x_i^2 = 185, \sum_{i=1}^{18} x_i y_i = 814.25, \sum_{i=1}^{18} y_i = 225, \sum_{i=1}^{18} x_i = 49.5$. Используя эти данные, оцените эту регрессию и на уровне значимости 5% проверьте гипотезу $H_0 : \beta_1 = 3.5$ против альтернативной $H_a : \beta_1 > 3.5$:
    (a) Приведите формулу для тестовой статистики
    (b) Укажите распределение тестовой статистики
    (c) Вычислите наблюдаемое значение тестовой статистики
    (d) Укажите границы области, где основная гипотеза не отвергается
    (e) Сделайте статистический вывод
13. Рассматривается модель $y_i = \mu + \varepsilon_i$, где $\mathbb{E}(\varepsilon_i) = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$ и $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$. При каких c_i несмещенная оцека
$$\hat{\mu} = \sum_{i=1}^{n} c_i y_i$$
имеет наименьшую дисперсию? Через теорему Гаусса–Маркова или через условную минимизацию, $c_i = 1/n$
14. Рассмотрим классическую линейную регрессионную модель, $y_t = \beta \cdot t + \varepsilon_t$. Какая из оценок, $\hat{\beta}$ или $\hat{\beta}'$ является более эффективной?
    (a) $\hat{\beta} = y_1$, $\hat{\beta}' = y_2/2$
    (b) $\hat{\beta} = y_1$, $\hat{\beta}' = 0.5 y_1 + 0.5 \frac{y_2}{2}$
    (c) $\hat{\beta} = \frac{1}{n}\left(y_1 + \frac{y_2}{2} + \frac{y_3}{3} + \ldots + \frac{y_n}{n}\right)$, $\hat{\beta}' = \frac{y_1 + 2y_2 + \ldots + n y_n}{1^2 + 2^2 + \ldots + n^2}$
15. Ошибки регрессии ε_i независимы и равновероятно принимают значения $+1$ и -1. Также известно, что $y_i = \beta \cdot i + \varepsilon_i$. Модель оценивается всего по двум наблюдениям.
    (a) Найдите закон распределения $\hat{\beta}$, RSS, ESS, TSS, R^2
    (b) Найдите $\mathbb{E}(\hat{\beta})$, $\mathrm{Var}(\hat{\beta})$, $\mathbb{E}(RSS)$, $\mathbb{E}(ESS)$, $\mathbb{E}(R^2)$
    (c) При каком β величина $\mathbb{E}(R^2)$ достигает максимума?
16. Рассмотрим модель с линейным трендом без свободного члена, $y_t = \beta t + \varepsilon_t$.
    (a) Найдите МНК оценку коэффициента β
    (b) Рассчитайте $\mathbb{E}(\hat{\beta})$ и $\mathrm{Var}(\hat{\beta})$ в предположениях теоремы Гаусса-Маркова
    (c) Верно ли, что оценка $\hat{\beta}$ состоятельна?

    (a) $\hat{\beta} = \frac{\sum y_t t}{\sum t^2}$
    (b) $\mathbb{E}(\hat{\beta}) = \beta$ и $\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{T} t^2}$
    (c) Да, состоятельна
17. В модели $y_t = \beta_1 + \beta_2 x_t$, где $x_t = \begin{cases} 2, \, t = 1 \\ 1, \, t > 1 \end{cases}$:
    (a) Найдите мнк-оценку $\hat{\beta}_2$
    (b) Рассчитайте $\mathbb{E}(\hat{\beta}_2)$ и $\mathrm{Var}(\hat{\beta}_2)$ в предположениях теоремы Гаусса-Маркова
    (c) Верно ли, что оценка $\hat{\beta}_2$ состоятельна?

    несостоятельна
18. В модели $y_t = \beta_1 + \beta_2 x_t$, где $x_t = \begin{cases} 1, \, t = 2k+1 \\ 0, \, t = 2k \end{cases}$:
    (a) Найдите мнк-оценку $\hat{\beta}_2$

(b) Рассчитайте $\mathbb{E}(\hat{\beta}_2)$ и $\mathrm{Var}(\hat{\beta}_2)$ в предположениях теоремы Гаусса-Маркова

(c) Верно ли, что оценка $\hat{\beta}_2$ состоятельна?

19. Априори известно, что парная регрессия должна проходить через точку (x_0, y_0).

    (a) Выведите формулы МНК оценок;

    (b) В предположениях теоремы Гаусса-Маркова найдите дисперсии и средние оценок

Вроде бы равносильно переносу начала координат и применению результата для регрессии без свободного члена. Должна остаться несмещенность.

20. Мы предполагаем, что y_t растёт с линейным трендом, т.е. $y_t = \beta_1 + \beta_2 t + \varepsilon_t$. Все предпосылки теоремы Гаусса-Маркова выполнены. В качестве оценки $\hat{\beta}_2$ предлагается $\hat{\beta}_2 = \frac{Y_T - 1}{T-1}$, где T — общее количество наблюдений.

    (a) Найдите $\mathbb{E}(\hat{\beta}_2)$ и $\mathrm{Var}(\hat{\beta}_2)$

    (b) Совпадает ли оценка $\hat{\beta}_2$ с классической мнк-оценкой?

    (c) У какой оценки дисперсия выше, у $\hat{\beta}_2$ или классической мнк-оценки?

21. Вася считает, что выборочная ковариация $\mathrm{sCov}(y, \hat{y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{y})}{n-1}$ это неплохая оценка для $\mathrm{Cov}(y_i, \hat{y}_i)$. Прав ли он? Не прав. Ковариация $\mathrm{Cov}(y_i, \hat{y}_i)$ зависит от i, это не одно неизвестное число, для которого можно предложить одну оценку.

22. В классической линейной регрессионной модели $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, дисперсия зависимой переменной не зависит от номера наблюдения, $\mathrm{Var}(y_i) = \sigma^2$. Почему для оценки σ^2 вместо известной из курса математической статистики формулы $\sum(y_i - \bar{y})^2/(n-1)$ используют $\sum \hat{\varepsilon}_i^2/(n-2)$? формула $\sum(y_i - \bar{y})^2/(n-1)$ неприменима так как $\mathbb{E}(y_i)$ не является константой

23. Оценка регрессии имеет вид $\hat{y}_i = 3 - 2x_i$. Выборочная дисперсия x равна 9, выборочная дисперсия y равна 40. Найдите R^2 и выборочные корреляции $\mathrm{sCorr}(x, y)$, $\mathrm{sCorr}(y, \hat{y})$. R^2 – это отношение выборочных дисперсий $\hat{y}$ и y.

24. Слитки-вариант. Перед нами два золотых слитка и весы, производящие взвешивания с ошибками. Взвесив первый слиток, мы получили результат 300 грамм, взвесив второй слиток — 200 грамм, взвесив оба слитка — 400 грамм. Предположим, что ошибки взвешивания — независимые одинаково распределенные случайные величины с нулевым средним.

    (a) Найдите несмещеную оценку веса первого слитка, обладающую наименьшей дисперсией.

    (b) Как можно проинтерпретировать нулевое математическое ожидание ошибки взвешивания?

Как отсутствие систематической ошибки.

25. Рассмотрим линейную модель $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, где ошибки ε_i нормальны $N(0; \sigma^2)$ и независимы.

    (a) Верно ли, что y_i одинаково распределены?

    (b) Верно ли, что $\bar{y}$ — это несмещенная оценка для $\mathbb{E}(y_i)$?

    (c) Верно ли, что $\sum(y_i - \bar{y})^2/(n-1)$ — несмещенная оценка для σ^2? Если да, то докажите, если нет, то определите величину смещения

нет, нет, нет

26. Рассмотрим модель регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, в которой ошибки ε_i независимы и нормальны $N(0; \sigma^2)$, оценивается по 22 наблюдениям. Найдите $\mathbb{E}(RSS)$, $\mathrm{Var}(RSS)$, $\mathbb{P}(10\sigma^2 < RSS < 30\sigma^2)$, $\mathbb{P}(10\hat{\sigma}^2 < RSS < 30\hat{\sigma}^2)$

$RSS/\sigma^2 \sim \chi^2_{n-k}$, $\mathbb{E}(RSS) = (n-k)\sigma^2$, $\mathrm{Var}(RSS) = 2(n-k)\sigma^4$, $\mathbb{P}(10\sigma^2 < RSS < 30\sigma^2) \approx 0.898$

3 Многомерный МНК без матриц

1. Эконометрэсса Ширли зашла в пустую аудиторию, где царил приятный полумрак, и увидела на доске до боли знакомую надпись:

$$\hat{y} = \underset{(2.37)}{1.1} - \underset{(-0.4)}{0.7} \cdot x_2 + \underset{(3.15)}{0.9} \cdot x_3 - \underset{(-0.67)}{19} \cdot x_4$$

Помогите эконометрэссе Ширли определить, что находится в скобках

(a) P-значения

(b) t-статистики

(c) стандартные ошибки коэффициентов

(d) R^2 скорректированный на номер коэффициента

(e) показатели VIF для каждого коэффициента

t-статистики

2. Для нормальной регрессии с 5-ю факторами (включая свободный член) известны границы симметричного по вероятности 80% доверительного интервала для дисперсии σ_ε^2: $A = 45$, $B = 87.942$.

(a) Определите количество наблюдений в выборке

(b) Вычислите $\hat{\sigma}_\varepsilon^2$

(a) Поскольку $\frac{\hat{\sigma}_\varepsilon^2(n-k)}{\sigma_\varepsilon^2} \sim \chi^2(n-k)$, где $\hat{\sigma}_\varepsilon^2 = \frac{RSS}{n-k}$, $k = 5$. $P(\chi_l^2 < \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} < \chi_u^2) = 0.8$. Преобразовав, получим $P(\frac{\hat{\sigma}_\varepsilon^2(n-5)}{\chi_u^2} < \sigma_\varepsilon^2 < \frac{\hat{\sigma}_\varepsilon^2(n-5)}{\chi_l^2}) = 0.8$, где $\chi_u^2 = \chi_{n-5;0.1}^2$, $\chi_l^2 = \chi_{n-5;0.9}^2$ — соответствующие квантили. По условию $\frac{\hat{\sigma}_\varepsilon^2(n-5)}{\chi_l^2} = A = 45$, $\frac{\hat{\sigma}_\varepsilon^2(n-5)}{\chi_u^2} = B = 87.942$. Поделим B на A, отсюда следует $\frac{\chi_u^2}{\chi_l^2} = 1.95426$. Перебором квантилей в таблице для хи-квадрат распределения мы находим, что $\frac{\chi_{30;0.1}^2}{\chi_{30;0.9}^2} = \frac{40.256}{20.599} = 1.95426$. Значит, $n - 5 = 30$, отсюда следует, что $n = 35$.

(b) $\hat{\sigma}_\varepsilon^2 = 45\frac{\chi_u^2}{n-5} = 45\frac{40.256}{30} = 60.384$

3. Рассмотрим следующую регрессионную модель зависимости логарифма заработной платы $\ln W$ от уровня образования Edu, опыта работы Exp, Exp^2 и уровня образования родителей $Fedu$, $Medu$:

$$\widehat{\ln W} = \hat{\beta}_1 + \hat{\beta}_2 Edu + \hat{\beta}_3 Exp + \hat{\beta}_4 Exp^2 + \hat{\beta}_5 Fedu + \hat{\beta}_6 Medu$$

Модель регрессии была отдельно оценена по выборкам из 35 мужчин и 23 женщин, и были получены остаточные суммы квадратов $RSS_1 = 34.4$ и $RSS_2 = 23.4$ соответственно. Остаточная сумма квадратов в регрессии, оценённой по объединённой выборке, равна 70.3. На уровне значимости 5% проверьте гипотезу об отсутствии дискриминации в оплате труда между мужчинами и женщинами.

Упорядочим нашу выборку таким образом, чтобы наблюдения с номерами с 1 по 35 относились к мужчинам, а наблюдения с номерами с 36 по 58 относились к женщинам. Тогда уравнение

$$\ln W_i = \beta_1 + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i, i = 1, \ldots, 35$$

соответствует регрессии, построенной для подвыборки из мужчин, а уравнение

$$\ln W_i = \gamma_1 + \gamma_2 Edu_i + \gamma_3 Exp_i + \gamma_4 Exp_i^2 + \gamma_5 Fedu_i + \gamma_6 Medu_i + \varepsilon_i, i = 36, \ldots, 58$$

соответствует регрессии, построенной для подвыборки из женщин. Введем следующие переменные:

$$d_i = \begin{cases} 1, & \text{если } i\text{–ое наблюдение соответствует мужчине,} \\ 0, & \text{в противном случае;} \end{cases}$$

$$dum_i = \begin{cases} 1, & \text{если } i\text{–ое наблюдение соответствует женщине,} \\ 0, & \text{в противном случае.} \end{cases}$$

Рассмотрим следующее уравнение регрессии:

$$\ln W_i = \beta_1 d_i + \gamma_1 dum_i + \beta_2 Edu_i d_i + \gamma_2 Edu_i dum_i + \beta_3 Exp_i d_i + \gamma_3 Exp_i dum_i + \beta_4 Exp_i^2 d_i +$$

$$+\gamma_4 Exp_i^2 dum_i + \beta_5 Fedu_i d_i + \gamma_5 Fedu_i dum_i + \beta_6 Medu_i d_i + \gamma_6 Medu_i dum_i + \varepsilon_i, i = 1, \ldots, 58$$

Гипотеза, которую требуется проверить в данной задаче, имеет вид

$$H_0 : \begin{cases} \beta_1 = \gamma_1, \\ \beta_2 = \gamma_2, \\ \ldots \\ \beta_6 = \gamma_6 \end{cases} \quad H_1 : |\beta_1 - \gamma_1| + |\beta_2 - \gamma_2| + \cdots + |\beta_6 - \gamma_6| > 0.$$

Тогда регрессия

$$\ln W_i = \beta_1 d_i + \gamma_1 dum_i + \beta_2 Edu_i d_i + \gamma_2 Edu_i dum_i + \beta_3 Exp_i d_i + \gamma_3 Exp_i dum_i + \beta_4 Exp_i^2 d_i +$$

$$+\gamma_4 Exp_i^2 dum_i + \beta_5 Fedu_i d_i + \gamma_5 Fedu_i dum_i + \beta_6 Medu_i d_i + \gamma_6 Medu_i dum_i + \varepsilon_i, i = 1, \ldots, 58$$

по отношению к основной гипотезе H_0 является регрессией без ограничений, а регрессия

$$\ln W_i = \beta_1 + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i, i = 1, \ldots, 58$$

является регрессией с ограничениями.
Кроме того, для решения задачи должен быть известен следующий факт:
$RSS_{UR} = RSS_1 + RSS_2$, где RSS_{UR} — это сумма квадратов остатков в модели:

$$\ln W_i = \beta_1 d_i + \gamma_1 dum_i + \beta_2 Edu_i d_i + \gamma_2 Edu_i dum_i + \beta_3 Exp_i d_i + \gamma_3 Exp_i dum_i + \beta_4 Exp_i^2 d_i +$$

$$+\gamma_4 Exp_i^2 dum_i + \beta_5 Fedu_i d_i + \gamma_5 Fedu_i dum_i + \beta_6 Medu_i d_i + \gamma_6 Medu_i dum_i + \varepsilon_i, i = 1, \ldots, 58$$

RSS_1 — это сумма квадратов остатков в модели:

$$\ln W_i = \beta_1 + \beta_2 Edu_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 Fedu_i + \beta_6 Medu_i + \varepsilon_i, i = 1, \ldots, 35$$

RSS_2 — это сумма квадратов остатков в модели:

$$\ln W_i = \gamma_1 + \gamma_2 Edu_i + \gamma_3 Exp_i + \gamma_4 Exp_i^2 + \gamma_5 Fedu_i + \gamma_6 Medu_i + \varepsilon_i, i = 36, \ldots, 58$$

(a) Тестовая статистика:

$$T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-m)},$$

где RSS_R – сумма квадратов остатков в модели с ограничениями;
RSS_{UR} – сумма квадратов остатков в модели без ограничений;
q – число линейно независимых уравнений в основной гипотезе H_0;
n – общее число наблюдений;
m – число коэффициентов в модели без ограничений

(b) Распределение тестовой статистики:

$$T \sim F(q, n-m)$$

(c) Наблюдаемое значение тестовой статистики:

$$T_{obs} = \frac{(70.3 - (34.4 + 23.4))/6}{(34.4 + 23.4)/(58 - 12)} = 1.66$$

(d) Область, в которой H_0 не отвергается:

$$[0; T_{cr}] = [0; 2.3]$$

(e) Статистический вывод:
Поскольку $T_{obs} \in [0; T_{cr}]$, то на основе имеющихся данных мы не можем отвергнуть гипотезу H_0 в пользу альтернативной H_1. Следовательно, имеющиеся данные не противоречат гипотезе об отсутствии дискриминации на рынке труда между мужчинами и женщинами.

4. Рассмотрим следующую регрессионную модель зависимости логарифма заработной платы $\ln W$ от уровня образования Edu, опыта работы Exp, Exp^2:

$$\widehat{\ln W} = \hat{\beta}_1 + \hat{\beta}_2 Edu + \hat{\beta}_3 Exp + \hat{\beta}_4 Exp^2$$

Модель регрессии была отдельно оценена по выборкам из 20 мужчин и 20 женщин, и были получены остаточные суммы квадратов $RSS_1 = 49.4$ и $RSS_2 = 44.1$ соответственно. Остаточная сумма квадратов в регрессии, оценённой по объединённой выборке, равна 105.5. На уровне 5% проверьте гипотезу об отсутствии дискриминации в оплате труда между мужчинами и женщинами.

5. Ниже приведены результаты оценивания спроса на молоко для модели $y_i = \beta_1 + \beta_2 I_i + \beta_3 P_i + \varepsilon_i$, где y_i – стоимость молока, купленного i–ой семьёй за последние 7 дней (в руб.), I_i – месячный доход i–ой семьи (в руб.), P_i – цена 1 литра молока (в руб.). Вычисления для общей выборки, состоящей из 2127 семей, дали $RSS = 8841601$. Для двух подвыборок, состоящих из 348 городских и 1779 сельских семей, соответствующие суммы квадратов

остатков оказались следующими: $RSS_1 = 1720236$ и $RSS_2 = 7099423$. Можно ли считать зависимость спроса на молоко от его цены и дохода единой для городской и сельской местности? Ответ обоснуйте подходящим тестом.

6. По 52 наблюдениям была оценена следующая зависимость цены квадратного метра квартиры $Price$ (в долларах) от площади кухни K (в квадратных метрах), времени в пути пешком до ближайшего метро M (в минутах), расстояния до центра города C (в км) и наличия рядом с домом лесопарковой зоны P (1 — есть, 0 — нет).

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.73)}{16.12} + \underset{(0.14)}{1.7}\,K - \underset{(0.03)}{0.35}M - \underset{(0.12)}{0.46}C + \underset{(0.98)}{2.22}P$$

$$R^2 = 0.78, \sum_{i=1}^{52}(Price_i - \overline{Price})^2 = 278$$

Предположим, что все квартиры в выборке можно отнести к двум категориям: квартиры на севере города (28 наблюдений) и квартиры на юге города (24 наблюдения). Модель регрессии была оценена отдельно только по квартирам на севере и только по квартирам на юге. Ниже приведены результаты оценивания.

Для квартир на севере:

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.3)}{14} + \underset{(0.23)}{1.6}\,K - \underset{(0.04)}{0.33}M - \underset{(0.22)}{0.4}\,C + \underset{(0.78)}{2.1}\,P, RSS = 21.8$$

Для квартир на юге:

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.9)}{16.8} + \underset{(0.4)}{1.62}K - \underset{(0.12)}{0.29}M - \underset{(0.23)}{0.51}C + \underset{(1.28)}{1.98}P, RSS = 19.2$$

На уровне значимости 5% проверьте гипотезу о различии в ценообразовании квартир на севере и на юге.

7. По 52 наблюдениям была оценена следующая зависимость цены квадратного метра квартиры $Price$ (в долларах) от площади кухни K (в квадратных метрах), времени в пути пешком до ближайшего метро M (в минутах), расстояния до центра города C (в км) и наличия рядом с домом лесопарковой зоны P (1 — есть, 0 — нет).

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.73)}{16.12} + \underset{(0.14)}{1.7}\,K - \underset{(0.03)}{0.35}M - \underset{(0.12)}{0.46}C + \underset{(0.98)}{2.22}P$$

$$R^2 = 0.78, \sum_{i=1}^{52}(Price_i - \overline{Price})^2 = 278$$

Предположим, что все квартиры в выборке можно отнести к двум категориям: квартиры на севере города (28 наблюдений) и квартиры на юге города (24 наблюдения). Пусть S — это фиктивная переменная, равная 1 для домов в южной части города и 0 для домов в северной части города. Используя эту переменную, была оценена следующая регрессия:

$$\underset{(s.e.)}{\widehat{Price}} = \underset{(3.13)}{14.12} + \underset{(0.11)}{0.25}S + \underset{(0.13)}{1.65}K + \underset{(0.14)}{0.17}K\cdot S - \underset{(0.039)}{0.37}\,M + \underset{(0.0012)}{0.05}\,M\cdot S - \underset{(0.13)}{0.44}C - \underset{(0.18)}{0.06}C\cdot S + \underset{(0.88)}{2.27}P - \underset{(0.08)}{0.23}P\cdot S$$

$$R^2 = 0.85$$

На уровне значимости 5% проверьте гипотезу о различии в ценообразовании квартир на севере и на юге.

8. На основе квартальных данных с 2003 по 2008 год было получено следующее уравнение

регрессии, описывающее зависимость цены на товар P от нескольких факторов:

$$P = 3.5 + 0.4X + 1.1W, ESS = 70.4, RSS = 40.5$$

Когда в уравнение были добавлены фиктивные переменные, соответствующие первым трем кварталам года Q_1, Q_2, Q_3, оцениваемая модель приобрела вид:

$$P_t = \beta + \beta_X X_t + \beta_W W_t + \beta_{Q_{1t}} Q_{1t} + \beta_{Q_{2t}} Q_{2t} + \beta_{Q_{3t}} Q_{3t} + \varepsilon_t$$

При этом величина ESS выросла до 86.4. Сформулируйте и на уровне значимости 5% проверьте гипотезу о наличии сезонности.

9. Рассмотрим следующую функцию спроса с сезонными переменными $SPRING$ (весна), $SUMMER$ (лето), $FALL$ (осень):

$$\widehat{\ln Q} = \hat{\beta}_1 + \hat{\beta}_2 \cdot \ln P + \hat{\beta}_3 \cdot SPRING + \hat{\beta}_4 \cdot SUMMER + \hat{\beta}_5 \cdot FALL$$

$$R^2 = 0.37, n = 20$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_3 = \beta_5$. Дайте интерпретацию проверяемой гипотезе. Пусть для регрессии с ограничениями был вычислен коэффициент $R^2_R = 0.23$. На уровне значимости 5% проверьте нулевую гипотезу.

10. Рассмотрим следующую функцию спроса с сезонными переменными $SPRING$ (весна), $SUMMER$ (лето), $FALL$ (осень):

$$\widehat{\ln Q} = \hat{\beta}_1 + \hat{\beta}_2 \cdot \ln P + \hat{\beta}_3 \cdot SPRING + \hat{\beta}_4 \cdot SUMMER + \hat{\beta}_5 \cdot FALL$$

$$R^2 = 0.24, n = 24$$

Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \begin{cases} \beta_3 = 0, \\ \beta_4 = \beta_5 \end{cases}$. Дайте интерпретацию проверяемой гипотезе. Пусть для регрессии с ограничениями был вычислен коэффициент $R^2_R = 0.13$. На уровне значимости 5% проверьте нулевую гипотезу.

11. Исследователь собирается по выборке, содержащей данные за 2 года, построить модель линейной регрессии с константой и 3-мя объясняющими переменными. В модель предполагается ввести 3 фиктивные сезонные переменные $SPRING$ (весна), $SUMMER$ (лето) и $FALL$ (осень) на все коэффициенты регрессии. Однако в процессе оценивания статистический пакет вывел на экран компьютера следующее сообщение “insufficient number of observations”. Объясните, почему имеющегося числа наблюдений не хватило для оценивания параметров модели.

12. По данным для 57 индивидов оценили зависимость длительности обучения индивида S от способностей индивида, описываемых обобщённой переменной IQ, и пола индивида, описываемого с помощью фиктивной переменной $MALE$ (равной 1 для мужчин и 0 для женщин), с помощью двух регрессий (в скобках под коэффициентами указаны оценки стандартных отклонений):

$$\underset{(s.e.)}{\hat{S}} = \underset{(0.44)}{6.12} + \underset{(0.088)}{0.147} \cdot IQ, RSS = 2758.6$$

$$\underset{(s.e.)}{\hat{S}} = \underset{(0.73)}{6.12} + \underset{(0.014)}{0.147} \cdot IQ - \underset{(0.933)}{1.035} \cdot MALE + \underset{(0.018)}{0.0166} \cdot (MALE \cdot IQ), RSS = 2090.98$$

Зависит ли длительность обучения от пола индивида и почему?

