Подборка экзаменов по эконометрике.
Факультет экономики, НИУ-ВШЭ

Коллектив кафедры
математической экономики и эконометрики,
фольклор,
очень умные студенты

21 ноября 2016 г.

Содержание

1. Описание

Свежая версия на https://github.com/bdemeshev/em301/tree/master/metrics_exams. Часть задач и многие решения придуманы студентами. Огромное спасибо всем тем, кто в мучениях, своей кровью дописывал этот документ :)

2. Вечное

2.1. Гимн-памятка для эконометриста

Эмилю Борисовичу Ершову посвящается

Ничего на свете лучше нету,
Чем оценивать параметр «бета»!
Лучшее оружие демократа —

Метод наименьшего квадрата!

Если вдруг подавит вас депрессия,
Виновата, значит, здесь дисперсия.
Убери гетероскедастичность,
Это придаёт оптимистичность.

Если в данных автокорреляция,
Всё, что посчитал ты, — профанация.
Применяй, не глядя исподлобия,
Максимальное правдоподобие.

Если ощутил ты свою бренность,
Не иначе это эндогенность.
Соглашайся выдать алименты
Тем, кто знает, где взять инструменты.

Где б ты ни был, в саклях и ярангах
Применяй везде условья ранга.
Помни также: лучшая зарядка —
Выполнить условие порядка.

Мы своё призванье не забудем!
BLUE-оценки мы предъявим людям!
Нам законов априорных своды
Не понизят степеней свободы!

2.2. Прошение о повышение оценки

От ....

Группа ......

Я считаю, что моя итоговая оценка по курсу .... должна быть исправлена с .... на ... по следующим причинам (обведите нужные).

1. Это единственная плохая оценка в моей зачетке

2. Тот, кто полностью списал мою работу, получил более высокую оценку

3. Тот, у кого я полностью списал работу, получил более высокую оценку

4. Из-за низкого рейтинга меня могут не взять в

   (a) РЭШ

   (b) СМЕРШ

   (c) МГУ

   (d) На Луну

   (e) ....

5. Мне нужно получить 10, чтобы компенсировать 4 по ....

6. Меня лишат стипендии

7. Я не успел договориться с тётечками из копировального отдела и раздобыть варианты контрольной, потому что ....

8. Я не посещал лекции, а тот, чьими конспектами я пользовался, не записал материал, необходимый для сдачи контрольных и домашек

9. Я отлично понимаю теорию, просто не умею решать задачи

10. Я умею решать все задачи, а на контрольной требовалось знание теории

11. У лектора/семинариста были предрассудки против ....

12. Все вопросы на экзамене допускали двойную трактовку. Я считаю, что не должен нести наказание за то, что мое мнение — особенное

13. Если я получу плохую оценку отец отберет у меня ключи от машины

14. Я не мог/могла заниматься из-за необходимости разгружать вагоны по ночам

15. Нам сказали использовать творческий подход, но не объяснили, что это означает

16. Я использовал в домашках творческий подход, но мне было сказано, что я несу всякую чушь

17. Все остальные преподаватели согласны повысить мою оценку

18. Семинары и лекции начинались:

    (a) слишком рано, я еще спал
    (b) слишком поздно, я уже спал
    (c) в обеденное время, я был голодный

19. Причина по которой я получил низкую оценку проста — я очень честный. Не хочу ничего говорить о моих одногруппниках

20. У меня нет особой причины, я просто хочу оценку повыше

Дата ....

Подпись ....

3. Немного теории

3.1. Конвенция об обозначениях

- y — вектор-столбец зависимых переменных размера $(n \times 1)$, наблюдаемый случайный
- β — вектор-столбец неизвестных коэффициентов размера $(k \times 1)$, ненаблюдаемый, неслучайный
- $\hat{y}$ — вектор столбец прогнозов для y, полученных по некоторой модели, размера $(n \times 1)$, наблюдаемый, случайный
- $\hat{\beta}$ — вектор-столбец оценок β размера $(k \times 1)$, наблюдаемый, случайный
- X — матрица всех объясняющих переменных, размера $(n \times k)$. Известная, стохастическая или детерминированная в зависимости от парадигмы.

- ε — вектор-столбец случайных ошибок размера $(n \times 1)$, ненаблюдаемый случайный
- $\hat{\varepsilon}$ — вектор-столбец остатков модели размера $(n \times 1)$, наблюдаемый случайный
- c — вектор из единиц

В некоторых учебниках используется обозначение Y для исходного вектора зависимых переменных, а y — для центрированного, т.е. $y = Y - \bar{Y}$. В этом документе y обозначает исходный вектор y.

3.2. Свойства ковариационных матриц

Здесь y — вектор-столбец $n \times 1$, z — вектор-столбец $k \times 1$, A — матрица констант подходящего размера, b — вектор констант подходящего размера.

1. $\mathbb{E}(Ay + b) = A\mathbb{E}(y) + b$, $\mathbb{E}(yA + b) = \mathbb{E}(y)A + b$
2. $\mathrm{Cov}(y, z) = \mathbb{E}(yz') - \mathbb{E}(y)\mathbb{E}(z')$
3. $\mathrm{Var}(y) = \mathbb{E}(yy') - \mathbb{E}(y)\mathbb{E}(y')$
4. $\mathrm{Cov}(Ay + b, z) = A\,\mathrm{Cov}(y, z)$, $\mathrm{Cov}(y, Az + b) = \mathrm{Cov}(y, z)A'$
5. $\mathrm{Var}(Ay + b) = A\,\mathrm{Var}(y)A'$
6. $\mathrm{Cov}(y, z) = \mathrm{Cov}(z, y)'$

3.3. ТГМ. Детерминированные регрессоры

3.4. ТГМ. Стохастические регрессоры

Если:

1. Истинная зависимость имеет вид $y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$

   В матричном виде: $y = X\beta + \varepsilon$
2. С помощью МНК оценивается регрессия y на константу, $x_{.2}$, $x_{.3}$, ..., $x_{.k}$

   В матричном виде: $\hat{\beta} = (X'X)^{-1}X'y$
3. Наблюдений больше, чем оцениваемых коэффициентов β: $n > k$
4. Строгая экзогенность: $E(\varepsilon_i |$ все $x_{ij}) = 0$

   В матричном виде: $E(\varepsilon_i | X) = 0$
5. Условная гомоскедастичность: $E(\varepsilon_i^2 |$ все $x_{ij}) = \sigma^2$

   В матричном виде: $E(\varepsilon_i^2 | X) = \sigma^2$
6. $Cov(\varepsilon_i, \varepsilon_j | X) = 0$ при $i \neq j$
7. вектора $(x_{i.}, y_i)$ — независимы и одинаково распределены
8. с вероятностью 1 среди регрессоров нет линейно зависимых $rank(X) = k$ $det(X'X) \neq 0$ $(X'X)^{-1}$ существует

То:

1. (тГМ) МНК оценки $\hat{\beta}$ линейны по y: $\hat{\beta}_j = c_1 y_1 + ... + c_n y_n$

2. (тГМ) МНК оценки несмещенные. А именно, $E(\hat{\beta}|X) = \beta$, и в частности $E(\hat{\beta}) = \beta$

3. (тГМ) МНК оценки эффективны среди линейных несмещённых оценок. Для любой альтернативной оценки $\hat{\beta}^{alt}$ удовлетворяющей свойствам 1 и 2: $Var(\hat{\beta}_j^{alt}|X) \geqslant Var(\hat{\beta}_j|X)$ $Var(\hat{\beta}_j^{alt}) \geqslant Var(\hat{\beta}_j)$

4. $Var(\hat{\beta}|X) = \sigma^2 (X'X)^{-1}$

5. $Cov(\hat{\beta}, \hat{\varepsilon}|X) = 0$

6. $E(\hat{\sigma}^2|X) = \sigma^2$, и $E(\hat{\sigma}^2) = \sigma^2$?остается ли при условной ГК?

Если дополнительно к предпосылкам теоремы Гаусса-Маркова известно, что $\varepsilon|X \sim N$, то:

1. МНК оценки эффективны среди всех несмещённых оценок.

2. $t|X \sim t_{n-k}$, $t \sim t_{n-k}$

3. $RSS/\sigma^2|X \sim \chi^2_{n-k}$, $RSS/\sigma^2 \sim \chi^2_{n-k}$

4. F тест $F|X \sim F$

Если дополнительно к предпосылкам теоремы Гаусса-Маркова известно, что $n \to \infty$, то:

1. $\hat{\beta} \to \beta$ по вероятности

2. $t \to N(0, 1)$

3. $rF \to \chi^2_r$, r — число ограничений

4. $nR^2 \to \chi^2_{k-1}$

5. $\frac{RSS}{n-k} \to \sigma^2$

3.5. Ликбез по линейной алгебре

Определение. *Неформально. Если матрица A квадратная, то её определителем называется площадь/объём параллелограмма/параллелепипеда образованного векторами-столбцами матрицы. Знак определителя задаётся порядком следования векторов.*

Свойства определителя:

1. $\det(AB) = \det(A)\det(B) = \det(BA)$, если A и B квадратные

2. $\det(A) = \prod \lambda_i$, где λ_i — собственное число матрицы A, возможно комплексное.

Определение. *Ненулевой вектор x называется собственным вектором матрицы A, если при умножении на матрицу A он остается на той же прямой, т.е. $Ax = \lambda x$.*

Определение. *Число λ называется собственным числом матрицы A, если существует вектор x, который при умножении на матрицу A изменяется в λ раз, т.е. $Ax = \lambda x$.*

Определение. *Если матрица A квадратная, то её следом называется сумма диагональных элементов,* $\operatorname{tr}(A) = \sum a_{ii}$.

Свойства следа:

1. $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$

2. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, если AB и BA существуют. При этом A и B могут не быть квадратными матрицами.

3. $\mathrm{tr}(A) = \sum \lambda_i$, где λ_i — собственное число матрицы A, возможно комплексное.

Смысл следа. Если умножение на матрицу A — это проецирование, то есть Ax — есть проекция вектора x на некоторое подпространство, то $\mathrm{tr}(A)$ — размерность этого подпространства. Действительно, если A — проектор, то $A^2 = A$ и собственные числа матрицы A равны нулю или единице. Поэтому $\mathrm{tr}(A)$ равен количеству собственных чисел равных единице. И, следовательно, $\mathrm{tr}(A)$ равен $\mathrm{rank}(A)$, то есть размерности пространства, на которое матрица A проецирует вектора. У следа матрицы есть и другие смыслы [1].

3.6. Ожидание от RSS

***Теорема* 1.** *След и математическое ожидание можно переставлять,* $\mathbb{E}(\mathrm{tr}(A)) = \mathrm{tr}(\mathbb{E}(A))$.

***Теорема* 2.** *Математическое ожидание квадратичной формы*

$$\mathbb{E}(x'Ax) = \mathrm{tr}(A\,\mathrm{Var}(x)) + \mathbb{E}(x')A\mathbb{E}(x) \tag{1}$$

Доказательство. Мы будем пользоваться простым приёмом. Если u — это скаляр, вектор размера 1 на 1, то $\mathrm{tr}(u) = u$.
Поехали,

$$\mathbb{E}(x'Ax) = \mathbb{E}(\mathrm{tr}(x'Ax)) = \mathbb{E}(\mathrm{tr}(Axx')) = \mathrm{tr}(\mathbb{E}(Axx')) = \mathrm{tr}(A\mathbb{E}(xx')) \tag{2}$$

По определению дисперсии, $\mathrm{Var}(x) = \mathbb{E}(xx') - \mathbb{E}(x)\mathbb{E}(x')$. Поэтому:

$$\mathrm{tr}(A\mathbb{E}(xx')) = \mathrm{tr}(A(\mathrm{Var}(x) + \mathbb{E}(x)\mathbb{E}(x'))) = \mathrm{tr}(A\,\mathrm{Var}(x)) + \mathrm{tr}(A\mathbb{E}(x)\mathbb{E}(x')) \tag{3}$$

И готовимся снова использовать приём $\mathrm{tr}(u) = u$:

$$\mathrm{tr}(A\,\mathrm{Var}(x)) + \mathrm{tr}(A\mathbb{E}(x)\mathbb{E}(x')) = \mathrm{tr}(A\,\mathrm{Var}(x)) + \mathrm{tr}(\mathbb{E}(x')A\mathbb{E}(x)) = \mathrm{tr}(A\,\mathrm{Var}(x)) + \mathbb{E}(x')A\mathbb{E}(x) \tag{4}$$

□

3.7. Устоявшиеся слова

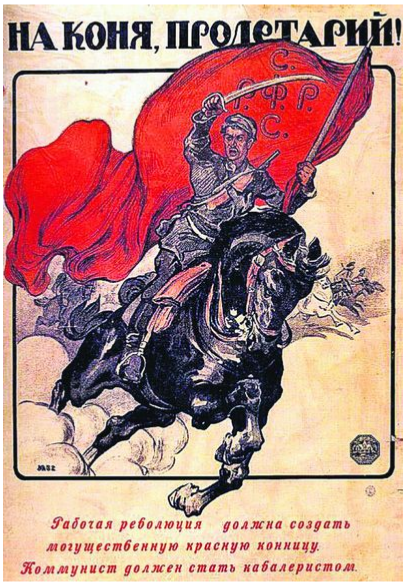
Выражение «гипотеза о значимости отдельного коэффициента» на самом деле означает «гипотеза о незначимости отдельного коэффициента», т.к. де-факто проверяется гипотеза H_0: $\beta_j = 0$.
Выражение «гипотеза о значимости регрессии в целом» или «гипотеза об адекватности регрессии» на самом деле означает «гипотеза о незначимости регрессии в целом», т.к. проверяется H_0: $\beta_2 = \ldots = \beta_k = 0$.
В некоторых источниках гипотезу об адекватности регрессии ошибочно обозначают H_0: $R^2 = 0$. Эту ошибку не нужно повторять.
Гипотезы имеет смысл проверять о ненаблюдаемых величинах, а величина R^2 является наблюдаемой. И если уж на то пошло, то проверить гипотезу о том, что $R^2 = 0$ тривиально. Для этого не нужно знать ничего из теории вероятностей, достаточно просто сравнить посчитанное значение R^2 с нулём.
Более того, даже корректировка H_0: $\mathbb{E}(R^2) = 0$ неверна. В модели, где в регрессоры включена только константа, величина R^2 тождественно равна нулю, поэтому $\mathbb{E}(R^2) = 0$ и проверять такую гипотезу бессмысленно. В модели, где в регрессоры включено что-то помимо константы, R^2 является неотрицательной случайной величиной с $\mathbb{P}(R^2 > 0) > 0$. Поэтому а-приори $\mathbb{E}(R^2) > 0$ и проверка гипотезы H_0: $\mathbb{E}(R^2) = 0$ снова бессмысленна.
Кстати, обозначение H_0 по-английски читается как «H naught», а не «H zero» или «H null». Также корректно говорить «the null hypothesis».

Список литературы

[1] *Смысл следа матрицы.* URL: http://mathoverflow.net/questions/13526/.

4. 2012-2013

4.1. Праздник 1. Пролетарий на коня!



1. Найдите длины векторов $a = (1, 2, 3)$ и $b = (1, 0, -1)$ и косинус угла между ними.

2. Сформулируйте теорему о трёх перпендикулярах.

3. Сформулируйте и докажите теорему Пифагора.

