Задачник по эконометрике-1

(с шахматами и поэтэссами)

Дмитрий Борзых, Борис Демешев

13 июля 2013 г.

Содержание

Todo list

1 Проверка гипотез строго по уставу!

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 Р-значение
7. Вывод

В качестве вывода допускается только одна из двух фраз:

- Гипотеза H_0 отвергается
- Гипотеза H_0 не отвергается

Остальные фразы считаются неуставными

2 Неклассифицировано

1. 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 $\hat{y}$.

2. 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.

3. 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 шара, обладающую наименьшей дисперсией.

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

   Как отсутствие систематической ошибки.

4. 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.

5. Сгенерировать набор данных, обладающий следующим свойством. Если попытаться сразу выкинуть регрессоры x и z, то гипотеза о их совместной незначимости отвергается. Если вместо этого попытаться выкинуть отдельно x, или отдельно z, то гипотеза о незначимости не отвергается. Сгенерировать сильно коррелированные x и z

6. Сгенерировать набор данных, обладающий следующим свойством. Если попытаться сразу выкинуть регрессоры x и z, то гипотеза о их совместной незначимости отвергается. Если вместо сначала выкинуть отдельно x, то гипотеза о незначимости не отвергается. Если затем выкинуть z, то гипотезы о незначимости тоже не отвергается. ??

7. 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

8. Как построить доверительный интервал для вершины параболы? ... bootstrap, дельта-метод

9. Про R^2_{adj}

   (a) Может ли в модели с константой R^2_{adj} быть отрицательным?

   (b) Что больше, R^2 или R^2_{adj} в модели с константой?

   (c) Вася оценил модель A, а затем выкинул из нее регрессор z и оценил получившуюся модель B. В моделях A и B оказались равные R^2_{adj}. Чему равна t-статистика коэффициента при z в модели A?

   (d) Есть две модели с одной и той же зависимой переменной, но с разными объясняющими переменными, модель A и модель B. В модели A коэффициент R^2_{adj} больше, чем в модели B. В какой из моделей больше коэффициент $\hat{\sigma^2}$?

   да, R^2, $t = 1$, B

10. 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.

11. 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
12. 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.
13. 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
14. 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, $\hat{\beta}_1 + \hat{\gamma}_1 = 1$, $\hat{\beta}_2 + \hat{\gamma}_2 = 0$

15. 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

16. 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
17. Эконометрист Вовочка оценил регрессию y только на константу. Какой коэффициент R^2 он получит? $R^2 = 0$
18. 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. $TSS_1 = TSS_2$, $R_2^2 \geqslant R_2^1$, $ESS_2 \geqslant ESS_1$, $RSS_2 \leqslant RSS_1$
19. Случайные величины w_1 и w_2 независимы и нормально распределены, $N(0,1)$. 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 $\mathrm{Var}(w)$, $\mathrm{Var}(z)$, $\mathrm{Cov}(w, z)$
    (d) Рассмотрим классическую линейную модель. Являются ли векторы $\hat{\varepsilon}$ и $\hat{y}$ перпендикулярными? Найдите $\mathrm{Cov}(\hat{\varepsilon}, \hat{y})$.
20. 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 $\sum w_i \neq 0$?
    (e) Чему в классической модели регрессии равны: $\mathbb{E}(\varepsilon)$ и $\sum \varepsilon_i$?
    (f) Чему в классической модели регрессии равны: $\mathbb{E}(\hat{\varepsilon})$ и $\sum \hat{\varepsilon}_i$?

    Каждый из вариантов возможен
21. 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?

3 МНК без матриц и вероятностей

1. 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. $\hat{\beta} = \sum x_i y_i / \sum x_i^2$
2. Даны n чисел: y_1, ..., y_n. Мы прогнозируем y_i по формуле $\hat{y}_i = \hat{\beta}$. Найдите $\hat{\beta}$ методом наименьших квадратов. $\hat{\beta} = \bar{y}$
3. 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. $\hat{\beta}_2 = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (x_i - \bar{x})^2$, $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}$
4. 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. $\hat{\beta} = \sum x_i (y_i - 1) / \sum x_i^2$
5. 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. $(300 - \hat{\beta}_1)^2 + (200 - \hat{\beta}_2)^2 + (400 - \hat{\beta}_1 - \hat{\beta}_2)^2 \to \min$