13. По данным, содержащим 30 наблюдений, построена регрессия:

$$\hat{y} = 1.3870 + 5.2587 \cdot x + 2.6259 \cdot d + 2.5955 \cdot x \cdot d,$$

где фиктивная переменная d определяется следующим образом:

$$d_i = \begin{cases} 1 & \text{при } i \in \{1, \ldots, 20\}, \\ 0 & \text{при } i \in \{21, \ldots, 30\}. \end{cases}$$

Найдите оценки коэффициентов в модели $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, построенной по первым 20-ти наблюдениям, т.е. при $i \in \{1, \ldots, 20\}$.

14. Выборка содержит 30 наблюдений зависимой переменной y и независимой переменной x. Ниже приведены результаты оценивания уравнения регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ по первым 20-ти и последним 10-ти наблюдениям соответственно:

$$\hat{y} = 4.0039 + 2.6632 \cdot x$$

$$\hat{y} = 1.3780 + 5.2587 \cdot x$$

По имеющимся данным найдите оценки коэффициентов модели, рассчитанной по 30-ти наблюдениям $y_i = \beta_1 + \beta_2 x_i + \Delta\beta_1 \cdot d_i + \Delta\beta_2 \cdot x_i \cdot d_i + \varepsilon_i$, где фиктивная переменная d определяется следующим образом:

$$d_i = \begin{cases} 1 & \text{при } i \in \{1, \ldots, 20\}, \\ 0 & \text{при } i \in \{21, \ldots, 30\}. \end{cases}$$

15. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \beta_2 = \beta_3 = \beta_4$. Запишите, какой вид имеет модель «с ограничением» для тестирования указанной гипотезы.

16. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \beta_3 = \beta_4 = 1$. Какая модель из приведённых ниже может выступать в качестве модели «с ограничением» для тестирования указанной гипотезы? Если ни одна из них, то запишите свою.

    (a) $y_i - (x_{i2} + x_{i3}) = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$
    (b) $y_i + (x_{i2} - x_{i3}) = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$
    (c) $y_i + x_{i2} + x_{i3} = \beta_1 + \beta_2 x_{i1} + \varepsilon_i$
    (d) $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 + \beta_4 + \varepsilon_i$

17. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \begin{cases} \beta_2 + \beta_3 + \beta_4 = 1, \\ \beta_3 + \beta_4 = 0. \end{cases}$ Какая модель из приведённых ниже может выступать в качестве модели «с ограничением» для тестирования указанной гипотезы? Если ни одна из них, то запишите свою.

    (a) $y_i - x_{i1} = \beta_1 + \beta_3(x_{i2} - x_{i3}) + \varepsilon_i$
    (b) $y_i - x_{i1} = \beta_1 + \beta_4(x_{i3} - x_{i2}) + \varepsilon_i$
    (c) $y_i + x_{i1} = \beta_1 + \beta_3(x_{i2} + x_{i3}) + \varepsilon_i$
    (d) $y_i + x_{i1} = \beta_1 + \beta_3(x_{i2} - x_{i3}) + \varepsilon_i$

18. Пусть регрессионная модель имеет вид $y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3} + \varepsilon_i, i = 1, \ldots, n$. Тестируемая гипотеза $H_0 : \begin{cases} \beta_2 - \beta_3 = 0, \\ \beta_3 + \beta_4 = 0. \end{cases}$ Какая модель из приведённых ниже может

выступать в качестве модели «с ограничением» для тестирования указанной гипотезы? Если ни одна из них, то запишите свою.

(a) $y_i = \beta_1 + \beta_3(x_{i2} - x_{i1} - x_{i3}) + \varepsilon_i$

(b) $y_i - x_{i1} = \beta_1 + \beta_4(x_{i3} - x_{i2}) + \varepsilon_i$

(c) $y_i = \beta_1 + \beta_3(x_{i1} + x_{i2} + x_{i3}) + \varepsilon_i$

(d) $y_i = \beta_1 + \beta_3(x_{i1} + x_{i2} - x_{i3}) + \varepsilon_i$

19. Известно, что P-значение для коэффициента регрессии равно 0.087, а уровень значимости 0.1. Является ли значимым данный коэффициент в регрессии?
20. Известно, что P-значение для коэффициента регрессии равно 0.078, а уровень значимости 0.05. Является ли значимым данный коэффициент в регрессии?
21. Известно, что P-значение для коэффициента регрессии равно 0.09. На каком уровне значимости данный коэффициент в регрессии будет признан значимым?
22. Ниже приведены результаты оценивания уравнения линейной регрессии зависимости количества смертей в автомобильных катастрофах от различных характеристик:

$$deaths_i = \beta_1 + \beta_2 drivers_i + \beta_3 popden_i + \beta_4 temp + \beta_5 fuel + \varepsilon_i$$

$$\widehat{deaths}_i = -\underset{(222.8803)}{27.1} + \underset{(0.3767)}{4.64} \cdot drivers_i - \underset{(0.0239)}{0.0228} \cdot popden_i + \underset{(4.6016)}{5.3} \cdot temp_i - \underset{(0.8679)}{0.663} \cdot fuel_i$$

	Estimate	St.Error	t value	P-value
Intercept	-27.10	222.88	-0.12	0.90
Drivers	4.64	0.38	12.30	0.00
Popden	-0.02	0.02	-0.95	0.35
Temp	5.30	4.60	1.15	0.26
Fuel	-0.66	0.87	-0.76	0.45

Перечислите, какие из переменных в регрессии являются значимыми и на каком уровне значимости.

23. Была оценена функция Кобба-Дугласа с учётом человеческого капитала H (K — физический капитал, L — труд):

$$\widehat{\ln Q} = 1.4 + 0.46 \ln L + 0.27 \ln H + 0.23 \ln K$$

$$ESS = 170.4, RSS = 80.3, n = 21$$

(a) Чему равен коэффициент R^2?

(b) На уровне значимости 1% проверьте гипотезу о значимости регрессии «в целом»

24. На основе опроса 25 человек была оценена следующая модель зависимости логарифма зарплаты $\ln W$ от уровня образования Edu (в годах) и возраста Age:

$$\widehat{\ln W} = 1.7 + 0.5Edu + 0.06Age - 0.0004Age^2$$

$$ESS = 90.3, RSS = 60.4$$

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

(a) Напишите спецификацию уравнения регрессии с учётом образования родителей

(b) Сформулируйте и на уровне значимости 5% проверьте гипотезу о значимом влиянии уровня образования родителей на заработную плату:

i. Сформулируйте гипотезу

ii. Приведите формулу для тестовой статистики

iii. Укажите распределение тестовой статистики

iv. Вычислите наблюдаемое значение тестовой статистики

v. Укажите границы области, где основная гипотеза не отвергается

vi. Сделайте статистический вывод

Ограниченная модель (Restricted model):

$$\ln W_i = \beta + \beta_{Edu} Edu_i + \beta_{Age} Age_i + \beta_{Age^2} Age_i^2 + \varepsilon_i$$

Неограниченная модель (Unrestricted model):

$$\ln W_i = \beta + \beta_{Edu} Edu_i + \beta_{Age} Age_i + \beta_{Age^2} Age_i^2 + \beta_{Fedu} Fedu_i + \beta_{Medu} Medu_i + \varepsilon_i$$

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:

$$\ln W_i = \beta + \beta_{Edu} Edu_i + \beta_{Age} Age_i + \beta_{Age^2} Age_i^2 + \beta_{Fedu} Fedu_i + \beta_{Medu} Medu_i + \varepsilon_i$$

(b) Проверка гипотезы

i. $H_0 : \begin{cases} \beta_{Fedu} = 0 \\ \beta_{Medu} = 0 \end{cases}$ $\quad H_a : |\beta_{Fedu}| + |\beta_{Medu}| > 0$

ii. $T = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k)}$, где $q = 2$ — число линейно независимых уравнений в основной гипотезе H_0, $n = 25$ — число наблюдений, $k = 6$ — число коэффициентов в модели без ограничения

iii. $T \sim F(q; n-k)$

iv. $T_{obs} = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k)} = \frac{(60.4 - 40.4)/2}{40.4/(25-6)} = 4.70$

v. Нижняя граница = 0, верхняя граница = 3.52

vi. Поскольку $T_{obs} = 4.70$, что не принадлежит промежутку от 0 до 3.52, то на основе имеющихся данных можно отвергнуть основную гипотезу на уровне значимости 5%. Таким образом, образование родителей существенно влияет на заработную плату.

25. Рассмотрим следующую модель зависимости цены дома $Price$ (в тысячах долларов) от его площади $Hsize$ (в квадратных метрах), площади участка $Lsize$ (в квадратных метрах), 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 статистистической гипотезы $H_0 : \beta_3 = 20\beta_4$. Дайте интерпретацию проверяемой гипотезе. Для регрессии с ограничением был вычислен коэффициент $R_R^2 = 0.136$. На уровне значимости 5% проверьте нулевую гипотезу.

26. Рассмотрим следующую модель зависимости почасовой оплаты труда W от уровня образования $Educ$, возраста Age, уровня образования родителей $Fathedu$ и $Mothedu$:

$$\widehat{\ln W} = \hat{\beta}_1 + \hat{\beta}_2 Educ + \hat{\beta}_3 Age + \hat{\beta}_4 Age^2 + \hat{\beta}_5 Fathedu + \hat{\beta}_6 Mothedu$$

$$R^2 = 0.341, n = 27$$

Напишите спецификацию регрессии с ограничениями для проверки статистистической гипотезы $H_0 : \beta_5 = 2\beta_4$. Дайте интерпретацию проверяемой гипотезе. Для регрессии с ограничением был вычислен коэффициент $R_R^2 = 0.296$. На уровне значимости 5% проверьте нулевую гипотезу.

27. По данным для 27 фирм исследователь оценил зависимость объёма выпуска y от труда l и капитала k с помощью двух моделей:

$$\ln y_i = \beta_1 + \beta_2 \ln l_i + \beta_3 \ln k_i + \varepsilon_i$$

$$\ln y_i = \beta_1 + \beta_2 \ln(l_i \cdot k_i) + \varepsilon_i$$

Он получил для этих двух моделей суммы квадратов остатков $RSS_1 = 0.851$ и $RSS_2 = 0.894$ соответственно. Сформулируйте гипотезу, которую хотел проверить исследователь. На уровне значимости 5% проверьте эту гипотезу и дайте экономическую интерпретацию.

28. Пусть задана линейная регрессионная модель:

$$y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \varepsilon_i, i = 1, \ldots, 20$$

По имеющимся данным оценены следующие регрессии:

$$\underset{(s.e.)}{\hat{y}_i} = \underset{(0.15)}{10.01} + \underset{(0.06)}{1.05}x_1 + \underset{(0.04)}{2.06}x_2 + \underset{(0.06)}{0.49}x_3 - \underset{(0.06)}{1.31}x_4, RSS = 6.85$$

$$\underset{(s.e.)}{y_i - \widehat{x_1 -} 2x_2} = \underset{(0.15)}{10.00} + \underset{(0.07)}{0.50}x_3 - \underset{(0.06)}{1.32}x_4, RSS = 8.31$$

$$\underset{(s.e.)}{y_i + \widehat{x_1 +} 2x_2} = \underset{(3.62)}{9.93} + \underset{(1.48)}{0.56}x_3 - \underset{(1.42)}{1.50}x_4, RSS = 4310.62$$

$$\underset{(s.e.)}{y_i - \widehat{x_1 +} 2x_2} = \underset{(3.26)}{10.71} + \underset{(1.33)}{0.09}x_3 - \underset{(1.28)}{1.28}x_4, RSS = 3496.85$$

$$\underset{(s.e.)}{y_i + \widehat{x_1 -} 2x_2} = \underset{(1.25)}{9.22} + \underset{(0.51)}{0.97}x_3 - \underset{(0.49)}{1.54}x_4, RSS = 516.23$$

На уровне значимости 5% проверьте гипотезу $H_0 : \begin{cases} \beta_2 = 1 \\ \beta_3 = 2 \end{cases}$ против альтернативной гипотезы $H_a : |\beta_2 - 1| + |\beta_3 - 2| \neq 0$.

29. Рассмотрим следующую модель зависимости расходов на образование на душу населения от дохода на душу населения, доли населения в возрасте до 18 лет, а также доли городского населения:

$$education_i = \beta_1 + \beta_2 income_i + \beta_3 young_i + \beta_4 urban_i + \varepsilon_i$$

Ниже приведены результаты оценивания уравнения этой линейной регрессии:

$$\widehat{education}_i = -\underset{(64.9199)}{287} + \underset{(0.0093)}{0.0807} \cdot income_i + \underset{(0.1598)}{0.817} \cdot young_i - \underset{(0.0343)}{0.106} \cdot urban_i$$

	Estimate	St.Error	t value	P-value
Intercept	-286.84	64.92	-4.42	0.00
Income	0.08	0.01	8.67	0.00
Young	0.82	0.16	5.12	0.00
Urban	-0.11	0.03	-3.09	0.00

(a) Сформулируйте основную и альтернативую гипотезы, которые соответствуют тесту на значимость коэффициента при переменной доход на душу населения в уравнении регрессии

(b) На уровне значимости 10% проверьте гипотезу о значимости коэффициента при переменной доход на душу населения в уравнении регрессии:

i. Приведите формулу для тестовой статистики

ii. Укажите распределение тестовой статистики

iii. Вычислите наблюдаемое значение тестовой статистики

iv. Укажите границы области, где основная гипотеза не отвергается

v. Сделайте статистический вывод

(c) На уровне значимости 5% проверьте гипотезу $H_0 : \beta_1 = 1$ против альтернативной $H_a : \beta_1 > 1$:

   i. Приведите формулу для тестовой статистики
   ii. Укажите распределение тестовой статистики
   iii. Вычислите наблюдаемое значение тестовой статистики
   iv. Укажите границы области, где основная гипотеза не отвергается
   v. Сделайте статистический вывод

(d) Сформулируйте основную гипотезу, которая соответствует тесту на значимость регрессии «в целом»

(e) На уровне значимости 1% проверьте гипотезу о значимости регрессии «в целом»,если известно, что $F-$статистика равна 34.81 со степенями свободы 3 и 47, $P-$значение равно $5.337e^{-12}$:

   i. Приведите формулу для тестовой статистики
   ii. Укажите распределение тестовой статистики
   iii. Вычислите наблюдаемое значение тестовой статистики
   iv. Укажите границы области, где основная гипотеза не отвергается
   v. Сделайте статистический вывод

(f) Далее приведены результаты оценивания уравнения регрессии без переменной, отражающей долю городского населения:

$$\widehat{education}_i = -\underset{(70.27134)}{301} + \underset{(0.00741)}{0.0612} \cdot income_i + \underset{(0.17327)}{0.836} \cdot young_i$$

	Estimate	St.Error	t value	P-value
Intercept	-301.09	70.27	-4.28	0.00
Income	0.06	0.01	8.25	0.00
Young	0.84	0.17	4.83	0.00

Также известно, что RSS для первой модели равен 33489.35, а для второй модели — 40276.61. На уровне значимости 5% проверьте гипотезу $H_0 : \beta_4 = 0$ против альтернативной $H_0 : \beta_4 \neq 0$:

   i. Приведите формулу для тестовой статистики
   ii. Укажите распределение тестовой статистики
   iii. Вычислите наблюдаемое значение тестовой статистики
   iv. Укажите границы области, где основная гипотеза не отвергается
   v. Сделайте статистический вывод

30. Рассмотрим следующую модель зависимости расходов на образование на душу населения от дохода на душу населения, доли населения в возрасте до 18 лет, а также доли городского населения:

$$education_i = \beta_1 + \beta_2 income_i + \beta_3 young_i + \beta_4 urban_i + \varepsilon_i$$

Ниже приведены результаты оценивания уравнения этой линейной регрессии:

$$\widehat{education}_i = -\underset{(64.9199)}{287} + \underset{(0.0093)}{0.0807} \cdot income_i + \underset{(0.1598)}{0.817} \cdot young_i - \underset{(0.0343)}{0.106} \cdot urban_i$$

(a) Сформулируйте основную и альтернативую гипотезы, которые соответствуют тесту на значимость коэффициента при переменной доля населения в возрасте до 18 лет в уравнении регрессии

	Estimate	St.Error	t value	P-value
Intercept	-286.84	64.92	-4.42	0.00
Income	0.08	0.01	8.67	0.00
Young	0.82	0.16	5.12	0.00
Urban	-0.11	0.03	-3.09	0.00

(b) На уровне значимости 10% проверьте гипотезу о значимости коэффициента при переменной доля населения в возрасте до 18 лет в уравнении регрессии:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

(c) Далее приведены результаты оценивания уравнения регрессии без переменной, отражающей долю населения в возрасте до 18 лет:

$$\widehat{education}_i = \underset{(27.3827)}{25.3} + \underset{(0.0114)}{0.0762} \cdot income_i - \underset{(0.0423)}{0.112} \cdot urban_i$$

	Estimate	St.Error	t value	P-value
Intercept	25.25	27.38	0.92	0.36
Income	0.08	0.01	6.67	0.00
Urban	-0.11	0.04	-2.66	0.01

Также известно, что RSS для первой модели равен 33489.35, а для второй модели — 52132.29. На уровне значимости 5% проверьте гипотезу $H_0 : \beta_3 = 0$ против альтернативной $H_0 : \beta_3 \neq 0$:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

31. Вася построил регрессию оценки за первую контрольную работу на константу, рост и вес студента, $\widehat{kr1}_i = \hat{\beta}_1 + \hat{\beta}_2 height_i + \hat{\beta}_3 weight_i$. Затем построил регрессию оценки за вторую контрольную работу на те же объясняющие переменные, $\widehat{kr2}_i = \hat{\beta}'_1 + \hat{\beta}'_2 height_i + \hat{\beta}'_3 weight_i$. Накопленная оценка считается по формуле $nak_i = 0.25 \cdot kr1_i + 0.75 \cdot kr2_i$. Чему равны оценки коэффициентов в регрессии накопленной оценки на те же объясняющие переменные? Ответ обоснуйте.

$0.25\hat{\beta}_1 + 0.75\hat{\beta}'_1$, $0.25\hat{\beta}_2 + 0.75\hat{\beta}'_2$ и $0.25\hat{\beta}_3 + 0.75\hat{\beta}'_3$

32. Истинная модель имеет вид $y_i = \beta x_i + \varepsilon_i$. Вася оценивает модель $\hat{y}_i = \hat{\beta} x_i$ по первой части выборки, получает $\hat{\beta}_a$, по второй части выборки — получает $\hat{\beta}_b$ и по всей выборке — $\hat{\beta}_{tot}$. Как связаны между собой $\hat{\beta}_a$, $\hat{\beta}_b$, $\hat{\beta}_{tot}$? Как связаны между собой дисперсии $\mathrm{Var}(\hat{\beta}_a)$, $\mathrm{Var}(\hat{\beta}_b)$ и $\mathrm{Var}(\hat{\beta}_{tot})$? Сами оценки коэффициентов никак детерминистически не связаны, но при большом размере подвыборок примерно равны. А дисперсии связаны соотношением $\mathrm{Var}(\hat{\beta}_a)^{-1} + \mathrm{Var}(\hat{\beta}_b)^{-1} = \mathrm{Var}(\hat{\beta}_{tot})^{-1}$

33. Сгенерируйте вектор y из 300 независимых нормальных $N(10, 1)$ случайных величин. Сгенерируйте 40 «объясняющих» переменных, по 300 наблюдений в каждой, каждое наблюдение — независимая нормальная $N(5, 1)$ случайная величина. Постройте регрессию y на все 40 регрессоров и константу.

(a) Сколько регрессоров оказалось значимо на 5% уровне?

(b) Сколько регрессоров в среднем значимо на 5% уровне?

(c) Эконометрист Вовочка всегда использует следующий подход: строит регрессию зависимой переменной на все имеющиеся регрессоры, а затем выкидывает из модели те регрессоры, которые оказались незначимы. Прокомментируйте Вовочкин эконометрический подход.

34. Мы попытаемся понять, как введение в регрессию лишнего регрессора влияет на оценки уже имеющихся. В регрессии будет 100 наблюдений. Возьмем $\rho = 0.5$. Сгенерим выборку совместных нормальных x_i и z_i с корреляцией ρ. Настоящий y_i задаётся формулой $y_i = 5 + 6x_i + \varepsilon_i$. Однако мы будем оценивать модель $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 z_i$.

(a) Повторите указанный эксперимент 500 раз и постройте оценку для функции плотности $\hat{\beta}_1$.

(b) Повторите указанный эксперимент 500 раз для каждого ρ от -1 до 1 с шагом в 0.05. Каждый раз сохраняйте полученные 500 значений $\hat{\beta}_1$. В осях $(\rho, \hat{\beta}_1)$ постройте 95%-ый предиктивный интервал для $\hat{\beta}_1$. Прокомментируйте.

35. Цель задачи — оценить модель CAPM несколькими способами.

(a) Соберите подходящие данные для модели CAPM. Нужно найти три временных ряда: ряд цен любой акции, любой рыночный индекс, безрисковый актив. Переведите цены в доходности.

(b) Постройте графики

(c) Оцените модель CAPM без свободного члена по всем наборам данных. Прокомментируйте смысл оцененного коэффициента

(d) Разбейте временной период на два участка и проверьте устойчивость коэффициента бета

(e) Добавьте в классическую модель CAPM свободный член и оцените по всему набору данных. Какие выводы можно сделать?

(f) Методом максимального правдоподобия оцените модель с ошибкой измерения $R^m - R^0$, т.е.

истинная зависимость имеет вид

$$(R^s - R^0) = \beta_1 + \beta_2(R^*_m - R^*_0) + \varepsilon \tag{2}$$

величины R^*_m и R^*_0 не наблюдаемы, но

$$R_m - R_0 = R^*_m - R^*_0 + u \tag{3}$$

36. По 47 наблюдениям оценивается зависимость доли мужчин занятых в сельском хозяйстве от уровня образованности и доли католического населения по Швейцарским кантонам в 1888 году.

$$Agriculture_i = \beta_1 + \beta_2 Examination_i + \beta_3 Catholic_i + \varepsilon_i$$

```
h <- swiss
model1 <- glm(Agriculture~Examination+Catholic,data=h)
coef.t <- coeftest(model1)
dimnames(coef.t)[[2]] <-
    c("Оценка","Ст. ошибка",  "t-статистика", "P-значение")
```

```r
coef.t <- coef.t[,-4]
coef.t[1,1] <- NA
coef.t[2,2] <- NA
coef.t[3,3] <- NA
```

```r
xtable(coef.t)
```

	Оценка	Ст. ошибка	t-статистика
(Intercept)		8.72	9.44
Examination	-1.94		-5.08
Catholic	0.01	0.07	

(a) Заполните пропуски в таблице

(b) Укажите коэффициенты, значимые на 10% уровне значимости.

(c) Постройте 99%-ый доверительный интервал для коэффициента при переменной Catholic

Набор данных доступен в пакете R:

```r
h <- swiss
```

37. Оценивается зависимость уровня фертильности всё тех же швейцарских кантонов в 1888 году от ряда показателей. В таблице представлены результаты оценивания двух моделей. Модель 1: $Fertility_i = \beta_1 + \beta_2 Agriculture_i + \beta_3 Education_i + \beta_4 Examination_i + \beta_5 Catholic_i + \varepsilon_i$ Модель 2: $Fertility_i = \gamma_1 + \gamma_2(Education_i + Examination_i) + \gamma_3 Catholic_i + u_i$

```r
m1 <- lm(Fertility~Agriculture+Education+Examination+Catholic,data=h)
m2 <- lm(Fertility~I(Education+Examination)+Catholic,data=h)
```

```r
apsrtable(m1,m2)
```

Таблица 1:

	Model 1	Model 2
(Intercept)	91.06*	80.52*
	(6.95)	(3.31)
Agriculture	−0.22*	
	(0.07)	
Education	−0.96*	
	(0.19)	
Examination	−0.26	
	(0.27)	
Catholic	0.12*	0.07*
	(0.04)	(0.03)
I(Education + Examination)		−0.48*
		(0.08)
N	47	47
R^2	0.65	0.55
adj. R^2	0.62	0.53
Resid. sd	7.74	8.56

Standard errors in parentheses

* indicates significance at $p < 0.05$

Набор данных доступен в пакете R:

```r
h <- swiss
```

(a) Проверьте гипотезу о том, что коэффициент при $Education$ в модели 1 равен -0.5.

(b) На 5% уровне значимости проверьте гипотезу о том, что переменные $Education$ и $Examination$ оказывают одинаковое влияние на $Fertility$.