4. Для матрицы
$$A = \begin{pmatrix} 2 & 3 & 0 \\ 3 & 10 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$
   (a) Найдите собственные числа и собственные векторы матрицы.
   (b) Найдите обратную матрицу, A^{-1}, ее собственные векторы и собственные числа.
   (c) Представьте матрицу A в виде $A = CDC^{-1}$, где D — диагональная матрица.
   (d) Представьте A^{2012} в виде произведения трёх матриц.

5. Вася и Петя независимо друг от друга решают тест по теории вероятностей. В тесте всего два вопроса. На каждый вопрос два варианта ответа. Петя знает решение каждого вопроса с вероятностью $0{,}7$. Если Петя не знает решения, то он отвечает равновероятно наугад. Вася знает решение каждого вопроса с вероятностью $0{,}5$. Если Вася не знает решения, то он отвечает равновероятно наугад.
   (a) Какова вероятность того, что Петя правильно ответил на оба вопроса?
   (b) Какова вероятность того, что Петя правильно ответил на оба вопроса, если его ответы совпали с Васиными?
   (c) Чему равно математическое ожидание числа Петиных верных ответов?

(d) Чему равно математическое ожидание числа Петиных верных ответов, если его ответы совпали с Васиными?

6. Для случайных величин X и Y заданы следующие значения: $\mathbb{E}(X) = 1, \mathbb{E}(Y) = 4, \mathbb{E}(XY) = 8$, $\text{Var}(X) = \text{Var}(Y) = 9$. Для случайных величин $U = X + Y$ и $V = X - Y$ вычислите:

   (a) $\mathbb{E}(U)$, $\text{Var}(U)$, $\mathbb{E}(V)$, $\text{Var}(V)$, $\text{Cov}(U, V)$

   (b) Можно ли утверждать, что случайные величины U и V независимы?

7. Вася ведёт блог. Обозначим X_i — количество слов в i–ой записи. После первого года он по своим записям обнаружил, что $\bar{X}_{200} = 95$ и выборочное стандартное отклонение равно 282 слова. На уровне значимости $\alpha = 0.10$ проверьте гипотезу о том, что $\mu = 100$ против альтернативной гипотезы $\mu \neq 100$. Найдите также точное P-значение.

4.2. Праздник 2. Базовая задача

1. Случайные величины Z_i независимы и нормально распределены $N(0, 1)$. Для их суммы $S = \sum_{i=1}^n Z_i$ найдите $\mathbb{E}(S)$ и $\text{Var}(S)$.

2. Социологическим опросам доверяют 70% жителей. Те, кто доверяют опросам, на все вопросы отвечают искренне; те, кто не доверяют, отвечают равновероятно наугад. Социолог Петя в анкету очередного опроса включил вопрос «Доверяете ли Вы социологическим опросам?»

   (a) Какова вероятность, что случайно выбранный респондент ответит «Да»?

   (b) Какова вероятность того, что он действительно доверяет, если известно, что он ответил «Да»?

3. 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

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

   Для удобства расчетов приведены матрицы

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

   (a) Укажите число наблюдений.

   (b) Укажите число регрессоров с учетом свободного члена.

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

   (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) Чему равен $\hat{\varepsilon}_4$, МНК-остаток регрессии, соответствующий 4-ому наблюдению?

   (f) Чему равен R^2 в модели?

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

(k) Найдите $\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\beta}_2)$

(l) Найдите $\widehat{\text{Corr}}(\hat{\beta}_1, \hat{\beta}_2)$, оценку коэффициента корреляции МНК-коэффициентов $\hat{\beta}_1$ и $\hat{\beta}_2$.

(m) Найдите $se(\hat{\beta}_1)$, стандартную ошибку МНК-коэффициента $\hat{\beta}_1$.

4. В классической линейной модели предполагается, что $\mathbb{E}(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2 I$. Найдите $\text{Cov}(y, \hat{\varepsilon})$, $\text{Cov}(\hat{y}, \hat{\varepsilon})$.

4.3. Праздник 3. Дню рождения буквы «ё» посвящается...

1. Выберите верные варианты.

(a) Побасёнка — Побасенка

(b) Вёдро — Ведро

(c) Гренадёр — Гренадер

(d) Новорождённый — Новорожденный

(e) Бытиё — Бытие

(f) Опёка — Опека

(g) Сёрфинг — Серфинг

(h) Пафнутий Львович Чебышёв — Пафнутий Львович Чебышев

(i) Лёв Николаевич Толстой — Лев Николаевич Толстой

2. 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 95%-ый доверительный интервал для коэффициента при переменной Catholic

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

	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 $p < 0.05$

Таблица 1

   (a) Посчитайте RSS для каждой модели.

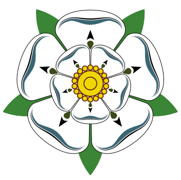
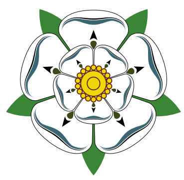
   (b) Какая модель является ограниченной (короткой), какая — неограниченной (длинной)?

   (c) Какие ограничения нужно добавить к неограниченной модели, чтобы получить ограниченную?

   (d) Найдите наблюдаемое значение F статистики.

   (e) Отвергается или не отвергается гипотеза об ограничениях?

4.4. Праздник 4, ML

 Версия Белой Розы 

1. Наблюдения $X_1, X_2, \ldots, X_n$ независимы и одинаково распределены с функцией плотности $f(x) = \frac{a(\ln(x))^{a-1}}{x}$ при $x \in [1; e]$. По 100 наблюдениям известно, что $\sum_{i=1}^{100} \ln(\ln(X_i)) = -20$

   (a) Оцените параметр a методом максимального правдоподобия

   (b) Проверьте гипотезу о том, что $a = 5$ против альтернативной $a \neq 5$ с помощью теста отношения правдоподобия, теста Вальда, теста множителей Лагранжа

   (c) Постройте 95%-ый доверительный интервал для параметра a

2. [R] Фактическое распределение часовой и десятиминутной скорости ветра хорошо приближается распределением Вейбулла. Случайная величина имеет распределение Вейбулла, если её функция плотности при $x > 0$ имеет вид

   $$f(x) = \frac{1}{\lambda^k} k x^{k-1} \exp(-x^k/\lambda^k)$$

   (a) Оцените параметры k и λ методом максимального правдоподобия

   (b) Постройте 95%-ые доверительные интервалы для k и λ

   Часовые данные я не нашёл, нашёл дневные. Данные по среднедневной скорости ветра содержатся в weather_nov_2012_moskow.csv в стобике wind. Данные взяты с сайта http://www.atlas-yakutia.ru/weather/climate_russia-I.html.

   Hint: read.csv("filename.csv")

Версия Алой Розы

1. Купив пачку мэндэмс я насчитал в ней 1 жёлтую, 7 зелёных, 4 оранжевых, 3 коричневых, 2 синих и 1 красную мэндэмсину. С помощью теста отношения правдоподобия проверьте гипотезу, что мэндэмсины всех цветов встречаются равновероятно.

2. [R] Фактическое распределение часовой и десятиминутной скорости ветра хорошо приближается распределением Вейбулла. Случайная величина имеет распределение Вейбулла, если её функция плотности при $x > 0$ имеет вид

   $$f(x) = \frac{1}{\lambda^k} k x^{k-1} \exp(-x^k/\lambda^k)$$

   (a) Найдите функцию распределения $F(x)$

   (b) Выразите медиану распределение Вейбулла, m, через параметры k и λ

   (c) Оцените параметры k и λ методом максимального правдоподобия

   (d) Постройте 95%-ые доверительные интервалы для k и λ

   (e) Выпишите функцию плотности распределения Вейбулла через m и k

   (f) Проверьте гипотезу о том, что медиана равна 1 м/сек с помощью трёх тестов

   Часовые данные я не нашёл, нашёл дневные. Данные по среднедневной скорости ветра содержатся в weather_nov_2012_moskow.csv в стобике wind. Данные взяты с сайта http://www.atlas-yakutia.ru/weather/climate_russia-I.html.

4.5. Праздник 5, 01.04.2013, Гетероскедастичность

С 1-м апреля!!!

1. Рождается старичком, умирает младенцем, сегодня празднует день рождения, но не Гоголь. Кто это? Опишите внешний вид, характер, или нарисуйте его :)

2. Для борьбы с гетероскедастичностью в модели $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ исследователь перешёл к модели $\tilde{y}_i = \beta_1 \frac{1}{z_i} + \beta_2 \tilde{x}_i + \tilde{\varepsilon}_i$, где $\tilde{x}_i = x_i/z_i$, $\tilde{y}_i = y_i/z_i$, $\tilde{\varepsilon}_i = \varepsilon_i/z_i$.

   Какой вид гетероскедастичности предполагался?

3. Василий Аспушкин провёл два разных теста на гетероскедастичность на одном уровне значимости. Оказалось, что в одном из них H_0 отвергается, а в другом — нет.

   (a) Почему это могло случиться?

   (b) Какой же вывод о гетероскедастичности следует сделать Василию? Что можно сказать об уровне значимости предложенного Вами способа сделать вывод?

4. Писатель Василий Аспушкин пишет Большой Роман. Количество страниц, которое он пишет ежедневно, зависит от количества съеденных пирожков, выпитого лимонада и числа посещений Музы.

$$Stranitsi_i = \beta_1 + \beta_2 Pirojki_i + \beta_3 Limonad_i + \beta_4 Musa_i + \varepsilon_i$$

   Когда идёт дождь, Василий Аспушкин очень волнуется: он ошибочно считает, что музы плохо летают в дождь. Поэтому в дождливые дни дисперсия ε_i может быть выше.

   (a) Отсортировав имеющиеся наблюдения по количеству осадков в день, Настойчивый издатель построил регрессию по 40 самым дождливым дням и получил $RSS = \sum_i (y_i - \hat{y}_i)^2 = 360$. В регрессии по 40 самым сухим дням $RSS = 252$. Всего имеется 100 наблюдений. Проверьте гипотезу о гомоскедастичности. Как называется соответствующий тест?

   (b) Василий Аспушкин оценил по 100 наблюдениям исходную модель с помощью МНК. А затем построил регрессию квадратов стьюдентизированных остатков на количество осадков и константу. Во второй регрессии $R^2 = 0.3$. Проверьте гипотезу о гомоскедастичности.

   (c) Предположим, что дисперсия ошибок линейно зависит от количества осадков.

      i. Как будет выглядеть функция максимального правдоподобия для оценивания коэффициентов исходной модели?

      ii. Опишите процедуру доступного обобщенного метода наименьших квадратов (FGLS, feasible generalized least squares) применительно к данной ситуации

   Hint: Функция плотности одномерного нормального распределения имеет вид

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

5. В курсе теории вероятностей изучался тест о равенстве математических ожиданий по двум нормальным выборкам при предпосылке о равенстве дисперсий. Предложите состоятельный способ тестировать гипотезу о равенстве математических ожиданий без предпосылки равенства дисперсий.

4.6. Домашнее задание 3. Знакомство с RLMS

1. Прочитайте про RLMS, http://www.hse.ru/rlms/

   Посмотрите описание проекта. Пролистайте вестник RLMS, чтобы иметь представление о том, какие исследования можно строить на основе RLMS.

2. Скачайте любую волну RLMS по своему выбору. Скачайте описание переменных.

   Пролистайте описание переменных. Там их больше тысячи. Попадаются довольно прикольные. Мне нравится pc9.6.5a, «У Вас есть GPRS навигатор?»

3. Загрузите данные в R.

   Данные RLMS выложены на сайте в формате SPSS. SPSS это потихоньку погибающий статистический пакет для домохозяек. Для чтения формата .sav в таблицу данных R можно сделать так

   Первая команда, library(foreign), подгружает библиотеку R, в которой содержатся команды для чтения вражеских форматов, spss, stata, etc

   Описания переменных при этом также загружаются в таблицу данных. Можно их выделить в отдельный вектор и прочитать, например, про переменную pc9.631a.

4. Выберите любую количественную переменную в качестве зависимой и несколько переменных в качестве объясняющей.

   Цель этой домашки скорее ознакомится с наличием мониторинга RLMS, поэтому можно не сильно заморачиваться с этим этапом. Хотя в реальности тут-то всё самое интересное и начинается. За оригинальные гипотезы будут плюшки.

5. Опишите выбранные переменные.

   Постройте симпатичные графики. Посчитайте описательные статистики. Много ли пропущенных наблюдений? Есть ли что-нибудь интересненькое?

6. Постройте регрессию зависимой переменной на объясняющие.

   Проверьте гипотезу о значимости каждого полученного коэффициента. Проверьте гипотезу о значимости регрессии в целом. Для нескольких коэффициентов (двух достаточно) постройте 95%-ый доверительный интервал.

7. Напишите свои пожелания и комментарии.

   Какие домашки хочется сделать? Что не ясно в курсе эконометрики? Содержательные комментарии позволяют получить бонус. Искусная лесть оценивается :)

4.7. Домашнее задание №$(n+1)$ по эконометрике-1.

Задача 1. «CAPM»
Оценим модель CAPM по реальным данным:

1. Коротко сформулируйте теоретические положения модели CAPM. За корректное отделение выводов от предпосылок — дополнительный бонус.

2. Соберите реальные данные по трём показателям: R_i — доходность некоей акции за i-ый период, $R_{m,i}$ — рыночная доходность за i-ый период, $R_{f,i}$ — безрисковая доходность за i-ый период. Статья quantile.ru/06/06-AT.pdf в помощь.

3. Представьте информацию графически

4. С помощью МНК оцените модель без константы, $R_i - R_{f,i} = \beta(R_{m,i} - R_{f,i}) + \varepsilon_i$. Предположим, что $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

5. Прокомментируйте результаты оценивания. В частности, проверьте гипотезы о значимости коэффициента и регрессии в целом.

6. С помощью МНК оцените модель с константой, $R_i - R_{f,i} = \beta_1 + \beta_2(R_{m,i} - R_{f,i}) + \varepsilon_i$. Предположим, что $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

7. Прокомментируйте результаты оценивания. В частности, проверьте гипотезы о значимости коэффициентов и регрессии в целом.

8. Труднее всего измерить безрисковую ставку процента. Поэтому предположим, что имеющиеся у нас наблюдения — это безрисковая ставка, измеренная с ошибкой. Т.е. имеющиеся у нас наблюдения $R_{f,i}$ представимы в виде $R_{f,i} = R_{f,i}^{true} + u_i$, где $u_i \sim N(0, \sigma_u^2)$. Величина $R_{f,i}^{true}$ ненаблюдаема, но именно она входит в модель CAPM. Получается, что оцениваемая модель имеет вид $R_i - R_{f,i}^{true} = \beta(R_{m,i} - R_{f,i}^{true}) + \varepsilon_i$.

   (a) Выпишите функцию правдоподобия для оценки данной модели

   (b) Найдите оценки $\hat{\beta}$, $\hat{s}2_u$, $\hat{s}2_\varepsilon$

   (c) Постройте 95%-ые доверительные интервалы

   (d) Проделайте аналогичные действия для модели с константой

   (e) Сделайте выводы

Задача 2. «Цифи́рьки на мониторе»
При входе на каждую станцию метро есть турникеты. Рядом с турникетами в будке сидит бабушка божий одуванчик. В будке у бабушки висит монитор. На этом мониторе — прямоугольники с цифи́рьками.

1. Понаблюдав за изменением цифи́рек, догадайтесь, что они означают.

2. Вечером какого-нибудь буднего дня запишите все цифирьки с монитора на своей родной станции метро.