6. 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. $2\cdot(10-\hat{\beta})^2+(3-\hat{\beta})^2 \to \min$
7. Регрессия на дамми-переменную...
8. 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

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

9. 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.
10. 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.
11. 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) $y_i = x_i^{\beta_2} e^{\beta_1+\varepsilon_i}$

4 Теорема Гаусса-Маркова и нормальность

1. Напишите формулу для оценок коэффициентов в парной регрессии без матриц
2. Напишите формулу для оценок коэффициентов в множественной регрессии с матрицами
3. (аналогично) для дисперсий
4. Сформулируйте теорему Гаусса-Маркова
5. 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?
6. 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")
coef.t <- coef.t[,-4]
coef.t[1,1] <- NA
coef.t[2,2] <- NA
coef.t[3,3] <- NA
```

```
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:

```
h <- swiss
```

7. 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: $Fertility_i = \gamma_1 + \gamma_2(Education_i + Examination_i) + \gamma_3 Catholic_i + u_i$

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

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

Набор данных доступен в пакете R:

```
h <- swiss
```

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

```
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)
```

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

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

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

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 $p < 0.05$

	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

(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.

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

```
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)
```

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

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

(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.

10. Рассмотрим модель с линейным трендом без свободного члена, $y_t = \beta t + \varepsilon_t$.

(a) Найдите МНК оценку коэффициента β

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

	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

(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

11. 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?

12. 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?

13. 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.

```
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)
```

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

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

14. 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 $\mathbb{P}(\hat{\beta}_3 > \sigma_{\hat{\beta}_3})$

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

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

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

5 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.

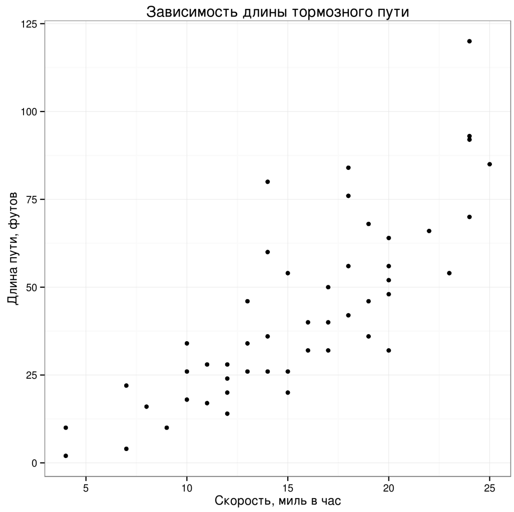
   (b) А теперь сгенерируйте данные так, чтобы корреляция между регрессорами была по-прежнему больше 0.9, но проблемы мультиколлинеарности бы не было, т.е. все коэффициенты были бы значимы.

   (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	-1.49	11.91	-0.12	0.90
speed	2.12	1.19	1.78	0.08
speed^3	0.00	0.00	1.63	0.11

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)
```

	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

(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 $speed = 10$

(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

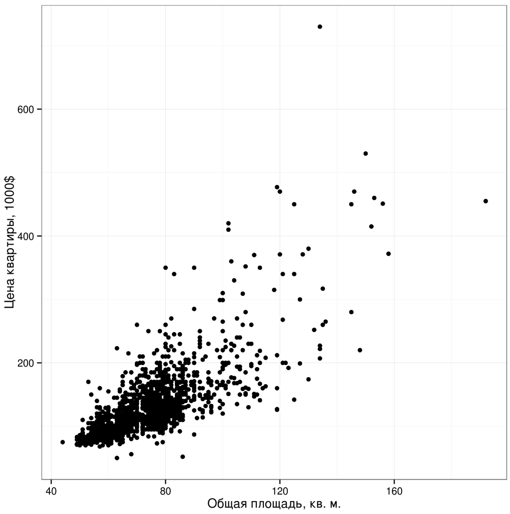
   (c) Как связаны между собой $\hat{\beta}_{rr}$ и $\hat{\beta}_{ols}$?

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

6 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 $\mathrm{Var}(\varepsilon_i) = \sigma^2 \cdot x_i^2$.

   (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, $\mathrm{Var}(\varepsilon_i)$? $\mathrm{Var}(\varepsilon_i) = cx_i^4$
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, $\mathrm{Var}(\varepsilon_i)$? $\mathrm{Var}(\varepsilon_i) = cx_i$
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 $GQ_{obs} \notin [0; 11.07]$, то на основании имеющихся наблюдений на уровне значимости 5% основная гипотеза H_0 отвергается. Таким образом, тест Уайта выявил гетероскедастичность.

7 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)")
```