38. По 2040 наблюдениям оценена модель зависимости стоимости квартиры в Москве (в 1000$) от общего метража и метража жилой площади.

```r
model1 <- lm(price~totsp+livesp,data=flats)
report <- summary(model1)
coef.table <- report$coefficients
rownames(coef.table) <- c("Константа","Общая площадь", "Жилая площадь")
xtable(coef.table)
```

	Estimate	Std. Error	t value	Pr(>|t|)
Константа	-88.81	4.37	-20.34	0.00
Общая площадь	1.70	0.10	17.78	0.00
Жилая площадь	1.99	0.18	10.89	0.00

Оценка ковариационной матрицы $\widehat{Var}(\hat{\beta})$ имеет вид

```r
var.hat <- vcov(model1)
xtable(var.hat)
```

	(Intercept)	totsp	livesp
(Intercept)	19.07	0.03	-0.45
totsp	0.03	0.01	-0.02
livesp	-0.45	-0.02	0.03

(a) Проверьте H_0: $\beta_{totsp} = \beta_{livesp}$. В чём содержательный смысл этой гипотезы?

(b) Постройте доверительный интервал дли $\beta_{totsp} - \beta_{livesp}$. В чём содержательный смысл этого доверительного интервала?

Из оценки ковариационной матрицы находим, что $se(\hat{\beta}_{totsp} = \hat{\beta}_{livesp}) = 0.2696$.
Исходя из $Z_{crit} = 1.96$ получаем доверительный интервал, $[-0.8221; 0.2348]$.

Вывод: при уровне значимости 5% гипотеза о равенстве коэффициентов не отвергается.

39. По 2040 наблюдениям оценена модель зависимости стоимости квартиры в Москве (в 1000$) от общего метража и метража жилой площади.

```r
model1 <- lm(price~totsp+livesp,data=flats)
report <- summary(model1)
coef.table <- report$coefficients
rownames(coef.table) <- c("Константа","Общая площадь", "Жилая площадь")
xtable(coef.table)
```

	Estimate	Std. Error	t value	Pr(>|t|)
Константа	-88.81	4.37	-20.34	0.00
Общая площадь	1.70	0.10	17.78	0.00
Жилая площадь	1.99	0.18	10.89	0.00

Оценка ковариационной матрицы $\widehat{Var}(\hat{\beta})$ имеет вид

```r
xtable(vcov(model1))
```

	(Intercept)	totsp	livesp
(Intercept)	19.07	0.03	-0.45
totsp	0.03	0.01	-0.02
livesp	-0.45	-0.02	0.03

(a) Постройте 95%-ый доверительный интервал для ожидаемой стоимости квартиры с жилой площадью 30 м^2 и общей площадью 60 м^2.

(b) Постройте 95%-ый прогнозный интервал для фактической стоимости квартиры с жилой площадью 30 м^2 и общей площадью 60 м^2.

40. По 2040 наблюдениям оценена модель зависимости стоимости квартиры в Москве (в 1000$) от общего метража, метража жилой площади и дамми-переменной, равной 1 для кирпичных домов.

```r
model1 <- lm(price~totsp+livesp+brick+brick:totsp+brick:livesp,data=flats)
report <- summary(model1)
coef.table <- report$coefficients
# rownames(coef.table) <- c("Константа","Общая площадь", "Жилая площадь")
xtable(coef.table)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-66.03	6.07	-10.89	0.00
totsp	1.77	0.12	14.98	0.00
livesp	1.27	0.25	5.05	0.00
brick	-19.59	9.01	-2.17	0.03
totsp:brick	0.42	0.20	2.10	0.04
livesp:brick	0.09	0.38	0.23	0.82

(a) Выпишите отдельно уравнения регрессии для кирпичных домов и для некирпичных домов

(b) Проинтерпретируйте коэффициент при $brick_i \cdot totsp_i$

41. По 20 наблюдениям оценивается линейная регрессия $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, причём истинная зависимость имеет вид $y = \beta_1 + \beta_2 x + \varepsilon$. Случайная ошибка ε_i имеет нормальное распределение $N(0,1)$.

(a) Найдите вероятность $\mathbb{P}(\hat{\beta}_3 > se(\hat{\beta}_3))$

(b) Найдите вероятность $\mathbb{P}(\hat{\beta}_3 > \sigma_{\hat{\beta}_3})$

(a) $\mathbb{P}(\hat{\beta}_3 > se(\hat{\beta}_3)) = \mathbb{P}(t_{17} > 1) = 0.1657$

(b) $\mathbb{P}(\hat{\beta}_3 > \sigma_{\hat{\beta}_3}) = \mathbb{P}(N(0,1) > 1) = 0.1587$

42. К эконометристу Вовочке в распоряжение попали данные с результатами контрольной работы студентов по эконометрике. В данных есть результаты по каждой задаче, переменные p_1, p_2, p_3, p_4 и p_5, и суммарный результат за контрольную, переменная kr. Чему будут равны оценки коэффициентов, их стандартные ошибки, t-статистики, P-значения, R^2, RSS, если

(a) Вовочка построит регрессию kr на константу, p_1, p_2, p_3, p_4 и p_5

(b) Вовочка построит регрессию kr на p_1, p_2, p_3, p_4 и p_5 без константы

43. Сгенерируйте данные так, чтобы при оценке линейной регрессионной модели оказалось, что скорректированный коэффициент детерминации, R^2_{adj}, отрицательный.

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

Следовательно, при R^2 близком к 0 и большом количестве регрессоров k может оказаться, что $R^2_{adj} < 0$.
Например,

```
set.seed(42)
y <- rnorm(200,sd=15)
X <- matrix(rnorm(2000),nrow=200)
model <- lm(y~X)
report <- summary(model)
report$adj.r.squared