3. Представьте информацию графически

4. Будем моделировать величину i-ой цифи́рьки пуассоновским распределением с математическим ожиданием λ_i. Предположим также, что $\lambda_i = \beta_1 + \beta_2 \cdot i$, где i — номер турникета считая от будки с бабушкой.

   (a) Выпишете функцию правдоподобия

   (b) Оцените параметры β_1 и β_2

   (c) Оцените ковариационную матрицу оценок $\hat{\beta}_1$ и $\hat{\beta}_2$

   (d) Постройте 95%-ые асимптотические доверительные интервалы для параметров

   (e) Проверьте гипотезу о том, что $\beta_2 = 0$. Альтернативную гипотезу сформулируйте самостоятельно.

PS. Своё смелое творчество в задачах поощряется!

4.8. Домашнее задание. Титаник.

Нужно зарегистрироваться на сайте www.kaggle.com и принять участие в конкурсе «Titanic: Machine Learning from Disaster». Крайний срок сдачи отчёта: в ночь с 14 на 15 апреля 2013 года.

Версия Белой Розы

1. Домашнее задание можно делать в одиночку или группой из двух человек.

2. А можно всё-таки группой из трёх человек? Нет :)

3. Письменный отчёт должен содержать как-минимум:

   (a) Логин группы

   (b) Графический анализ имеющихся данных

   (c) Результаты оценивания logit и probit моделей

   (d) Графический анализ logit и probit моделей

   (e) «Если бы я был пассажиром Титаника, то я спасся бы с вероятностью...».
   С помощью logit и probit моделей необходимо построить 95%-ый доверительный интервал для вероятности спасения каждого из участников группы, сдающей домашку. Пол и возраст взять фактические, а остальные объясняющие переменные — по своему желанию.

Версия Алой Розы

1. Домашнее задание можно делать только в одиночку :)

2. Нет, нельзя :)

3. Письменный отчёт должен содержать как-минимум:

   (a) Логин

   (b) Графический анализ имеющихся данных

   (c) Результаты оценивания logit и probit моделей

   (d) Прогнозирование с использованием Random Forest

   (e) Прогнозирование с использованием метода опорных векторов (SVM)

   (f) Графический анализ оценённых моделей

   (g) «Если бы я был пассажиром Титаника, то я спасся бы с вероятностью...».
   С помощью логит и пробит моделей необходимо построить 95%-ый доверительный интервал для своей вероятности спасения. Для Random Forest требуется только точечная оценка вероятности спасения. Пол и возраст взять фактические, а остальные объясняющие переменные — по своему желанию.

5. 2013-2014

5.1. Праздник 1. Вперед в рукопашную!

1. Найдите длины векторов $a = (2, 1, 1)$ и $b = (-2, 0, 1)$ и косинус угла между ними.

2. Сформулируйте теорему о трёх перпендикулярах

3. Для матрицы

$$A = \begin{pmatrix} -2 & 0 & 0 \\ 0 & 3 & 4 \\ 0 & 4 & 9 \end{pmatrix}$$

   (a) Найдите собственные числа и собственные векторы матрицы.

   (b) Найдите обратную матрицу, A^{-1}, ее собственные векторы и собственные числа.

   (c) Представьте матрицу A в виде $A = CDC^{-1}$, где D — диагональная матрица.

   (d) Представьте A^{2013} в виде произведения трёх матриц.

4. Матрицы A и B таковы, что $\det(AB)$, $\det(BA)$, $\mathrm{tr}(AB)$ и $\mathrm{tr}(BA)$ определены. Возможно ли что $\det(AB) \neq \det(BA)$? Возможно ли, что $\mathrm{tr}(AB) \neq \mathrm{tr}(BA)$? Если неравенство возможно, то приведите пример.

5. Вася и Петя независимо друг от друга решают тест по теории вероятностей. В тесте всего два вопроса. На каждый вопрос два варианта ответа. Петя знает решение каждого вопроса с вероятностью $0{,}4$. Если Петя не знает решения, то он отвечает равновероятно наугад. Вася знает решение каждого вопроса с вероятностью $0{,}7$. Если Вася не знает решения, то он отвечает равновероятно наугад.

   (a) Какова вероятность того, что Петя правильно ответил на оба вопроса?

   (b) Какова вероятность того, что Петя правильно ответил на оба вопроса, если его ответы совпали с Васиными?

   (c) Чему равно математическое ожидание числа Петиных верных ответов?

   (d) Чему равно математическое ожидание числа Петиных верных ответов, если его ответы совпали с Васиными?

6. Для случайных величин X и Y заданы следующие значения: $\mathbb{E}(X) = 1, \mathbb{E}(Y) = 4, \mathbb{E}(XY) = 8$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 9$. Для случайных величин $U = X + Y$ и $V = X - Y$ вычислите:

   (a) $\mathbb{E}(U)$, $\mathrm{Var}(U)$, $\mathbb{E}(V)$, $\mathrm{Var}(V)$, $\mathrm{Cov}(U, V)$

   (b) Можно ли утверждать, что случайные величины U и V независимы?

7. Вася ведёт блог. Обозначим X_i — количество слов в i–ой записи. После первого года он по своим записям обнаружил, что $\bar{X}_{200} = 95$ и выборочное стандартное отклонение равно 282 слова. На уровне значимости $\alpha = 0.10$ проверьте гипотезу о том, что $\mu = 100$ против альтернативной гипотезы $\mu \neq 100$. Найдите также точное P-значение.

5.2. Праздник 2. Мегаматрица

В рамках классической линейной модели с детерминистическими регрессорами найдите $\mathrm{Var}(\hat{\beta})$, $\mathrm{Cov}(\hat{\varepsilon}, \hat{\beta})$, $\mathrm{Cov}(\hat{\varepsilon}, \hat{y})$.

5.3. Праздник 3. Базовая задача

Пусть регрессионная модель $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$, $i = 1, \ldots, n$, задана в матричном виде при помощи уравнения $y = X\beta + \varepsilon$, где $\beta = \begin{pmatrix} \beta_1 & \beta_2 & \beta_3 \end{pmatrix}^T$. Известно, что ошибки ε нормально распределены с $\mathbb{E}\varepsilon = 0$ и $\mathrm{Var}(\varepsilon) = \sigma^2 \cdot I$. Известно также, что:

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}$$

Для удобства расчётов ниже приведены матрицы:

$X^T X = \begin{pmatrix} 5 & 3 & 1 \\ 3 & 3 & 1 \\ 1 & 1 & 1 \end{pmatrix}$ и $(X^T X)^{-1} = \begin{pmatrix} 0.5 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 1.5 \end{pmatrix}$.

1. Оценки $\hat{\beta}$
2. Спрогнозируйте y, если $x_2 = 1$ и $x_3 = -2$
3. TSS, ESS, RSS, R^2
4. $\mathbb{E}(\hat{\sigma}^2)$
5. $\hat{\sigma}^2$
6. $\mathrm{Var}(\varepsilon_1)$
7. $\mathrm{Var}(\beta_1)$
8. $\mathrm{Var}(\hat{\beta}_1)$
9. $\widehat{\mathrm{Var}}(\hat{\beta}_1)$
10. $\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_3)$
11. $\widehat{\mathrm{Cov}}(\hat{\beta}_2, \hat{\beta}_3)$
12. $\mathrm{Var}(\hat{\beta}_2 - \hat{\beta}_3)$
13. $\widehat{\mathrm{Var}}(\hat{\beta}_2 - \hat{\beta}_3)$
14. $\mathrm{Var}(\beta_2 - \beta_3)$
15. Проверьте гипотезу H_0: $\beta_1 = 1$ против гипотезы H_a: $\beta_1 \neq 1$ на уровне значимости 5%
16. Проверьте гипотезу H_0: $\beta_2 = 0$ против гипотезы H_a: $\beta_2 \neq 0$ на уровне значимости 10%
17. Проверьте гипотезу H_0: $\beta_2 = \beta_3$ против гипотезы H_a: $\beta_2 \neq \beta_3$ на уровне значимости 5%

5.4. Праздник 4

1. Пусть $y = X\beta + \varepsilon$ — регрессионная модель, где $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$. Пусть $Z = XD$, где $D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$. Рассмотрите «новую» регрессионную модель $y = Z\alpha + u$, где $\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}$. Определите, как выражаются «новые» МНК-коэффициенты через «старые».

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

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

3. 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

4. Рассмотрим модель: $y_i = \beta_1 + \beta_2 x_{1i} + \beta_3 x_{2i} + \beta_4 x_{3i} + \beta_5 x_{4i} + \varepsilon_i$. По 20 наблюдениям оценены следующие регрессии:

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

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

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

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

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

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

5.5. Праздник 5. Максимальное правдоподобие

1. Случайные величины $X_1, \ldots, X_n$ — независимы и одинаково распределены с функцией плотности $f(t) = \frac{\theta \cdot (\ln t)^{\theta - 1}}{t}$ при $t \in [1; e]$. По выборке из 100 наблюдений оказалось, что $\sum \ln(\ln(X_i)) = -30$

(a) Найдите ML оценку параметра θ

(b) Постройте 95% доверительный интервал для θ

(c) С помощью LR, LM и W теста проверьте гипотезу о том, что $\theta = 1$.

2. Величины X_1, ..., X_n — независимы и нормально распределены, $N(\mu, \sigma^2)$. По 100 наблюдениям $\sum X_i = 100$ и $\sum X_i^2 = 900$.

   (a) Найдите ML оценки неизвестных параметров μ и σ^2.

   (b) Постройте 95%-ые доверительные интервалы для μ и σ^2

   (c) С помощью LR, LM и W теста проверьте гипотезу о том, что $\sigma^2 = 1$.

   (d) С помощью LR, LM и W теста проверьте гипотезу о том, что $\sigma^2 = 1$ и одновременно $\mu = 2$.

Всех участников правдоподобной контрольной с древнерусским эконометрическим праздником!

Сегодня **Аксинья-полухлебница**.

«На Аксинью гадали о ценах на хлеб в ближайшее время и на будущий урожай: брали печёный хлеб и взвешивали его сначала вечером, а потом утром. Коли вес оставался неизменным — цена на хлеб не изменится. Если за ночь вес уменьшался — значит, хлеб подешевеет, а если увеличивался, то подорожает»

Wikipedia

5.6. Переписывание кр 5. Максимальное правдоподобие

1. По совету Лисы Волк опустил в прорубь хвост и поймал 100 чудо-рыб. Веса рыбин независимы и имеют распределение Вейбулла, $f(x) = 2\exp(-x^2/a^2) \cdot x/a^2$ при $x \geqslant 0$. Известно, что $\sum x_i^2 = 120$.

   (a) Найдите ML оценку параметра a

   (b) Постройте 95% доверительный интервал для a

   (c) С помощью LR, LM и W теста проверьте гипотезу о том, что $a = 1$.

2. Как известно, Фрекен-Бок пьет коньяк по утрам и иногда видит привидения. За 110 дней имеются следующие статистические данные

Рюмок	1	2	3
Дней с привидениями	10	25	20
Дней без привидений	20	25	10

   Вероятность увидеть привидение зависит от того, сколько рюмок коньяка было выпито утром, а именно, $p = \exp(a+bx)/(1+\exp(a+bx))$, где x — количество рюмок, а a и b — неизвестные параметры.

   (a) Найдите[1] ML оценки неизвестных параметров a и b.

   (b) Постройте 95%-ые доверительные интервалы для a и b

   (c) С помощью LR, LM и W теста проверьте гипотезу о том, что $b = 0$.

   (d) С помощью LR, LM и W теста проверьте гипотезу о том, что $a = 0$ и одновременно $b = 0$.

[1]Здесь потребуется максимизировать функцию в R. Если этот пункт не получился, то в последующих пунктах можно считать, что $\hat{a} = -1.5$, а $\hat{b} = 0.5$. Это сильно округленные значения коэффициентов.

Всем участникам переписывания правдоподобной контрольной счастья! Много!

Сегодня, 20 марта, **Международный День счастья**.

5.7. Праздник 6. Гетероскедастичность

1. Желая протестировать наличие гетероскедастичности в модели $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \beta_4 w_i + \varepsilon_i$, эконометресса Глафира решила провести тест Уайта и получила во вспомогательной регрессии $R^2 = 0.50$. Глафира строит модель удоя по 200 коровам. Помогите ей провести тест на уровне значимости 5%.

2. На всякий случай эконометресса Глафира решила подстраховаться и провести тест Голдфельда-Квандта. Но она совсем забыла, как его делать. Напомните Глафире, как провести тест Голдфельда-Квандта, если она подозревает, что дисперсия $Var(\varepsilon_i)$ возрастает с ростом z_i. Чётко напишите гипотезы H_0, H_a, методику проведения теста, правило согласно которому отвергается или не отвергается H_0.

3. Имеются три наблюдения, $x = (1, 2, 2)'$, $y = (2, 1, 0)'$. Предполагая, что в модели $y_i = \beta x_i + \varepsilon_i$ имеется гетероскедастичность вида $Var(\varepsilon_i) = \sigma^2 x_i^4$ найдите:

   (a) Обычную МНК-оценку параметра β

   (b) Самую эффективную среди несмещенных оценку параметра β

   (c) Во сколько раз отличается истинная дисперсия этих двух оценок?

   (d) Во сколько раз отличаются оценки дисперсий этих оценок, если дисперсии оценивается без поправки на гетероскедастичность в обоих случаях?

5.8. Большой Устный ЗАчёт

1. Метод Наименьших Квадратов.

   (a) МНК-картинка

   (b) Нахождение всего-всего, если известен вектор y и матрица X

2. Теорема Гаусса-Маркова

   (a) Формулировка с детерминистическими регрессорами

   (b) Доказательство с детерминистическими регрессорами

   (c) Формулировки со стохастическими регрессорами

   (d) Что даёт дополнительное предположение о нормальности ε?

3. Проверка гипотез о линейных ограничениях

   (a) Проверка гипотезы о значимости коэффициента

   (b) Проверка гипотезы о значимости регрессии в целом

   (c) Проверка гипотезы об одном линейном соотношении с помощью ковариационной матрицы

   (d) Ограниченная и неограниченная модель

   (e) Тест Чоу на стабильность коэффициентов

   (f) Тест Чоу на прогнозную силу

4. Метод максимального правдоподобия
   (a) Свойства оценок
   (b) Два способа получения оценки дисперсии
   (c) Три теста (LM, Wald, LR)
   (d) Выписать функцию ML для обычной регрессии
   (e) для AR(1) процесса
   (f) для MA(1) процесса
   (g) для логит модели
   (h) для пробит модели
   (i) для модели с заданным видом гетероскедастичности
5. Мультиколлинеарность
   (a) Определение, последствия
   (b) Величины, измеряющие силу мультиколлинеарности
   (c) Методы борьбы
   (d) Сюда же: метод главных компонент, хотя он используется и для других целей
6. Гетероскедастичность
   (a) Определение, последствия
   (b) Тесты, график
   (c) Стьюдентизированные остатки
   (d) HC оценки ковариации
   (e) GLS и FGLS
7. Временные ряды
   (a) Стационарный временной ряд
   (b) ACF, PACF
   (c) Модель ARMA
   (d) Модель GARCH (не будет, не успели)
8. Логит и пробит
   (a) Описание моделей
   (b) Предельные эффекты
   (c) Чувствительность, специфичность
   (d) Кривая ROC
9. Эндогенность
   (a) Три примера: одновременность, пропущенные переменные, ошибки измерения
   (b) IV, двухшаговый МНК
10. Модели панельных данных

(a) RE, FE, сквозная регрессии

(b) Тест Хаусмана

11. Альтернативные методы. Уметь объяснить суть метода. Уметь реализовать его в R.