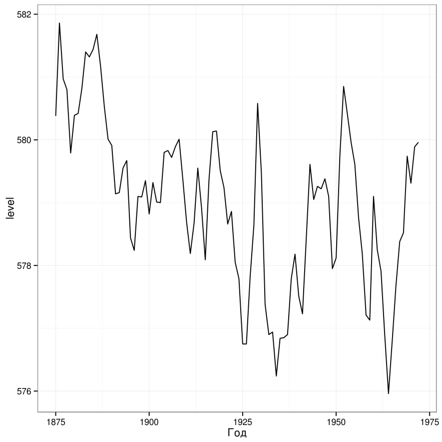
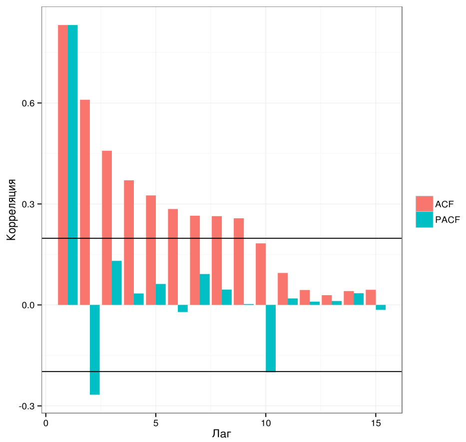