## [1] -0.02745
```

Косяк. Почему-то книтр внутри solution ругается на доллар.

44. Для коэффициентов регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 w_i + \varepsilon_i$ даны 95%-ые доверительные интервалы: $\beta_2 \in (0.16; 0.66)$, $\beta_3 \in (-0.33; 0.93)$ и $\beta_4 \in (-1.01; 0.54)$.
   (a) Найдите $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$
   (b) Определите, какие из переменных в регрессии значимы на уровне значимости 5%.

$\hat{\beta}_2 = 0.41$, $\hat{\beta}_3 = 0.3$, $\hat{\beta}_4 = -0.235$, переменная x значима

45. Для коэффициентов регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 w_i + \varepsilon_i$ даны 95%-ые доверительные интервалы: $\beta_2 \in (-0.15; 1.65)$, $\beta_3 \in (0.32; 0.93)$ и $\beta_4 \in (0.14; 1.55)$.
   (a) Найдите $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$
   (b) Определите, какие из переменных в регрессии значимы на уровне значимости 5%.

$\hat{\beta}_2 = 0.75$, $\hat{\beta}_3 = 0.625$, $\hat{\beta}_4 = 0.845$, переменные z и w значимы

46. Эконометрэсса Мырли очень суеверна и поэтому оценила три модели:
   M1 $y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \varepsilon_i$ по всем наблюдениям.
   M2 $y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \beta_4 d_i + \varepsilon_i$ по всем наблюдениям, где d_i — дамми-переменная равная 1 для 13-го наблюдения и нулю иначе.
   M3 $y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \varepsilon_i$ по всем наблюдениям, кроме 13-го.
   (a) Сравните между собой RSS во всех трёх моделях
   (b) Есть ли совпадающие оценки коэффициентов в этих трёх моделях? Если есть, то какие?
   (c) Может ли Мырли не выполняя вычислений узнать ошибку прогноза для 13-го наблюдения при использовании третьей модели? Если да, то как?

$RSS_1 > RSS_2 = RSS_3$, в моделях два и три, ошибка прогноза равна $\hat{\beta}_4$

47. Рассмотрим модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \beta_4 z_i + \varepsilon_i$. При оценке модели по 24 наблюдениям оказалось, что $RSS = 15$, $\sum(y_i - \bar{y} - w_i + \bar{w})^2 = 20$. На уровне значимости 1% протестируйте гипотезу

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

48. Модель регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$, в которой ошибки ε_i независимы и нормальны $N(0;\sigma^2)$, оценивается по 13 наблюдениям. Найдите $\mathbb{E}(RSS)$, $\mathrm{Var}(RSS)$, $\mathbb{P}(5\sigma^2 < RSS < 10\sigma^2)$, $\mathbb{P}(5\hat\sigma^2 < RSS < 10\hat\sigma^2)$

$RSS/\sigma^2 \sim \chi^2_{n-k}$, $\mathbb{E}(RSS) = (n-k)\sigma^2$, $\mathrm{Var}(RSS) = 2(n-k)\sigma^4$, $\mathbb{P}(5\sigma^2 < RSS < 10\sigma^2) \approx 0.451$

4 МНК с матрицами и вероятностями

1. Пусть $y = X\beta + \varepsilon$ — регрессионная модель.
   (a) Сформулируйте теорему Гаусса-Маркова
   (b) Верно ли, что оценка $\hat\beta = (X'X)^{-1}X'y$ несмещённая?
   (c) В условиях теоремы Гаусса-Маркова найдите ковариационную матрицу $\hat\beta$
2. Пусть $y = X\beta + \varepsilon$ — регрессионная модель и $\tilde\beta = ((X'X)^{-1}X' + A)y$ — несмещённая оценка вектора неизвестных параметров β. Верно ли, что $AX = 0$?
3. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, $X = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}$, $y = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$, $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 I$. Найдите коэффициент корреляции $\mathrm{Corr}(\hat\beta_1, \hat\beta_2)$.
4. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, где $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$. Пусть $Z = XD$, где $D = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. Рассмотрите «новую» регрессионную модель $y = Z\alpha + u$, где $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$. Определите, как выражаются «новые» МНК-коэффициенты через «старые».
5. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, где $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$. Пусть $Z = XD$, где $D = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. Рассмотрите «новую» регрессионную модель $y = Z\alpha + u$, где $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$. Определите, как выражаются «новые» МНК-коэффициенты через «старые».
6. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, где $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$. Пусть $Z = XD$, где $D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. Рассмотрите «новую» регрессионную модель $y = Z\alpha + u$, где $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$. Определите, как выражаются «новые» МНК-коэффициенты через «старые».
7. Пусть $y = X\beta + \varepsilon$ — регрессионная модель. Верно ли, что $\hat\varepsilon'\hat y = 0$ и $\hat y'\hat\varepsilon = 0$? да, да
8. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, где $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2_\varepsilon I$. Пусть A — неслучайная матрица размера $k \times k$, $\det(A) \neq 0$. Совершается преобразование регрессоров по правилу $Z = XA$. В преобразованных регрессорах уравнение выглядит так: $y = Z\gamma + u$, где $\mathbb{E}(u) = 0$, $\mathrm{Var}(u) = \sigma^2_u I$.
   (a) Как связаны между собой МНК-оценки $\hat\beta$ и $\hat\gamma$?
   (b) Как связаны между собой векторы остатков регрессий?
   (c) Как связаны между собой прогнозные значения, полученные по двум регрессиям?

   (a) $\hat\gamma = (Z'Z)^{-1}Z'y = A^{-1}(X'X)^{-1}(A')^{-1}A'X'y = A^{-1}(X'X)^{-1}X'y = A^{-1}\hat\beta$
   (b) $\hat u = y - Z\hat\gamma = y - XAA^{-1}\hat\beta = y - X\hat\beta = \hat\varepsilon$

(c) Пусть $z^0 = \begin{pmatrix} 1 & z_1^0 & \ldots & z_{k-1}^0 \end{pmatrix}$ — вектор размера $1 \times k$ и $x^0 = \begin{pmatrix} 1 & x_1^0 & \ldots & x_{k-1}^0 \end{pmatrix}$ — вектор размера $1 \times k$. Оба эти вектора представляют собой значения факторов. Тогда $z^0 = x^0 A$ и прогнозное значение для регрессии с преобразованными факторами равно $z^0 \hat{\gamma} = x^0 A A^{-1} \hat{\beta} = x^0 \hat{\beta}$ прогнозному значению для регрессии с исходными факторами.

9. Рассмотрим оценку вида $\tilde{\beta} = ((X'X)^{-1} + \gamma I)X'y$ для вектора коэффициентов регрессионного уравнения $y = X\beta + \varepsilon$, удовлетворяющего условиям классической регрессионной модели. Найдите $\mathbb{E}(\tilde{\beta})$ и $\mathrm{Var}(\tilde{\beta})$.

   (a) $\mathbb{E}(\tilde{\beta}) = ((X'X)^{-1} + \gamma I)X'\mathbb{E}(y) = ((X'X)^{-1} + \gamma I)X'X\beta = \beta + \gamma X'X\beta$

   (b) $\mathrm{Var}(\tilde{\beta}) = \mathrm{Var}(((X'X)^{-1} + \gamma I)X'y) = \mathrm{Var}(((X'X)^{-1} + \gamma I)X'\varepsilon) =$
   $= (((X'X)^{-1} + \gamma I)X')\,\mathrm{Var}(\varepsilon)(((X'X)^{-1} + \gamma I)X')' =$
   $= (((X'X)^{-1} + \gamma I)X')\sigma_\varepsilon^2 I(((X'X)^{-1} + \gamma I)X')' = \sigma_\varepsilon^2((X'X)^{-1} + \gamma I)X'X((X'X)^{-1} + \gamma I) =$
   $= \sigma_\varepsilon^2((X'X)^{-1} + \gamma I)(I + \gamma X'X) = \sigma_\varepsilon^2((X'X)^{-1} + 2\gamma I + \gamma^2 X'X)$

10. Верно ли, что при невырожденном преобразовании факторов R^2 не меняется? А именно, пусть заданы две регрессионные модели: $y = X\beta + \varepsilon$ и $y = Z\alpha + u$, где y — вектор размера $n \times 1$, X и Z — матрицы размера $n \times k$, β и α — вектора рамзера $k \times 1$, ε и u — вектора размера $n \times 1$, а также $Z = XD$, $\det(D) \neq 0$. Верно ли, что коэффициенты детерминации представленных выше моделей равны между собой?

11. Верно ли, что при невырожденном преобразовании факторов RSS не меняется. А именно, пусть заданы две регрессиионные модели: $y = X\beta + \varepsilon$ и $y = Z\alpha + u$, где y — вектор размера $n \times 1$, X и Z — матрицы размера $n \times k$, β и α — вектора рамзера $k \times 1$, ε и u — вектора размера $n \times 1$, а также $Z = XD$, $\det(D) \neq 0$. Верно ли, что сумма квадратов остатков в представленных выше моделях равны между собой?

12. 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 $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}^T$. Известно, что $\mathbb{E}\varepsilon = 0$ и $\mathrm{Var}(\varepsilon) = 4 \cdot I$. Известно также, что:
$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$, $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$
Для удобства расчётов ниже приведены матрицы:
$X^T X = \begin{pmatrix} 5 & 3 & 1 \\ 3 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ и $(X^T X)^{-1} = \begin{pmatrix} 0.5 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$.
Найдите:

    (a) $\mathrm{Var}(\varepsilon_1)$
    (b) $\mathrm{Var}(\beta_1)$
    (c) $\mathrm{Var}(\hat{\beta}_1)$
    (d) $\widehat{\mathrm{Var}}(\hat{\beta}_1)$
    (e) $\mathbb{E}(\hat{\beta}_1^2) - \beta_1^2$
    (f) $\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_3)$
    (g) $\widehat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3)$
    (h) $\mathrm{Var}(\hat{\beta}_2 - \hat{\beta}_3)$
    (i) $\widehat{\mathrm{Var}}(\hat{\beta}_2 - \hat{\beta}_3)$
    (j) $\mathrm{Var}(\beta_2 - \beta_3)$
    (k) $\mathrm{Corr}(\hat{\beta}_2, \hat{\beta}_3)$
    (l) $\widehat{\mathrm{Corr}}(\hat{\beta}_2, \hat{\beta}_3)$
    (m) $\mathbb{E}(\hat{\sigma}^2)$

(n) $\hat{\sigma}^2$

13. Пусть $y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \varepsilon_i$ — регрессионная модель, где $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, $y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$,

$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$, $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$, ошибки ε_i независимы и нормально распределены с $\mathbb{E}(\varepsilon) = 0$,

$Var(\varepsilon) = \sigma^2 I$. Для удобства расчётов даны матрицы: $X'X = \begin{pmatrix} 5 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ и $(X'X)' =$

$\begin{pmatrix} 0.3333 & -0.3333 & 0.0000 \\ -0.3333 & 1.3333 & -1.0000 \\ 0.0000 & -1.0000 & 2.0000 \end{pmatrix}$

(a) Укажите число наблюдений

(b) Укажите число регрессоров в модели, учитывая свободный член

(c) Найдите $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$

(d) Найдите $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

(e) Методом МНК найдите оценку для вектора неизвестных коэффициентов

(f) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оценённого уравнения регрессии

(g) Сформулируйте основную и альтернативную гипотезы, которые соответствуют тесту на значимость переменной x_1 в уравнении регрессии

(h) Протестируйте на значимость переменную x_1 в уравнении регрессии на уровне значимости 10%:

  i. Приведите формулу для тестовой статистики
  ii. Укажите распределение тестовой статистики
  iii. Вычислите наблюдаемое значение тестовой статистики
  iv. Укажите границы области, где основная гипотеза не отвергается
  v. Сделайте статистический вывод о значимости переменной x_1

(i) Найдите $P-$значение, соответствующее наблюдаемому значению тестовой статистики (T_{obs}) из предыдущего пункта. На основе полученного $P-$значения сделайте вывод о значимости переменной x_1

(j) На уровне значимости 10% проверьте гипотезу $H_0 : \beta_1 = 1$ против альтернативной $H_a : \beta_1 \neq 1$:

  i. Приведите формулу для тестовой статистики
  ii. Укажите распределение тестовой статистики
  iii. Вычислите наблюдаемое значение тестовой статистики
  iv. Укажите границы области, где основная гипотеза не отвергается
  v. Сделайте статистический вывод

(k) На уровне значимости 10% проверьте гипотезу $H_0 : \beta_1 = 1$ против альтернативной $H_a : \beta_1 > 1$:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

(l) На уровне значимости 10% проверьте гипотезу $H_0 : \beta_1 = 1$ против альтернативной $H_a : \beta_1 < 1$:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

(m) Сформулируйте основную гипотезу, которая соответствует тесту на значимость регрессии «в целом»

(n) На уровне значимости 5% проверьте гипотезу о значимости регрессии «в целом»:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

(o) Найдите $P-$значение, соответствующее наблюдаемому значению тестовой статистики (T_{obs}) из предыдущего пункта. На основе полученного $P-$значения сделайте вывод о значимости регрессии «в целом»

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

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

(q) На уровне значимости 5% проверьте гипотезу $H_0 : \beta_1 + \beta_2 = 2$ против альтернативной $H_a : \beta_1 + \beta_2 > 2$:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается
v. Сделайте статистический вывод

(r) На уровне значимости 5% проверьте гипотезу $H_0 : \beta_1 + \beta_2 = 2$ против альтернативной $H_a : \beta_1 + \beta_2 < 2$:

i. Приведите формулу для тестовой статистики
ii. Укажите распределение тестовой статистики
iii. Вычислите наблюдаемое значение тестовой статистики
iv. Укажите границы области, где основная гипотеза не отвергается

v. Сделайте статистический вывод

(a) $n = 5$

(b) $k = 3$

(c) $TSS = 10$

(d) $RSS = 2$

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

(f) $R^2 = 1 - \frac{RSS}{TSS} = 0.8$. R^2 высокий, построенная эконометрическая модель «хорошо» описывает данные

(g) Основная гипотеза — $H_0 : \beta_1 = 0$, альтернативная гипотеза — $H_a : \beta_1 \neq 0$

(h) Проверка гипотезы

i. $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}}; n = 5; k = 3$

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

iii. $T_{obs} = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 0}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}} = \frac{2-0}{\sqrt{\frac{2}{5-3}1.3333}} = 1.7321$

iv. Нижняя граница $= -2.920$, верхняя граница $= 2.920$

v. Поскольку $T_{obs} = 1.7321$, что принадлежит промежутку от -2.920 до 2.920, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10%

(i) $p-value(T_{obs}) = \mathbb{P}(|T| > |T_{obs}|) = 2F_T(|T_{obs}|)$, где $F_T(|T_{obs}|)$ — функция распределения $t-$распределения с $n-k = 5-3 = 2$ степенями свободы в точке $|T_{obs}|$. $p-value(T_{obs}) = 2tcdf(-|T_{obs}|, n-k) = 2tcdf(-1.7321, 2) = 0.2253$. Поскольку $P-$значение превосходит уровень значимости 10%, то основная гипотеза — $H_0 : \beta_1 = 0$ не может быть отвергнута

(j) Проверка гипотезы

i. $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}}; n = 5; k = 3$

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

iii. $T_{obs} = \frac{\hat{\beta}_1 - 1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}} = \frac{2-1}{\sqrt{\frac{2}{5-3}1.3333}} = 0.8660$

iv. Нижняя граница $= -2.920$, верхняя граница $= 2.920$

v. Поскольку $T_{obs} = 0.8660$, что принадлежит промежутку от -2.920 до 2.920, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10%

(k) Проверка гипотезы

i. $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}}; n = 5; k = 3$

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

iii. $T_{obs} = \frac{\hat{\beta}_1 - 1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}} = \frac{2-1}{\sqrt{\frac{2}{5-3}1.3333}} = 0.8660$

iv. Нижняя граница $= -\infty$, верхняя граница $= 1.8856$

v. Поскольку $T_{obs} = 0.8660$, что принадлежит промежутку от $-\infty$ до 1.8856, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10%

(l) Проверка гипотезы

i. $T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}}; n = 5; k = 3$

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

iii. $T_{obs} = \frac{\hat{\beta}_1 - 1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1 - 1}{\sqrt{\frac{RSS}{n-k}[(X'X)^{-1}]_{22}}} = \frac{2-1}{\sqrt{\frac{2}{5-3}1.3333}} = 0.8660$

iv. Нижняя граница $= -1.8856$, верхняя граница $= +\infty$

v. Поскольку $T_{obs} = 0.8660$, что принадлежит промежутку от -1.8856 до $+\infty$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 10%

(m) Основная гипотеза — $H_0 : \beta_1 = \beta_2 = 0$, альтернативная гипотеза — $H_a : |\beta_1| + |\beta_2| > 0$

(n) Проверка гипотезы

i. $T = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k}; n = 5; k = 3$

ii. $T \sim F(n-k); n = 5; k = 3$

iii. $T_{obs} = \frac{R^2}{1-R^2} \cdot \frac{n-k}{k} = \frac{0.8}{1-0.8} \cdot \frac{5-3}{2} = 4$

iv. Нижняя граница $= 0$, верхняя граница $= 19$

v. Поскольку $T_{obs} = 4$, что принадлежит промежутку от 0 до 19, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5%. Следовательно, регрессия в целом незначима. Напомним, что $R^2 = 0.8$, то есть он высокий. Но при этом регрессия «в целом» незначима. Такой эффект может возникать при малом объёме выборки, например, таком, как в данной задаче

(o) $p-value(T_{obs}) = \mathbb{P}(|T| > |T_{obs}|) = 2F_T(|T_{obs}|)$, где $F_T(|T_{obs}|)$ — функция распределения $F-$распределения с $k = 3$ и $n-k = 5-3 = 2$ степенями свободы в точке T_{obs}. $p-value(T_{obs}) = 1 - fcdf(-|T_{obs}|, n-k) = 1 - fcdf(4, 2) = 0.2$. Поскольку $P-$значение превосходит уровень значимости 10%, то основная гипотеза — $H_0 : \beta_1 = \beta_2 = 0$ не может быть отвергнута. Таким образом, регрессия «в целом» незначима

(p) Проверка гипотезы

i. $T = \frac{\hat{\beta}_1 + \hat{\beta}_2 - (\beta_1 + \beta_2)}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1 + \hat{\beta}_2)}}$, где $\widehat{\mathrm{Var}}(\hat{\beta}_1 + \hat{\beta}_2) = \widehat{\mathrm{Var}}(\hat{\beta}_1) + \widehat{\mathrm{Var}}(\hat{\beta}_2) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1; \hat{\beta}_2) = \hat{\sigma}^2[(X'X)^{-1}]_{22} + 2\hat{\sigma}^2[(X'X)^{-1}]_{23} + \hat{\sigma}^2[(X'X)^{-1}]_{33} = \frac{RSS}{n-k}([(X'X)^{-1}]_{22} + 2[(X'X)^{-1}]_{23} + [(X'X)^{-1}]_{33})$

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

iii. $\widehat{\mathrm{Var}}(\hat{\beta}_1 + \hat{\beta}_2) = \frac{RSS}{n-k}([(X'X)^{-1}]_{22} + 2[(X'X)^{-1}]_{23} + [(X'X)^{-1}]_{33}) = \frac{2}{5-3}(1.3333 + 2(-1.0000) + 2.0000) = 1.3333$. Тогда $T_{obs} = \frac{\hat{\beta}_1 + \hat{\beta}_2 - 2}{\sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1 + \hat{\beta}_2)}} = \frac{2+1-2}{\sqrt{1.3333}} = 0.8660$

iv. Нижняя граница $= -4.3027$, верхняя граница $= 4.3027$

v. Поскольку $T_{obs} = 0.8660$, что принадлежит промежутку от -4.3027 до 4.3027, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5%

(q) Проверка гипотезы

i. $T = \frac{\hat\beta_1+\hat\beta_2-(\beta_1+\beta_2)}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2)}}$, где $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2) = \widehat{\mathrm{Var}}(\hat\beta_1)+\widehat{\mathrm{Var}}(\hat\beta_2)+2\widehat{\mathrm{Cov}}(\hat\beta_1;\hat\beta_2) = \hat\sigma^2[(X'X)^{-1}]_{22}+2\hat\sigma^2[(X'X)^{-1}]_{23}+\hat\sigma^2[(X'X)^{-1}]_{33} = \frac{RSS}{n-k}([(X'X)^{-1}]_{22}+2[(X'X)^{-1}]_{23}+[(X'X)^{-1}]_{33})$

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

iii. $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2) = \frac{RSS}{n-k}([(X'X)^{-1}]_{22}+2[(X'X)^{-1}]_{23}+[(X'X)^{-1}]_{33}) = \frac{2}{5-3}(1.3333+2(-1.0000)+2.0000) = 1.3333$. Тогда $T_{obs} = \frac{\hat\beta_1+\hat\beta_2-2}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2)}} = \frac{2+1-2}{\sqrt{1.3333}} = 0.8660$

iv. Нижняя граница $= -\infty$, верхняя граница $= 2.9200$

v. Поскольку $T_{obs} = 0.8660$, что принадлежит промежутку от $-\infty$ до 2.9200, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5%

(r) Проверка гипотезы

i. $T = \frac{\hat\beta_1+\hat\beta_2-(\beta_1+\beta_2)}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2)}}$, где $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2) = \widehat{\mathrm{Var}}(\hat\beta_1)+\widehat{\mathrm{Var}}(\hat\beta_2)+2\widehat{\mathrm{Cov}}(\hat\beta_1;\hat\beta_2) = \hat\sigma^2[(X'X)^{-1}]_{22}+2\hat\sigma^2[(X'X)^{-1}]_{23}+\hat\sigma^2[(X'X)^{-1}]_{33} = \frac{RSS}{n-k}([(X'X)^{-1}]_{22}+2[(X'X)^{-1}]_{23}+[(X'X)^{-1}]_{33})$

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

iii. $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2) = \frac{RSS}{n-k}([(X'X)^{-1}]_{22}+2[(X'X)^{-1}]_{23}+[(X'X)^{-1}]_{33}) = \frac{2}{5-3}(1.3333+2(-1.0000)+2.0000) = 1.3333$. Тогда $T_{obs} = \frac{\hat\beta_1+\hat\beta_2-2}{\sqrt{\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2)}} = \frac{2+1-2}{\sqrt{1.3333}} = 0.8660$

iv. Нижняя граница $= -2.9200$, верхняя граница $= +\infty$

v. Поскольку $T_{obs} = 0.8660$, что принадлежит промежутку от -2.9200 до $+\infty$, то на основе имеющихся данных нельзя отвергнуть основную гипотезу на уровне значимости 5%

14. Пусть $y = X\beta+\varepsilon$ — регрессионная модель, где $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, $y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$,

$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{pmatrix}$, $\mathbb{E}(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2 I$.

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

(c) Вычислите наблюдаемое значение тестовой статистики

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

(e) Сделайте статистический вывод

15. По 13 наблюдениям Вася оценил модель со свободным членом, пятью количественными регрессорами и двумя качественными. Качественные регрессоры Вася правильно закодировал с помощью дамми-переменных. Одна качественная переменная принимала четыре значения, другая — пять.

(a) Найдите SSR, R^2

(b) Как выглядит матрица $X(X'X)^{-1}X'$?

(c) Почему 13 — несчастливое число?

16. В рамках классической линейной модели найдите все математические ожидания и все ковариационные матрицы всех пар случайных векторов: ε, y, $\hat y$, $\hat\varepsilon$, $\hat\beta$. Т.е. найдите $\mathbb{E}(\varepsilon)$, $\mathbb{E}(y)$, ... и $\mathrm{Cov}(\varepsilon, y)$, $\mathrm{Cov}(\varepsilon, \hat y)$, ... $\mathrm{Var}(\hat\beta) = \sigma^2(X'X)^{-1}$

17. Найдите $\mathbb{E}(\sum(\varepsilon_i-\bar\varepsilon)^2)$, $\mathbb{E}(RSS)$ $(n-1)\sigma^2$, $(n-k)\sigma^2$

18. Используя матрицы $P = X(X'X)^{-1}X'$ и $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$ запишите RSS, TSS и ESS в матричной форме $TSS = y'(I-\pi)y$, $RSS = y'(I-P)y$, $ESS = y'(P-\pi)y$

19. $\mathbb{E}(TSS)$, $\mathbb{E}(ESS)$ — громоздкие $\mathbb{E}(TSS) = (n-1)\sigma^2 + \beta' X'(I-\pi)X\beta$
20. Вася строит регрессию y на некий набор объясняющих переменных и константу. А на самом деле $y_i = \beta_1 + \varepsilon_i$. Чему равно $\mathbb{E}(TSS)$, $\mathbb{E}(RSS)$, $\mathbb{E}(ESS)$ в этом случае? $(n-1)\sigma^2$, $(n-k)\sigma^2$, $(k-1)\sigma^2$
21. Рассмотрим классическую линейную модель. Являются ли векторы $\hat{\varepsilon}$ и $\hat{y}$ перпендикулярными? Найдите $\mathrm{Cov}(\hat{\varepsilon}, \hat{y})$
22. Чему в классической модели регрессии равны $\mathbb{E}(\varepsilon)$, $\mathbb{E}(\hat{\varepsilon})$? Верно ли что $\sum \varepsilon_i$ равна 0? Верно ли что $\sum \hat{\varepsilon}_i$ равна 0?
23. Найдите на Картинке все перпендикулярные векторы. Найдите на Картинке все прямоугольные треугольники. Сформулируйте для них теоремы Пифагора. $\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{\varepsilon}_i^2$, $TSS = ESS + RSS$,
24. Покажите на Картинке TSS, ESS, RSS, R^2, $\mathrm{sCov}(\hat{y}, y)$
25. Предложите аналог R^2 для случая, когда константа среди регрессоров отсутствует. Аналог должен быть всегда в диапазоне $[0;1]$, совпадать с обычным R^2, когда среди регрессоров есть константа, равняться единице в случае нулевого $\hat{\varepsilon}$. Спроецируем единичный столбец на «плоскость», обозначим его $1'$. Делаем проекцию y на «плоскость» и на $1'$. Далее аналогично.
26. Вася оценил регрессию y на константу, x и z. А затем, делать ему нечего, регрессию y на константу и полученный $\hat{y}$. Какие оценки коэффициентов у него получатся? Чему будет равна оценка дисперсии коэффицента при $\hat{y}$? Почему оценка коэффициента неслучайна, а оценка её дисперсии положительна? Проекция y на $\hat{y}$ это $\hat{y}$, поэтому оценки коэффициентов будут 0 и 1. Оценка дисперсии $\frac{RSS}{(n-2)ESS}$. Нарушены предпосылки теоремы Гаусса-Маркова, например, ошибки новой модели в сумме дают 0, значит коррелированы.
27. При каких условиях $TSS = ESS + RSS$? Либо в регрессию включена константа, либо единичный столбец (тут была опечатка, столбей) можно получить как линейную комбинацию регрессоров, например, включены дамми-переменные для каждого возможного значения качественной переменной.
28. Истинная модель имеет вид $y = X\beta + \varepsilon$. Вася оценивает модель $\hat{y} = X\hat{\beta}$ по первой части выборки, получает $\hat{\beta}_a$, по второй части выборки — получает $\hat{\beta}_b$ и по всей выборке — $\hat{\beta}_{tot}$. Как связаны между собой $\hat{\beta}_a$, $\hat{\beta}_b$, $\hat{\beta}_{tot}$? Как связаны между собой ковариационные матрицы $\mathrm{Var}(\hat{\beta}_a)$, $\mathrm{Var}(\hat{\beta}_b)$ и $\mathrm{Var}(\hat{\beta}_{tot})$? Сами оценки коэффициентов никак детерминистически не связаны, но при большом размере подвыборок примерно равны. А ковариационные матрицы связаны соотношением $\mathrm{Var}(\hat{\beta}_a)^{-1} + \mathrm{Var}(\hat{\beta}_b)^{-1} = \mathrm{Var}(\hat{\beta}_{tot})^{-1}$
29. Модель линейной регрессии имеет вид $y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + u_i$. Сумма квадратов остатков имеет вид $Q\left(\hat{\beta}_1, \hat{\beta}_2\right) = \sum_{i=1}^n (y_1 - \hat{\beta}_1 x_{i,1} - \hat{\beta}_2 x_{i,2})^2$.

    (a) Выпишите необходимые условия минимума суммы квадратов остатков

    (b) Найдите матрицу $X'X$ и вектор $X'y$ если матрица X имеет вид $X = \begin{pmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{pmatrix}$, а вектор y имеет вид $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$

    (c) Докажите, что необходимые условия равносильны матричному уравнению $X'X\hat{\beta} = X'y$, где $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$

    (d) Предполагая, что матрица $X'X$ обратима, найдите $\hat{\beta}$

30. Вася оценил исходную модель:
$$y_i = \beta_1 + \beta_2 x_i + u_i$$
Для надежности Вася стандартизировал переменные, т.е. перешёл к $y_i^* = (y_i - \bar{y})/s_y$ и $x_i^* = (x_i - \bar{x})/s_x$. Затем Вася оценил ещё две модели:

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

и

$$y_i^* = \beta_2'' x_i^* + u_i''$$

В решении можно считать s_x и s_y известными.

(a) Найдите $\hat{\beta}_1'$

(b) Как связаны между собой $\hat{\beta}_2$, $\hat{\beta}_2'$ и $\hat{\beta}_2''$?

(c) Как связаны между собой $\hat{u}_i$, $\hat{u}_i'$ и $\hat{u}_i''$?

(d) Как связаны между собой $\widehat{\mathrm{Var}}\left(\hat{\beta}_2\right)$, $\widehat{\mathrm{Var}}\left(\hat{\beta}_2'\right)$ и $\widehat{\mathrm{Var}}\left(\hat{\beta}_2''\right)$?

(e) Как выглядит матрица $\widehat{\mathrm{Var}}\left(\hat{\beta}'\right)$?

(f) Как связаны между собой t-статистики $t_{\hat{\beta}_2}$, $t_{\hat{\beta}_2'}$ и $t_{\hat{\beta}_2''}$?

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

(h) В нескольких предложениях прокомментируйте последствия перехода к стандартизированным переменным

31. Регрессионная модель задана в матричном виде при помощи уравнения $y = X\beta + \varepsilon$, где $\beta = (\beta_1, \beta_2, \beta_3)'$. Известно, что $\mathbb{E}(\varepsilon) = 0$ и $\mathrm{Var}(\varepsilon) = \sigma^2 \cdot I$. 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 расчетов приведены матрицы

$$X'X = \begin{pmatrix} 5 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 1 \end{pmatrix} \text{ и } (X'X)^{-1} = \frac{1}{3}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 4 & -3 \\ 0 & -3 & 6 \end{pmatrix}.$$

(a) Укажите число наблюдений.

(b) Укажите число регрессоров с учетом свободного члена.

(c) Запишите модель в скалярном виде

(d) Рассчитайте $TSS = \sum (y_i - \bar{y})^2$, $RSS = \sum (y_i - \hat{y}_i)^2$ и $ESS = \sum (\hat{y}_i - \bar{y})^2$.

(e) Рассчитайте при помощи метода наименьших квадратов $\hat{\beta}$, оценку для вектора неизвестных коэффициентов.

(f) Чему равен $\hat{\varepsilon}_5$, МНК-остаток регрессии, соответствующий 5-ому наблюдению?

(g) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оцененного уравнения регрессии.

(h) Используя приведенные выше данные, рассчитайте несмещенную оценку для неизвестного параметра σ^2 регрессионной модели.

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

(j) Найдите $\widehat{\mathrm{Var}}(\hat{\beta}_1)$, несмещенную оценку дисперсии МНК-коэффициента $\hat{\beta}_1$.

(k) Найдите $\widehat{\mathrm{Var}}(\hat{\beta}_2)$, несмещенную оценку дисперсии МНК-коэффициента $\hat{\beta}_2$.

(l) Найдите $\widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$, несмещенную оценку ковариации МНК-коэффициентов $\hat{\beta}_1$ и $\hat{\beta}_2$.

(m) Найдите $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2)$, $\widehat{\mathrm{Var}}(\hat\beta_1-\hat\beta_2)$, $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2+\hat\beta_3)$, $\widehat{\mathrm{Var}}(\hat\beta_1+\hat\beta_2-2\hat\beta_3)$

(n) Найдите $\widehat{\mathrm{Corr}}(\hat\beta_1,\hat\beta_2)$, оценку коэффициента корреляции МНК-коэффициентов $\hat\beta_1$ и $\hat\beta_2$.

(o) Найдите $s_{\hat\beta_1}$, стандартную ошибку МНК-коэффициента $\hat\beta_1$.

(p) Рассчитайте выборочную ковариацию y и $\hat y$.

(q) Найдите выборочную дисперсию y, выборочную дисперсию $\hat y$.