    (a) Метод опорных векторов (не будет, не успели)

    (b) Классификационные деревья и случайный лес

12. R. Можно принести файл со своей заготовкой, можно пользоваться Интернетом для поиска информации, но не для общения.

    (a) Загрузить данные из .csv файла в R

    (b) Посчитать описательные статистики (среднее, мода, медиана и т.д.)

    (c) Построить подходящие описательные графики для переменных

    (d) Оценить линейную регрессию с помощью МНК. Провести диагностику на что-нибудь (гетероскедастичность, автокорреляцию, мультиколлинеарность).

    (e) Оценить logit, probit модели, посчитать предельные эффекты

    (f) Оценить ARMA модель

    (g) Выделить главные компоненты

5.9. Экзамен.

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

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

   Для удобства расчетов приведены матрицы

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

   (a) Найдите вектор МНК-оценок коэффициентов $\hat{\beta}$.

   (b) Найдите несмещенную оценку для неизвестного параметра σ^2.

   (c) Проверьте гипотезу $\beta_2 = 0$ против альтернативной о неравенстве на уровне значимости 5%

2. По данным о пассажирах Титаника оценивается логит-модель. Зависимая переменная survived равна 1, если пассажир выжил. Объясняющая переменная sexmale равна 1 для мужчин.

   (a) Оцените вероятность выжить для женщины 20 лет

   (b) Оцените предельный эффект увеличения возраста для женщины 20 лет

   (c) С помощью какого метода оценивается логит-модель? Каким образом при этом получаются оценки стандартных ошибок коэффициентов?

3. Теорема Гаусса-Маркова.

	Model 1
(Intercept)	1.92^{***}
	(0.28)
age	-0.01
	(0.01)
sexmale	-2.84^{***}
	(0.21)
AIC	633.45
BIC	646.80
Log Likelihood	-313.72
Deviance	627.45
Num. obs.	633

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Таблица 2: Statistical models

(a) Аккуратно сформулируйте теорему Гаусса-Маркова для нестохастических регрессоров.

(b) Поясните каждое из свойств оценок, фигурирующих в теореме.

(c) Как меняются свойства оценок МНК при нарушении предпосылки теоремы о том, что дисперсия ε_i постоянна?

4. Рассмотрим временной ряд, описываемый MA(2) моделью,

$$y_t = \gamma + \varepsilon_t + \alpha_1 \varepsilon_{t-1} + \alpha_2 \varepsilon_{t-2},$$

где ε_t — белый шум с $\mathrm{Var}(\varepsilon_t) = \sigma^2$.

(a) Является ли данный процесс стационарным? Что такое стационарный процесс?

(b) Найдите автокорреляционную функцию данного процесса, $\rho(k) = \mathrm{Corr}(y_t, y_{t-k})$.

(c) Выпишите функцию правдоподобия для данной модели в предположении нормальности ε_t.

5. Рассмотрите модель $y_i = \beta x_i + \varepsilon_i$. Предположим, что все предпосылки классической линейной регрессионной модели выполнены. Модель оценивается с помощью МНК и получается оценка $\hat{\beta}_{OLS}$. В условиях мультиколлинеарности для снижения дисперсии оценки $\hat{\beta}$ можно применять ряд методов, например, алгоритм «ridge regression». Он состоит в том, что при некотором фиксированном $\lambda \geqslant 0$ минимизируется по $\hat{\beta}$ величина

$$Q(\hat{\beta}) = \sum_i (y_i - \hat{\beta} x_i)^2 + \lambda \hat{\beta}^2$$

(a) Как выглядит МНК оценка $\hat{\beta}_{OLS}$?

(b) Как выглядит оценка методом «ridge regression», $\hat{\beta}_{RR}$?

(c) Верно ли, что оценка $\hat{\beta}_{RR}$ является несмещенной только при $\lambda = 0$?

(d) (*) Верно ли, что всегда найдется такое λ, что среднеквадратичная ошибка оценки $\hat{\beta}_{RR}$ будет меньше, т.е. $\mathbb{E}((\hat{\beta}_{RR} - \beta)^2) < \mathbb{E}((\hat{\beta}_{OLS} - \beta)^2)$?

5.10. Пересдача экзамена

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

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

   Для удобства расчетов приведены матрицы

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

   (a) Найдите вектор МНК-оценок коэффициентов $\hat{\beta}$.

   (b) Найдите несмещенную оценку для неизвестного параметра σ^2.

   (c) Проверьте гипотезу $\beta_2 = 0$ против альтернативной о неравенстве на уровне значимости 5%

2. По данным о пассажирах Титаника оценивается логит-модель. Зависимая переменная survived равна 1, если пассажир выжил. Объясняющая переменная sexmale равна 1 для мужчин.

	Model 1
(Intercept)	1.92^{***}
	(0.28)
age	-0.01
	(0.01)
sexmale	-2.84^{***}
	(0.21)
AIC	633.45
BIC	646.80
Log Likelihood	-313.72
Deviance	627.45
Num. obs.	633

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Таблица 3: Statistical models

   (a) Оцените вероятность выжить для женщины 20 лет

   (b) Оцените предельный эффект увеличения возраста для женщины 20 лет

   (c) С помощью какого метода оценивается логит-модель? Каким образом при этом получаются оценки стандартных ошибок коэффициентов?

3. Теорема Гаусса-Маркова.

   (a) Аккуратно сформулируйте теорему Гаусса-Маркова для нестохастических регрессоров.

   (b) Поясните каждое из свойств оценок, фигурирующих в теореме.

(c) Как меняются свойства оценок МНК при нарушении предпосылки теоремы о том, что дисперсия ε_i постоянна?

4. 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.

(a) Предполагая гомоскедастичность остатков на уровне значимости 5% проверьте гипотезу, что исследуемая зависимость одинакова на всех трёх частях всей выборки.

(b) Протестируйте ошибки на гетероскедастичность на уровне значимости 5%.

(c) Какой тест можно на гетероскедастичность можно было бы использовать, если бы не было уверенности в нормальности остатков? Опишите пошагово процедуру этого теста.

5.11. Домашняя работа 1. RLMS и гетероскедастичность

1. Прочитайте про RLMS, http://www.hse.ru/rlms/

   Посмотрите описание проекта. Пролистайте вестник RLMS, чтобы иметь представление о том, какие исследования можно строить на основе RLMS.

2. Скачайте любую волну RLMS по своему выбору. Скачайте описание переменных.

   Пролистайте описание переменных. Там их больше тысячи. Попадаются довольно прикольные. Мне нравится pc9.6.5a, «У Вас есть GPRS навигатор?»

3. Загрузите данные в R.

   Данные RLMS выложены на сайте в формате SPSS. SPSS это потихоньку погибающий статистический пакет для домохозяек. Для чтения формата .sav в таблицу данных R можно сделать так

```
library(foreign)
file.name <- "/home/boris/downloads/r20hall23c.sav"
h <- read.spss(file.name, to.data.frame = TRUE)
```

   Первая команда, library(foreign), подгружает библиотеку R, в которой содержатся команды для чтения вражеских форматов, spss, stata, etc

   Описания переменных при этом также загружаются в таблицу данных. Можно их выделить в отдельный вектор и прочитать, например, про переменную pc9.631a.

```
var.labels <- attr(h,"variable.labels")
var.labels["pc9.631a"]
```

4. Выберите любую количественную переменную в качестве зависимой и несколько переменных в качестве объясняющей.

   Цель этой домашки скорее ознакомится с наличием мониторинга RLMS, поэтому можно не сильно заморачиваться с этим этапом. Хотя в реальности тут-то всё самое интересное и начинается. За оригинальные гипотезы будут плюшки. Кстати, неплохо бы дать выбранным переменным понятные названия.

5. Опишите выбранные переменные.

   Постройте симпатичные графики. Посчитайте описательные статистики. Много ли пропущенных наблюдений? Есть ли что-нибудь интересненькое?

6. Постройте регрессию зависимой переменной на объясняющие.

   Проверьте гипотезу о значимости каждого полученного коэффициента. Проверьте гипотезу о значимости регрессии в целом. Для нескольких коэффициентов (двух достаточно) постройте 95%-ый доверительный интервал.

7. Разберитесь с возможным наличием гетероскедастичности в данных.

   С какой переменной может быть связана дисперсия $\mathrm{Var}(\varepsilon_i)$? Проведите визуальный анализ на гетероскедастичность. Проведите формальные тесты на гетероскедастичность. Примените оценки дисперсии $\hat{\beta}$ устойчивые к гетероскедастичности. Прокомментируйте. Может помочь http://bdemeshev.github.io/r_cycle/cycle_files/12_hetero.html

8. Покажите буйство своей фантазии и аккуратность!

   Не стоит думать, что побуквенное выполнение этих инструкций гарантирует оценку в десять баллов. Эконометрика — это не ремесло, а искусство! Фантазируйте! Убедите меня в работе, что вы были на лекциях, даже если это так :) Аккуратность в виде подписанных осей на графиках, указанных единицах измерения также не повредит.

9. Срок сдачи — 27 февраля 2014 года.

   Работа принимается исключительно в печатном виде с применением грамотного программирования R + LaTeX. Каждый день более поздней сдачи умножает оценку за работу на 0.8. Работа должна представлять слитный текст, код скрывать не нужно. В конце должна быть команда sessionInfo().

5.12. Домашняя работа 2. Титаник

1. Зарегистрируйтесь на сайте www.kaggle.com в конкурсе «Titanic: Machine Learning from Disaster». В работе укажите login, использованный при регистрации.

2. Проанализируйте данные графически и с помощью описательных статистик (среднее, мода, медиана и т.д.)

   Прокомментируйте графики, обратите внимание на количество пропущенных значений.

3. Оцените logit и probit модели.

   Приведите оценки моделей. Какие коэффициенты значимы? Прокомментируйте знак коэффициентов. Посчитайте и сравните предельные эффекты.

4. Оцените random forest и SVM модели.

   Параметры методов подберите с помощью кросс-валидации. Можно применять любые другие подходы, не только random forest и SVM. Другой подход следует описать в тексте.

5. «Если бы я был пассажиром Титаника, то я спасся бы с вероятностью...».

   С помощью логит и пробит моделей постройте 95%-ый доверительный интервал для вероятности своего спасения. Для random forest — только точечный прогноз вероятности, для svm — только прогноз типа «да»/«нет».

6. Подумайте, чем можно заполнить пропущенные значения. Заполните пропущенные значения и заново оцените logit, random forest и svm. Насколько сильно меняется качество оцененных моделей?

7. Сравните все использованные подходы по прогнозной силе на тестовой выборке с сайта. Какой оказался наилучшим?

8. При прогнозировании и расчете предельных эффектов используйте свои фактические пол и возраст, а остальные объясняющие переменные — выбирайте согласно своей фантазии :)

9. Срок сдачи — 30 апреля 2014 года.

   Работа принимается исключительно в печатном виде с применением грамотного программирования R + LaTeX. Каждый день более поздней сдачи умножает оценку за работу на 0.8. Работа должна представлять слитный текст, код скрывать не нужно. В конце должна быть команда sessionInfo().

10. Популярные ошибки прошлой домашки будут караться со всей строгостью военного времени!

    Цикл заметок про R в помощь http://bdemeshev.github.io/r_cycle/.

6. 2014-2015

6.1. Праздник номер 1

Вперёд, в рукопашную!

1. Сформулируйте теорему о трёх перпендикулярах и обратную к ней.

2. Для матрицы

   $A = \begin{pmatrix} 3 & 4 \\ 4 & 9 \end{pmatrix}$

   (a) Найдите собственные числа и собственные векторы матрицы.

   (b) Найдите обратную матрицу, A^{-1}, ее собственные векторы и собственные числа.

   (c) Представьте матрицу A в виде $A = CDC^{-1}$, где D — диагональная матрица.

   (d) Найдите A^{42}

   (e) Не находя A^{100} найдите tr(A^{100}) и det(A^{100})

3. Игрок получает случайным образом 13 карт из колоды в 52 карты.

   (a) Какова вероятность, что у него как минимум два туза?

   (b) Каково ожидаемое количество тузов у игрока?

   (c) Какова вероятность, что у него как минимум два туза, если известно, что у него есть хотя бы один туз?

(d) Каково ожидаемое количество тузов у игрока, если известно, что у него на руках хотя бы один туз?

4. В ходе анкетирования 100 сотрудников банка «Омега» ответили на вопрос о том, сколько времени они проводят на работе ежедневно. Среднее выборочное оказалось равно 9.5 часам при выборочном стандартном отклонении 0.5 часа.

   (a) Постройте 95% доверительный интервал для математического ожидания времени проводимого сотрудниками на работе

   (b) Проверьте гипотезу о том, что в среднем люди проводят на работе 10 часов, против альтернативной гипотезы о том, что в среднем люди проводят на работе меньше 10 часов, укажите точное P-значение.

6.2. Праздник номер 2

Паниковать на контрольной строго воспрещается! :)

1. 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

2. В рамках классической линейной модели с неслучайными регрессорами найдите $\mathrm{Var}(\hat{\varepsilon})$, $\mathrm{Cov}(\hat{\beta}, \hat{\varepsilon})$. Верно ли, что $\mathrm{Cov}(\hat{\varepsilon}_1, \hat{\varepsilon}_2) = 0$?

3. Эконометресса Ефросинья оценивала модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$. Найдя матрицы $X'X$ и $(X'X)^{-1}$, она призадумалась...

$$X'X = \begin{bmatrix} 47 & 775 & 1934 \\ 775 & 15707 & 23121 \\ 1934 & 23121 & 159570 \end{bmatrix}, (X'X)^{-1} = \begin{bmatrix} 0.26653 & -0.01067 & -0.00168 \\ -0.01067 & 0.00051 & 0.00006 \\ -0.00168 & 0.00006 & 0.00002 \end{bmatrix}$$

   (a) Помогите Ефросинье найти количество наблюдений, $\bar{z}$, $\sum x_i z_i$, $\sum (x_i - \bar{x})(z_i - \bar{z})$

   (b) (*) Ефросинья решила зачем-то также оценить модель $x_i = \gamma_1 + \gamma_2 z_i + u_i$. Как она может найти RSS в новой модели в одно арифметическое действие?

4. 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

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 2 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

Для удобства расчетов приведены матрицы

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

(a) Найдите вектор МНК-оценок коэффициентов $\hat{\beta}$.

(b) Найдите несмещенную оценку для неизвестного параметра σ^2.

(c) Проверьте гипотезу $\beta_2 = 0$ против альтернативной о неравенстве на уровне значимости 5%

6.3. Праздник номер 3

Примечание: во всех задачах, если явно не сказано обратное, предполагается, что выполнены стандартные предпосылки классической линейной регрессионной модели.

