

# Задачник по эконометрике-1

(с шахматами и поэтэссами)

Дмитрий Борзых, Борис Демешев

16 января 2013 г.

## Содержание

|   |    |
|---|----|
| 1 Проверка гипотез строго по уставу!    | 2  |
| 2 Неклассифицировано                    | 2  |
| 3 МНК без матриц и вероятностей         | 5  |
| 4 Теорема Гаусса-Маркова и нормальность | 6  |
| 5 Мультиколлинеарность                  | 8  |
| 6 Гетероскедастичность                  | 8  |
| 7 Временные ряды                        | 10 |
| 8 Функциональная форма                  | 11 |
| 9 Инструментальные переменные           | 11 |
| 10 Проекция, Картинка                   | 11 |
| 11 МЕГАМАТРИЦА                          | 12 |
| 12 Метод максимального правдоподобия    | 12 |
| 13 Голая линейная алгебра               | 13 |
| 14 Парадигма случайных величин          | 13 |
| 15 Компьютерные упражнения              | 13 |

## Todo list

|   |    |
|---|----|
| Опция tidy=FALSE не работает. Почему? . . . . .                                 | 1  |
| Переделать классификацию. Каждый раздел делится на компьютерные и ручные задачи | 13 |

Опция tidy=FALSE не работает. Почему?

# 1 Проверка гипотез строго по уставу!

1. Условия применимости теста
2. Формулировка  $H_0$ ,  $H_a$  и уровня значимости  $\alpha$
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  $\text{Var}(\varepsilon) = \sigma^2 \cdot I$ . 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  $\sigma^2$  регрессионной модели.
- (i) Рассчитайте  $\widehat{\text{Var}}(\hat{\beta})$ , оценку для ковариационной матрицы вектора МНК-коэффициентов  $\hat{\beta}$ .
- (j) Найдите  $\widehat{\text{Var}}(\hat{\beta}_1)$ , несмещенную оценку дисперсии МНК-коэффициента  $\hat{\beta}_1$ .
- (k) Найдите  $\widehat{\text{Var}}(\hat{\beta}_2)$ , несмещенную оценку дисперсии МНК-коэффициента  $\hat{\beta}_2$ .
- (l) Найдите  $\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$ , несмещенную оценку ковариации МНК-коэффициентов  $\hat{\beta}_1$  и  $\hat{\beta}_2$ .
- (m) Найдите  $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2)$ ,  $\widehat{\text{Var}}(\hat{\beta}_1 - \hat{\beta}_2)$ ,  $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)$ ,  $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2 - 2\hat{\beta}_3)$
- (n) Найдите  $\widehat{\text{Corr}}(\hat{\beta}_1, \hat{\beta}_2)$ , 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
  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?
  4. Вася считает, что выборочная ковариация  $s\text{Cov}(y, \hat{y}) = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{n-1}$  это неплохая оценка для  $\text{Cov}(y_i, \hat{y}_i)$ . Прав ли он?
  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.
  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? ...
  9. Про  $R_{adj}^2$ 
    - (a) Может ли в модели с константой  $R_{adj}^2$  быть отрицательным?
    - (b) Что больше,  $R^2$  или  $R_{adj}^2$  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  $R_{adj}^2$ . 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  $R_{adj}^2$  больше, чем в модели  $B$ . В какой из моделей больше коэффициент  $\hat{\sigma}^2$ ?
  10. Сгенерируйте данные так, чтобы при оценке линейной регрессионной модели оказалось, что скорректированный коэффициент детерминации,  $R_{adj}^2$ , отрицательный.
  11. В классической линейной регрессионной модели  $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ , дисперсия зависимой переменной не зависит от номера наблюдения,  $\text{Var}(y_i) = \sigma^2$ . Почему для оценки  $\sigma^2$  вместо известной из курса математической статистики формулы  $\sum (y_i - \bar{y})^2 / (n - 1)$  используют  $\sum \hat{\varepsilon}_i^2 / (n - 2)$ ?