5 Метод максимального правдоподобия — общая теория

Пусть
$X=(X_1,\ldots,X_n)$ — случайная выборка
$x=(x_1,\ldots,x_n)$ — реализация данной случайной выборки
$f_{X_i}(x_i,\theta)$ — плотность распределения случайной величины X_i, $i=1,\ldots,n$
$\theta=(\theta_1,\ldots,\theta_k)$ — вектор неизвестных параметров
$\Theta\subseteq\mathbb{R}^k$ — множество допустимых значений вектора неизвестных параметров
$\mathrm{L}(\theta)=\prod_{i=1}^n f_{X_i}(x_i,\theta)$ — функция правдоподобия
$l(\theta):=\ln\mathrm{L}(\theta)$ — логарифмическая функция правдоподобия
Пусть требуется протестировать систему (нелинейных) ограничений относительно вектора неизвестных параметров

$$H_0:\begin{cases} g_1(\theta)=0\\ g_2(\theta)=0\\ \ldots\\ g_r(\theta)=0\end{cases}$$

где $g_i(\theta)$ — функция, которая задаёт i-ое ограничение на вектор параметров θ, $i=1,\ldots,r$.

$$\frac{\partial g}{\partial\theta'}=\begin{bmatrix}\partial g_1/\partial\theta'\\ \partial g_2/\partial\theta'\\ \vdots\\ \partial g_r/\partial\theta'\end{bmatrix}=\begin{bmatrix}\frac{\partial g_1}{\partial\theta_1} & \frac{\partial g_1}{\partial\theta_2} & \cdots & \frac{\partial g_1}{\partial\theta_k}\\ \frac{\partial g_2}{\partial\theta_1} & \frac{\partial g_2}{\partial\theta_2} & \cdots & \frac{\partial g_2}{\partial\theta_k}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial g_r}{\partial\theta_1} & \frac{\partial g_r}{\partial\theta_2} & \cdots & \frac{\partial g_r}{\partial\theta_k}\end{bmatrix}$$

$$\frac{\partial g'}{\partial\theta}=\begin{bmatrix}\frac{\partial g_1'}{\partial\theta} & \frac{\partial g_2'}{\partial\theta} & \cdots & \frac{\partial g_r'}{\partial\theta}\end{bmatrix}=\begin{bmatrix}\frac{\partial g_1}{\partial\theta_1} & \frac{\partial g_2}{\partial\theta_1} & \cdots & \frac{\partial g_r}{\partial\theta_1}\\ \frac{\partial g_2}{\partial\theta_2} & \frac{\partial g_2}{\partial\theta_2} & \cdots & \frac{\partial g_r}{\partial\theta_2}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial g_1}{\partial\theta_k} & \frac{\partial g_2}{\partial\theta_k} & \cdots & \frac{\partial g_r}{\partial\theta_k}\end{bmatrix}$$

$$I(\theta)=-\mathbb{E}\left(\frac{\partial^2 l}{\partial\theta\partial\theta'}\right)=-\mathbb{E}\begin{bmatrix}\frac{\partial^2 l}{\partial\theta_1\partial\theta_1} & \frac{\partial^2 l}{\partial\theta_1\partial\theta_2} & \cdots & \frac{\partial^2 l}{\partial\theta_1\partial\theta_k}\\ \frac{\partial^2 l}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 l}{\partial\theta_2\partial\theta_2} & \cdots & \frac{\partial^2 l}{\partial\theta_2\partial\theta_k}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{\partial^2 l}{\partial\theta_k\partial\theta_1} & \frac{\partial^2 l}{\partial\theta_k\partial\theta_2} & \cdots & \frac{\partial^2 l}{\partial\theta_k\partial\theta_k}\end{bmatrix}$$ — информационная матрица Фишера

$$\frac{\partial l}{\partial\theta}=\begin{bmatrix}\frac{\partial l}{\partial\theta_1}\\ \frac{\partial l}{\partial\theta_2}\\ \vdots\\ \frac{\partial l}{\partial\theta_k}\end{bmatrix}$$

$\Theta_{UR}:=\Theta$ — множество допустимых значений вектора неизвестных параметров без учёта ограничений
$\Theta_R:=\{\theta\in\Theta:g(\theta)=0\}$ — множество допустимых значений вектора неизвестных параметров с учётом ограничений
$\hat\theta_{UR}\in\Theta_{UR}$ — точка максимума функции l на множестве Θ_{UR}

$\hat{\theta}_R \in \Theta_R$ — точка максимума функции l на множестве Θ_R

Тогда для тестирования гипотезы H_0 можно воспользоваться одной из следующих ниже статистик.

$LR := -2(l(\hat{\theta}_R) - l) \overset{a}{\sim} \chi^2_r$ — статистика отношения правдоподобия

$W := g'(\hat{\theta}_{UR}) \cdot \left[\frac{\partial g}{\partial \theta'}(\hat{\theta}_{UR}) \cdot I^{-1}(\hat{\theta}_{UR}) \cdot \frac{\partial g'}{\partial \theta}(\hat{\theta}_{UR})\right]^{-1} g(\hat{\theta}_{UR}) \overset{a}{\sim} \chi^2_r$ — статистика Вальда

$LM := \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right]' \cdot I^{-1}(\hat{\theta}_R) \cdot \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right] \overset{a}{\sim} \chi^2_r$ — статистика множителей Лагранжа

1. Пусть p — неизвестная вероятность выпадения орла при бросании монеты. Из 100 испытаний 42 раза выпал «Орел» и 58 — «Решка».
   (a) Найдите оценку $\hat{p}$ методом максимального правдоподобия
   (b) Постройте 95% доверительный интервал для p
   (c) Протестируйте на 5%-ом уровне значимости гипотезу о том, что монетка — «правильная» с помощью теста Вальда, теста множителей Лагранжа, теста отношения правдоподобия
2. Дядя Вова (Владимир Николаевич) и Скрипач (Гедеван) зарабатывают на Плюке чатлы, чтобы купить гравицапу. Число заработанных за i-ый день чатлов имеет пуассоновское распределение, заработки за разные дни независимы. За прошедшие 100 дней они заработали 250 чатлов.
   (a) Оцените параметр λ пуассоновского распределения методом максимального правдоподобия
   (b) Сколько дней им нужно давать концерты, чтобы оценка вероятности купить гравицапу составила 0.99? Гравицапа стоит пол кц или 2200 чатлов.
   (c) Постройте 95% доверительный интервал для λ
   (d) Проверьте гипотезу о том, что средний дневной заработок равен 2 чатла с помощью теста отношения правдоподобия, теста Вальда, теста множителей Лагранжа
3. Инопланетянин Кapп совершил вынужденную посадку на Землю. Каждый день он выходит на связь со своей далёкой планетой. Продолжительность каждого сеанса связи имеет экспоненциальное распределение с параметром λ. Прошедшие 100 сеансов связи в сумме длились 11 часов.
   (a) Оцените параметр λ экспоненциального распределения методом максимального правдоподобия
   (b) Постройте 95% доверительный интервал для λ
   (c) Проверьте гипотезу о том, что средняя продолжительность сеанса связи равна 5 минутам с помощью теста отношения правдоподобия, теста Вальда, теста множителей Лагранжа
4. [R] По ссылке http://people.reed.edu/~jones/141/Coal.html скачайте данные о количестве крупных аварий на английских угольных шахтах.
   (a) Методом максимального правдоподобия оцените две модели:
      i. Пуассоновская модель: количества аварий независимы и имеют Пуассоновское распределение с параметром λ.
      ii. Модель с раздутым нулём «zero inflated poisson model»: количества аварий независимы, с вероятностью p аварий не происходит вообще, с вероятностью $(1-p)$ количество аварий имеет Пуассоновское распределение с параметром λ. Смысл этой модели в том, что по сравнению с Пуассоновским распределением у события $\{X_i = 0\}$ вероятность выше, а пропорции вероятностей положительных количеств

аварий сохраняются. В модели с раздутым нулём дисперсия и среднее количества аварий отличаются. Чему в модели с раздутым нулём равна $\mathbb{P}(X_i = 0)$?

(b) С помощью тестов множителей Лагранжа, Вальда и отношения правдоподобия проверьте гипотезу H_0: верна пуассоновская модель против H_a: верна модель с раздутым нулём

(c) Постройте доверительные интервалы для оценённых параметров в обоих моделях

(d) Постройте доверительный интервал для вероятности полного отсутствия аварий по обеим моделям

5. Совместное распределение величин X и Y задано функцией

$$f(x, y) = \frac{\theta(\beta y)^x e^{-(\theta+\beta)y}}{x!}$$

Величина X принимает целые неотрицательные значения, а величина Y — действительные неотрицательные. Имеется случайная выборка (X_1, Y_1), $\ldots (X_n, Y_n)$.

(a) С помощью метода максимального правдоподобия оцените θ и β

(b) С помощью метода максимального правдоподобия оцените $a = \theta/(\beta + \theta)$

$\hat{\theta} = 1/\bar{Y}$, $\hat{\beta} = \bar{X}/\bar{Y}$, $\hat{a} = 1/(1 + \bar{X})$

6. Пусть $X = (X_1, \ldots, X_n)$ — случайная выборка из нормального распределения с математическим ожиданием μ и дисперсией ν; $\mu \in \mathbb{R}$ и $\nu > 0$ — неизвестные параметры. Реализация случайной выборки $x = (x_1, \ldots, x_n)$ приведена ниже:

-2.80	-1.12	-2.27	-1.31	-0.98	-2.15	-1.52	-2.82	-1.19	0.87

При помощи теста отношения правдоподобия, теста Вальда и теста множителей Лагранжа протестируйте гипотезу:

$$H_0 : \begin{cases} \mu = 0 \\ \nu = 1 \end{cases}$$

на уровне значимости 5%.

В данном примере мы имеем
$\theta = [\mu \quad \nu]'$ — вектор неизвестных параметров
$\Theta = \mathbb{R} \times (0; +\infty)$ — множество допустимых значений вектора неизвестных параметров
Функция правдоподобия имеет вид:

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} f_{X_i}(x_i, \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\nu}} \cdot \exp\left\{-\frac{(x_i - \mu)^2}{2\nu}\right\} = (2\pi)^{-n/2} \cdot \nu^{-n/2} \cdot \exp\left\{-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\nu}\right\}$$

Логарифмическая функция правдоподобия:

$$l(\theta) := \ln \mathrm{L}(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\nu - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\nu}$$

$\Theta_{UR} = \Theta$
$\Theta_R = \{(0, 1)\}$
Из системы уравнений

$$\begin{cases} \frac{\partial l}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i - \mu)}{\nu} = 0 \\ \frac{\partial l}{\partial \nu} = -\frac{n}{2\nu} + \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\nu^2} = 0 \end{cases}$$

находим
$\hat{\theta}_{UR} = (\hat{\mu}_{UR}, \hat{\nu}_{UR})$, где $\hat{\mu}_{UR} = \overline{x} = -1.5290$, $\hat{\nu}_{UR} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = 1.0603$
$\hat{\theta}_R = (\hat{\mu}_R, \hat{\nu}_R) = (0, 1)$
По имеющимся данным находим
$l(\hat{\theta}_R) = -\frac{10}{2}\ln(2\pi) - \frac{10}{2}\ln 1 - \frac{\sum_{i=1}^{n}(x_i - 0)^2}{2 \cdot 1} = -26.1804$
$l = -\frac{10}{2}\ln(2\pi) - \frac{10}{2}\ln(1.0603) - \frac{\sum_{i=1}^{n}(x_i + 1.5290)^2}{2 \cdot 1.0603} = -14.4824$
$LR_{\text{набл}} = -2(l(\hat{\theta}_R) - l) = -2 \cdot (-26.1804 + 14.4824) = 23.3959$
Критическое значение χ^2 распределения с двумя степенями свободы, отвечающее уровню значимости 5%, равно 5.9915. Следовательно, тест отношения правдоподобия говорит о том, что гипотеза H_0 должна быть отвергнута.
Для выполнения тестов Вальда и множителей Лагранжа нам понадобится информационная матрица Фишера
$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{v}$, $\frac{\partial^2 l}{\partial \nu \partial \mu} = -\frac{\sum_{i=1}^{n}(x_i - \mu)}{\nu^2}$, $\frac{\partial^2 l}{\partial \nu^2} = \frac{n}{2\nu^2} - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{\nu^3}$
$\mathbb{E}\frac{\partial^2 l}{\partial \nu \partial \mu} = -\frac{\sum_{i=1}^{n}\mathbb{E}(x_i - \mu)}{\nu^2} = 0$, $\mathbb{E}\frac{\partial^2 l}{\partial \nu^2} = \frac{n}{2\nu^2} - \frac{\sum_{i=1}^{n}\mathbb{E}(x_i - \mu)^2}{\nu^3} = \frac{n}{2\nu^2} - \frac{n\nu}{nu^3} = -\frac{n}{2\nu^2}$

$I(\theta) = -\mathbb{E}\begin{bmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \nu \partial \mu} \\ \frac{\partial^2 l}{\partial \nu \partial \mu} & \frac{\partial^2 l}{\partial \nu^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\nu} & 0 \\ 0 & \frac{n}{2\nu^2} \end{bmatrix}$

$I(\hat{\theta}_{UR}) = \begin{bmatrix} \frac{n}{\hat{\nu}_{UR}} & 0 \\ 0 & \frac{n}{2\cdot \hat{\nu}_{UR}^2} \end{bmatrix} = \begin{bmatrix} \frac{10}{1.0603} & 0 \\ 0 & \frac{10}{2\cdot 1.0603^2} \end{bmatrix} = \begin{bmatrix} 9.4307 & 0 \\ 0 & 4.4469 \end{bmatrix}$

$g(\hat{\theta}_{UR}) = \begin{bmatrix} \hat{\mu}_{UR} - 0 \\ \hat{\nu}_{UR} - 1 \end{bmatrix} = \begin{bmatrix} -1.5290 - 0 \\ 1.0603 - 1 \end{bmatrix} = \begin{bmatrix} -1.5290 \\ 0.0603 \end{bmatrix}$

$\frac{\partial g}{\partial \theta'} = \begin{bmatrix} \frac{\partial c_1}{\partial \mu} & \frac{\partial c_1}{\partial \nu} \\ \frac{\partial c_2}{\partial \mu} & \frac{\partial c_2}{\partial \nu} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\frac{\partial g'}{\partial \theta} = \begin{bmatrix} \frac{\partial c_1}{\partial \mu} & \frac{\partial c_2}{\partial \mu} \\ \frac{\partial c_1}{\partial \nu} & \frac{\partial c_2}{\partial \nu} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

$W_{\text{набл}} = g'(\hat{\theta}_{UR}) \cdot \left[\frac{\partial g}{\partial \theta'}(\hat{\theta}_{UR}) \cdot I^{-1}(\hat{\theta}_{UR}) \cdot \frac{\partial g'}{\partial \theta}(\hat{\theta}_{UR})\right]^{-1} g(\hat{\theta}_{UR}) =$

$\begin{bmatrix} -1.5290 & 0.0603 \end{bmatrix} \cdot \left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 9.4307 & 0 \\ 0 & 4.4469 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right]^{-1} \cdot \begin{bmatrix} -1.5290 \\ 0.0603 \end{bmatrix} = 22.0635$

Тест Вальда также говорит о том, что на основании имеющихся наблюдений гипотеза H_0 должна быть отвергнута.

$I(\hat{\theta}_R) = \begin{bmatrix} \frac{n}{\hat{\nu}_R} & 0 \\ 0 & \frac{n}{2\cdot \hat{\nu}_R^2} \end{bmatrix} = \begin{bmatrix} \frac{10}{1} & 0 \\ 0 & \frac{10}{2\cdot 1^2} \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}$

$\frac{\partial l}{\partial \theta}(\hat{\theta}_R) = \begin{bmatrix} \frac{\sum_{i=1}^n (x_i - \hat{\mu}_R)}{\hat{\nu}_R} \\ -\frac{n}{2\cdot \hat{\nu}_R} + \frac{\sum_{i=1}^n (x_i - \hat{\mu}_R)^2}{2\cdot \hat{\nu}_R^2} \end{bmatrix} = \begin{bmatrix} \frac{\sum_{i=1}^n (x_i - 0)}{1} \\ -\frac{10}{2\cdot 1} + \frac{\sum_{i=1}^n (x_i - 0)^2}{2\cdot 1^2} \end{bmatrix} = \begin{bmatrix} -15.29 \\ 11.9910 \end{bmatrix}$

$LM_{\text{набл}} = \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right]' \cdot I^{-1}(\hat{\theta}_R) \cdot \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right] = \begin{bmatrix} -15.29 & 11.9910 \end{bmatrix} = \cdot \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix}^{-1} \cdot \begin{bmatrix} -15.29 \\ 11.9910 \end{bmatrix} = 52.1354$

Тест множителей Лагранжа также указывает на то, что гипотеза H_0 должна быть отвергнута.

7. Пусть p — неизвестная вероятность выпадения орла при бросании монеты. Из 100 испытаний 42 раза выпал «Орел» и 58 — «Решка». Протестируйте на 5%-ом уровне значимости гипотезу о том, что монетка — «правильная» с помощью:

   (a) теста отношения правдоподобия

   (b) теста Вальда

   (c) теста множителей Лагранжа

В данной задаче мы имеем:
$\theta = p$ — вектор неизвестных параметров
$\Theta = (0, 1)$ — множество допустимых значений вектора неизвестных параметров
Функция правдоподобия имеет вид:

$$\mathrm{L}(\theta) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n p^{x_i} \cdot (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} \cdot (1-p)^{n - \sum_{i=1}^n x_i}$$

Логарифмическая функция правдоподобия:

$$l(\theta) := ln\mathrm{L}(\theta) = \left(\sum_{i=1}^n x_i\right) \cdot \ln p + \left(n - \sum_{i=1}^n x_i\right) \cdot \ln(1-p)$$

$\Theta_{UR} = \Theta$
$\Theta_R = \{0.5\}$
Решая уравнение правдоподобия

$$\frac{\partial l}{\partial p} = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p} = 0$$

получаем
$\hat{\theta}_{UR} = \hat{p}_{UR}$, где $\hat{p}_{UR} = \overline{x} = 0.42$
$\hat{\theta}_R = \hat{p}_R = 0.5$
По имеющимся данным находим
$l(\hat{\theta}_R) = 42 \cdot \ln(0.5) + (100 - 42) \cdot \ln(1 - 0.5) = -69.3147$
$l(\hat{\theta}_{UR} = 42 \cdot \ln(0.42) + (100 - 42) \cdot \ln(1 - 0.42) = -68.0292$
$LR_{\text{набл}} = -2(l(\hat{\theta}_R) - l) = -2 \cdot (-69.3147 + 68.0292) = 2.5710$
Критическое значение χ^2 распределения с одной степенью свободы, отвечающее за 5% уровень значимости, равно 3.8414. Следовательно, тест отношения правдоподобия говорит о том, что на основании имеющихся данных, основная гипотеза $H_0 : p = 0.5$ не может быть отвергнута.
Для выполнения тестов Вальда и множителей Лагранжа нам понадобится информационная матрица Фишера
$\frac{\partial^2 l}{\partial p^2} = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{n - \sum_{i=1}^n x_i}{(1-p)^2}$
$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 l}{\partial p^2}\right] = -\mathbb{E}\left[-\frac{\sum_{i=1}^n x_i}{p^2} - \frac{n - \sum_{i=1}^n x_i}{(1-p)^2}\right] = -\left(-\frac{np}{p^2} - \frac{n - np}{(1-p)^2}\right) = \frac{n}{p(1-p)}$
$I(\hat{\theta}_{UR}) = \frac{n}{\hat{p}_{UR}(1 - \hat{p}_{UR})} = \frac{100}{0.42 \times (1 - 0.42)} = 172.4138$
$g(\hat{\theta}_{UR}) = \hat{\theta}_{UR} - 0.5 = 0.42 - 0.5 = -0.08$
$\frac{\partial g}{\partial \theta'} = 1'$, $\frac{\partial g'}{\partial \theta} = 1$
$W_{\text{набл}} = g'(\hat{\theta}_{UR}) \cdot \left[\frac{\partial g}{\partial \theta'}(\hat{\theta}_{UR}) \cdot I^{-1}(\hat{\theta}_{UR}) \cdot \frac{\partial g'}{\partial \theta}(\hat{\theta}_{UR})\right]^{-1} g(\hat{\theta}_{UR}) = [-0.08]' \cdot [1' \cdot 172.4138^{-1} \cdot 1]^{-1} \cdot [-0.08] = 2.6272$
Тест Вальда также говорит о том, что гипотеза H_0 не отвергается.
$I(\hat{\theta}_R) = \frac{n}{\hat{p}_R(1 - \hat{p}_R)} = \frac{100}{0.5 \times (1 - 0.5)} = 400$
$\frac{\partial l}{\partial \theta}(\hat{\theta}_R) = \frac{\sum_{i=1}^n x_i}{\hat{p}_R} - \frac{n - \sum_{i=1}^n x_i}{1 - \hat{p}_R} = \frac{42}{0.5} - \frac{100 - 42}{1 - 0.5} = -32$
$LM_{\text{набл}} = \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right]' \cdot I^{-1}(\hat{\theta}_R) \cdot \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right] = [-32]' \cdot [400]^{-1} \cdot [-32] = 2.56$

Согласно тесту множителей Лагранжа, основная гипотеза H_0 не может быть отвергнута.

8. Пусть $x = (x_1, \ldots, x_n)$ — реализация случайной выборки из распределения Пуассона с неизвестным параметром $\lambda > 0$. Известно, что выборочное среднее $\overline{x}$ по 80 наблюдениям равно 1.7. Протестируйте на 5%-ом уровне значимости гипотезу $H_0 : \lambda = 2$ с помощью

   (a) теста отношения правдоподобия

(b) теста Вальда

(c) теста множителей Лагранжа

В данной задаче мы имеем
$\theta = \lambda$ — вектор неизвестных параметров
$\Theta = (0, +\infty)$ — множество допустимых значений вектора неизвестных параметров
Функция правдоподобия имеет вид:

$$\mathrm{L}(\theta) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{x_1! \ldots x_n!} e^{-\lambda n}$$

Логарифмическая функция правдоподобия:

$$l(\theta) := ln\mathrm{L}(\theta) = \left(\sum_{i=1}^{n} x_i \right) \cdot \ln \lambda - \sum_{i=1}^{n} \ln(x_i!) - \lambda n$$

$\Theta_{UR} = \Theta$
$\Theta_R = \{2\}$
Решая уравнение правдоподобия

$$\frac{\partial l}{\partial p} = \frac{\sum_{i=1}^n x_i}{\lambda} - n = 0$$

получаем
$\hat{\theta}_{UR} = \hat{\lambda}_{UR}$, где $\hat{\lambda}_{UR} = \overline{x} = 1.7$
$\hat{\theta}_R = \hat{p}_R = 2$
По имеющимся данным находим
$l(\hat{\theta}_R) = (80 \cdot 1.7) \cdot \ln(2) - \sum_{i=1}^n \ln(x_i!) - 2 \cdot 80 = -65.7319$
$l(\hat{\theta}_{UR} = (80 \cdot 1.7) \cdot \ln(1.7) - \sum_{i=1}^n \ln(x_i!) - 1.7 \cdot 80 = -63.8345$
$LR_{\text{набл}} = -2(l(\hat{\theta}_R) - l) = -2 \cdot (-65.7319 + 63.8345) = 3.7948$
Критическое значение χ^2 распределения с одной степенью свободы, отвечающее за 5% уровень значимости, равно 3.8414. Следовательно, тест отношения правдоподобия говорит о том, что на основании имеющихся данных, основная гипотеза $H_0 : \lambda = 2$ не может быть отвергнута.
Для выполнения тестов Вальда и множителей Лагранжа нам понадобится информационная матрица Фишера
$\frac{\partial^2 l}{\partial p^2} = -\frac{\sum_{i=1}^n x_i}{\lambda^2}$
$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 l}{\partial p^2}\right] = -\mathbb{E}\left[-\frac{\sum_{i=1}^n x_i}{\lambda^2}\right] = -\left(-\frac{n\lambda}{\lambda^2}\right) = \frac{n}{\lambda}$
$I(\hat{\theta}_{UR}) = \frac{n}{\hat{\lambda}_{UR}} = \frac{80}{1.7} = 47.0588$
$g(\hat{\theta}_{UR}) = \hat{\theta}_{UR} - 2 = 1.7 - 2 = -0.3$
$\frac{\partial g}{\partial \theta'} = 1'$, $\frac{\partial g'}{\partial \theta} = 1$
$W_{\text{набл}} = g'(\hat{\theta}_{UR}) \cdot \left[\frac{\partial g}{\partial \theta'}(\hat{\theta}_{UR}) \cdot I^{-1}(\hat{\theta}_{UR}) \cdot \frac{\partial g'}{\partial \theta}(\hat{\theta}_{UR})\right]^{-1} g(\hat{\theta}_{UR}) = [-0.3]' \cdot [1' \cdot 47.0588^{-1} \cdot 1]^{-1} \cdot [-0.3] = 4.2352$
Поскольку наблюдаемое значение статистики Вальда превосходит критическое значение 3.8414, то гипотеза H_0 должна быть отвергнута.
$I(\hat{\theta}_R) = \frac{n}{\hat{\lambda}_R} = \frac{80}{2} = 40$
$\frac{\partial l}{\partial \theta}(\hat{\theta}_R) = \frac{\sum_{i=1}^n x_i}{\hat{\lambda}_R} - n = \frac{80 \cdot 1.7}{2} - 80 = -12$
$LM_{\text{набл}} = \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right]' \cdot I^{-1}(\hat{\theta}_R) \cdot \left[\frac{\partial l}{\partial \theta}(\hat{\theta}_R)\right] = [-12]' \cdot [40]^{-1} \cdot [-12] = 3.6$

Согласно тесту множителей Лагранжа, основная гипотеза H_0 не может быть отвергнута.

6 Логит и пробит

1. Случайная величина X имеет логистическое распределение, если её функция плотности имеет вид $f(x) = e^{-x}/(1 + e^{-x})^2$.

   (a) Является ли $f(x)$ чётной?

   (b) Постройте график $f(x)$

   (c) Найдите функцию распределения, $F(x)$

   (d) Найдите $\mathbb{E}(X)$, $\mathrm{Var}(X)$

   (e) На какое известный закон распределения похож логистический?

   $f(x)$ чётная, $\mathbb{E}(X) = 0$, $\mathrm{Var}(X) = \pi^2/3$, логистическое похоже на $N(0, \pi^2/3)$

2. Логит модель часто формулируют в таком виде:

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

   где ε_i имеет логистическое распределение, и

$$y_i = \begin{cases} 1, \ y_i^* \geqslant 0 \\ 0, \ y_i^* < 0 \end{cases}$$

   (a) Выразите $\mathbb{P}(y_i = 1)$ с помощью логистической функции распределения

(b) Найдите $\ln\left(\frac{\mathbb{P}(y_i=1)}{\mathbb{P}(y_i=0)}\right)$

$\ln\left(\frac{\mathbb{P}(y_i=1)}{\mathbb{P}(y_i=0)}\right) = \beta_1 + \beta_2 x_i$.

3. [R] Сравните на одном графике
   (a) Функции плотности логистической и нормальной $N(0, \pi^2/3)$ случайных величин
   (b) Функции распределения логистической и нормальной $N(0, \pi^2/3)$ случайных величин

4. Как известно, Фрекен Бок любит пить коньяк по утрам. За прошедшие 4 дня она записала, сколько рюмочек коньяка выпила утром, x_i, и видела ли она в этот день привидение, y_i,

y_i	1	0	1	0
x_i	2	1	3	0

Зависимость между y_i и x_i описывается логит-моделью,

$$\ln\left(\frac{\mathbb{P}(y_i=1)}{\mathbb{P}(y_i=0)}\right) = \beta_1 + \beta_2 x_i$$

   (a) Выпишите в явном виде логарифмическую функцию максимального правдоподобия
   (b) [R] Найдите оценки параметров β_1 и β_2

5. При оценке логит модели

$$\mathbb{P}(y_i = 1) = \Lambda(\beta_1 + \beta_2 x_i)$$

оказалось, что $\hat{\beta}_1 = 0.7$ и $\hat{\beta}_2 = 3$. Найдите максимальный предельный эффект роста x_i на вероятность $\mathbb{P}(y_i = 1)$.

6. Винни-Пух знает, что мёд бывает правильный, $honey_i = 1$, и неправильный, $honey_i = 0$. Пчёлы также бывают правильные, $bee_i = 1$, и неправильные, $bee_i = 0$. По 100 своим попыткам добыть мёд Винни-Пух составил таблицу сопряженности:

	$honey_i = 1$	$honey_i = 0$
$bee_i = 1$	12	36
$bee_i = 0$	32	20

Используя метод максимального правдоподобия Винни-Пух хочет оценить логит-модель для прогнозирования правильности мёда с помощью правильности пчёл:

$$\ln\left(\frac{\mathbb{P}(honey_i=1)}{\mathbb{P}(honey_i=0)}\right) = \beta_1 + \beta_2 bee_i$$

   (a) Выпишите функцию правдоподобия для оценки параметров β_1 и β_2
   (b) Оцените неизвестные параметры
   (c) С помощью теста отношения правдоподобия проверьте гипотезу о том, правильность пчёл не связана с правильностью мёда на уровне значимости 5%.
   (d) Держась в небе за воздушный шарик, Винни-Пух неожиданно понял, что перед ним неправильные пчёлы. Помогите ему оценить вероятность того, что они делают неправильный мёд.

Для краткости введем следующие обозначения: $y_i = honey_i$, $d_i = bee_i$[1].

(a) Функция правдоподобия имеет следующий вид:

$$\mathrm{L}(\beta_1, \beta_2) = \prod_{i=1}^{n} \mathbb{P}_{\beta_1,\beta_2}(\{Y_i = y_i\}) = \prod_{i:y_i=0} \mathbb{P}_{\beta_1,\beta_2}(\{Y_i = 1\}) \cdot \prod_{i:y_i=1} \mathbb{P}_{\beta_1,\beta_2}(\{Y_i = 0\}) =$$

$$\prod_{i:y_i=1} \Lambda(\beta_1 + \beta_2 d_i) \cdot \prod_{i:y_i=0} [1 - \Lambda(\beta_1 + \beta_2 d_i)] =$$

$$\prod_{i:y_i=1,d_i=1} \Lambda(\beta_1 + \beta_2) \cdot \prod_{i:y_i=1,d_i=0} \Lambda(\beta_1) \cdot \prod_{i:y_i=0,d_i=1} [1 - \Lambda(\beta_1 + \beta_2)] \cdot \prod_{i:y_i=0,d_i=0} [1 - \Lambda(\beta_1)] =$$

$$\Lambda(\beta_1 + \beta_2)^{\#\{i:y_i=1,d_i=1\}} \cdot \Lambda(\beta_1)^{\#\{i:y_i=1,d_i=0\}} \cdot [1 - \Lambda(\beta_1 + \beta_2)]^{\#\{i:y_i=0,d_i=1\}} \cdot [1 - \Lambda(\beta_1)]^{\#\{i:y_i=0,d_i=0\}}$$

[1] Y_i — случайный Мёд, y_i — реализация случайного Мёда (наблюдаемый Мёд)

где
$$\Lambda(x) = \frac{e^x}{1+e^x} \tag{4}$$
логистическая функция распределения, $\#A$ означает число элементов множества A.

(b) Введём следующие обозначения:
$$a := \Lambda(\beta_1) \tag{5}$$
$$b := \Lambda(\beta_1 + \beta_2) \tag{6}$$

Тогда с учетом имеющихся наблюдений функция правдоподобия принимает вид:
$$\mathrm{L}(a, b) = b^{12} \cdot a^{32} \cdot [1-b]^{36} \cdot [1-a]^{20}$$

Логарифмическая функция правдоподобия:
$$l(a, b) = \ln \mathrm{L}(a, b) = 12 \ln b + 32 \ln a + 36 \ln[1-b] + 20 \ln[1-a]$$

Решая систему уравнений правдоподобия
$$\begin{cases} \frac{\partial l}{\partial a} = \frac{32}{a} - \frac{20}{1-a} = 0 \\ \frac{\partial l}{\partial b} = \frac{12}{b} - \frac{36}{1-b} = 0 \end{cases}$$

получаем $\hat{a} = \frac{8}{13}$, $\hat{b} = \frac{1}{4}$. Из формул (4) и (5), находим $\hat{\beta}_{1,UR} = \ln\left(\frac{\hat{a}}{1-\hat{a}}\right) = \ln\left(\frac{8}{5}\right) = 0.47$. Далее, из (4) и (6) имеем $\hat{\beta}_{1,UR} + \hat{\beta}_{2,UR} = \ln\left(\frac{\hat{b}}{1-\hat{b}}\right)$. Следовательно, $\hat{\beta}_{2,UR} = \ln\left(\frac{\hat{b}}{1-\hat{b}}\right) - \hat{\beta}_{1,UR} = \ln\left(\frac{1}{3}\right) - \ln\left(\frac{8}{5}\right) = -1.57$.

(c) Гипотеза, состоящая в том, что «правильность Мёда не связана с правильностью пчёл» формализуется как $H_0 : \beta_2 = 0$. Протестируем данную гипотезу при помощи теста отношения правдоподобия. Положим в функции правдоподобия $\mathrm{L}(\beta_1, \beta_2)$ $\beta_2 = 0$. Тогда с учетом (5) и (6) получим
$$\mathrm{L}(a, b = a) = a^{32+12} \cdot [1-a]^{20+36}$$

В этом случае логарифмическая функция правдоподобия имеет вид:
$$l(a, b = a) := \mathrm{L}(a, b = a) = 44 \ln a + 56 \ln[1-a]$$

Решаем уравнение правдоподобия
$$\frac{\partial l}{\partial a} = \frac{44}{a} - \frac{56}{1-a} = 0$$

и получаем $\hat{a} = \frac{11}{25}$. Следовательно, согласно (4) и (5), $\hat{\beta}_{1,R} = -0.24$ и $\hat{\beta}_{2,R} = 0$.
Статистика отношения правдоподобия имеет вид:
$$LR = -2(l(\hat{\beta}_{1,R}, \hat{\beta}_{2,R}) - l(\hat{\beta}_{1,UR}, \hat{\beta}_{2,UR}))$$

и имеет асимптотическое χ^2 распределение с числом степеней свободы, равным числу ограничений, составляющих гипотезу H_0, т.е. в данном случае $LR \overset{a}{\sim} \chi^2_1$.
Находим наблюдаемое значение статистики отношения правдоподобия:
$$l(\hat{\beta}_{1,R}, \hat{\beta}_{2,R}) = l(\hat{a}_R, \hat{b}_R = \hat{a}_R) = 44 \ln \hat{a}_R + 56 \ln[1-\hat{a}_R] = 44 \ln\left[\frac{11}{25}\right] + 56 \ln\left[1 - \frac{11}{25}\right] = -68.59$$
$$l(\hat{\beta}_{1,UR}, \hat{\beta}_{2,UR}) = l(\hat{a}_{UR}, \hat{b}_{UR}) = 12 \ln \hat{b}_{UR} + 32 \ln \hat{a}_{UR} + 36 \ln[1-\hat{b}_{UR}] + 20 \ln[1-\hat{a}_{UR}] =$$
$$12 \ln\left[\frac{1}{4}\right] + 32 \ln\left[\frac{8}{13}\right] + 36 \ln\left[1 - \frac{1}{4}\right] + 20 \ln\left[1 - \frac{8}{13}\right] = -61.63$$

Следовательно, $LR_{\text{набл}} = -2(-68.59 + 61.63) = 13.92$, при этом критическое значение χ^2 распределения с одной степенью свободы для 5% уровня значимости равна 3.84. Значит, на основании теста отношения правдоподобия гипотеза $H_0 : \beta_2 = 0$ должна быть отвергнута. Таким образом, данные показывают, что, в действительности, правильность мёда связана с правильностью пчёл.

(d)
$$\hat{\mathbb{P}}\{honey = 0 | bee = 0\} = 1 - \hat{\mathbb{P}}\{honey = 1 | bee = 0\} = 1 - \frac{\exp\{\hat{\beta}_{1,UR} + \hat{\beta}_{2,UR} \cdot 0\}}{1 + \exp\{\hat{\beta}_{1,UR} + \hat{\beta}_{2,UR} \cdot 0\}} =$$
$$1 - \frac{\exp\{\ln\left(\frac{8}{5}\right)\}}{1 + \exp\{\ln\left(\frac{8}{5}\right)\}} = 1 - 0.62 = 0.38$$

7 Мультиколлинеарность

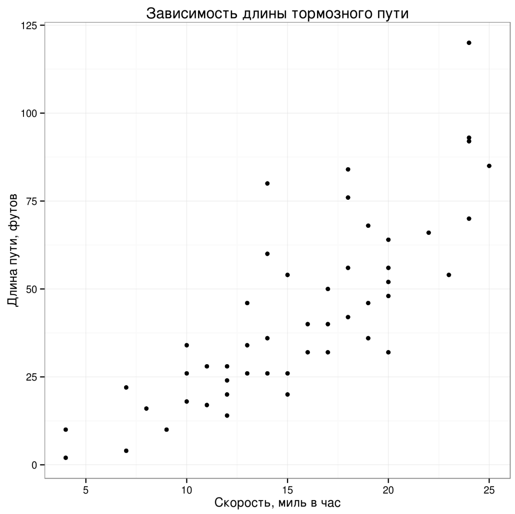
1. Сгенерируйте данные так, чтобы при оценке модели $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$ оказывалось, что по отдельности оценки коэффициентов $\hat{\beta}_2$ и $\hat{\beta}_3$ незначимы, но модель в целом — значима.
2. В этом задании нужно сгенерировать зависимую переменную y и два регрессора x и z.
   (a) Сгенерируйте данные так, чтобы корреляция между регрессорами x и z была больше 0.9, и проблема мультиколлинеарности есть, т.е. по отдельности регрессоры не значимы, но регрессия в целом — значима.
   (b) А теперь сгенерируйте данные так, чтобы корреляция между регрессорами была по-прежнему больше 0.9, но проблемы мультиколлинеарности бы не было, т.е. все коэффициенты были бы значимы.

(c) Есть несколько способов, как изменить генерации случайных величин, чтобы перейти от ситуации «a» к ситуации «b». Назовите хотя бы два.

увеличить количество наблюдений, уменьшить дисперсию случайной ошибки

3. Исследуем зависимость длины тормозного пути автомобиля от скорости по историческим данным 1920-х годов.