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

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

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

   (a) Напишите спецификацию регрессии с ограничениями для проверки статистической гипотезы $H_0 : \beta_5 = 2\beta_4$

   (b) Дайте интерпретацию проверяемой гипотезе

   (c) Для регрессии с ограничением был вычислен коэффициент $R^2_R = 0.296$. На уровне значимости 5% проверьте нулевую гипотезу

2. По ежегодным данным с 2002 по 2009 год оценивался тренд в динамике общей стоимости экспорта из РФ: $Exp_t = \beta_1 + \beta_2 t + \varepsilon_t$, где t — год ($t = 0$ для 2002 г., $t = 1$ для 2003 г., ..., $t = 7$ для 2009 г.), Exp_t — стоимость экспорта из РФ во все страны в млрд. долл. Оценённое уравнение выглядит так: $\widehat{Exp}_t = 111.9 + 43.2t$. Получены также оценки дисперсии случайной ошибки $\hat{\sigma}^2 = 4009$ и ковариационной матрицы оценок коэффициентов:

$$\widehat{Var}(\hat{\beta}) = \begin{pmatrix} 1671 & -334 \\ -334 & 95 \end{pmatrix}$$

   (a) Постройте 95%-ый доверительный интервал для коэффициента β_2

   (b) Спрогнозируйте стоимость экспорта на 2010 год и постройте 90%-ый предиктивный интервал для прогноза.

3. Имеется 100 наблюдений. Исследователь Вениамин предполагает, что дисперсия случайной ошибки в последних 50-ти наблюдениях в 4 раза выше, чем в первых 50-ти, в частности $\mathrm{Var}(\varepsilon_1) = \sigma^2$, а $\mathrm{Var}(\varepsilon_{100}) = 4\sigma^2$. Вениамин оценивает модель $y_i = \beta x_i + \varepsilon_i$ с помощью МНК.

   (a) Найдите истинную дисперсию МНК оценки коэффициента β

(b) Предложите более эффективную оценку $\hat{\beta}^{alt}$

(c) Чему равна истинная дисперсия новой оценки?

(d) Подробно опишите любой способ, который позволяет протестировать гипотезу о гомоскедастичности против предположения Вениамина о дисперсии.

4. Закон больших чисел гласит, что если z_i независимы и одинаково распределены, то $\operatorname{plim} \bar{z}_n = \mathbb{E}(z_1)$. Предположим, что регрессоры — стохастические, а именно, наблюдения являются случайной выборкой (то есть отдельные наблюдения независимы и одинаково распределены), и $\mathbb{E}(\varepsilon|X) = 0$. Модель имеет вид:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 w_i + \varepsilon_i$$

(a) Найдите $\mathbb{E}(\varepsilon)$, $\mathbb{E}(x_1 \cdot \varepsilon_1)$

(b) Найдите $\operatorname{plim} \frac{1}{n} X'\varepsilon$

(c) Найдите $\operatorname{plim} \frac{1}{n} X'X$

(d) Докажите, что вектор МНК оценок $\hat{\beta}$ является состоятельным

5. Эконометресса Эвридика хочет оценить модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$. К сожалению, она измеряет зависимую переменную с ошибкой. Т.е. вместо y_i она знает значение $y_i^* = y_i + u_i$ и использует его в качестве зависимой переменной при оценке регрессии. Ошибки измерения u_i некоррелированы между собой и с ε_i, имеют нулевое математическое ожидание и постоянную дисперсию σ_u^2.

(a) Будут ли оценки Эвридики несмещенными?

(b) Могут ли дисперсии оценок Эвридики быть ниже чем дисперсии МНК оценок при использовании настоящего y_i?

(c) Могут ли оценки дисперсий оценок Эвридики быть ниже чем оценок дисперсий МНК оценок при использовании настоящего y_i?

6.4. Миникр

Миникр 5

1. Как проверить гипотезу об одновременной незначимости всех коэффициентов регрессии кроме свободного члена? Укажите H_0, H_a, тестовую статистику и её распределение при верной H_0.

2. Рассмотрим модель $y = X\beta + \varepsilon$, где n — количество наблюдений, k — количество коэффициентов и ε_i — одинаково распределены и независимы.

(a) Укажите вид матрицы X

(b) Выпишите формулу для МНК оценки $\hat{\beta}$

(c) Выпишите формулу для ковариационной матрицы оценок $\hat{\beta}$

3. Опишите подробно тест Чоу на стабильность коэффициентов по двум наборам данных из n_1 и n_2 наблюдений соответственно. Число оцениваемых коэффициентов равно k.

4. Рассмотрим модель со свободным членом. Как вычисляются R^2 и скорректированный R^2_{adj}? Что может произойти с этими величинами при увеличении количества регрессоров? При уменьшении?

5. Какую гипотезу можно проверить, зная отношение ESS/RSS? Укажите H_0, H_a, тестовую статистику и её распределение при верной H_0.

6. Опишите подробно тест Чоу на прогнозную силу по двум наборам данных из n_1 и n_2 наблюдений соответственно. Число оцениваемых коэффициентов равно k.

6.5. Зачет. Базовый поток

1. Сформулируйте теорему Гаусса-Маркова применительно к модели $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$. Поясните смысл каждого используемого термина.

2. Как проверить гипотезу о том, что эксцесс случайной выборки совпадает с эксцессом нормально распределенной случайной величины? Аккуратно укажите проверяемые H_0, H_a, используемую статистику и её асимптотический закон распределения при верной H_0.

3. Рассмотрим модель спроса на продукцию трёх фирм $y_i = \beta_1 + \beta_2 x_i + \beta_3 d_{i1} + \beta_4 d_{i2} + \varepsilon_i$. Здесь x_i — цена, а y_i — величина спроса.

   Дамми переменные определены следующим образом:

	d_{i1}	d_{i2}
Фирма 1	0	1
Фирма 2	0	0
Фирма 3	1	0

   (a) Как проверить гипотезу, что спрос на продукцию трёх фирм совпадает? Укажите H_0, H_a, тестовую статистику, закон распределения статистики при верной H_0.

   (b) Дамми-переменные d_{i1} и d_{i2} заменяют на d_{i3} и d_{i4}:

	d_{i3}	d_{i4}
Фирма 1	0	0
Фирма 2	1	0
Фирма 3	1	1

   В новых переменных модель имеет вид $y_i = \beta_1' + \beta_2' x_i + \beta_3' d_{i1} + \beta_4' d_{i2} + \varepsilon_i$. Как новые коэффициенты β' выражаются через старые коэффициенты β?

4. Что можно сказать об оценке МНК $\hat{\beta}_2$ в модели $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ при наличии ошибок измерений x_i? А при наличии ошибок измерений y_i?

5. Рассмотрим модель $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, где $\varepsilon_i \sim N(0; \sigma^2)$.

   (a) Напишите выражения для оценок дисперсий и ковариации коэффициентов, т.е. $\widehat{\mathrm{Var}}(\hat{\beta}_1)$, $\widehat{\mathrm{Var}}(\hat{\beta}_2)$, $\widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2)$

   (b) Найдите математическое ожидание и дисперсию каждой из выписанных оценок

   (c) Какой закон распределения с точностью до масштабирования имеют эти оценки?

6. Для модели данных $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$ по 100 наблюдениям получены результаты:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.9681	0.0934	21.07	0.0000
x	2.9063	0.1037	28.02	0.0000
z	-0.1531	0.0968	-1.58	0.1172

(a) Выпишите полученное уравнение регрессии

(b) Укажите, какие коэффициенты значимы при $\alpha = 0.05$

(c) Проверьте H_0: $\beta_2 - \beta_3 = 3$ предполагая, что оценки коэффициентов β_2 и β_3 независимы

7. Рассмотрим модель данных $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, где $\varepsilon_i \sim N(0;\sigma^2)$.

   (a) Выпишите формулу для оценок коэффициентов и оценки дисперсии ошибок

   (b) Укажите математическое ожидание и дисперсию выписанных оценок

   (c) Для оценок коэффициентов укажите закон распределения

   (d) Для оценки дисперсии ошибок укажите закон распределения с точностью до масштабирования

6.6. Зачет, 26.12.2014. Ликвидация безграмотности

В этот день, 26 декабря 1919 года, совнарком РСФСР принял декрет «О ликвидации безграмотности в РСФСР». Всем желаю отметить этот день написанием грамотного зачета по эконометрике! Удачи!

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) Найдите вектор МНК-оценок коэффициентов $\hat{\beta}$.

   (b) Найдите коэффициент детерминации R^2

   (c) Предполагая нормальное распределение вектора ε, проверьте гипотезу H_0: $\beta_2 = 0$ против альтернативной H_a: $\beta_2 \neq 0$

2. 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.

   (a) Предполагая гомоскедастичность остатков на уровне значимости 5% проверьте гипотезу, что исследуемая зависимость одинакова на всех трёх частях всей выборки.

(b) Протестируйте ошибки на гетероскедастичность на уровне значимости 5%.

(c) Какой тест можно на гетероскедастичность можно было бы использовать, если бы не было уверенности в нормальности остатков? Опишите пошагово процедуру этого теста.

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

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

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

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

Оценка стандартной ошибки случайной составляющей, $\hat{\sigma} = 33.0252513$.

(a) Можно ли интерпретировать коэффициент при переменной $totsp$ как стоимость одного метра нежилой площади?

(b) Проверьте гипотезу о том, что коэффициенты при регрессорах $totsp$ и $livesp$ равны.

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

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

4. Аккуратно сформулируйте теорему Гаусса-Маркова

(a) для нестохастических регрессоров

(b) для стохастических регрессоров в предположении, что наблюдения являются случайной выборкой

6.7. Домашняя работа 1. RLMS и гетероскедастичность

1. Прочитайте про RLMS, http://www.hse.ru/rlms/

   Посмотрите описание проекта. Пролистайте вестник RLMS, чтобы иметь представление о том, какие исследования можно строить на основе RLMS.

2. Скачайте любую волну RLMS по своему выбору. Скачайте описание переменных.

   Пролистайте описание переменных. Там их больше тысячи. Попадаются довольно прикольные. Мне нравится pc9.6.5a, «У Вас есть GPRS навигатор?»

3. Загрузите данные в R.

   Данные RLMS выложены на сайте в формате SPSS. SPSS это потихоньку погибающий статистический пакет для домохозяек. Для удобства можно воспользоваться готовой функцией для чтения данных RLMS в пакете rlms.

```
library("rlms")
h <- read.rlms("/home/boris/downloads/r20hall23c.sav")
```

Про установку пакета rlms можно прочитать на страничке https://github.com/bdemeshev/rlms

Описания переменных при этом также загружаются в таблицу данных. Можно их посмотреть:

```
var_meta <- attr(h,"var_meta")
var_meta
```

4. Выберите любую количественную переменную в качестве зависимой и несколько переменных в качестве объясняющих.

   Цель этой домашки скорее ознакомится с наличием мониторинга RLMS, поэтому можно не сильно заморачиваться с этим этапом. Хотя в реальности тут-то всё самое интересное и начинается. За оригинальные гипотезы будут плюшки. Кстати, неплохо бы дать выбранным переменным понятные названия.

5. Опишите выбранные переменные.

   Постройте симпатичные графики. Посчитайте описательные статистики. Много ли пропущенных наблюдений? Есть ли что-нибудь интересненькое?

6. Постройте регрессию зависимой переменной на объясняющие.

   Проверьте гипотезу о значимости каждого полученного коэффициента. Проверьте гипотезу о значимости регрессии в целом. Для нескольких коэффициентов (двух достаточно) постройте 95%-ый доверительный интервал.

7. Разберитесь с возможным наличием гетероскедастичности в данных.

   С какой переменной может быть связана дисперсия $\mathrm{Var}(\varepsilon_i)$? Проведите визуальный анализ на гетероскедастичность. Проведите формальные тесты на гетероскедастичность. Примените оценки дисперсии $\hat{\beta}$ устойчивые к гетероскедастичности. Прокомментируйте. Может помочь http://bdemeshev.github.io/r_cycle/cycle_files/12_hetero.html

8. Покажите буйство своей фантазии и аккуратность!

   Не стоит думать, что побуквенное выполнение этих инструкций гарантирует оценку в десять баллов. Эконометрика — это не ремесло, а искусство! Фантазируйте! Убедите меня в работе, что вы были на лекциях, даже если это не так :) Аккуратность в виде подписанных осей на графиках, указанных единицах измерения также не повредит.

9. Срок сдачи — 12 января 2015 года.

   Работа принимается исключительно в печатном виде с применением грамотного программирования R + LaTeXили markdown. Каждый день более поздней сдачи умножает оценку за работу на 0.8. Работа должна представлять слитный текст, код скрывать не нужно. В конце должна быть команда sessionInfo().

6.8. Экзамен. 15.06.15

Больше двух столетий назад, 15 июня 1763 года Екатерина II издала манифест, запрещающий произнесение необдуманных речей, опасных для общественного спокойствия. Давайте применим этот манифест к экзамену по эконометрике!

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

   $$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}.$$

   Для удобства расчетов приведены матрицы

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

   (a) Найдите вектор МНК-оценок коэффициентов $\hat{\beta}$.

   (b) Найдите несмещенную оценку для неизвестного параметра σ^2.

   (c) Предполагая нормальность ε_i проверьте гипотезу $\beta_2 = 0$ против альтернативной о неравенстве на уровне значимости 5%

2. Аккуратно сформулируйте теорему Гаусса-Маркова в парадигме стохастических регрессоров для ситуации случайной выборки

3. Немного вопросов про гетероскедастичность:

   (a) Что такое гетероскедастичность?

   (b) К каким последствиям она приводит?

   (c) Что можно предпринять в условиях гетероскедастичности и что эта мера даёт?

4. Для MA(1) модели $y_t = 3 + \varepsilon_t + 0.5\varepsilon_{t-1}$ посчитайте автокорреляционную и частную автокорреляционную функцию.

5. По 100 наблюдениям была оценена логит модель, $\hat{y}_i^* = 2.3 - 2x_i$.

   (a) Оцените вероятность $P(y_i = 1)$ для $x_i = 1$

   (b) Оцените предельный эффект $dP(y_i = 1)/dx$ для $x_i = 1$

6. Рассмотрим классическую модель парной регрессии $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ с нестохастическими регрессорами и нормальными остатками. Модель оценивается по 100 наблюдениям. Известно, что F-статистика, проверяющая гипотезу о незначимости регрессии в целом, равна 50.

   Рассчитайте значения статистики множителей Лагранжа, LM, статистики Вальда, W, для проверки гипотезы о незначимости регрессии в целом.

7. Иван Андреевич Крылов хочет оценить, как зависит количество написанных им за день строчек басни, y_i, от количества съеденных булочек, x_i. То есть его интересует коэффициент β_2 в уравнении

   $$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Однако когда его окрыляет вдохновение, он очень невнимательно считает, поэтому величина x_i ненаблюдаема. Кухарка и жена всегда готовы сказать, сколько булочек Крылов якобы съел за день. Эти величины содержат ошибку измерения, то есть $x_i^A = x_i + u_i^A$ и $x_i^B = x_i + u_i^B$. Ошибки измерения u_i^A, u_i^B, случайная составляющая ε_i независимы, $u_i^A \sim N(0, \sigma_A^2)$, $u_i^B \sim N(0, \sigma_B^2)$, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Иван Андреевич располагает информацией о 100 случайно выбранных днях.

(a) Рассмотрим оценку $\hat{\beta}_2$, получаемую с помощью обычного МНК в регрессии $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i^A$. Будет ли она состоятельной?

(b) Рассмотрим двухшаговый МНК. На первом шаге строим регрессию x_i^A на x_i^B и получаем прогнозные значения $\hat{x}_i^A$. На втором шаге оцениваем регрессию $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 \hat{x}_i^A$. Будет ли оценка $\hat{\beta}_2$ состоятельной?