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  $\text{sCorr}(x, y)$ ,  $\text{sCorr}(y, \hat{y})$ .
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 А  $m_i = \beta_1 + \beta_2 r_i + \varepsilon_i$

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

- (a) Как связаны между собой оценки  $\hat{\beta}_1$  и  $\hat{\gamma}_1$ ?
  - (b) Как связаны между собой оценки  $\hat{\beta}_2$  и  $\hat{\gamma}_2$ ?
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?
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  $\hat{\gamma} < 0$ ?
17. Эконометрист Вовочка оценил регрессию  $y$  только на константу. Какой коэффициент  $R^2$  он получит?
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  $R^2$ .
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  $\text{Var}(w)$ ,  $\text{Var}(z)$ ,  $\text{Cov}(w, z)$
  - (d) Рассмотрим классическую линейную модель. Являются ли векторы  $\hat{\varepsilon}$  и  $\hat{y}$  перпендикулярными? Найдите  $\text{Cov}(\hat{\varepsilon}, \hat{y})$ .
20. Есть случайный вектор  $w = (w_1, w_2, \dots, w_n)'$ .
- (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  $E(w) \neq 0$  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  $\sum \hat{\varepsilon}_i$ ?

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.
- Найдите  $\mathbb{E}(\hat{\beta}_2)$  и  $\text{Var}(\hat{\beta}_2)$
  - Совпадает ли оценка  $\hat{\beta}_2$  с классической мнк-оценкой?
  - У какой оценки дисперсия выше, у  $\hat{\beta}_2$  или классической мнк-оценки?

### 3 МНК без матриц и вероятностей

- Даны  $n$  пар чисел:  $(x_1, y_1), \dots, (x_n, y_n)$ . Мы прогнозируем  $y_i$  по формуле  $\hat{y}_i = \hat{\beta} x_i$ . Найдите  $\hat{\beta}$  методом наименьших квадратов.
- Даны  $n$  чисел:  $y_1, \dots, y_n$ . Мы прогнозируем  $y_i$  по формуле  $\hat{y}_i = \hat{\beta}$ . Найдите  $\hat{\beta}$  методом наименьших квадратов.
- Даны  $n$  пар чисел:  $(x_1, y_1), \dots, (x_n, y_n)$ . 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.
- Даны  $n$  пар чисел:  $(x_1, y_1), \dots, (x_n, y_n)$ . Мы прогнозируем  $y_i$  по формуле  $\hat{y}_i = 1 + \hat{\beta} x_i$ . Найдите  $\hat{\beta}$  методом наименьших квадратов.
- 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.
- 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.
- Регрессия на дамми-переменную...
- Функция  $f(x)$  дифференцируема 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

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

- 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  $\hat{y} = \hat{\eta}_1 + \hat{\eta}_2 x$  по всем наблюдениям.
  - Возможно ли, что  $\hat{\beta}_2 > 0$ ,  $\hat{\gamma}_2 > 0$ , но  $\hat{\eta}_2 < 0$ ?
  - Возможно ли, что  $\hat{\beta}_1 > 0$ ,  $\hat{\gamma}_1 > 0$ , но  $\hat{\eta}_1 < 0$ ?
  - Возможно ли одновременное выполнение всех упомянутых условий?
- 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?
- Какие из указанные моделей можно представить в линейном виде?
  - $y_i = \beta_1 + \frac{\beta_2}{x_i} + \varepsilon_i$
  - $y_i = \exp(\beta_1 + \beta_2 x_i + \varepsilon_i)$
  - $y_i = 1 + \frac{1}{\exp(\beta_1 + \beta_2 x_i + \varepsilon_i)}$
  - $y_i = \frac{1}{1 + \exp(\beta_1 + \beta_2 x_i + \varepsilon_i)}$
  - $y_i = x_i^{\beta_2} e^{\beta_1 + \varepsilon_i}$

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

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

|             | 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
```

6. 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:  $Fertility_i = \gamma_1 + \gamma_2(Education_i + Examination_i) + \gamma_3 Catholic_i + u_i$  Набор данных доступен в пакете R:

```
h <- swiss
```

7. По 2040 наблюдениям оценена модель зависимости стоимости квартиры в Москве (в 1000\$) от общего метража и метража жилой площади. Оценка ковариационной матрицы  $\widehat{Var}(\hat{\beta})$  имеет вид
  - (a) Проверьте  $H_0: \beta_{totsp} = \beta_{livesp}$ . В чём содержательный смысл этой гипотезы?
  - (b) Постройте доверительный интервал для  $\beta_{totsp} - \beta_{livesp}$ . В чём содержательный смысл этого доверительного интервала?
8. По 2040 наблюдениям оценена модель зависимости стоимости квартиры в Москве (в 1000\$) от общего метража и метража жилой площади. Оценка ковариационной матрицы  $\widehat{Var}(\hat{\beta})$  имеет вид
  - (a) Постройте 95%-ый доверительный интервал для ожидаемой стоимости квартиры с жилой площадью 30 м<sup>2</sup> и общей площадью 60 м<sup>2</sup>.
  - (b) Постройте 95%-ый прогнозный интервал для фактической стоимости квартиры с жилой площадью 30 м<sup>2</sup> и общей площадью 60 м<sup>2</sup>.
9. Рассмотрим модель с линейным трендом без свободного члена,  $y_t = \beta t + \varepsilon_t$ .
  - (a) Найдите МНК оценку коэффициента  $\beta$
  - (b) Рассчитайте  $E(\hat{\beta})$  и  $Var(\hat{\beta})$  в предположениях теоремы Гаусса-Маркова

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     |

(с) Верно ли, что оценка  $\hat{\beta}$  состоятельна?10. 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  $\text{Var}(\hat{\beta}_2)$  в предположениях теоремы Гаусса-Маркова(с) Верно ли, что оценка  $\hat{\beta}_2$  состоятельна?11. 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  $\text{Var}(\hat{\beta}_2)$  в предположениях теоремы Гаусса-Маркова(с) Верно ли, что оценка  $\hat{\beta}_2$  состоятельна?

12. 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.

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

(b) Проинтерпретируйте коэффициент при  $brick_i \cdot totsp_i$ 13. 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  $\varepsilon_i$  имеет нормальное распределение  $N(0, 1)$ .

|             | (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     |

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

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

## 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. 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  $s\text{Corr}(g, h) = -0.9$ . Гена и Чебурашка собирали независимо от Пионеров, поэтому выборочные корреляции  $s\text{Corr}(g, x) = 0$ ,  $s\text{Corr}(h, x) = 0$ . 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.

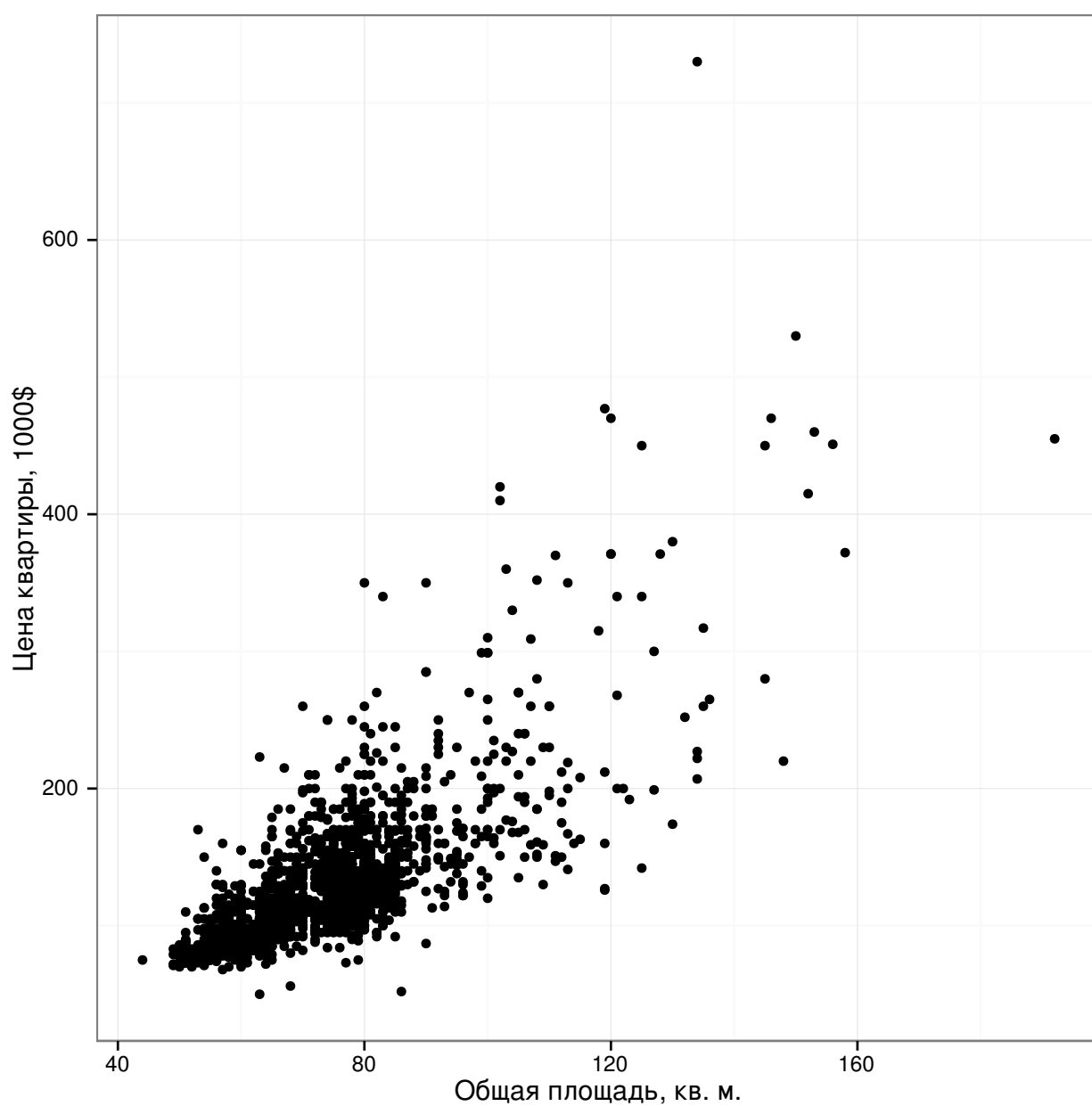
## 6 Гетероскедастичность

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



|             | (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     |



Какие подходы к оцениванию зависимости имеет смысл посоветовать исходя из данного

графика?

3. По наблюдениям  $x = (1, 2, 3)'$ ,  $y = (2, -1, 3)'$  оценивается модель  $y = \beta_1 + \beta_2 x + \varepsilon$ . Ошибки  $\varepsilon$  гетероскедастичны и известно, что  $\text{Var}(\varepsilon_i) = \sigma^2 \cdot x_i^2$ .
