

Задачник по эконометрике-1

(с шахматами и поэтэссами)

Дмитрий Борзых, Борис Демешев

29 октября 2012 г.

1 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) Рассчитайте $TSS = \sum (y_i - \bar{y})^2$, $RSS = \sum (y_i - \hat{y}_i)^2$ и $ESS = \sum (\hat{y}_i - \bar{y})^2$.
- (d) Рассчитайте при помощи метода наименьших квадратов $\hat{\beta}$, оценку для вектора неизвестных коэффициентов.
- (e) Чему равен $\hat{\varepsilon}_5$, МНК-остаток регрессии, соответствующий 5-ому наблюдению?
- (f) Чему равен R^2 в модели? Прокомментируйте полученное значение с точки зрения качества оцененного уравнения регрессии.
- (g) Используя приведенные выше данные, рассчитайте несмещенную оценку для неизвестного параметра σ^2 регрессионной модели.
- (h) Рассчитайте $\widehat{\text{Var}}(\hat{\beta})$, оценку для ковариационной матрицы вектора МНК-коэффициентов $\hat{\beta}$.
- (i) Найдите $\widehat{\text{Var}}(\hat{\beta}_1)$, несмещенную оценку дисперсии МНК-коэффициента $\hat{\beta}_1$.
- (j) Найдите $\widehat{\text{Var}}(\hat{\beta}_2)$, несмещенную оценку дисперсии МНК-коэффициента $\hat{\beta}_2$.
- (k) Найдите $\widehat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$, несмещенную оценку ковариации МНК-коэффициентов $\hat{\beta}_1$ и $\hat{\beta}_2$.
- (l) Найдите $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2)$, $\widehat{\text{Var}}(\hat{\beta}_1 - \hat{\beta}_2)$, $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3)$, $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2 - 2\hat{\beta}_3)$
- (m) Найдите $\widehat{\text{Corr}}(\hat{\beta}_1, \hat{\beta}_2)$, оценку коэффициента корреляции МНК-коэффициентов $\hat{\beta}_1$ и $\hat{\beta}_2$.
- (n) Найдите $s_{\hat{\beta}_1}$, стандартную ошибку МНК-коэффициента $\hat{\beta}_1$.
- (o) Рассчитайте выборочную ковариацию y и \hat{y} .

- (р) Найдите выборочную дисперсию y , выборочную дисперсию \hat{y} .
2. Априори известно, что парная регрессия должна проходить через точку (x_0, y_0) .
 - (а) Выведите формулы МНК оценок;
 - (б) 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.
 - (а) Найдите несмещеную оценку веса первого шара, обладающую наименьшей дисперсией.
 - (б) Как можно проинтерпретировать нулевое математическое ожидание ошибки взвешивания?
 4. Вася считает, что $sCov(y, \hat{y}) = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{n-1}$ это неплохая оценка для $Cov(y_i, \hat{y}_i)$. 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
 - (а) Вовочка построит регрессию kr на константу, p_1, p_2, p_3, p_4 и p_5
 - (б) 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
 - (а) Может ли в модели с константой R_{adj}^2 быть отрицательным?
 - (б) Что больше, R^2 или R_{adj}^2 в модели с константой?
 - (с) 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. 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 σ^2 вместо известной из курса математической статистики формулы $\sum (y_i - \bar{y})^2 / (n - 1)$ используют $\sum \hat{\varepsilon}_i^2 / (n - 2)$?
 11. 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 $sCorr(x, y)$, $sCorr(y, \hat{y})$.

12. 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

13. 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 $\hat{\gamma}_2$?

14. 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?

15. 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 $\hat{\gamma} < 0$?

16. Эконометрист Вовочка оценил регрессию y только на константу. Какой коэффициент R^2 он получит?

2 МНК без матриц и вероятностей

1. 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.
2. 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.
3. 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.
4. 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.
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.
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.
7. Регрессия на дамми-переменную...
8. 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

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

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

3 Инструментальные переменные

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

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

- Табличка 2 на 2. Найдите $\mathbb{E}(\varepsilon)$, $\mathbb{E}(\varepsilon|x)$, $\text{Cov}(\varepsilon, x)$.
- Приведите примеры дискретных случайных величин ε и x , таких, что
 - $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon | x) = 0$, но величины зависимы. Чему в этом случае равно $\text{Cov}(\varepsilon, x)$?
 - $\mathbb{E}(\varepsilon) = 0$, $\text{Cov}(\varepsilon, x) = 0$, но $\mathbb{E}(\varepsilon | x) \neq 0$. Зависимы ли эти случайные величины?
- Все предпосылки классической линейной модели выполнены, $y = \beta_1 + \beta_2 x + \varepsilon$. Рассмотрим альтернативную оценку коэффициента β_2 ,

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

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

4.

4 Проекция, Картинка

- Найдите на Картинке четыре прямоугольных треугольника. Сформулируйте четыре теоремы Пифагора.
- Покажите на Картинке TSS, ESS, RSS, R^2 , $\text{sCov}(\hat{y}, y)$
- 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 $TSS = ESS + RSS$?

5 МЕГАМАТРИЦА

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 $\text{Cov}(\varepsilon, y)$, $\text{Cov}(\varepsilon, \hat{y})$, ...
2. Найдите $\mathbb{E}(\sum(\varepsilon_i - \bar{\varepsilon})^2)$, $\mathbb{E}(RSS)$
3. $\mathbb{E}(TSS)$, $\mathbb{E}(ESS)$ — громоздкие
4. 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 Голая линейная алгебра

Здесь будет собран минимум задач по линейной алгебре.

1. Приведите пример таких A и B , что $\det(AB) \neq \det(BA)$.
2. Для матриц-проекторов $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$ и $P = X(X'X)^{-1}X'$ найдите $\text{tr}(\pi)$, $\text{tr}(P)$, $\text{tr}(I - \pi)$, $\text{tr}(I - P)$.
3. Выпишите в явном виде матрицы $X'X$, $(X'X)^{-1}$ и $X'y$, если

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \text{ и } X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$
4. Выпишите в явном виде матрицы π , πy , $\pi \varepsilon$, $I - \pi$, если $\pi = \vec{1}(\vec{1}'\vec{1})^{-1}\vec{1}'$.

7 Компьютерные упражнения

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 переукладывания. На выходе функция возвращает количества стаканчиков в каждой стопке после одного переукладывания.

- (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 $\hat{\beta}$, $\widehat{\text{Var}}(\hat{\beta})$, \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.
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.

8 Вопросы теоретического характера

1. Что означают слова автокорреляция, гетероскедастичность, гомоскедастичность?
2. Напишите формулу для оценок коэффициентов в парной регрессии без матриц
3. Напишите формулу для оценок коэффициентов в множественной регрессии
4. Аналогично для дисперсий