7. 2015-2016

7.1. Праздник номер 1. Вспомнить всё! 15.09.2015

1. Найдите длины векторов $a = (1, 1, 1, 1)$ и $b = (1, 2, 3, 4)$ и косинус угла между ними. Найдите один любой вектор, перпенидкулярный вектору a.

2. Сформулируйте теорему о трёх перпендикулярах

3. Для матрицы

$$A = \begin{pmatrix} 10 & 15 \\ 15 & 26 \end{pmatrix}$$

   (a) Найдите собственные числа и собственные векторы матрицы

   (b) Найдите $\det(A)$, $\mathrm{tr}(A)$

   (c) Найдите обратную матрицу, A^{-1}, ее собственные векторы и собственные числа

4. Известно, что X — матрица размера $n \times k$ и $n > k$, известно, что $X'X$ обратима. Рассмотрим матрицу $H = X(X'X)^{-1}X'$. Укажите размер матрицы H, найдите H^{2015}, $\mathrm{tr}(H)$, $\det(H)$, собственные числа матрицы H. Штрих означает транспонирование.

5. Для случайных величин X и Y заданы следующие значения: $\mathbb{E}(X) = 1$, $\mathbb{E}(Y) = 4$, $\mathbb{E}(XY) = 8$, $\mathrm{Var}(X) = \mathrm{Var}(Y) = 16$. Для случайных величин $U = X + Y$ и $V = X - Y$ вычислите:

   (a) $\mathbb{E}(U)$, $\mathrm{Var}(U)$, $\mathbb{E}(V)$, $\mathrm{Var}(V)$, $\mathrm{Cov}(U, V)$

   (b) Можно ли утверждать, что случайные величины U и V независимы?

6. Вася ведёт блог. Обозначим X_i — количество слов в i–ой записи. После первого года он по 200 своим записям обнаружил, что $\bar{X}_{200} = 95$ и выборочное стандартное отклонение равно 300 слов. На уровне значимости $\alpha = 0.15$ проверьте гипотезу о том, что $\mu = 100$ против альтернативной гипотезы $\mu \neq 100$. Постройте 85-ти процентный доверительный интервал для μ.

7. Саша и Маша решали одну и ту же задачу. Саша правильно решает задачу с вероятностью 0.8, Маша, независимо от Саши (!), с вероятностью 0.7. Какова вероятность того, что Маша верно решила задачу, если задачу верно решил только кто-то один из них?

7.2. Праздник номер 1. Вспомнить всё! 15.09.2015, решение

Автор решения: Кирилл Пономарёв

1. Ну тут все понятно

2. Правильная формулировка: «Если прямая, проведенная на плоскости через основание наклонной, перпендикулярна её проекции, то она перпендикулярна и самой наклонной»

3. (a) Собственные значения λ_i:

$$\begin{vmatrix} 10-\lambda & 15 \\ 15 & 26-\lambda \end{vmatrix} = \lambda^2 - 36\lambda + 35 = 0 \Rightarrow \lambda_1 = 1,\ \lambda_2 = 35$$

Собственный вектор h соответствующий собственному значению λ по определению:

$$Ah = \lambda h \Rightarrow (A - \lambda \cdot I)h = 0$$

То есть для λ_1:

$$\begin{pmatrix} 9 & 15 \\ 15 & 25 \end{pmatrix} \cdot \begin{pmatrix} h_{11} \\ h_{21} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Откуда:

$$h_{11} = -\frac{5}{3} \cdot h_{21} \quad \text{то есть, например, вектор:} \begin{pmatrix} 5 \\ -3 \end{pmatrix}$$

Аналогично находится собственный вектор, соответствующий $\lambda_2 = 35$

(b) Подставив $\lambda = 0$ в первый определитель, получим $\text{det}(A) = 35$, а $\text{tr}(A) = \lambda_1 + \lambda_2 = 36$

(c) По определению $Ah = \lambda h$. Домножив на A^{-1} (если она существует) обе части, получим:

$$h = \lambda A^{-1} h \;\Rightarrow\; A^{-1}h = \frac{1}{\lambda} h$$

То есть собственные значения для A^{-1} это $1/\lambda_i$ а собственные векторы такие же, как и у матрицы A.

4. Размер матрицы H:

$$\underset{n\times k}{X}\underset{k\times k}{(X'X)^{-1}}\underset{k\times n}{X'} = \underset{n\times n}{H}$$

Заметим что $H^k = H$ для $k = 1, 2, \ldots$

$$H^2 = X\underbrace{(X'X)^{-1}X'X}_{=I}(X'X)^{-1}X' = X(X'X)^{-1}X' = H$$

Такие матрицы называются идемпотентными, а для них: $\text{tr}(A) = \text{rank}(A)$. Найдем след, пользуясь свойством что $\text{tr}(AB) = \text{tr}(BA)$.

$$\text{tr}(H) = \text{tr}(X(X'X)^{-1}X') = \text{tr}((X'X)^{-1}XX') = \text{tr}(I_k) = k$$

То есть в данном случае $\text{rank}(H) = k$. Так как $\text{rank}(H) < n$, определитель этой матрицы равен 0.

Для того чтобы найти собственные числа (λ), снова вспомним определение:

$$Hv = \lambda v$$

Домножив обе части этого уравнения на H, получим:

$$H^2 v = \lambda \cdot Hv \;\Leftrightarrow\; Hv = \lambda^2 v$$

Можем делать так сколько угодно раз, то есть если λ – собственное число матрицы H, то и λ^k тоже. Значит это могут быть только 0 или 1. Тот факт, что 0 является собственным значением сразу вытекает из неполного ранга.

5. (a)
$$\mathbb{E}(U) = \mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 5$$
$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = 8 - 4 = 4$$
$$\mathrm{Var}(U) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\cdot\mathrm{Cov}(X,Y) = 16 + 16 + 2\cdot 4 = 40$$
$$\mathbb{E}(V) = \mathbb{E}(X-Y) = \mathbb{E}(X) - \mathbb{E}(Y) = -3$$
$$\mathrm{Var}(V) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\cdot\mathrm{Cov}(X,Y) = 16 + 16 - 2\cdot 4 = 24$$
$$\mathrm{Cov}(U,V) = \mathrm{Cov}(X+Y, X-Y) = \mathrm{Cov}(X,X) - \mathrm{Cov}(X,Y) + \mathrm{Cov}(Y,X) - \mathrm{Cov}(Y,Y) = \\ = \mathrm{Var}(X) - \mathrm{Var}(Y) = 0 \quad (5)$$

(b) Нельзя, это не следует из равенства ковариации нулю. Можно было бы утверждать про независимость, если бы величины имели совместное нормальное распределение.

6. Ликбез: если известно, что сами X_i нормальные, а истинной дисперсии нет, то используется распределение Стьюдента, и только в этом случае! Здесь же про распределение X_i ничего не известно, но асимптотически (по ЦПТ) можно использовать нормальное.
$$\sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma} \sim \mathcal{N}(0,1)$$
Проверяем нулевую гипотезу против двухсторонней альтернативной:
$$\begin{cases} H_0: & \mu = 100 \\ H_{al}: & \mu \neq 100 \end{cases}$$
Это значит что критические 15 процентов должны быть распределены по двум хвостам, поэтому нас интересуют $z_{0.075}$ и $z_{0.925}$. Рассчетная статистика:
$$T = \sqrt{200}\cdot\frac{95-100}{300} = -0.24 > -1.44 = z_{0.075}$$
Значит нулевая гипотеза не отвергается. Строим доверительный интервал:
$$\mathbb{P}\left(z_{0.925} < \sqrt{n}\cdot\frac{\overline{X}-\mu}{\sigma} < z_{0.925}\right) = 0.85$$
Так как распределение Стьюдента симметрично:
$$\mathbb{P}\left(\overline{X} - z_{0.925}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{0.925}\cdot\frac{\sigma}{\sqrt{n}}\right) = 0.85$$
Таким образом:
$$\mathbb{P}(64.46 < \mu < 125.43) = 0.85$$

7. Пусть событие A: «Маша верно решила задачу», а событие B: «Задачу решил только кто-то один». По формуле Байеса:
$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\cdot\mathbb{P}(A)}{\mathbb{P}(B)}$$
Здесь:
$$\begin{aligned} \mathbb{P}(B|A) &= \mathbb{P}(\text{Саша не решил}) = 0.2 \\ \mathbb{P}(A) &= 0.7 \\ \mathbb{P}(B) &= 0.8\cdot 0.3 + 0.7\cdot 0.2 = 0.38 \end{aligned}$$
Поэтому:
$$\mathbb{P}(A|B) = \frac{0.2\cdot 0.7}{0.38} = \frac{7}{19}$$

7.3. Праздник номер 2, 10 ноября 2015

1. В рамках классической линейной регрессионной модели $y = X\beta + \varepsilon, \mathbb{E}(\varepsilon) = 0, \mathrm{Var}(\varepsilon) = \sigma^2 \cdot I$, найдите: $\mathbb{E}(\hat{\varepsilon})$, $\mathrm{Var}(\hat{\varepsilon})$, $\mathrm{Cov}(\hat{\varepsilon}, y)$

2. Имеются данные:

y_i	x_i	z_i
1	2	1
2	-1	2
3	-3	-3
4	2	0

Предположим, что ошибки нормальны $N(0; \sigma^2)$ и независимы.

   (a) Оцените с помощью МНК модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

   (b) Найдите RSS, TSS, ESS и R^2

   (c) Проверьте гипотезу о незначимости коэффициента $\hat{\beta}_2$ на уровне значимости 5%.

   (d) Найдите оценку ковариационной матрицы коэффициентов $\widehat{\mathrm{Var}}(\hat{\beta})$

   (e) Проверьте гипотезу $H_0 : \beta_2 = \beta_3$ на уровне значимости 5%.

   (f) Для четвёртого наблюдения постройте прогноз и найдите ошибку прогноза.

3. Как могут измениться (могут ли увеличиться? уменьшиться?) RSS, TSS и ESS при добавлении дополнительного наблюдения? При добавлении дополнительного регрессора?

4. Эконометресса Агнесса оценила множественную регрессию $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$. Потом она добавила два наблюдения к своей выборке: $y_{n+1} = 1$, $x_{n+1} = 2$, $z_{n+1} = 3$ и $y_{n+2} = -1$, $x_{n+2} = 1$, $z_{n+2} = 2$. Как при этом изменились матрицы $X'X$ и $X'y$?

5. По 47 наблюдениям оценивается зависимость фертильности женщин от доли мужчин занятых в сельском хозяйстве и доли католического населения по Швейцарским кантонам в 1888 году.

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

	Оценка	Ст. ошибка	t-статистика
(Intercept)		4.98	16.68
Examination	-0.89		-4.08
Catholic	0.04	0.04	

   (a) Заполните пропуски в таблице.

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

   (c) Постройте 95%-ый доверительный интервал для коэффициента при Examination

6. Аккуратно сформулируйте (с «Если» и «то») теорему Гаусса-Маркова для случая нестохастических регрессоров.

7. Нарисуйте Самую Главную Картинку, иллюстрирующую метод наименьших квадратов для множественной регрессии. Отметьте на картинке RSS, ESS, TSS и R^2

8. Эконометресса Ефросинья оценивала модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$. Найдя матрицы $X'X$ и $(X'X)^{-1}$, она призадумалась...

$$X'X = \begin{bmatrix} 47 & 775 & 1934 \\ 775 & 15707 & 23121 \\ 1934 & 23121 & 159570 \end{bmatrix}, (X'X)^{-1} = \begin{bmatrix} 0.26653 & -0.01067 & -0.00168 \\ -0.01067 & 0.00051 & 0.00006 \\ -0.00168 & 0.00006 & 0.00002 \end{bmatrix}$$

   (a) Помогите Ефросинье найти количество наблюдений, $\bar{z}$, $\sum x_i z_i$, $\sum (x_i - \bar{x})(z_i - \bar{z})$

   (b) Ефросинья решила зачем-то также оценить модель $x_i = \gamma_1 + \gamma_2 z_i + u_i$. Как выглядят матрицы $X'X$ и $X'y$ для новой модели?

   (c) (*) Как Ефросинья может найти RSS в новой модели в одно арифметическое действие?

9. Как известно, $\hat{y} = Hy$, где матрица-шляпница H задаётся формулой $H = X(X'X)^{-1}X'$.

   (a) Является ли вектор остатков $\hat{\varepsilon}$ собственным вектором матрицы H? Если да, то какое собственное число ему соответствует?

   (b) Является ли вектор прогнозов $\hat{y}$ собственным вектором матрицы H? Если да, то какое собственное число ему соответствует?

   (c) Является ли регрессор z (скажем, второй столбец матрицы X) собственным вектором матрицы H? Если да, то какое собственное число ему соответствует?

10. Задача на компе с R.

    Подключите библиотеку ggplot2 командой library("ggplot2"). Рассмотрим набор данных по цене бриллиантов diamonds. Оцените линейную модель зависимости цены бриллианта price от массы carat и глубины depth. После оценки модели:

    (a) Поместите R^2 в переменную r_sq

    (b) Поместите RSS в переменную rss

    (c) Поместите оценку коэффициента при carat в переменную hb_carat

    (d) Поместите прогнозы в вектор y_hat

7.4. Блокбастер, 28-12-2015

В этот день, 28 декабря 1895 года, в индийском салоне «Гран-кафе» на бульваре Капуцинок в Париже состоялся публичный показ «Синематографа братьев Люмьер» :)

1. Регрессионная модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$ задана в матричном виде $y = X\beta + \varepsilon$, 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

$$y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

   Для удобства расчетов приведены матрицы

$$X'X = \begin{pmatrix} 5 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix} \text{ и } (X'X)^{-1} = \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1.5 & 1 \\ -1 & 1 & 1.5 \end{pmatrix}.$$

   (a) Найдите вектор МНК-оценок коэффициентов $\hat{\beta}$.

(b) Найдите коэффициент детерминации R^2

(c) Предполагая нормальное распределение вектора ε, проверьте гипотезу H_0: $\beta_3 = 0$ против альтернативной H_a: $\beta_3 \neq 0$ на уровне значимости 5%.

(d) Постройте точечный прогноз и 95%-ый предиктивный интервал для $x_6 = 2$ и $z_6 = 0$.

2. Рассмотрим модель со стохастическими регрессорами $y = X\beta + \varepsilon$. При этом $\mathbb{E}(\varepsilon|X) = 0$, как и положено, однако ошибки ε хитро зависят друг от друга, и поэтому $\mathrm{Var}(\varepsilon|X)$ есть некоторая известная недиагональная матрица V. Несмотря на нарушение предпосылок теоремы Гаусса-Маркова Чак Норрис использует обычный МНК для получения оценок $\hat{\beta}$.

   Найдите $\mathbb{E}(\hat{\beta}|X)$, $\mathrm{Var}(\hat{\beta}|X)$ и $\mathrm{Cov}(\hat{y}, \hat{\varepsilon}|X)$

3. Рассмотрим классическую линейную регрессионную модель:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$$

   (a) Огюст Люмьер утверждает, что при нестохастических регрессорах математические ожидания $\mathbb{E}(y_i)$ различны. Луи Люмьер утверждает, что при стохастических регрессорах и предпосылке о том, что наблюдения являются случайной выборкой, все $\mathbb{E}(y_i)$ равны между собой. Кто из них прав?