  - (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  $\text{Var}(\varepsilon_i) = \sigma^2 x_i^2$ . Как надо преобразовать исходные регрессоры и зависимую переменную, чтобы устранить гетероскедастичность?

## 7 Временные ряды

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

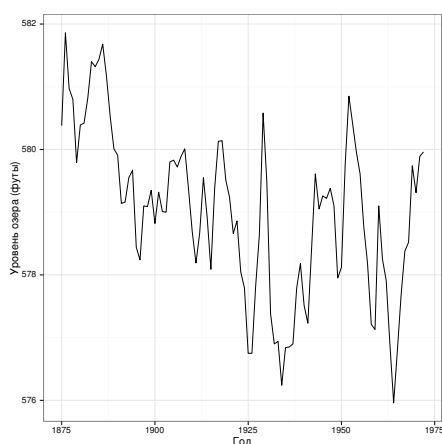
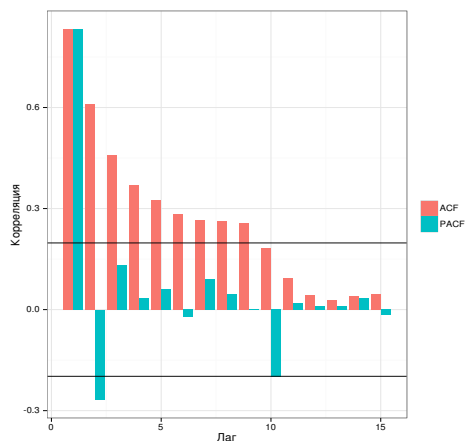


График автокорреляционной и частной автокорреляционной функций:



- (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,  $\mathbb{E}(\varepsilon | x) = 0$

Предопределённость,  $\mathbb{E}(\varepsilon_t | x_t) = 0$  для всех  $t$

1. Табличка 2 на 2. Найдите  $\mathbb{E}(\varepsilon)$ ,  $\mathbb{E}(\varepsilon|x)$ ,  $\text{Cov}(\varepsilon, x)$ .
2. Приведите примеры дискретных случайных величин  $\varepsilon$  и  $x$ , таких, что
  - (a)  $\mathbb{E}(\varepsilon) = 0$ ,  $\mathbb{E}(\varepsilon | x) = 0$ , но величины зависимы. Чему в этом случае равно  $\text{Cov}(\varepsilon, x)$ ?
  - (b)  $\mathbb{E}(\varepsilon) = 0$ ,  $\text{Cov}(\varepsilon, x) = 0$ , но  $\mathbb{E}(\varepsilon | x) \neq 0$ . Зависимы ли эти случайные величины?
3. Все предпосылки классической линейной модели выполнены,  $y = \beta_1 + \beta_2 x + \varepsilon$ . Рассмотрим альтернативную оценку коэффициента  $\beta_2$ ,

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

- (a) Является ли оценка несмещенной?
- (b) Любые ли  $z_i$  можно брать?
- (c) Найдите  $\text{Var}(\hat{\beta}_{2,IV})$
- 4.

## 10 Проекция, Картинка

1. Найдите на Картинке все перпендикулярные векторы. Найдите на Картинке все прямоугольные треугольники. Сформулируйте для них теоремы Пифагора.
2. Покажите на Картинке TSS, ESS, RSS,  $R^2$ ,  $\text{sCov}(\hat{y}, y)$
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  $\hat{\varepsilon}$ .
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 коэффициента при  $\hat{y}$ ? Почему оценка коэффициента неслучайна, а оценка её дисперсии положительна?

5. При каких условиях  $TSS = ESS + RSS$ ?

## 11 МЕГАМАТРИЦА

1. В рамках классической линейной модели найдите все математические ожидания и все ковариационные матрицы всех пар случайных векторов:  $\varepsilon$ ,  $y$ ,  $\hat{y}$ ,  $\hat{\varepsilon}$ ,  $\hat{\beta}$ . Т.е. найдите  $\mathbb{E}(\varepsilon)$ ,  $\mathbb{E}(y)$ ,  $\dots$  и  $\text{Cov}(\varepsilon, y)$ ,  $\text{Cov}(\varepsilon, \hat{y})$ ,  $\dots$
2. Найдите  $\mathbb{E}(\sum(\varepsilon_i - \bar{\varepsilon})^2)$ ,  $\mathbb{E}(RSS)$
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
4.  $\mathbb{E}(TSS)$ ,  $\mathbb{E}(ESS)$  — громоздкие
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?
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  $\varepsilon' A \varepsilon$ ?
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  $u = A^{-1/2}\varepsilon$ ?

## 12 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:  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  при  $i \neq j$ , и  $\text{Var}(\varepsilon_i) = \sigma^2 x_i^2$ .
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:  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  при  $i \neq j$ , и  $\text{Var}(\varepsilon_i) = \sigma^2 |x_i|$ .