```
h <- cars
ggplot(h,aes(x=speed,y=dist))+geom_point()+
    labs(title="Зависимость длины тормозного пути",
    x="Скорость, миль в час",y="Длина пути, футов")
```




```
speed.mean <- mean(h$speed)
```

Построим результаты оценивания нецентрированной регрессии:

```
cars.model <- lm(dist~speed+I(speed^2)+I(speed^3),data=h)
cars.table <- as.table(coeftest(cars.model))
rownames(cars.table) <-c("Константа","speed","speed^2","speed^3")
```

с тремя переменными руками громоздко делать, а с двумя вроде не видно мультик.

```
xtable(cars.table)
```

	Estimate	Std. Error	t value	Pr(>|t|)
Константа	-19.51	28.41	-0.69	0.50
speed	6.80	6.80	1.00	0.32
speed^2	-0.35	0.50	-0.70	0.49
speed^3	0.01	0.01	0.91	0.37

Ковариационная матрица коэффициентов имеет вид:

```
cars.vcov <- vcov(cars.model)
rownames(cars.vcov) <-c("Константа","speed","speed^2","speed^3")
colnames(cars.vcov) <-c("Константа","speed","speed^2","speed^3")
xtable(cars.vcov)
```

(a) Проверьте значимость всех коэффициентов и регрессии в целом

(b) Постройте 95%-ый доверительный интервал для $\mathbb{E}(dist)$ при $speed = 10$

(c) Постройте 95%-ый доверительный интервал для $\mathbb{E}(ddist/dspeed)$ при $speed = 10$

	Константа	speed	speed^2	speed^3
Константа	806.86	-186.20	12.88	-0.27
speed	-186.20	46.26	-3.35	0.07
speed^2	12.88	-3.35	0.25	-0.01
speed^3	-0.27	0.07	-0.01	0.00

(d) Как выглядит уравнение регрессии, если вместо *speed* использовать центрированную скорость? Известно, что средняя скорость равна 15.4

(e) С помощью регрессии с центрированной скоростью ответьте на предыдущие вопросы.

4. Пионеры, Крокодил Гена и Чебурашка собирали металлолом несколько дней подряд. В распоряжение иностранной шпионки, гражданки Шапокляк, попали ежедневные данные по количеству собранного металлолома: вектор g — для Крокодила Гены, вектор h — для Чебурашки и вектор x — для Пионеров. Гена и Чебурашка собирали вместе, поэтому выборочная корреляция $\mathrm{sCorr}(g,h) = -0.9$. Гена и Чебурашка собирали независимо от Пионеров, поэтому выборочные корреляции $\mathrm{sCorr}(g,x) = 0$, $\mathrm{sCorr}(h,x) = 0$. Если регрессоры g, h и x центрировать и нормировать, то получится матрица $\tilde{X}$.

   (a) Найдите параметр обусловленности матрицы $(\tilde{X}'\tilde{X})$

   (b) Вычислите одну или две главные компоненты (выразите их через вектор-столбцы матрицы $\tilde{X}$), объясняющие не менее 70% общей выборочной дисперсии регрессоров

   (c) Шпионка Шапокляк пытается смоделировать ежедневный выпуск танков, y. Выразите коэффициенты регрессии $y = \beta_1 + \beta_2 g + \beta_3 h + \beta_4 x + \varepsilon$ через коэффициенты регрессии на главные компоненты, объясняющие не менее 70% общей выборочной дисперсии.

5. Для модели $y_i = \beta x_i + \varepsilon$ рассмотрите модель Ridge regression с коэффициентом λ.

   (a) Выведите формулу для $\hat{\beta}_{RR}$

   (b) Найдите $\mathbb{E}(\hat{\beta}_{RR})$, смещение оценки $\hat{\beta}_{RR}$,

   (c) Найдите $\mathrm{Var}(\hat{\beta}_{RR})$, $MSE(\hat{\beta}_{RR})$

   (d) Всегда ли оценка $\hat{\beta}_{RR}$ смещена?

   (e) Всегда ли оценка $\hat{\beta}_{RR}$ имеет меньшую дисперсию, чем $\hat{\beta}_{ols}$?

   (f) Найдите такое λ, что $MSE(\hat{\beta}_{RR}) < MSE(\hat{\beta}_{ols})$

6. Известно, что в модели $y = X\beta + \varepsilon$ все регрессоры ортогональны.

   (a) Как выглядит матрица $X'X$ в случае ортогональных регрессоров?

   (b) Выведите $\hat{\beta}_{rr}$ в явном виде

   (c) Как связаны между собой $\hat{\beta}_{rr}$ и $\hat{\beta}_{ols}$?

7. Для модели $y_i = \beta x_i + \varepsilon_i$ выведите в явном виде $\hat{\beta}_{lasso}$.

8. Предположим, что для модели $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$ выборочная корреляционная матрица регрессоров x_2, x_3, x_4 имеет вид

$$C = \begin{pmatrix} 1 & r & r \\ r & 1 & r \\ r & r & 1 \end{pmatrix}$$

   (a) Найдите такое значение $r^* \in (-1;1)$ коэффициента корреляции, при котором $\det C = 0$.

   (b) Найдите собственные значения и собственные векторы матрицы C при корреляции равной найденному r^*.

   (c) Найдите число обусловленности матрицы C при корреляции равной найденному r^*.

(d) Сделайте вывод о наличии мультиколлинеарности в модели при корреляции равной найденному r^*.

$r^* = -1/2$

9. Предположим, что для модели $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$ выборочная корреляционная матрица регрессоров x_2, x_3, x_4 и x_5 имеет вид

$$C = \begin{pmatrix} 1 & r & r & r \\ r & 1 & r & r \\ r & r & 1 & r \\ r & r & r & 1 \end{pmatrix}$$

(a) Найдите такое значение $r^* \in (-1; 1)$ коэффициента корреляции, при котором $\det C = 0$.

(b) Найдите собственные значения и собственные векторы матрицы C при корреляции равной найденному r^*.

(c) Найдите число обусловленности матрицы C при корреляции равной найденному r^*.

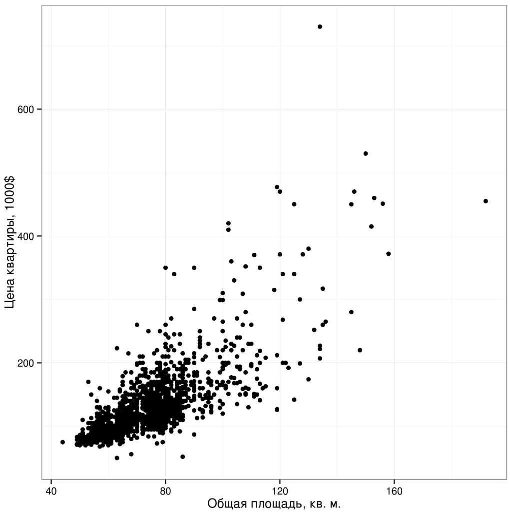
(d) Сделайте вывод о наличии мультиколлинеарности в модели при корреляции равной найденному r^*.

$r^* = -1/3$

8 Гетероскедастичность

1. Что такое гетероскедастичность? Гомоскедастичность?
2. Диаграмма рассеяния стоимости квартиры в Москве (в 1000$) и общей площади квартиры имеет вид:

```
ggplot(flats,aes(x=totsp,y=price))+geom_point()+
    labs(x="Общая площадь, кв. м.",y="Цена квартиры, 1000$")
```



Какие подходы к оцениванию зависимости имеет смысл посоветовать исходя из данного графика?

По графику видно, что с увеличением общей площади увеличивается разброс цены. Поэтому разумно, например, рассмотреть следующие подходы:

(a) Перейти к логарифмам, т.е. оценивать модель $\ln price_i = \beta_1 + \beta_2 \ln totsp_i + \varepsilon_i$

(b) Оценивать квантильную регрессию. В ней угловые коэффициенты линейной зависимости будут отличаться для разных квантилей переменной $price$.

(c) Обычную модель линейной регрессии с гетероскедастичностью вида $Var(\varepsilon_i) = \sigma^2 totsp_i^2$

3. По наблюдениям $x = (1, 2, 3)'$, $y = (2, -1, 3)'$ оценивается модель $y = \beta_1 + \beta_2 x + \varepsilon$. Ошибки ε гетероскедастичны и известно, что $\operatorname{Var}(\varepsilon_i) = \sigma^2 \cdot x_i^2$.

(a) Найдите оценки $\hat{\beta}_{ols}$ с помощью МНК и их ковариационную матрицу

(b) Найдите оценки $\hat{\beta}_{gls}$ с помощью обобщенного МНК и их ковариационную матрицу

4. В модели $y = \hat{\beta}_1 + \hat{\beta}_2 x + \varepsilon$ присутствует гетероскедастичность вида $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$. Как надо преобразовать исходные регрессоры и зависимую переменную, чтобы устранить гетероскедастичность? Поделить зависимую переменную и каждый регрессор, включая единичный столбец, на $|x_i|$.
5. В модели $y = \hat{\beta}_1 + \hat{\beta}_2 x + \varepsilon$ присутствует гетероскедастичность вида $\mathrm{Var}(\varepsilon_i) = \lambda |x_i|$. Как надо преобразовать исходные регрессоры и зависимую переменную, чтобы устранить гетероскедастичность? Поделить зависимую переменную и каждый регрессор, включая единичный столбец, на $\sqrt{|x_i|}$.
6. Известно, что после деления каждого уравнения регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ на x_i^2 гетероскедастичность ошибок была устранена. Какой вид имела дисперсия ошибок, $\mathrm{Var}(\varepsilon_i)$?

   $\mathrm{Var}(\varepsilon_i) = c x_i^4$
7. Известно, что после деления каждого уравнения регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ на $\sqrt{x_i}$ гетероскедастичность ошибок была устранена. Какой вид имела дисперсия ошибок, $\mathrm{Var}(\varepsilon_i)$?

   $\mathrm{Var}(\varepsilon_i) = c x_i$
8. Для линейной регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$ была выполнена сортировка наблюдений по возрастанию переменной x. Исходная модель оценивалась по разным частям выборки:

Выборка	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	RSS
$i = 1, \ldots, 30$	1.21	1.89	2.74	48.69
$i = 1, \ldots, 11$	1.39	2.27	2.36	10.28
$i = 12, \ldots, 19$	0.75	2.23	3.19	5.31
$i = 20, \ldots, 30$	1.56	1.06	2.29	14.51

   Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Протестируйте ошибки на гетероскедастичность на уровне значимости 5%.

   Протестируем гетероскедастичность ошибок при помощи теста Голдфельда- Квандта. $H_0 : \mathrm{Var}(\varepsilon_i) = \sigma^2$, $H_a : \mathrm{Var}(\varepsilon_i) = f(x_i)$

   (a) Тестовая статистика $GQ = \frac{RSS_3/(n_3-k)}{RSS_1/(n_1-k)}$, где $n_1 = 11$ — число наблюдений в первой подгруппе, $n_3 = 11$ — число наблюдений в последней подгруппе, $k = 3$ — число факторов в модели, считая единичный столбец.

   (b) Распределение тестовой статистики при верной H_0: $GQ \sim F_{n_3-k, n_1-k}$

   (c) Наблюдаемое значение $GQ_{obs} = 1.41$

   (d) Область в которой H_0 не отвергается: $GQ \in [0; 3.44]$

   (e) Статистический вывод: поскольку $GQ_{obs} \in [0; 3.44]$, то на основании имеющихся наблюдений на уровне значимости 5% основная гипотеза H_0 не может быть отвергнута. Таким образом, тест Голдфельда-Квандта не выявил гетероскедастичность.

9. Для линейной регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$ была выполнена сортировка наблюдений по возрастанию переменной x. Исходная модель оценивалась по разным частям выборки:

Выборка	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	RSS
$i = 1, \ldots, 50$	1.16	1.99	2.97	174.69
$i = 1, \ldots, 21$	0.76	2.25	3.18	20.41
$i = 22, \ldots, 29$	0.85	1.81	3.32	3.95
$i = 30, \ldots, 50$	1.72	1.41	2.49	130.74

   Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Протестируйте ошибки на гетероскедастичность на уровне значимости 1%.

   Протестируем гетероскедастичность ошибок при помощи теста Голдфельда- Квандта. $H_0 : \mathrm{Var}(\varepsilon_i) = \sigma^2$, $H_a : \mathrm{Var}(\varepsilon_i) = f(x_i)$

   (a) Тестовая статистика $GQ = \frac{RSS_3/(n_3-k)}{RSS_1/(n_1-k)}$, где $n_1 = 21$ — число наблюдений в первой подгруппе, $n_3 = 21$ — число наблюдений в последней подгруппе, $k = 3$ — число факторов в модели, считая единичный столбец.

   (b) Распределение тестовой статистики при верной H_0: $GQ \sim F_{n_3-k, n_1-k}$

   (c) Наблюдаемое значение $GQ_{obs} = 6.49$

   (d) Область в которой H_0 не отвергается: $GQ \in [0; 3.12]$

   (e) Статистический вывод: поскольку $GQ_{obs} \notin [0; 3.12]$, то на основании имеющихся наблюдений на уровне значимости 1% основная гипотеза H_0 отвергается. Таким образом, тест Голдфельда-Квандта выявил гетероскедастичность.

10. Для линейной регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$ была выполнена сортировка наблюдений по возрастанию переменной x. Исходная модель оценивалась по разным частям выборки:

Выборка	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	RSS
$i = 1, \ldots, 30$	0.96	2.25	3.44	52.70
$i = 1, \ldots, 11$	1.07	2.46	2.40	5.55
$i = 12, \ldots, 19$	1.32	1.01	2.88	11.69
$i = 20, \ldots, 30$	1.04	2.56	4.12	16.00

Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Протестируйте ошибки на гетероскедастичность на уровне значимости 5%.

Протестируем гетероскедастичность ошибок при помощи теста Голдфельда- Квандта. $H_0 : \mathrm{Var}(\varepsilon_i) = \sigma^2$, $H_a : \mathrm{Var}(\varepsilon_i) = f(x_i)$

(a) Тестовая статистика $GQ = \frac{RSS_3/(n_3-k)}{RSS_1/(n_1-k)}$, где $n_1 = 11$ — число наблюдений в первой подгруппе, $n_3 = 11$ — число наблюдений в последней подгруппе, $k = 3$ — число факторов в модели, считая единичный столбец.

(b) Распределение тестовой статистики при верной H_0: $GQ \sim F_{n_3-k,n_1-k}$

(c) Наблюдаемое значение $GQ_{obs} = 2.88$

(d) Область в которой H_0 не отвергается: $GQ \in [0; 3.44]$

(e) Статистический вывод: поскольку $GQ_{obs} \in [0; 3.44]$, то на основании имеющихся наблюдений на уровне значимости 5% основная гипотеза H_0 не может быть отвергнута. Таким образом, тест Голдфельда-Квандта не выявил гетероскедастичность.

11. Для линейной регрессии $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$ была выполнена сортировка наблюдений по возрастанию переменной x. Исходная модель оценивалась по разным частям выборки:

Выборка	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	RSS
$i = 1, \ldots, 50$	0.93	2.02	3.38	145.85
$i = 1, \ldots, 21$	1.12	2.01	3.32	19.88
$i = 22, \ldots, 29$	0.29	2.07	2.24	1.94
$i = 30, \ldots, 50$	0.87	1.84	3.66	117.46

Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Протестируйте ошибки на гетероскедастичность на уровне значимости 5%.

Протестируем гетероскедастичность ошибок при помощи теста Голдфельда- Квандта. $H_0 : \mathrm{Var}(\varepsilon_i) = \sigma^2$, $H_a : \mathrm{Var}(\varepsilon_i) = f(x_i)$

(a) Тестовая статистика $GQ = \frac{RSS_3/(n_3-k)}{RSS_1/(n_1-k)}$, где $n_1 = 21$ — число наблюдений в первой подгруппе, $n_3 = 21$ — число наблюдений в последней подгруппе, $k = 3$ — число факторов в модели, считая единичный столбец.

(b) Распределение тестовой статистики при верной H_0: $GQ \sim F_{n_3-k,n_1-k}$

(c) Наблюдаемое значение $GQ_{obs} = 5.91$

(d) Область в которой H_0 не отвергается: $GQ \in [0; 2.21]$

(e) Статистический вывод: поскольку $GQ_{obs} \notin [0; 2.21]$, то на основании имеющихся наблюдений на уровне значимости 5% основная гипотеза H_0 отвергается. Таким образом, тест Голдфельда-Квандта выявил гетероскедастичность.

12. Рассмотрим линейную регрессию $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$. При оценивании с помощью МНК были получены результаты: $\hat{\beta}_1 = 1.21$, $\hat{\beta}_2 = 1.11$, $\hat{\beta}_3 = 3.15$, $R^2 = 0.72$.
Оценена также вспомогательная регрессия: $\hat{\varepsilon}_i = \delta_1 + \delta_2 x_i + \delta_3 z_i + \delta_4 x_i^2 + \delta_5 z_i^2 + \delta_6 x_i z_i + u_i$.
Результаты оценивания следующие: $\hat{\delta}_1 = 1.50$, $\hat{\delta}_2 = -2.18$, $\hat{\delta}_3 = 0.23$, $\hat{\delta}_4 = 1.87$, $\hat{\delta}_5 = -0.56$, $\hat{\delta}_6 = -0.09$, $R^2_{aux} = 0.36$
Известно, что ошибки в модели являются независимыми нормальными случайными величинами с нулевым математическим ожиданием. Протестируйте ошибки на гетероскедастичность на уровне значимости 5%.

Протестируем гетероскедастичность ошибок при помощи теста Уайта. $H_0 : \mathrm{Var}(\varepsilon_i) = \sigma^2$, $H_a : \mathrm{Var}(\varepsilon_i) = \delta_1 + \delta_2 x_i + \delta_3 z_i + \delta_4 x_i^2 + \delta_5 z_i^2 + \delta_6 x_i z_i$.

(a) Тестовая статистика $W = n \cdot R^2_{aux}$, где n — число наблюдений, R^2_{aux} — коэффициент детерминации для вспомогательной регрессии.

(b) Распределение тестовой статистики при верной H_0: $W \sim \chi^2_{k_{aux}-1}$, где $k_{aux} = 6$ — число регрессоров во вспомогательной регрессии, считая константу.

(c) Наблюдаемое значение тестовой статистики: $W_{obs} = 18$

(d) Область в которой H_0 не отвергается: $W \in [0; W_{crit}] = [0; 11.07]$

(e) Статистический вывод: поскольку $W_{obs} \notin [0; 11.07]$, то на основании имеющихся наблюдений на уровне значимости 5% основная гипотеза H_0 отвергается. Таким образом, тест Уайта выявил гетероскедастичность.

13. Объясните, с какой целью используются стандартные ошибки в форме Уайта. Приведите развернутый ответ. Верно ли, что стандартные ошибки в форме Уайта позволяют

(a) устранить гетероскедастичность?

(b) корректно тестировать гипотезы относительно коэффициентов регрессии в условиях гетероскедастичности?

14. Объясните, с какой целью используются стандартные ошибки в форме Невье–Веста. Приведите развернутый ответ. Верно ли, что стандартные ошибки в форме Невье–Веста позволяют
    (a) устранить гетероскедастичность?
    (b) корректно тестировать гипотезы относительно коэффициентов регрессии в условиях гетероскедастичности?
15. Рассматривается модель $y_t = \beta_1 + \varepsilon_t$, где ошибки ε_t — независимые случайные величины с $\mathbb{E}(\varepsilon_t) = 0$ и $\mathrm{Var}(\varepsilon_t) = t$. Найдите наиболее эффективную оценку неизвестного параметра β_1 в классе линейных по y и несмещенных оценок.
16. Рассматривается модель $y_t = \beta_1 + \varepsilon_t$, где ошибки ε_t — независимые случайные величины с $\mathbb{E}(\varepsilon_t) = 0$ и $\mathrm{Var}(\varepsilon_t) = t^2$. Найдите наиболее эффективную оценку неизвестного параметра β_1 в классе линейных по y и несмещенных оценок.
17. Рассматривается модель $y_t = \beta_1 x_t + \varepsilon_t$, где ошибки ε_t — независимые случайные величины с $\mathbb{E}(\varepsilon_t) = 0$ и $\mathrm{Var}(\varepsilon_t) = t$. Найдите наиболее эффективную оценку неизвестного параметра β_1 в классе линейных по y и несмещенных оценок.
18. Рассматривается модель $y_t = \beta_1 x_t + \varepsilon_t$, где ошибки ε_t — независимые случайные величины с $\mathbb{E}(\varepsilon_t) = 0$ и $\mathrm{Var}(\varepsilon_t) = t^2$. Найдите наиболее эффективную оценку неизвестного параметра β_1 в классе линейных по y и несмещенных оценок.

9 Ошибки спецификации

1. По 25 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, для которой $RSS = 73$. При помощи вспомогательной регрессии $\hat{\hat{y}} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z + \hat{\gamma}_4 \hat{y}^2$, для которой $RSS = 70$, выполните тест Рамсея на уровне значимости 5%.
2. По 20 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, для которой $R^2 = 0.7$. При помощи вспомогательной регрессии $\hat{\hat{y}} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z + \hat{\gamma}_4 \hat{y}^2$, для которой $R^2 = 0.75$, выполните тест Рамсея на уровне значимости 5%.
3. По 30 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, для которой $RSS = 150$. При помощи вспомогательной регрессии $\hat{\hat{y}} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z + \hat{\gamma}_4 \hat{y}^2 + \hat{\gamma}_5 \hat{y}^3$, для которой $RSS = 120$, выполните тест Рамсея на уровне значимости 5%.
4. По 35 наблюдениям при помощи метода наименьших квадратов оценена модель $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, для которой $R^2 = 0.7$. При помощи вспомогательной регрессии $\hat{\hat{y}} = \hat{\gamma}_1 + \hat{\gamma}_2 x + \hat{\gamma}_3 z + \hat{\gamma}_4 \hat{y}^2 + \hat{\gamma}_5 \hat{y}^3$, для которой $R^2 = 0.8$, выполните тест Рамсея на уровне значимости 5%.
5. Используя 80 наблюдений, исследователь оценил две конкурирующие модели: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_1 = 36875$ и $\widehat{\ln y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_2 = 122$.
   Выполнив преобразование $y_i^* = y_i / \sqrt[n]{\prod y_i}$, исследователь также оценил две вспомогательные регрессии: $\hat{y}^* = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_1^* = 239$ и $\widehat{\ln y^*} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_2^* = 121$.
   Завершите тест Бокса-Кокса на уровне значимости 5%.
6. Используя 40 наблюдений, исследователь оценил две конкурирующие модели: $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_1 = 250$ и $\widehat{\ln y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_2 = 12$.
   Выполнив преобразование $y_i^* = y_i / \sqrt[n]{\prod y_i}$, исследователь также оценил две вспомогательные регрессии: $\hat{y}^* = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_1^* = 20$ и $\widehat{\ln y^*} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$, в которой $RSS_2^* = 25$.
   Завершите тест Бокса-Кокса на уровне значимости 5%.

7. Почему при реализации теста Бокса-Кокса на компьютере предпочтительнее использовать формулу $y_i^* = \exp(\ln y_i - \sum \ln y_i/n)$, а не формулу $y_i^* = y_i/\sqrt[n]{\prod y_i}$? чтобы избежать переполнения при подсчете произведения всех y_i

10 Временные ряды

1. Что такое автокорреляция?
2. На графике представлены данные по уровню озера Гурóн в футах в 1875-1972 годах:

```
ggplot(df,aes(x=obs,y=level))+geom_line()+
    labs(x="Год",ylab="Уровень озера (футы)")
```

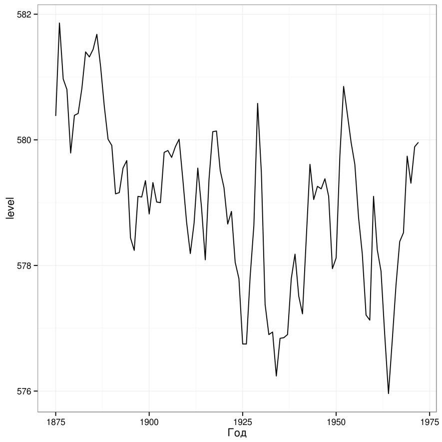
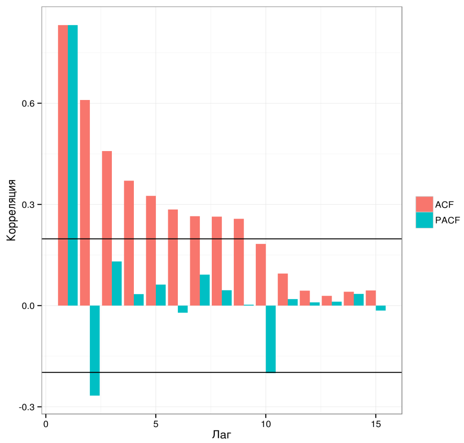


График автокорреляционной и частной автокорреляционной функций:

```
ggplot(acfs.df,aes(x=lag,y=acf,fill=acf.type))+
    geom_histogram(position="dodge",stat="identity")+
  xlab("Лаг")+ylab("Корреляция") +
  guides(fill=guide_legend(title=NULL))+
  geom_hline(yintercept=1.96/sqrt(nrow(df)))+
  geom_hline(yintercept=-1.96/sqrt(nrow(df)))
```



(a) Судя по графикам, какие модели класса ARMA или ARIMA имеет смысл оценить?

(b) По результатам оценки некоей модели ARMA с двумя параметрами, исследователь посчитал оценки автокорреляционной функции для остатков модели. Известно, что для остатков модели первые три выборочные автокорреляции равны соответственно 0.0047, -0.0129 и -0.063. С помощью подходящей статистики проверьте гипотезу о том, что первые три корреляции ошибок модели равны нулю.

3. Винни-Пух пытается выявить закономерность в количестве придумываемых им каждый день ворчалок. Винни-Пух решил разобраться, является ли оно стационарным процессом, для этого он оценил регрессию

$$\Delta\hat{y}_t = \underset{(0.5)}{4.5} - \underset{(0.1)}{0.4}y_{t-1} + \underset{(0.5)}{0.7}\Delta y_{t-1}$$

Из-за опилок в голове Винни-Пух забыл, какой тест ему нужно провести, то ли Доктора Ватсона, то ли Дикого Фуллера.

   (a) Аккуратно сформулируйте основную и альтернативную гипотезы
   (b) Проведите подходящий тест на уровне значимости 5%
   (c) Сделайте вывод о стационарности ряда
   (d) Почему Сова не советовала Винни-Пуху пользоваться широко применяемым в Лесу t-распределением?

4. Рассматривается модель $y_t = \beta_1 + \beta_2 x_{t1} + \ldots + \beta_k x_{tk} + \varepsilon_t$. Ошибки ε_t гомоскедастичны, но в них возможно присутствует автокорреляция первого порядка, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. При известном числе наблюдений T на уровне значимости 5% сделайте статистический вывод о наличии автокорреляции.

   (a) $T = 25$, $k = 2$, $DW = 0.8$
   (b) $T = 30$, $k = 3$, $DW = 1.6$
   (c) $T = 50$, $k = 4$, $DW = 1.8$
   (d) $T = 100$, $k = 5$, $DW = 1.1$

5. По 100 наблюдениям была оценена модель линейной регрессии $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$. Оказалось, что $RSS = 120$, $\hat{\varepsilon}_1 = -1$, $\hat{\varepsilon}_{100} = 2$, $\sum_{t=2}^{100}\hat{\varepsilon}_t\hat{\varepsilon}_{t-1} = -50$. Найдите DW и ρ.
6. Применяется ли статистика Дарбина-Уотсона для выявления автокорреляции в следующих моделях

   (a) $y_t = \beta_1 x_t + \varepsilon_t$
   (b) $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$
   (c) $y_t = \beta_1 + \beta_2 y_{t-1} + \varepsilon_t$
   (d) $y_t = \beta_1 + \beta_2 t + \beta_3 y_{t-1} + \varepsilon_t$
   (e) $y_t = \beta_1 t + \beta_2 x_t + \varepsilon_t$
   (f) $y_t = \beta_1 + \beta_2 t + \beta_3 x_t + \beta_4 x_{t-1} + \varepsilon_t$

7. По 21 наблюдению была оценена модель линейной регрессии $\underset{(se)}{\hat{y}} = \underset{(0.3)}{1.2} + \underset{(0.18)}{0.9} \cdot y_{t-1} + \underset{(0.01)}{0.1} \cdot t$, $R^2 = 0.6$, $DW = 1.21$. Протестируйте гипотезу об отсутствии автокорреляции ошибок на уровне значимости 5%.
8. По 24 наблюдениям была оценена модель линейной регрессии $\underset{(se)}{\hat{y}} = \underset{(0.01)}{0.5} + \underset{(0.02)}{2} \cdot t$, $R^2 = 0.9$, $DW = 1.3$. Протестируйте гипотезу об отсутствии автокорреляции ошибок на уровне значимости 5%.
9. По 32 наблюдениям была оценена модель линейной регрессии $\underset{(se)}{\hat{y}} = \underset{(2.5)}{10} + \underset{(0.5)}{2.5} \cdot t - \underset{(0.01)}{0.1} \cdot t^2$, $R^2 = 0.75$, $DW = 1.75$. Протестируйте гипотезу об отсутствии автокорреляции ошибок на уровне значимости 5%.
10. Рассмотрим модель $y_t = \beta_1 + \beta_2 x_{t1} + \ldots + \beta_k x_{tk} + \varepsilon_t$, где ε_t подчиняются автокорреляционной схеме первого порядка, т.е.

   (a) $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, $-1 < \rho < 1$

(b) $\mathrm{Var}(\varepsilon_t) = const$, $\mathbb{E}(\varepsilon_t) = const$

(c) $\mathrm{Var}(u_t) = \sigma^2$, $\mathbb{E}(u_t) = 0$

(d) Величины u_t независимы между собой

(e) Величины u_t и ε_s независимы, если $t \geqslant s$

Найдите:

(a) $\mathbb{E}(\varepsilon_t)$, $\mathrm{Var}(\varepsilon_t)$

(b) $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h})$

(c) $\mathrm{Corr}(\varepsilon_t, \varepsilon_{t+h})$

(a) $\mathbb{E}(\varepsilon_t) = 0$, $\mathrm{Var}(\varepsilon_t) = \sigma^2/(1-\rho^2)$

(b) $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t+h}) = \rho^h \cdot \sigma^2/(1-\rho^2)$

(c) $\mathrm{Corr}(\varepsilon_t, \varepsilon_{t+h}) = \rho^h$

11. Ошибки в модели $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$ являются автокоррелированными первого порядка, $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$. Шаман-эконометрист Ойуун выполняет два камлания-преобразования. Поясните смысл камланий:

    (a) Камлание А, при $t \geqslant 2$, Ойуун преобразует уравнение к виду $y_t - \rho y_{t-1} = \beta_1(1-\rho) + \beta_2(x_t - \rho x_{t-1}) + \varepsilon_t - \rho\varepsilon_{t-1}$

    (b) Камлание Б, при $t = 1$, Ойуун преобразует уравнение к виду $\sqrt{1-\rho^2}y_1 = \sqrt{1-\rho^2}\beta_1 + \sqrt{1-\rho^2}\beta_2 x_1 + \sqrt{1-\rho^2}\varepsilon_1$.

12. Пусть y_t — стационарный процесс. Верно ли, что стационарны:

    (a) $z_t = 2y_t$

    (b) $z_t = y_t + 1$

    (c) $z_t = \Delta y_t$

    (d) $z_t = 2y_t + 3y_{t-1}$

    все линейные комбинации стационарны

13. Известно, что временной ряд y_t порожден стационарным процессом, задаваемым соотношением $y_t = 1 + 0.5y_{t-1} + \varepsilon_t$. Имеется 1000 наблюдений. Вася построил регрессию y_t на константу и y_{t-1}. Петя построил регрессию на константу и y_{t+1}. Какие примерно оценки коэффициентов они получат? Они будут примерно одинаковы. Оценка наклона определяется автоковариационной функцией.

14. Рассмотрим следующий AR(1)-ARCH(1) процесс, $y_t = 1 + 0.5y_{t-1} + \varepsilon_t$, $\varepsilon_t = \nu_t \cdot \sigma_t$
    ν_t независимые $N(0;1)$ величины.
    $\sigma_t^2 = 1 + 0.8\varepsilon_{t-1}^2$
    Также известно, что $y_{100} = 2$, $y_{99} = 1.7$

    (a) Найдите $\mathbb{E}_{100}(\varepsilon_{101}^2)$, $\mathbb{E}_{100}(\varepsilon_{102}^2)$, $\mathbb{E}_{100}(\varepsilon_{103}^2)$, $\mathbb{E}(\varepsilon_t^2)$

    (b) $\mathrm{Var}(y_t)$, $\mathrm{Var}(y_t|\mathcal{F}_{t-1})$

    (c) Постройте доверительный интервал для y_{101}:

        i. проигнорировав условную гетероскедастичность

        ii. учтя условную гетерескедастичность

15. Пусть x_t, $t = 0, 1, 2, ...$ - случайный процесс и $y_t = (1+\mathrm{L})^t x_t$. Выразите x_t с помощью y_t и оператора лага L. $x_t = (1-\mathrm{L})^t y_t$

16. Пусть F_n - последовательность чисел Фибоначчи. Упростите величину

$$F_1 + C_5^1 F_2 + C_5^2 F_3 + C_5^3 F_4 + C_5^4 F_5 + C_5^5 F_6$$

$F_n = \mathrm{L}(1+\mathrm{L})F_n$, значит $F_n = \mathrm{L}^k(1+\mathrm{L})^k F_n$ или $F_{n+k} = (1+\mathrm{L})^k F_n$

17. Пусть y_t, $t = ... - 2, -1, 0, 1, 2, ...$ - случайный процесс. И $y_t = x_{-t}$. Являются ли верными рассуждения?

   (a) $\mathrm{L}y_t = \mathrm{L}x_{-t} = x_{-t-1}$

   (b) $\mathrm{L}y_t = y_{t-1} = x_{-t+1}$

   а - неверно, б - верно.

18. Представьте процесс AR(1), $y_t = 0.9y_{t-1} - 0.2y_{t-2} + \varepsilon_t$, ε ~WN(0;1) в виде модели состояние-наблюдение.

   а) Выбрав в качестве состояний вектор $\begin{pmatrix} y_t \\ y_{t-1} \end{pmatrix}$

   б) Выбрав в качестве состояний вектор $\begin{pmatrix} y_t \\ \hat{y}_{t,1} \end{pmatrix}$

   Найдите дисперсии ошибок состояний

19. Представьте процесс MA(1), $y_t = \varepsilon_t + 0.5\varepsilon_{t-1}$, ε ~WN(0;1) в виде модели состояние-наблюдение.

   a) $\begin{pmatrix} \varepsilon_t \\ \varepsilon_{t-1} \end{pmatrix}$

   b) $\begin{pmatrix} \varepsilon_t + 0.5\varepsilon_{t-1} \\ 0.5\varepsilon_t \end{pmatrix}$

20. Представьте процесс ARMA(1,1), $y_t = 0.5y_{t-1} + \varepsilon_t + \varepsilon_{t-1}$, ε ~WN(0;1) в виде модели состояние-наблюдение.

   Вектор состояний имеет вид x_t, x_{t-1}, где $x_t = \frac{1}{1-0.5L}\varepsilon_t$

21. Рекурсивные коэффициенты

   (a) Оцените модель вида $y_t = a + b_t x_t + \varepsilon_t$, где $b_t = b_{t-1}$.

   (b) Сравните графики filtered state и smoothed state.

   (c) Сравните финальное состояние b_T с коэффициентом в обычной модели линейной регрессии, $y_t = a + bx_t + \varepsilon_t$.

22. Пусть u_t — независимые нормальные случайные величины с математическим ожиданием 0 и дисперсией σ^2. Известно, что $\varepsilon_1 = u_1$, $\varepsilon_t = u_1 + u_2 + \ldots + u_t$. Рассмотрим модель $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$.

   (a) Найдите $\mathrm{Var}(\varepsilon_t)$, $\mathrm{Cov}(\varepsilon_t, \varepsilon_s)$, $\mathrm{Var}(\varepsilon)$

   (b) Являются ли ошибки ε_t гетероскедастичными?

   (c) Являются ли ошибки ε_t автокоррелированными?

   (d) Предложите более эффективную оценку вектора коэффициентов регрессии по сравнению МНК-оценкой.

   (e) Результаты предыдущего пункта подтвердите симуляциями Монте-Карло на компьютере.

23. Найдите безусловная дисперсия GARCH-процессов

   (a) $\varepsilon_t = \sigma_t \cdot z_t$, $\sigma_t^2 = 0.1 + 0.8\sigma_{t-1}^2 + 0.1\varepsilon_{t-1}^2$

   (b) $\varepsilon_t = \sigma_t \cdot z_t$, $\sigma_t^2 = 0.4 + 0.7\sigma_{t-1}^2 + 0.1\varepsilon_{t-1}^2$

   (c) $\varepsilon_t = \sigma_t \cdot z_t$, $\sigma_t^2 = 0.2 + 0.8\sigma_{t-1}^2 + 0.1\varepsilon_{t-1}^2$

   1, 2, 2

24. Являются ли верными следующие утверждения?

   (a) GARCH-процесс является процессом белого шума, условная дисперсия которого изменяется во времени

(b) Модель GARCH(1,1) предназначена для прогнозирования меры изменчивости цены финансового инструмента, а не для прогнозирования самой цены инструмента

(c) При помощи GARCH-процесса можно устранять гетероскедастичность

(d) Безусловная дисперсия GARCH-процесса изменяется во времени

(e) Модель GARCH(1,1) может быть использована для прогнозирования волатильности финансовых инструментов на несколько торговых недель вперёд

25. Рассмотрим GARCH-процесс $\varepsilon_t = \sigma_t \cdot z_t$, $\sigma_t^2 = k + g_1\sigma_{t-1}^2 + a_1\varepsilon_{t-1}^2$. Найдите

(a) $\mathbb{E}(z_t)$, $\mathbb{E}(z_t^2)$, $\mathbb{E}(\varepsilon_t)$, $\mathbb{E}(\varepsilon_t^2)$

(b) $\mathrm{Var}(z_t)$, $\mathrm{Var}(\varepsilon_t)$, $\mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1})$

(c) $\mathbb{E}(\varepsilon_t \mid \mathcal{F}_{t-1})$, $\mathbb{E}(\varepsilon_t^2 \mid \mathcal{F}_{t-1})$, $\mathbb{E}(\sigma_t^2 \mid \mathcal{F}_{t-1})$

(d) $\mathbb{E}(z_t z_{t-1})$, $\mathbb{E}(z_t^2 z_{t-1}^2)$, $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t-1})$, $\mathrm{Cov}(\varepsilon_t^2, \varepsilon_{t-1}^2)$

(e) $\lim_{h\to\infty} \mathbb{E}(\sigma_{t+h}^2 \mid \mathcal{F}_t)$

26. Используя 500 наблюдений дневных логарифмических доходностей y_t , была оценена GARCH(1,1)-модель: $\hat{y}_t = -0.000708 + \hat{\varepsilon}_t$, $\varepsilon_t = \sigma_t \cdot z_t$, $\sigma_t^2 = 0.000455 + 0.6424\sigma_{t-1}^2 + 0.2509\varepsilon_{t-1}^2$. Также известно, что $\hat{\sigma}_{499}^2 = 0.002568$, $\hat{\varepsilon}_{499}^2 = 0.000014$, $\hat{\varepsilon}_{500}^2 = 0.002178$. Найдите

(a) $\hat{\sigma}_{500}^2$, $\hat{\sigma}_{501}^2$, $\hat{\sigma}_{502}^2$

(b) Волатильность в годовом выражении в процентах, соответствующую наблюдению с номером $t = 500$

11 SVM

1. Имеются три наблюдения A, B и C:

	x	y
A	1	-2
B	2	1
C	3	0

(a) Найдите расстояние AB и косинус угла ABC

(b) Найдите расстояние AB и косинус угла ABC в расширенном пространстве с помощью гауссовского ядра с $\sigma = 1$.

(c) Найдите расстояние AB и косинус угла ABC в расширенном пространстве с помощью полиномиального ядра второй степени

2. Переход из двумерного пространства в расширяющее задан функцией

$$f : (x_1, x_2) \to (1, x_1, x_2, 3x_1x2, 2x_1^2, 4x_2^2)$$

Найдите соответствующую ядерную функцию

3. Ядерная функция имеет вид

$$K(x, y) = x_1^2y_1^2 + x_2^2y_2^2 + 2x_1x_2y_1y_2$$

Как может выглядеть функция $f : \mathbb{R}^2 \to \mathbb{R}^3$ переводящие исходные векторы в расширенное пространство? $f(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$

4. Дана плоскость. На ней точки. Симметрично ox. Найдите разделяющую гиперплоскость при разных C.

12 Деревья и Random Forest

1. Для случайных величин X и Y найдите индекс Джини и энтропию
$\begin{array}{ccc} X & 0 & 1 \\ \hline \mathbb{P}() & 0.2 & 0.8 \end{array}$, $\begin{array}{cccc} Y & 0 & 1 & 5 \\ \hline \mathbb{P}() & 0.2 & 0.3 & 0.5 \end{array}$
2. Случайная величина X принимает значение 1 с вероятностью p и значение 0 с вероятностью $1-p$.
   (a) Постройте график зависимости индекса Джини и энтропии от p
   (b) При каком p энтропия и индекс Джини будут максимальны?
3. табличка с тремя признаками...
   (a) Какой фактор нужно использовать при прогнозировании y, чтобы минимизировать энтропию?
   (b) Какой фактор нужно использовать при прогнозировании y, чтобы минимизировать индекс Джини?

13 Линейная алгебра

1. Найдите каждую из следующих матриц в каждой из следующих степеней $\frac{1}{2}$, $\frac{1}{3}$, $-\frac{1}{2}$, $-\frac{1}{3}$, -1, 100.
   (a) $\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$
   (b) $\begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix}$
2. Найдите ортогональную проекцию и ортогональную составляющую (перпендикуляр) вектора u_1 на линейное подпространство $L = \mathcal{L}(u_2)$, порождённое вектором u_2, если
   (a) $u_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}, u_2 = \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix}$
   (b) $u_1 = \begin{pmatrix} 2 & 2 & 2 & 2 \end{pmatrix}, u_2 = \begin{pmatrix} 1 & 0 & 0 & 1 \end{pmatrix}$
   (c) $u_1 = \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix}, u_2 = \begin{pmatrix} 7 & 0 & 0 & 7 \end{pmatrix}$
3. Найдите обратные матрицы ко всем матрицам, представленным ниже.
   (a) $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$
   (b) $\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$
   (c) $\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
   (d) $\begin{pmatrix} 0 & 0 & a \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
4. Найдите ранг следующих матриц в зависимости от значений параметра λ.
   (a) $\begin{pmatrix} \lambda & 1 & 1 \\ 1 & \lambda & 1 \\ 1 & 1 & \lambda \end{pmatrix}$

(b) $\begin{pmatrix} 1-\lambda & 1-2\lambda \\ 1+\lambda & 1+3\lambda \end{pmatrix}$

(c) $\begin{pmatrix} 1 & \lambda & -1 & 2 \\ 2 & -1 & \lambda & 5 \\ 1 & 10 & -6 & 1 \end{pmatrix}$

(d) $\begin{pmatrix} \lambda & 1 & -1 & -1 \\ 1 & \lambda & -1 & -1 \\ 1 & 1 & -\lambda & -1 \\ 1 & 1 & -1 & -\lambda \end{pmatrix}$

5. Пусть $i = (1, \ldots, 1)'$ — вектор из n единиц и $\pi = i(i'i)^{-1}i'$. Найдите:
   (a) $\mathrm{tr}(\pi)$ и $\mathrm{rk}(\pi)$
   (b) $\mathrm{tr}(I-\pi)$ и $\mathrm{rk}(I-\pi)$
6. Пусть X — матрица размера $n \times k$, где $n > k$, и пусть $\mathrm{rk}(X) = k$. Верно ли, что матрица $P = X(X'X)^{-1}X'$ симметрична и идемпотентна?
7. Пусть X — матрица размера $n \times k$, где $n > k$, и пусть $\mathrm{rk}(X) = k$. Верно ли, что каждый столбец матрицы $P = X(X'X)^{-1}X'$ является собственным вектором матрицы P, отвечающим собственному значению 1?
8. Пусть X — матрица размера $n \times k$, где $n > k$, пусть $\mathrm{rk}(X) = k$ и $P = X(X'X)^{-1}X'$. Верно ли, что каждый вектор-столбец u, такой что $X'u = 0$, является собственным вектором матрицы P, отвечающим собственному значению 0?
9. Верно ли, что для любых матриц A размера $m \times n$ и матриц B размера $n \times m$ выполняется равенство $\mathrm{tr}(AB) = \mathrm{tr}(BA)$?
10. Верно ли, что собственные значения симметричной и идемпотентной матрицы могут быть только нулями и единицами?
11. Пусть P — матрица размера $n \times n$, $P' = P$, $P^2 = P$. Верно ли, что $\mathrm{rk}(P) = \mathrm{tr}(P)$?
12. Верно ли, что для симметричной матрицы собственные векторы, отвечающие различным собственным значениям, ортогональны?
13. Найдите собственные значения и собственные векторы матрицы $P = X(X'X)^{-1}X'$, если
    (a) $X = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$
    (b) $X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$
    (c) $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$
    (d) $X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}$
14. Приведите пример таких A и B, что $\det(AB) \neq \det(BA)$. Например, $A = (1, 2, 3)$, $B = (1, 0, 1)'$
15. Для матриц-проекторов $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$ и $P = X(X'X)^{-1}X'$ найдите $\mathrm{tr}(\pi)$, $\mathrm{tr}(P)$, $\mathrm{tr}(I-\pi)$, $\mathrm{tr}(I-P)$. $\mathrm{tr}(I) = n$, $\mathrm{tr}(\pi) = 1$, $\mathrm{tr}(P) = k$

16. Выпишите в явном виде матрицы $X'X$, $(X'X)^{-1}$ и $X'y$, если

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ и } X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

17. Выпишите в явном виде матрицы π, πy, $\pi\varepsilon$, $I - \pi$, если $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$.
18. Формула Фробениуса. Матрицу A размера $(n+m) \times (n+m)$ разрезали на 4 части: $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$. Кусок A_{11} имеет размер $n \times n$, кусок A_{22} имеет размер $m \times m$. Известно, что A — обратима и $A^{-1} = B$. На аналогичные по размеру и расположению части разрезали матрицу $B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$.
    (a) Каковы размеры кусков A_{12} и A_{21}?
    (b) Может ли кусок A_{11} быть необратимой матрицей?
    (c) Чему равно $B_{22}(A_{22} - A_{21}A_{11}^{-1}A_{12})$?

    $n \times m$, $m \times n$, A_{11} обратима, I
19. Спектральное разложение. Симметричная матрица A размера $n \times n$ имеет n собственных чисел $\lambda_1, \ldots, \lambda_n$ с собственными векторами $u_1, \ldots, u_n$. Докажите, что A можно представить в виде $A = \sum \lambda_i u_i u_i'$.
20. Найдите определитель, собственные значения, собственные векторы и число обусловленности матрицы A. Также найдите A^{-1}, $A^{-1/2}$ и $A^{1/2}$.
    (a) $A = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.1 \end{pmatrix}$
    (b) $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$
    (c) $A = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$
    (d) $A = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$
    (e) $A = \begin{pmatrix} 3 & 2 & 1 \\ 2 & 3 & 2 \\ 1 & 2 & 3 \end{pmatrix}$
    (f) $A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$
    (g) $A = \begin{pmatrix} 4 & -1 & 1 \\ -1 & 4 & -1 \\ 1 & -1 & 4 \end{pmatrix}$
    (h) $A = \begin{pmatrix} 6 & -2 & 2 \\ -2 & 5 & 0 \\ 2 & 0 & 7 \end{pmatrix}$

14 Случайные вектора

1. Пусть $y = (y_1, y_2, y_3, y_4, y_5)'$ — случайный вектор доходностей пяти ценных бумаг. Известно, что $\mathbb{E}(y') = (5, 10, 20, 30, 40)$, $\mathrm{Var}(y_1) = 0$, $\mathrm{Var}(y_2) = 10$, $\mathrm{Var}(y_3) = 20$, $\mathrm{Var}(y_4) = 40$, $\mathrm{Var}(y_5) =$

40 и

$$\mathrm{Corr}(y) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.3 & -0.2 & 0.1 \\ 0 & 0.3 & 1 & 0.3 & -0.2 \\ 0 & -0.2 & 0.3 & 1 & 0.3 \\ 0 & 0.1 & -0.2 & 0.3 & 1 \end{pmatrix}$$

С помощью компьютера найдите ответы на вопросы:

(a) Какая ценная бумага является безрисковой?

(b) Найдите ковариационную матрицу $\mathrm{Var}(y)$

(c) Найдите ожидаемую доходность и дисперсию доходности портфеля, доли ценных бумаг в котором равны соответственно:

i. $\alpha = (0.2, 0.2, 0.2, 0.2, 0.2)'$

ii. $\alpha = (0.0, 0.1, 0.2, 0.3, 0.4)'$

iii. $\alpha = (0.0, 0.4, 0.3, 0.2, 0.1)'$

(d) Составьте из данных бумаг пять некоррелированных портфелей

2. Пусть $i = (1, \ldots, 1)'$ — вектор из n единиц, $\pi = i(i'i)^{-1}i'$ и $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)' \sim N(0, I)$.

(a) Найдите $\mathbb{E}(\varepsilon'\pi\varepsilon)$, $\mathbb{E}(\varepsilon'(I-\pi)\varepsilon)$ и $\mathbb{E}(\varepsilon\varepsilon')$

(b) Как распределены случайные величины $\varepsilon'\pi\varepsilon$ и $\varepsilon'(I-\pi)\varepsilon$?

(c) Запишите выражения $\varepsilon'\pi\varepsilon$ и $\varepsilon'(I-\pi)\varepsilon$, используя знак суммы

3. Пусть $X = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$, $P = X(X'X)^{-1}X'$, случайные величины $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ независимы и одинаково распределены $\sim N(0, 1)$.

(a) Найдите распределение случайной величины $\varepsilon' P\varepsilon$, где $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}'$

(b) Найдите $\mathbb{E}(\varepsilon' P\varepsilon)$

(c) При помощи таблиц найдите такое число q, что $\mathbb{P}(\varepsilon' P\varepsilon > q) = 0.1$

4. Пусть $X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$, $P = X(X'X)^{-1}X'$, случайные величины $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ независимы и одинаково распределены $\sim N(0, 1)$.

(a) Найдите распределение случайной величины $\varepsilon' P\varepsilon$, где $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}'$

(b) Найдите $\mathbb{E}(\varepsilon' P\varepsilon)$

(c) При помощи таблиц найдите такое число q, что $\mathbb{P}(\varepsilon' P\varepsilon > q) = 0.1$

5. Пусть $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, $P = X(X'X)^{-1}X'$, случайные величины $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ независимы и одинаково распределены $\sim N(0, 1)$.

(a) Найдите распределение случайной величины $\varepsilon' P\varepsilon$, где $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}'$.

(b) Найдите $\mathbb{E}(\varepsilon' P\varepsilon)$.

(c) При помощи таблиц найдите такое число q, что $\mathbb{P}(\varepsilon' P\varepsilon > q) = 0.1$.

6. Пусть $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\mathbb{E}(x) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $\operatorname{Var}(x) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Найдите $\mathbb{E}(y)$, $\operatorname{Var}(y)$ и $\mathbb{E}(z)$, если
   (a) $y = x - \mathbb{E}(x)$
   (b) $y = \operatorname{Var}(x)x$
   (c) $y = \operatorname{Var}(x)(x - \mathbb{E}(x))$
   (d) $y = \operatorname{Var}(x)^{-1}(x - \mathbb{E}(x))$
   (e) $y = \operatorname{Var}(x)^{-1/2}(x - \mathbb{E}(x))$
   (f) $z = (x - \mathbb{E}(x))' \operatorname{Var}(x)(x - \mathbb{E}(x))$
   (g) $z = (x - \mathbb{E}(x))' \operatorname{Var}(x)^{-1}(x - \mathbb{E}(x))$
   (h) $z = x' \operatorname{Var}(x)x$
   (i) $z = x' \operatorname{Var}(x)^{-1}x$
7. Пусть $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\mathbb{E}(x) = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$, $\operatorname{Var}(x) = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}$. Найдите $\mathbb{E}(y)$, $\operatorname{Var}(y)$ и $\mathbb{E}(z)$, если
   (a) $y = x - \mathbb{E}(x)$
   (b) $y = \operatorname{Var}(x)x$
   (c) $y = \operatorname{Var}(x)(x - \mathbb{E}(x))$
   (d) $y = \operatorname{Var}(x)^{-1}(x - \mathbb{E}(x))$
   (e) $y = \operatorname{Var}(x)^{-1/2}(x - \mathbb{E}(x))$
   (f) $z = (x - \mathbb{E}(x))' \operatorname{Var}(x)(x - \mathbb{E}(x))$
   (g) $z = (x - \mathbb{E}(x))' \operatorname{Var}(x)^{-1}(x - \mathbb{E}(x))$
   (h) $z = x' \operatorname{Var}(x)x$
   (i) $z = x' \operatorname{Var}(x)^{-1}x$
8. Известно, что случайные величины x_1, x_2 и x_3 имеют следующие характеристики:
   (a) $\mathbb{E}(x_1) = 5$, $\mathbb{E}(x_2) = 10$, $\mathbb{E}(x_3) = 8$
   (b) $\operatorname{Var}(x_1) = 6$, $\operatorname{Var}(x_2) = 14$, $\operatorname{Var}(x_3) = 1$
   (c) $\operatorname{Cov}(x_1, x_2) = 3$, $\operatorname{Cov}(x_1, x_3) = 1$, $\operatorname{Cov}(x_2, x_3) = 0$

   Пусть случайные величины y_1, y_2 и y_3, представляют собой линейные комбинации случайных величин X_1, X_2 и X_3:
   $$y_1 = x_1 + 3x_2 - 2x_3$$
   $$y_2 = 7x_1 - 4x_2 + x_3$$
   $$y_3 = -2x_1 - x_2 + 4x_3$$
   (a) Выпишите математическое ожидание и ковариационную матрицу случайного вектора $x = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}^T$
   (b) Напишите матрицу A, которая позволяет перейти от случайного вектора $x = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix}^T$ к случайному вектору $y = \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix}^T$
   (c) С помощью матрицы A найдите математическое ожидание и ковариационную матрицу случайного вектора $y = \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix}^T$
9. Пусть ξ_1, ξ_2, ξ_3 — случайные величины, такие что $\operatorname{Var}(\xi_1) = 2$, $\operatorname{Var}(\xi_2) = 3$, $\operatorname{Var}(\xi_3) = 4$, $\operatorname{Cov}(\xi_1, \xi_2) = 1$, $\operatorname{Cov}(\xi_1, \xi_3) = -1$, $\operatorname{Cov}(\xi_2, \xi_3) = 0$. Пусть $\xi = \begin{pmatrix} \xi_1 & \xi_2 & \xi_3 \end{pmatrix}^T$. Найдите $\operatorname{Var}(\xi)$ и $\operatorname{Var}(\xi_1 + \xi_2 + \xi_3)$.
   По определению ковариационной матрицы:

$\operatorname{Var}(\xi) = \begin{pmatrix} \operatorname{Var}(\xi_1) & \operatorname{Cov}(\xi_1,\xi_2) & \operatorname{Cov}(\xi_1,\xi_3) \\ \operatorname{Cov}(\xi_2,\xi_1) & \operatorname{Var}(\xi_2) & \operatorname{Cov}(\xi_2,\xi_3) \\ \operatorname{Cov}(\xi_3,\xi_1) & \operatorname{Cov}(\xi_3,\xi_2) & \operatorname{Var}(\xi_3) \end{pmatrix} = \begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 4 \end{pmatrix}$

$\operatorname{Var}(\xi_1+\xi_2+\xi_3) = \operatorname{Var}\left(\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix}\right) = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \operatorname{Var}\begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -1 \\ 1 & 3 & 0 \\ -1 & 0 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 9$

10. Пусть $h = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$; $\mathbb{E}(h) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $\operatorname{Var}(h) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$; $z_1 = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$. Найдите $\mathbb{E}(z_1)$ и $\operatorname{Var}(z_1)$.

$\mathbb{E}(z_1) = \mathbb{E}\left(\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \mathbb{E}\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$

$\operatorname{Var}(z_1) = \operatorname{Var}\left(\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \operatorname{Var}\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}^T = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}^T =$
$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$

11. Пусть $h = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$; $\mathbb{E}(h) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $\operatorname{Var}(h) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$; $z_2 = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. Найдите $\mathbb{E}(z_2)$ и $\operatorname{Var}(z_2)$

$\mathbb{E}(z_2) = \mathbb{E}\left(\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \mathbb{E}\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$

Поскольку $z_2 = z_1 + \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, где z_1 — случайный вектор из предыдущей задачи, то $\operatorname{Var}(z_2) = \operatorname{Var}(z_1)$. Сдвиг случайного вектора на вектор-константу не меняет его ковариационную матрицу.

12. Пусть $h = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$; $\mathbb{E}(h) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $\operatorname{Var}(h) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$; $z_3 = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} - \begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix}$. Найдите $\mathbb{E}(z_3)$ и $\operatorname{Var}(z_3)$

В данном примере проиллюстрирована процедура центрирования случайного вектора — процедура вычитания из случайного вектора его математического ожидания.

$\mathbb{E}(z_3) = \mathbb{E}\left(\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} - \begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix}\right) = \mathbb{E}\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} - \mathbb{E}\begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix} = \begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix} - \begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

Заметим, что вектор z_3 отличается от вектора z_1 (из задачи 15) сдвигом на вектор-константу $\begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix}$, поэтому $\operatorname{Var}(z_3) = \operatorname{Var}(z_1)$.

13. Пусть r_1, r_2 и r_3 — годовые доходности трёх рисковых финансовых инструментов. Пусть α_1, α_2 и α_3 — доли, с которыми данные инструменты входят в портфель инвестора. Считаем, что $\sum_{i=1}^{3} \alpha_i = 1$ и $\alpha_i \geqslant 0$ для всех $i = 1, 2, 3$. Пусть $r = \begin{pmatrix} r_1 & r_2 & r_3 \end{pmatrix}^T$, $\mathbb{E}(r) = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix}^T$, $\operatorname{Var}(r) = \begin{pmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{pmatrix}$. Параметры $\{a_i\}$ и $\{c_i\}$ известны.

   (a) Найдите годовую доходность портфеля П инвестора

   (b) Докажите, что дисперсия доходности портфеля П равна $\sum_{i=1}^{3} \sum_{j=1}^{3} \alpha_i c_{ij} \alpha_j$

   (c) Для случая $\alpha_1 = 0.1$, $\alpha_2 = 0.5$, $\alpha 3 = 0.4$, $\mathbb{E}(r) = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix}^T = \begin{pmatrix} 0.10 & 0.06 & 0.05 \end{pmatrix}^T$, $\operatorname{Var}(r) = \begin{pmatrix} 0.04 & 0 & -0.005 \\ 0 & 0.01 & 0 \\ -0.005 & 0 & 0.0025 \end{pmatrix}$ найдите $\mathbb{E}(\Pi)$ и $\operatorname{Var}(\Pi)$

14. Пусть $h = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$; $\mathbb{E}(h) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $\operatorname{Var}(h) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$; $z_3 = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} - \begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix}$; $z_4 = \operatorname{Var}(h)^{-1/2} z_3$. Найдите $\mathbb{E}(z_4)$ и $\operatorname{Var}(z_4)$

15. Пусть $h = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix}$; $\mathbb{E}(h) = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$; $\operatorname{Var}(h) = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$; $z_3 = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} - \begin{pmatrix} \mathbb{E}\xi_1 \\ \mathbb{E}\xi_2 \end{pmatrix}$; $z_4 = \operatorname{Var}(h)^{-1/2} z_3$. Найдите $\mathbb{E}(z_4)$ и $\operatorname{Var}(z_4)$

16. Случайные величины w_1 и w_2 независимы с нулевым ожиданием и единичной дисперсией. Из них составлено два вектора, $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ и $z = \begin{pmatrix} -w_2 \\ w_1 \end{pmatrix}$

   (a) Являются ли векторы w и z перпендикулярными?

   (b) Найдите $\mathbb{E}(w)$, $\mathbb{E}(z)$

   (c) Найдите $\operatorname{Var}(w)$, $\operatorname{Var}(z)$, $\operatorname{Cov}(w, z)$

17. Есть случайный вектор $w = (w_1, w_2, \ldots, w_n)'$.

   (a) Возможно ли, что $E(w) = 0$ и $\sum w_i = 0$?

(b) Возможно ли, что $E(w) \neq 0$ и $\sum w_i = 0$?

(c) Возможно ли, что $E(w) = 0$ и $\sum w_i \neq 0$?

(d) Возможно ли, что $E(w) =\neq$ и $\sum w_i \neq 0$?

Каждый из вариантов возможен

18. Известна ковариационная матрица вектора $\varepsilon = (\varepsilon_1, \varepsilon_2)$,

$$\mathrm{Var}(\varepsilon) = \begin{pmatrix} 9 & -1 \\ -1 & 9 \end{pmatrix}$$

Найдите четыре различных матрицы A, таких что вектор $v = A\varepsilon$ имеет некоррелированные компоненты с единичной дисперсией, то есть $\mathrm{Var}(A\varepsilon) = I$.

15 Многомерное нормальное и квадратичные формы

1. Пусть $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)' \sim N(0, I)$ и матрица A представлена ниже. Найдите $\mathbb{E}(\varepsilon' A \varepsilon)$ и распределение случайной величины $\varepsilon' A \varepsilon$.

(a) $\begin{pmatrix} 2/3 & -1/3 & 1/3 \\ -1/3 & 2/3 & 1/3 \\ 1/3 & 1/3 & 2/3 \end{pmatrix}$

(b) $\begin{pmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{pmatrix}$

(c) $\begin{pmatrix} 1/3 & 1/3 & -1/3 \\ 1/3 & 1/3 & -1/3 \\ -1/3 & -1/3 & 1/3 \end{pmatrix}$

(d) $\begin{pmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix}$

(e) $\begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix}$

(f) $\begin{pmatrix} 1/2 & 0 & -1/2 \\ 0 & 1 & 0 \\ -1/2 & 0 & 1/2 \end{pmatrix}$

(g) $\begin{pmatrix} 1/2 & -1/2 & 0 \\ -1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

(h) $\begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

(i) $\begin{pmatrix} 0.8 & 0.4 & 0 \\ 0.4 & 0.2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

(j) $\begin{pmatrix} 0.2 & -0.4 & 0 \\ -0.4 & 0.8 & 0 \\ 0 & 0 & 0 \end{pmatrix}$

2. Пусть $i = (1, \ldots, 1)'$ — вектор из n единиц, $\pi = i(i'i)^{-1}i'$ и $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)' \sim N(0, I)$.
   (a) Найдите $\mathbb{E}(\varepsilon'\pi\varepsilon)$, $\mathbb{E}(\varepsilon'(I-\pi)\varepsilon)$ и $\mathbb{E}(\varepsilon\varepsilon')$
   (b) Как распределены случайные величины $\varepsilon'\pi\varepsilon$ и $\varepsilon'(I-\pi)\varepsilon$?
   (c) Запишите выражения $\varepsilon'\pi\varepsilon$ и $\varepsilon'(I-\pi)\varepsilon$, используя знак суммы

3. Пусть $X = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}$, $P = X(X'X)^{-1}X'$, случайные величины $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ независимы и одинаково распределены $\sim N(0, 1)$.
   (a) Найдите распределение случайной величины $\varepsilon'P\varepsilon$, где $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}'$
   (b) Найдите $\mathbb{E}(\varepsilon'P\varepsilon)$
   (c) При помощи таблиц найдите такое число q, что $\mathbb{P}(\varepsilon'P\varepsilon > q) = 0.1$

4. Пусть $X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{pmatrix}$, $P = X(X'X)^{-1}X'$, случайные величины $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ независимы и одинаково распределены $\sim N(0, 1)$.
   (a) Найдите распределение случайной величины $\varepsilon'P\varepsilon$, где $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}'$
   (b) Найдите $\mathbb{E}(\varepsilon'P\varepsilon)$
   (c) При помощи таблиц найдите такое число q, что $\mathbb{P}(\varepsilon'P\varepsilon > q) = 0.1$

5. Пусть $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$, $P = X(X'X)^{-1}X'$, случайные величины $\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ независимы и одинаково распределены $\sim N(0, 1)$.
   (a) Найдите распределение случайной величины $\varepsilon'P\varepsilon$, где $\varepsilon = \begin{pmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & \varepsilon_4 \end{pmatrix}'$.
   (b) Найдите $\mathbb{E}(\varepsilon'P\varepsilon)$.
   (c) При помощи таблиц найдите такое число q, что $\mathbb{P}(\varepsilon'P\varepsilon > q) = 0.1$.

6. Пусть $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)' \sim N(0, I)$. Найдите $\mathbb{E}(\varepsilon'P\varepsilon)$ и распределение случайной величины $\varepsilon'P\varepsilon$, если $P = X(X'X)^{-1}X'$ и матрица X' представлена ниже.
   (a) $\begin{pmatrix} 1 & 1 & 1 \end{pmatrix}$
   (b) $\begin{pmatrix} 1 & 2 & 3 \end{pmatrix}$
   (c) $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$
   (d) $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{pmatrix}$
   (e) $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$

7. Пусть $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} \sim N(0, \sigma^2 I)$, $i = (1, \ldots, 1)'$ — вектор из n единиц, $\pi = i(i'i)^{-1}i'$, X — матрица размера $n \times k$, $P = X(X'X)^{-1}X'$. Найдите:

(a) $\mathbb{E}(\varepsilon'(P-\pi)\varepsilon)$

(b) $\mathbb{E}(\varepsilon'(I-\pi)\varepsilon)$

(c) $\mathbb{E}(\varepsilon' P\varepsilon)$

(d) $\mathbb{E}(\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2)$

8. Пусть $\varepsilon=(\varepsilon_1,\varepsilon_2,\varepsilon_3)'\sim N(0,4I)$, $A=\begin{pmatrix}4&1&1\\1&3&1\\1&1&2\end{pmatrix}$. Найдите:

   (a) $\mathbb{E}(\varepsilon' A\varepsilon)$

   (b) $\mathbb{E}(\varepsilon'(I-A)\varepsilon)$

9. Пусть $x=\begin{bmatrix}x_1 & x_2\end{bmatrix}^T$ — случайный вектор, имеющий двумерное нормальное распределение с математическим ожиданием $\mu=\begin{bmatrix}1 & 2\end{bmatrix}^T$ и ковариационной матрицей $\Sigma=\begin{bmatrix}2&1\\1&2\end{bmatrix}$.

   (a) Найдите Σ^{-1}

   (b) Найдите $\Sigma^{-1/2}$

   (c) Найдите математическое ожидание и ковариационную матрицу случайного вектора $y=\Sigma^{-1/2}\cdot(x-\mu)$

   (d) Какое распределение имеет вектор y из предыдущего пункта?

   (e) Найдите распределение случайной величины $q=(x-\mu)^T\cdot\Sigma^{-1}\cdot(x-\mu)$

10. Пусть $z=\begin{bmatrix}z_1 & z_2 & z_3\end{bmatrix}^T\sim N(0,I_{3x3})$, $b=\begin{bmatrix}1 & 2 & 3\end{bmatrix}^T$,
$A=\begin{bmatrix}1&1&1\\0&1&1\\0&0&1\end{bmatrix}$, $K=\begin{bmatrix}1&0&0\\0&1/2&1/2\\0&1/2&1/2\end{bmatrix}$.

   (a) Найдите $\mathbb{E}x$ и $\mathrm{Var}(x)$ случайного вектора $x=A\cdot z+b$

   (b) Найдите распределение случайного вектора x

   (c) Найдите $\mathbb{E}q$ случайной величины $q=z^T\cdot K\cdot z$

   (d) Найдите распределение случайной величины q

11. Известно, что $\varepsilon\sim N(0,I)$, $\varepsilon=(\varepsilon_1,\varepsilon_2,\varepsilon_3)'$. Матрица $A=\begin{pmatrix}2/3&-1/3&-1/3\\-1/3&2/3&-1/3\\-1/3&-1/3&2/3\end{pmatrix}$.

   (a) Найдите $\mathbb{E}(\varepsilon' A\varepsilon)$

   (b) Как распределена случайная величина $\varepsilon' A\varepsilon$?

по χ^2-распределению

12. Известно, что $\varepsilon\sim N(0,A)$, $\varepsilon=(\varepsilon_1,\varepsilon_2)'$. Матрица $A=\begin{pmatrix}4&1\\1&4\end{pmatrix}$, матрица $B=\begin{pmatrix}-1&3\\2&1\end{pmatrix}$

   (a) Как распределен вектор $h=B\varepsilon$?

   (b) Найдите $A^{-1/2}$

   (c) Как распределен вектор $u=A^{-1/2}\varepsilon$?

$u\sim N(0,I)$

16 Задачи по программированию

Все наборы данных доступны по ссылке https://github.com/bdemeshev/em301/wiki/Datasets.

1. Начиная с какого знака в числе $\pi = 3.1415\ldots$ можно обнаружить твой номер телефона? Первый 10 миллионов знаков числа π можно найти на сайте http://code.google.com/p/pc2012-grupo-18-turma-b/downloads/list. Если не хватает, то миллиард знаков, файл размера примерно в 1 гигабайт, доступен по ссылке http://stuff.mit.edu/afs/sipb/contrib/pi/. Настоящие челябинцы рассчитывают π самостоятельно. Краткая история о том, как маньяки считали π до 10 миллиардов знаков и потеряли полгода из-за сбоев компьютерного железа, http://www.numberworld.org/misc_runs/pi-10t/details.html.
2. Отряд Иосифа Флавия из 40 воинов, защищающий город Йодфат, блокирован в пещере превосходящими силам римлян. Чтобы не сдаться врагу, воины стали по кругу и договорились, что сами будут убивать каждого третьего, пока не погибнут все. При этом двое воинов, оставшихся последними в живых, должны были убить друг друга. Хитренький Иосиф Флавий, командующий этим отрядом, хочет определить, где нужно встать ему и его товарищу, чтобы остаться последними. Не для того, чтобы убить друг друга, а чтобы сдать крепость римлянам. Напишите программу, которая для n воинов вставших в круг определяет, какие двое останутся последними, если будут убивать каждого k-го.
3. Напишите программу, которая печатает сама себя.
4. Задача Макар-Лиманова. У торговца 55 пустых стаканчиков, разложенных в несколько стопок. Пока нет покупателей он развлекается: берет верхний стаканчик из каждой стопки и формирует из них новую стопку. Потом снова берет верхний стаканчик из каждой стопки и формирует из них новую стопку и т.д.
   (a) Напишите функцию ‘makar_step‘. На вход функции подаётся вектор количества стаканчиков в каждой стопке до перекладывания. На выходе функция возвращает количества стаканчиков в каждой стопке после одного перекладывания.
   (b) Изначально стаканчики были разложены в две стопки, из 25 и 30 стаканчиков. Как разложатся стаканчики если покупателей не будет достаточно долго?
5. Напишите программу, которая находит сумму элементов побочной диагонали квадратной матрицы.

17 Устав проверки гипотез

1. Условия применимости теста
2. Формулировка H_0, H_a и уровня значимости α
3. Формула расчета и наблюдаемое значения статистики, S_{obs}
4. Закон распределения S_{obs} при верной H_0
5. Область в которой H_0 не отвергается
6. Точное P-значение
7. Статистический вывод о том, отвергается ли H_0 или нет.

В качестве статистического вывода допускается только одна из двух фраз:

- Гипотеза H_0 отвергается
- Гипотеза H_0 не отвергается

Остальные фразы считаются неуставными