   (b) Помогите Луи Люмьеру найти $\mathrm{plim}\,\hat{\varepsilon}_1$ и $\mathrm{plim}\,\hat{y}_1$

4. Рассмотрим классическую линейную регрессионную модель:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$$

   Наблюдения является случайной выборкой. Истинная коварация $\mathrm{Cov}(x_i, z_i)$ равна нулю. Мы оцениваем с помощью МНК две регрессии.

   Регрессия 1:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\beta}_3 z_i$$

   Регрессия 2:

$$\hat{y}_i = \hat{\gamma}_1 + \hat{\gamma}_2 x_i$$

   (a) Верно ли, что $\hat{\beta}_2$ совпадает с $\hat{\gamma}_2$?

   (b) Верно ли, что $\mathrm{plim}\,\hat{\beta}_2 = \mathrm{plim}\,\hat{\gamma}_2$?

5. Аккуратно опишите процедуру сравнения с помощью F-теста двух вложенных (ограниченной и неограниченной) линейных моделей:

   (a) Сформулируйте H_0 и H_a

   (b) Сформулируйте все предпосылки теста

   (c) Укажите способ подсчёта тестовой статистики

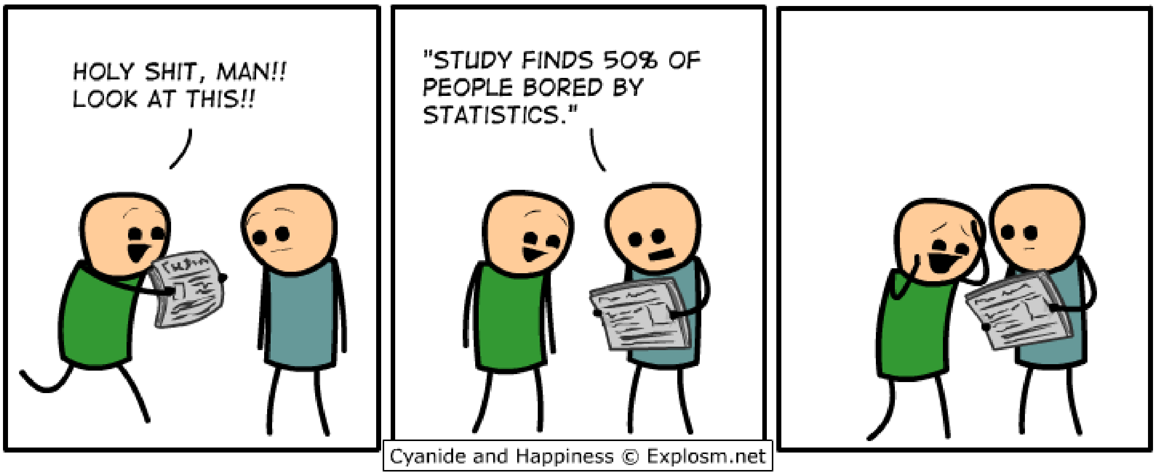
   (d) Укажите закон распределения тестовой статистики при верной H_0

   (e) Сформулируйте правило, по которому делается вывод об H_0

6. Чтобы не выдать себя, Джеймс Бонд оценивает с помощью МНК только однопараметрические регрессии вида $y_i = \beta x_i + \varepsilon_i$. Однако он знаком с теоремой Фриша-Вау.

   (a) Сколько подобных однопараметрических регрессий ему придется оценить, чтобы получить оценку коэффициента β_3 в множественной регрессии $y_i = \beta_1 w_i + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$?

(b) Укажите, какие именно регрессии нужно построить для данной цели



7.5. Максимально правдоподобно, 25-02-2016

1. Как известно, Фрекен Бок любит пить коньяк по утрам. За прошедшие пять дней 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	0
x_i	2	1	3	1	0

Зависимость между y_i и x_i описывается пробит-моделью, $\mathbb{P}(y_i = 1) = F(\beta_1 + \beta_2 x_i)$.

   (a) Выпишите логарифмическую функцию правдоподобия

   (b) Выпишите условия первого порядка для оценки β_1 и β_2

2. Приведите пример небольшого набора данных для которого оценки логит модели $\mathbb{P}(y_i = 1) = F(\beta_1 + \beta_2 x_i)$ не существуют. В наборе данных должны присутствовать хотя бы одно наблюдение $y_i = 0$ и хотя бы одно наблюдение $y_i = 1$.

3. Почему в пробит-модели предполагается, что $\varepsilon_i \sim \mathcal{N}(0; 1)$, а не $\varepsilon_i \sim \mathcal{N}(0; \sigma^2)$ как в линейной регрессии?

4. Исследователь Вениамин пытается понять, как логарифм количества решённых им по эконометрике задач зависит от количества съеденных им пирожков. Для этого он собрал 100 наблюдений. Первые 50 наблюдений — относятся к пирожкам с мясом, а последние 50 наблюдений — к пирожкам с повидлом. Вениамин считает, что ожидаемое количество решённых задач не зависит от начинки пирожков, а только от их количества, т.е. $y_i = \beta x_i + u_i$. Однако он полагает, что для пирожков с мясом — $u_i \sim \mathcal{N}(0; \sigma_M^2)$, а для пирожков с повидлом — $u_i \sim \mathcal{N}(0; \sigma_J^2)$.

   (a) Выпишите логарифмическую функцию правдоподобия

   (b) Выпишите условия первого порядка для оценки β, σ_M^2, σ_J^2

5. При оценке логит модели $\mathbb{P}(y_i = 1) = \Lambda(\beta_1 + \beta_2 x_i)$ по 500 наблюдениям оказалось, что $\hat{\beta}_1 = 0.7$ и $\hat{\beta}_2 = 3$. Оценка ковариационной матрицы коэффициентов имеет вид

$$\begin{pmatrix} 0.04 & 0.01 \\ 0.01 & 0.09 \end{pmatrix}$$

(a) Проверьте гипотезу о незначимости коэффициента $\hat{\beta}_2$

(b) Найдите предельный эффект роста x_i на вероятность $\mathbb{P}(y_i = 1)$ при $x_i = -0.5$

(c) Найдите максимальный предельный эффект роста x_i на вероятность $\mathbb{P}(y_i = 1)$

(d) Постройте точечный прогноз вероятности $\mathbb{P}(y_i = 1)$ если $x_i = -0.5$

(e) Найдите стандартную ошибку построенного прогноза

6. После долгих изысканий Вениамин пришёл к выводу, что $\beta = 0$, т.е. что логарифм количества решенных им по эконометрике за вечер задач имеет нормальное распределение y_i с математическим ожиданием ноль. Однако он по прежнему уверен, что дисперсия y_i зависит от того, какие пирожки он ел в этом вечер. Для пирожков с повидлом $y_i \sim \mathcal{N}(0; \sigma_J^2)$, а для пирожков с мясом — $y_i \sim \mathcal{N}(0; \sigma_M^2)$. Всего 100 наблюдений. Первые 50 вечеров относятся к пирожкам с мясом, последние 50 вечеров — к пирожкам с повидлом:

$$\sum_{i=1}^{50} y_i = 10,\ \sum_{i=1}^{50} y_i^2 = 100,\ \sum_{i=51}^{100} y_i = -10,\ \sum_{i=51}^{100} y_i^2 = 300$$

(a) Найдите оценки σ_M^2, σ_J^2, которые получит Вениамин.

(b) Помогите Вениамину проверить гипотезу $\sigma_M^2 = \sigma_J^2$ с помощью тестов отношения правдоподобия, множителей Лагранжа и Вальда.

7.6. Решение задач КР по Эконометрике, 2015-2016

1. Функция максимального правдоподобия

$$L = \prod_{i=1}^{5} \mathbb{P}(Y_i = j),\ \text{где } j \in \{0, 1\} \implies \tag{6}$$

$$L = \mathbb{P}(Y_1 = 1) \times \mathbb{P}(Y_2 = 0) \times \mathbb{P}(Y_3 = 1) \times \mathbb{P}(Y_4 = 0) \times \mathbb{P}(Y_5 = 0) \implies \tag{7}$$

$$l = \log L = \sum_{i=1,3} \log(F(\beta_1 + x_i\beta_2)) + \sum_{i=2,4,5} \log(1 - F(\beta_1 + x_i\beta_2)) \tag{8}$$

Найдем теперь условия первого порядка

$$\frac{\partial L}{\partial \beta_1} = \sum_{i=1,3} \frac{f(\beta_1 + x_i\beta_2)}{F(\beta_1 + x_i\beta_2)} - \sum_{i=2,4,5} \frac{f(\beta_1 + x_i\beta_2)}{1 - F(\beta_1 + x_i\beta_2)} = 0, \tag{9}$$

$$\frac{\partial L}{\partial \beta_2} = \sum_{i=1,3} \frac{x_i f(\beta_1 + x_i\beta_2)}{F(\beta_1 + x_i\beta_2)} - \sum_{i=2,4,5} \frac{x_i f(\beta_1 + x_i\beta_2)}{1 - F(\beta_1 + x_i\beta_2)} = 0. \tag{10}$$

2. Пример.
Пусть $x_1 = 12, x_2 = 23, x_3 = 1223$ и $y_1 = y_2 = 0, y_3 = 1$.

3. Краткий ответ. Потому что y^* сравнивается с 0.
Немного более развернутый ответ. Пробит модель в общем случае выглядит такие образом:

$$y_i^* = \beta_1 + \beta_2 x_i + \epsilon_i, \tag{11}$$

где $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, а $y = 1$, когда $y^* \geqslant c$ и $y = 0$ иначе. Заметим, что замена $\tilde{y}^* = \sigma^{-1}(y^* - c)$ приводит к случаю, когда $\tilde{y}^* \sim \mathcal{N}(0, 1)$, следовательно всегда можно перейти к $\epsilon_i \sim \mathcal{N}(0, 1)$ и сравнивать y^* с нулем.

4. Функция максимального правдоподобия

$$L = \prod_{i=1}^{50} \frac{1}{\sqrt{2\pi\sigma_M^2}} \exp\left\{-\frac{(y_i - \beta x_i)^2}{2\sigma_M^2}\right\} \times \prod_{i=51}^{100} \frac{1}{\sqrt{2\pi\sigma_J^2}} \exp\left\{-\frac{(y_i - \beta x_i)^2}{2\sigma_J^2}\right\} \implies \tag{12}$$

$$L = \frac{1}{(2\pi \cdot \sigma_M \cdot \sigma_J)^{50}} \exp\left\{-\sum_{i=1}^{50} \frac{(y_i - \beta x_i)^2}{2\sigma_M^2} - \sum_{i=51}^{100} \frac{(y_i - \beta x_i)^2}{2\sigma_J^2}\right\} \tag{13}$$

$$l = \log L = -50 \log(2\pi \cdot \sigma_M \cdot \sigma_J) - \sum_{i=1}^{50} \frac{(y_i - \beta x_i)^2}{2\sigma_M^2} - \sum_{i=51}^{100} \frac{(y_i - \beta x_i)^2}{2\sigma_J^2} \tag{14}$$

Запишем теперь условия первого порядка

$$\frac{\partial L}{\partial \sigma_M^2} = -\frac{25}{\sigma_M^2} + \frac{\sum_{i=1}^{50}(y_i - \beta x_i)^2}{2(\sigma_M^2)^2} = 0, \tag{15}$$

$$\frac{\partial L}{\partial \sigma_J^2} = -\frac{25}{\sigma_J^2} + \frac{\sum_{i=51}^{100}(y_i - \beta x_i)^2}{2(\sigma_J^2)^2} = 0, \tag{16}$$

$$\frac{\partial L}{\partial \beta} = \frac{\sum_{i=1}^{50} 2(y_i - \beta x_i)x_i}{2\sigma_M^2} + \frac{\sum_{i=51}^{100} 2(y_i - \beta x_i)x_i}{2\sigma_J^2} = 0. \tag{17}$$

5. (a)

$$H_0 : \hat{\beta}_2 = 0 \tag{18}$$
$$H_a : \hat{\beta}_2 \neq 0 \tag{19}$$

$t_{\text{obs}} = 3/\sqrt{0.09} = 10 \implies$, гипотеза о незначимости коэффициента отвергается, т.е. коэффициент β_2 значим.

(b)

$$\frac{\partial \mathbb{P}(y_i = 1)}{\partial x_i}|_{x_i=-0.5} = \frac{e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}}{(1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i})^2}\hat{\beta}_2 = \frac{e^{0.8}}{(1 + e^{0.8})^2} \cdot 3 = 0.64 \tag{20}$$

(c)

$$\frac{\partial \mathbb{P}(y_i = 1)}{\partial x_i} = \frac{e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}}{(1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i})^2}\hat{\beta}_2 \to \max_{x_i} \implies \tag{21}$$

$$\frac{\hat{\beta}_2 e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}(1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i})^2 - 2\hat{\beta}_2(1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i})(e^{\hat{\beta}_1 + \hat{\beta}_2 x_i})^2}{(1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i})^4}\hat{\beta}_2 = 0, \tag{22}$$

$$1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i} - 2e^{\hat{\beta}_1 + \hat{\beta}_2 x_i} = 0 \implies x_i = -\frac{0.7}{3}. \tag{23}$$

(d)

$$\mathbb{P}(y_i = 1) = \Lambda(\hat{\beta}_1 + \hat{\beta}_2 x_i) = \frac{1}{1 + \exp\{-0.7 + 1.5\}} = 0.3100255 \tag{24}$$

(e) Здесь нужно использовать Дельта-метод. Линеаризуем функцию $F(\hat{\beta}) = \hat{\mathbb{P}}(y_i = 1)$ в окрестности точки $\beta = (\beta_1, \beta_2)'$:

$$F(\hat{\beta}) \approx F(\beta) + \nabla_F(\beta)' \cdot (\hat{\beta} - \beta)$$

Где все векторы — столбцы. Тогда оценка дисперсии прогноза:

$$\widehat{Var}(\hat{P}(y_i = 1)) = \nabla_F(\hat{\beta})' \cdot \widehat{\text{Var}}(\hat{\beta}) \cdot \nabla_F(\hat{\beta})$$

Градиент функции F:

$$\nabla_F(\hat{\beta}) = \begin{pmatrix} \frac{e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}}{1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}} & x_i \cdot \frac{e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}}{1 + e^{\hat{\beta}_1 + \hat{\beta}_2 x_i}} \end{pmatrix}' = (0.69 \quad -0.34)'$$

Тогда стандартная ошибка прогноза:

$$s.e(\hat{P}(y_i = 1)) = (0.69^2 \cdot 0.04 - 2 \cdot 0.01 \cdot 0.69 \cdot 0.34 + 0.34^2 \cdot 0.09)^{1/2} = 0.18$$

6. Используя формулы (15) и (16) получаем

$$25 = \frac{\sum_{i=1}^{50} y_i^2}{2\sigma_M^2} \implies \hat{\sigma}_M^2 = 2, \tag{25}$$
$$25 = \frac{\sum_{i=51}^{100} y_i^2}{2\sigma_J^2} \implies \hat{\sigma}_J^2 = 6. \tag{26}$$

Проверим следующую гипотезу

$$H_0 : \sigma_M^2 = \sigma_J^2 \tag{27}$$
$$H_a : \sigma_M^2 \neq \sigma_J^2 \tag{28}$$

Тест отношения правдоподобия

$$LR = 2(\hat{l}_{UR} - \hat{l}_R) \sim \chi_1^2, \tag{29}$$

где $\hat{l}_{UR}$ и $\hat{l}_R$ значение логарифмиеских функций правдоподобия в точках максимума для неограниченной и ограниченной моделей соответственно. Решим ограниченную модель. Функция правдоподобия:

$$L(\sigma^2) = \prod_{i=1}^{100} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{y_i^2}{2\sigma^2}\right\} \to \max_{\sigma^2}$$