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  $\beta_1$  и  $\beta_2$
- (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». 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  $f(x) = e^{-x}/(1 + e^{-x})^2$ .
- Является ли  $f(x)$  чётной?
  - Постройте график  $f(x)$
  - Найдите функцию распределения,  $F(x)$
  - Найдите  $\mathbb{E}(X)$ ,  $\text{Var}(X)$
  - На какое известный закон распределения похож логистический?
6. Логит модель часто формулируют в таком виде:

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

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

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

## 13 Голая линейная алгебра

Здесь будет собран минимум задач по линейной алгебре.

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

## 14 Парадигма случайных величин

- Найдите  $E(Y|X)$
- Про многомерное нормальное распределение
-

## 15 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>.

- 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.
  - Какая задача из первой контрольной работы наиболее существенно влияет на результат второй контрольной?
  - Влияет ли пол на результат второй контрольной?
  - Влияет ли редкость имени на результат второй контрольной?
  - Что можно сказать про влияние группы, в которой учится студент?
- 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.
  - Напишите функцию 'makar\_step'. На вход функции подаётся вектор количества стаканчиков в каждой стопке до переукладывания. На выходе функция возвращает количества стаканчиков в каждой стопке после одного переукладывания.
  - Изначально стаканчики были разложены в две стопки, из 25 и 30 стаканчиков. Как разложатся стаканчики если покупателей не будет достаточно долго?
- 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  $\hat{\beta}$ ,  $\widehat{\text{Var}}(\hat{\beta})$ ,  $\hat{y}$ ,  $\hat{\varepsilon}$ ,  $ESS$ ,  $RSS$  и  $TSS$ . По возможности функция должна проверять корректность аргументов, например, что в  $y$  и  $X$  одинаковое число наблюдений и т.д.
- 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.
  - Сколько регрессоров оказалось значимо на 5% уровне?
  - Сколько регрессоров в среднем значимо на 5% уровне?
  - 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.
- (?) 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.
- (?) 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?
- 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  $\rho$ . 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  $\rho$  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 САРМ несколькими способами.
- (a) Соберите подходящие данные для модели САРМ. Нужно найти три временных ряда: ряд цен любой акции, любой рыночный индекс, безрисковый актив. Переведите цены в доходности.
  - (b) Постройте графики
  - (c) Оцените модель САРМ 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 САРМ 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

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

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

$$R_m - R_0 = R_m^* - R_0^* + u \quad (4)$$