График автокорреляционной и частной автокорреляционной функций:

```
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 применяемой в Лесу t-статистикой?

8 Функциональная форма

1. Сгенерируйте данные так, чтобы при оценке модели $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x + \hat{\beta}_3 z$ оказывалось, что $\hat{\beta}_2 > 0$, а при оценке модели $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x$ оказывалось, что $\hat{\beta}_2 < 0$.

9 Инструментальные переменные

Экзогенность, $\mathbb{E}(\varepsilon \mid x) = 0$
Предопределённость, $\mathbb{E}(\varepsilon_t \mid x_t) = 0$ для всех t

1. Табличка 2 на 2. Найдите $\mathbb{E}(\varepsilon)$, $\mathbb{E}(\varepsilon|x)$, $\mathrm{Cov}(\varepsilon, x)$.
2. Приведите примеры дискретных случайных величин ε и x, таких, что

   (a) $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon \mid x) = 0$, но величины зависимы. Чему в этом случае равно $\mathrm{Cov}(\varepsilon, x)$?

   (b) $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Cov}(\varepsilon, x) = 0$, но $\mathbb{E}(\varepsilon \mid x) \neq 0$. Зависимы ли эти случайные величины?

3. Все предпосылки классической линейной модели выполнены, $y = \beta_1 + \beta_2 x + \varepsilon$. Рассмотрим альтернативную оценку коэффициента β_2,

$$\hat{\beta}_{2,IV} = \frac{\sum z_i(y_i - \bar{y})}{\sum z_i(x_i - \bar{x})} \tag{2}$$

   (a) Является ли оценка несмещенной?

   (b) Любые ли z_i можно брать?

   (c) Найдите $\mathrm{Var}(\hat{\beta}_{2,IV})$

   Да, является. Любые, кроме констант. $\mathrm{Var}(\hat{\beta}_{2,IV}) = \sigma^2 \sum (z_i - \bar{z})^2 / \left(\sum (z_i - \bar{z}) x_i\right)^2$.

4.

10 Проекция, Картинка

1. Найдите на Картинке все перпендикулярные векторы. Найдите на Картинке все прямоугольные треугольники. Сформулируйте для них теоремы Пифагора. $\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{\varepsilon}_i^2$, $TSS = ESS + RSS$,
2. Покажите на Картинке TSS, ESS, RSS, R^2, $\mathrm{sCov}(\hat{y}, y)$

3. 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.
4. 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.
5. 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.

11 Деревья и Random Forest

1. Для случайных величин X и Y найдите индекс Джини и энтропию

X	0	1
$\mathbb{P}()$	0.2	0.8

Y	0	1	5
$\mathbb{P}()$	0.2	0.3	0.5

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?

12 МЕГАМАТРИЦА (операции со случайными векторами)

1. 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 $\mathrm{Cov}(\varepsilon, y)$, $\mathrm{Cov}(\varepsilon, \hat{y})$, ... $\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$
2. Найдите $\mathbb{E}(\sum(\varepsilon_i - \bar{\varepsilon})^2)$, $\mathbb{E}(RSS)$ $(n-1)\sigma^2$, $(n-k)\sigma^2$
3. 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 $TSS = y'(I-\pi)y$, $RSS = y'(I-P)y$, $ESS = y'(P-\pi)y$
4. $\mathbb{E}(TSS)$, $\mathbb{E}(ESS)$ — громоздкие $\mathbb{E}(TSS) = (n-1)\sigma^2 + \beta'X'(I-\pi)X\beta$
5. 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? $(n-1)\sigma^2$, $(n-k)\sigma^2$, $(k-1)\sigma^2$
6. 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

7. 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 $u = A^{-1/2}\varepsilon$?

   $u \sim N(0, I)$

8. 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 $\mathrm{Var}(A\varepsilon) = I$.

13 Метод максимального правдоподобия

1. Выпишите в явном виде функцию максимального правдоподобия для модели $y = \beta_1 + \beta_2 x + \varepsilon$, если $\varepsilon \sim N(0, A)$. Матрица A устроена по принципу: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$, и $\mathrm{Var}(\varepsilon_i) = \sigma^2 x_i^2$.
2. Выпишите в явном виде функцию максимального правдоподобия для модели $y = \beta_1 + \beta_2 x + \varepsilon$, если $\varepsilon \sim N(0, A)$. Матрица A устроена по принципу: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ при $i \neq j$, и $\mathrm{Var}(\varepsilon_i) = \sigma^2 |x_i|$.
3. 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.

4. 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

5. 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 $N(0,1)$

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

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

7. [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

8. 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

9. 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

10. Инопланетянин Капп 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

11. Предположим, что в классической линейной модели ошибки имеют нормальное распределение, т.е.

$$y_i = \beta_1 + \beta_2 x_{2,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

   где ε_i нормальны $N(0, \sigma^2)$ и независимы

   (a) Найдите оценки для β и σ^2 методом максимального правдоподобия.

   (b) Являются ли полученные оценки $\hat{\beta}_{ML}$ и $\hat{s}^2_{ML}$ несмещенными?

   (c) Выведите формулу LR-статистики у теста отношения правдоподобия для тестирования гипотезы об адекватности регрессии H_0: $\beta_2 = \beta_3 = \ldots = \beta_k = 0$.

12. Наблюдения $X_1, \ldots, X_n$ независимы и нормальны $N(\mu, 1)$. По 100 наблюдениям оказалось, что $\sum x_i = 200$, $\sum x_i^2 = 900$.

   (a) Оцените μ методом максимального правдоподобия

   (b) Постройте 95% доверительный интервал для μ

   (c) Проверьте гипотезу о том, что $\mu = 3$ против альтернативной $\mu \neq 3$ с помощью тестов Вальда, множителей Лагранжа и отношения правдоподобия

   (d) Постройте 95% доверительный интервал для неизвестной величины $\mathbb{P}(X_i > 2.5)$

13. Наблюдения $X_1, \ldots, X_n$ независимы и нормальны $N(0, \sigma^2)$. По 100 наблюдениям оказалось, что $\sum x_i = 200$, $\sum x_i^2 = 900$.

   (a) Оцените σ^2 методом максимального правдоподобия

   (b) Постройте 95% доверительный интервал для σ^2

   (c) Проверьте гипотезу о том, что $\sigma^2 = 4$ против альтернативной $\sigma^2 \neq 4$ с помощью тестов Вальда, множителей Лагранжа и отношения правдоподобия

   (d) Постройте 95% доверительный интервал для неизвестной величины $\mathbb{P}(X_i > 2.5)$

14. Наблюдения $X_1, \ldots, X_n$ независимы и нормальны $N(\mu, \sigma^2)$. По 100 наблюдениям оказалось, что $\sum x_i = 200$, $\sum x_i^2 = 900$.

   (a) Оцените μ и σ^2 методом максимального правдоподобия

   (b) Постройте 95% доверительный интервал для μ, σ^2

   (c) [R] Проверьте гипотезу о том, что $\sigma^2 = 4$ против альтернативной $\sigma^2 \neq 4$ с помощью тестов Вальда, множителей Лагранжа и отношения правдоподобия

   (d) [R] Проверьте гипотезу о том, что $\mu = 3$ против альтернативной $\mu \neq 3$ с помощью тестов Вальда, множителей Лагранжа и отношения правдоподобия

(e) [R] Постройте 95% доверительный интервал для неизвестной величины $\mathbb{P}(X_i > 2.5)$

(f) [R] На графике постройте двумерную 95% доверительную область для вектора (μ, σ^2)

15. [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

14 Голая линейная алгебра

Здесь будет собран минимум задач по линейной алгебре.

1. Приведите пример таких A и B, что $\det(AB) \neq \det(BA)$. Например, $A = (1, 2, 3)$, $B = (1, 0, 1)'$
2. Для матриц-проекторов $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$ и $P = X(X'X)^{-1}X'$ найдите $\mathrm{tr}(\pi)$, $\mathrm{tr}(P)$, $\mathrm{tr}(I-\pi)$, $\mathrm{tr}(I-P)$. $\mathrm{tr}(I) = n$, $\mathrm{tr}(\pi) = 1$, $\mathrm{tr}(P) = k$
3. Выпишите в явном виде матрицы $X'X$, $(X'X)^{-1}$ и $X'y$, если
   $y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ и $X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$
4. Выпишите в явном виде матрицы π, πy, $\pi\varepsilon$, $I - \pi$, если $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$.

15 Парадигма случайных величин

1. Найдите $E(Y|X)$
2. Про многомерное нормальное распределение
3. Известна совместная функция плотности пары величин X_i, Y_i

$$f(x, y) =$$

   Найдите

   (a) $\mathbb{E}(X_i)$, $\mathbb{E}(Y_i)$, $\mathrm{Var}(X_i)$, $\mathrm{Var}(Y_i)$, $\mathrm{Cov}(X_i, Y_i)$

   (b) $\mathbb{E}(Y_i \mid X_i)$, $\mathbb{E}(X_i \mid Y_i)$

(c) Вася оценивает модель $y_i = \beta_1 + \beta_2 x_i + \epsilon_i$ по огромному количеству наблюдений, $n >> 0$. Чему примерно у него окажутся равны $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{s}^2$, $\widehat{\mathrm{Var}}(\hat{\beta}_2)$? Чему равно $\mathbb{E}(\hat{\beta}_2)$? (или оно не будет браться???)

(d) Петя оцениваем модель $y_i = \beta_1 + \beta_2 x_i + \beta_2 x_i^2 + \epsilon_i$. Найдите $\mathbb{E}(\hat{\beta}_1)$, $\mathbb{E}(\hat{\beta}_2)$, $\mathbb{E}(\hat{\beta}_3)$, $\mathrm{Var}(\hat{\beta})$ (?)

16 Компьютерные упражнения

Переделать классификацию. Каждый раздел делится на компьютерные и ручные задачи

Все наборы данных доступны по ссылке https://github.com/bdemeshev/em301/wiki/Datasets.

1. Скачайте результаты двух контрольных работ по теории вероятностей, с описанием данных, . Наша задача попытаться предсказать результат второй контрольной работы зная позадачный результат первой контрольной, пол и группу студента.
   (a) Какая задача из первой контрольной работы наиболее существенно влияет на результат второй контрольной?
   (b) Влияет ли пол на результат второй контрольной?
   (c) Влияет ли редкость имени на результат второй контрольной?
   (d) Что можно сказать про влияние группы, в которой учится студент?
2. 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?
3. Напишите функцию, которая бы оценивала регрессию методом наименьших квадратов. На вход функции должны подаваться вектор зависимых переменных y и матрица регрессоров X. На выходе функция должна выдавать список из $\hat{\beta}$, $\widehat{\mathrm{Var}}(\hat{\beta})$, $\hat{y}$, $\hat{\varepsilon}$, ESS, RSS и TSS. По возможности функция должна проверять корректность аргументов, например, что в y и X одинаковое число наблюдений и т.д.
4. 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.
5. (?) 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.

6. (?) У меня есть набор данных с выборочным средним $\bar{y}$ и выборочной дисперсией s^2. Как нужно преобразовать данные, чтобы выборочное среднее равнялось 7, а выборочная дисперсия — 9? $y_i^* = 7 + 3(y_i - \bar{y})/s$
7. 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.
8. 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

$$(R^s - R^0) = \beta_1 + \beta_2(R_m^* - R_0^*) + \varepsilon \tag{3}$$

   величины R_m^* и R_0^* не наблюдаемы, но

$$R_m - R_0 = R_m^* - R_0^* + u \tag{4}$$