Логарифмическая функция правдоподобия:

$$l(\sigma^2) = -50ln(2\pi) - 50ln(\sigma^2) - \sum_{i=1}^{100} \frac{y_i^2}{2\sigma^2} \to \max_{\sigma^2}$$

Откуда:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{100} y_i^2}{100} = 4$$

Тогда:

$$LR = 2(-25ln(12) - 50 + 50ln(4) + 50) = 14.38 > \chi_{crit}^2 = 3.84$$

То есть нулевая гипотеза отвергается!
Тест Вальда

$$W = g(\hat{\theta}_{UR})'(\nabla_g \hat{I}(\hat{\theta}_{UR})^{-1}\nabla_g')^{-1}g(\hat{\theta}_{UR})$$

где градиент $\nabla_g = (1 \quad -1)$ — вектор строка. Нам пригодится информационная матрица Фишера в задачи без ограничений. Найти ее можно взяв вторые производные от функции правдоподобия в четвертой задаче по σ_M^2 и σ_J^2, положив $\beta = 0$ и домножив на -1.

$$\hat{I}(\hat{\theta}_{UR}) = \begin{pmatrix} -\frac{25\hat{\sigma}_M^2 - 100}{\left(\hat{\sigma}_M^2\right)^3} & 0 \\ 0 & -\frac{25\hat{\sigma}_J^2 - 300}{\left(\hat{\sigma}_J^2\right)^3} \end{pmatrix} = \begin{pmatrix} \frac{50}{8} & 0 \\ 0 & \frac{150}{216} \end{pmatrix}$$

Тогда:

$$W = (6-2)^2 \left((1 \quad -1) \begin{pmatrix} \frac{150}{216} & 0 \\ 0 & \frac{50}{8} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \right)^{-1} = 10$$

Как и ранее, нулевая гипотеза отвергается!
Тест множителей Лагранжа
Нужно посчитать статистику:

$$LM = \left(\frac{\partial l(\hat{\theta}_R)}{\partial \theta} \right)' \hat{I}(\hat{\theta}_R)^{-1} \left(\frac{\partial l(\hat{\theta}_R)}{\partial \theta} \right) \sim \chi_r^2,$$

При верной H_0 логарифмическая функция правдоподобия будет выглядеть следующим образом:

$$l(\sigma^2) = -50ln(2\pi) - 50ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{100} y_i^2$$

Тогда:

$$\frac{\partial l(\hat{\sigma^2})}{\partial \sigma^2} = -\frac{50}{\hat{\sigma}^2} + \frac{1}{2}\sum_{i=1}^{100} y_i^2 \frac{1}{(\hat{\sigma}^2)^2} = 0$$

Откуда: $\hat{\sigma}_R^2 = 4$. Заметим, что если подставить эту оценку в матрицу $\hat{I}(\theta)$, она не будет иметь обратной, что эквивалентно бесконечной LM статистике. Гипотеза H_0, конечно, отвергается!

7.7. Большой Устный ЗАчёт - 2016

1. Метод Наименьших Квадратов.
   (a) МНК-картинка
   (b) Нахождение всего-всего, если известен вектор y и матрица X
2. Теорема Гаусса-Маркова
   (a) Формулировка с детерминистическими регрессорами
   (b) Доказательство с детерминистическими регрессорами
   (c) Формулировки со стохастическими регрессорами
   (d) Что даёт дополнительное предположение о нормальности ε?
   (e) Теорема Фриша-Вау

3. Проверка гипотез о линейных ограничениях
   (a) Проверка гипотезы о значимости коэффициента
   (b) Проверка гипотезы о значимости регрессии в целом
   (c) Проверка гипотезы об одном линейном соотношении с помощью ковариационной матрицы
   (d) Ограниченная и неограниченная модель
   (e) Тест Чоу на стабильность коэффициентов
   (f) Тест Чоу на прогнозную силу
4. Метод максимального правдоподобия
   (a) Свойства оценок
   (b) Два способа получения оценки дисперсии
   (c) Три теста (LM, Wald, LR)
   (d) Выписать функцию ML для обычной регрессии
   (e) для AR(1) процесса
   (f) для MA(1) процесса
   (g) для логит модели
   (h) для пробит модели
   (i) для модели с заданным видом гетероскедастичности
5. Мультиколлинеарность
   (a) Определение, последствия
   (b) Величины, измеряющие силу мультиколлинеарности
   (c) Методы борьбы
   (d) Сюда же: метод главных компонент, хотя он используется и для других целей
6. Гетероскедастичность
   (a) Определение, последствия
   (b) Тесты, график
   (c) Стьюдентизированные остатки
   (d) HC оценки ковариации
   (e) GLS и FGLS
7. Временные ряды
   (a) Стационарный временной ряд
   (b) ACF, PACF
   (c) Модель ARMA
   (d) ARIMA-SARIMA
   (e) Модель GARCH (не будет, не успели)
8. Логит и пробит

(a) Описание моделей
(b) Предельные эффекты
(c) Чувствительность, специфичность (не будет, не успели)
(d) Кривая ROC (не будет, не успели)

9. Эндогенность
   (a) Три примера: одновременность, пропущенные переменные, ошибки измерения
   (b) IV, двухшаговый МНК

10. Модели панельных данных (не будет, не успели)
    (a) RE, FE, сквозная регрессии
    (b) Тест Хаусмана

11. И ещё алгоритмы. Уметь объяснить суть метода. Уметь реализовать его в R.
    (a) Метод опорных векторов
    (b) Классификационные деревья и случайный лес
    (c) Ridge regression
    (d) LASSO
    (e) Квантильная регрессия
    (f) Байесовский подход к регрессии (не будет, не успели)

12. R. Можно принести файл со своей заготовкой, можно пользоваться Интернетом для поиска информации, но не для общения.
    (a) Загрузить данные из .csv файла в R
    (b) Посчитать описательные статистики (среднее, мода, медиана и т.д.)
    (c) Построить подходящие описательные графики для переменных
    (d) Оценить линейную регрессию с помощью МНК. Провести диагностику на что-нибудь (гетероскедастичность, автокорреляцию, мультиколлинеарность).
    (e) Оценить logit, probit модели, посчитать предельные эффекты
    (f) Оценить ARMA модель
    (g) Выделить главные компоненты

7.8. Экзамен 20.06.2016. Вариант 1

1. На основании опроса была оценена следующая модель:

$$\ln(wage_i) = \beta_1 + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 married_i + \beta_5 educ_i + \beta_6 black_i + \varepsilon_i$$

где:

- $wage_i$ — величина заработной платы в долларах
- $exper_i$ — опыт работы в годах
- $educ_i$ — количество лет обучения
- $married_i$ — наличие супруга/супруги (1 — есть, 0 — нет)

- $black_i$ — принадлежность к негроидной расе (1 — да, 0 — нет)

Показатель	Значение
R^2	**B7**
Скорректированный R^2	0.219
Стандартная ошибка регрессии	**B6**
Количество наблюдений	**B2**

Результаты дисперсионного анализа:

	df	SS	MS	F	P-значение
Регрессия	**B1**	5.993	1.199	**B5**	0.000
Остаток	134	18.240	0.136		
Итого	**B3**	**B4**			

Коэффициент	Оценка	$se(\hat{\beta})$	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Константа	4.529	0.331	13.688	0.000	3.874	5.183
$exper$	0.090	0.037	2.419	0.017	0.016	0.164
$exper^2$	-0.003	0.002	-1.790	0.076	-0.006	0.000
$married$	0.240	0.079	3.045	0.003	**B8**	**B9**
$educ$	0.078	0.017	**B10**	0.000	0.045	0.111
$black$	0.073	0.171	0.424	0.672	-0.266	0.411

Найдите пропущенные числа **B1–B10**.

Ответ округляйте до 3-х знаков после запятой. Кратко поясняйте, например, формулой, как были получены результаты.

2. На основании данных по ценам на квартиры в Москве были построена модель

$$\ln(price_i) = \beta_1 + \beta_2 totsp_i + \beta_3 metrdist_i + \beta_4 dist_i + \beta_5 floor_i + \varepsilon_i,$$

где:

- $\ln(price_i)$ — логарифм цены квартиры в тысячах долларов
- $totsp_i$ — общая площадь квартиры в кв. м.
- $metrdist_i$ — расстояние до метров в минутах
- $dist_i$ — расстояние до центра города в км
- $floor_i$ — дамми-переменная (1 — если квартира не на первом и последнем этажах, 0 — иначе)

Модели были оценены на пяти разных выборках, результаты представлены в таблице:

Коэффициент	Выборка A	Выборка B	Выборка C	Выборка D	Выборка E
Константа	3.980***	3.926***	3.929***	3.719***	4.224***
$totsp$	0.0155***	0.0148***	0.0163***	0.0179***	0.0139***
$metrdist$	-0.00858***	-0.0169***	-0.00566**	-0.0108***	-0.0077
$dist$	-0.0267***	-0.0186***	-0.0253***	-0.0150***	-0.0350***
$floor$	0.0419**	0.0633*	0.0224	0.0225	0.0228
Наблюдений	460	145	315	150	150
R^2	0.693	0.684	0.723	0.328	0.520
RSS	15.120	4.503	9.408	2.163	8.545

* — значимость на 10%, ** — значимость на 5%, *** — значимость на 1%.

(a) Для всей выборки (выборка A) проинтерпретируйте коэффициент при переменной $dist_i$.

(b) Определите на 5%-ом уровне значимости, можно ли использовать одну модель для квартир, находящихся в пешей доступности от метро (выборка C), и квартир, находящихся в транспортной доступности (выборка B).

(c) Исследователь предположил, что дисперсия ошибок модели возрастает с увеличением площади квартиры. Проверьте, есть ли в модели гетероскедастичность на 10% уровне значимости на основании соответствующего теста. В выборку D включены 150 квартир с наименьшей общей площадью, в выборку E — 150 квартир с наибольшей общей площадью.

При проверке гипотез: выпишите H_0, H_a, найдите значение тестовой статистики, укажите её распределение, найдите критическое значение, сделайте выводы

3. По ежемесячным данным, 146 наблюдений, была оценена зависимость

$$\widehat{credit}_t = 362.21 - 7.50 r\_cred_t - 13.09 ipc_t,\ R^2 = 0.44$$

где:

- $credit_t$ — объём потребительских кредитов, выданных домашним хозяйствам РФ
- r_credit_t — ставка процента по кредитам
- ipc_t — индекс потребительских цен

Известно, что $\sum_{t=2}^{146}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2 = 266491$, $\sum_{t=1}^{146}\hat{\varepsilon}_t^2 = 438952$, $\sum_{t=2}^{146}|\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}| = 3617$, $\sum_{t=1}^{146}|\hat{\varepsilon}_t| = 6382$.

Кроме того, была оценена вспомогательная модель для остатков исходной модели:

$$\hat{\hat{\varepsilon}}_t = 15.67 - 0.75 r\_credit_t + 0.02 ipc_t + 0.39\hat{\varepsilon}_{t-1} + 0.21\hat{\varepsilon}_{t-2} + 0.24\hat{\varepsilon}_{t-3},\ R^2 = 0.56$$

(a) На 1%-ом уровне значимости проверьте гипотезу об адекватности исходной регрессии

(b) Проведите тест Дарбина-Уотсона на 5%-ом уровне значимости

(c) Проведите тест Бройша-Годфри на 5%-ом уровне значимости

При проверке гипотез: выпишите H_0, H_a, найдите значение тестовой статистики, укажите её распределение, найдите критическое значение, сделайте выводы

4. Домохозяйка Глаша очень любит читать романы Л.Н. Толстого и смотреть сериалы. Её сын Петя учится на третьем курсе ВШЭ. Последние 30 дней он записывал, сколько Глаша прочитала страниц «Анны Карениной», $pages_t$, и посмотрела серий «Доктора Хауса», $series_t$. На основании этих наблюдений при помощи МНК Петя оценил следующую модель:

$$\widehat{pages}_t = 100 - 3 series_t$$

Оценка ковариационной матрицы коэффициентов, $\widehat{\mathrm{Var}}(\hat{\beta}) = \begin{pmatrix} 11 & 0.5 \\ 0.5 & 1 \end{pmatrix}$

Оценка дисперсии ошибок равна $\hat{s}^2 = 323$.

Завтра Глаша собирается посмотреть 10 серий «Доктора Хауса».

(a) Постройте точечный прогноз количества прочитанных Глашей страниц романа

(b) Постройте 95%-ый доверительный интервал для $\mathbb{E}(pages_t | series_t = 10)$, ожидаемого количества прочитанных страниц

(c) Постройте 95%-ый предиктивный интервал для фактического количества прочитанных страниц

5. Опишите МНК для парной регрессии: выпишите целевую функцию, систему нормальных уравнений, оценки коэффициентов, оценки дисперсий коэффициентов.

6. Сформулируйте теорему Гаусса-Маркова для детерминированных регрессоров.

7. Опишите тест Уайта: сформулируйте нулевую и альтернативную гипотезы, способ получения тестовой статистики, её распределение при верной нулевой гипотезе, вид критической области.

7.9. Экзамен 20.06.2016. Вариант 2

1. На основании опроса была оценена следующая модель:

$$\ln(wage_i) = \beta_1 + \beta_2 exper_i + \beta_3 exper_i^2 + \beta_4 married_i + \beta_5 educ_i + \beta_6 black_i + \varepsilon_i$$

где:

- $wage_i$ — величина заработной платы в долларах
- $exper_i$ — опыт работы в годах
- $educ_i$ — количество лет обучения
- $married_i$ — наличие супруга/супруги (1 — есть, 0 — нет)
- $black_i$ — принадлежность к негроидной расе (1 — да, 0 — нет)

Показатель	Значение
R^2	**B6**
Скорректированный R^2	0.164
Стандартная ошибка регрессии	**B7**
Количество наблюдений	**B1**

Результаты дисперсионного анализа:

	df	SS	MS	F	P-значение
Регрессия	**B2**	**B4**	7.425	**B5**	0.000
Остаток	**B3**	184.954	0.145		
Итого	1279	222.079			

Коэффициент	Оценка	$se(\hat{\beta})$	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Константа	4.906	0.106	46.129	0.000	4.698	5.115
$exper$	0.095	0.011	8.956	0.000	0.074	0.115
$exper^2$	-0.003	0.001	-5.437	0.000	-0.004	-0.002
$married$	**B8**	**B9**	**B10**	0.234	-0.018	0.074
$educ$	0.064	0.006	11.582	0.000	0.053	0.075
$black$	-0.183	0.028	-6.490	0.000	-0.238	-0.127

Найдите пропущенные числа **B1–B10**.

Ответ округляйте до 3-х знаков после запятой. Кратко поясняйте, например, формулой, как были получены результаты.

2. Винни-Пух и Пятачок попробовали очень странный мёд. После его употребления, к ним пришли слоники в количестве 100 штук и начали водить вокруг них хороводы. Винни-Пух смог на глазок оценить вес и рост каждого слоника, а Пятачок — его возраст. Эти данные позволили им оценить следующую модель:

$$weight_i = \beta_1 + \beta_2 \ln(height_i) + \beta_3 \ln(age_i) + \varepsilon_i$$

где:

- $weight_i$ — вес слоника
- $\ln(height_i)$ — логарифм роста слоника
- $\ln(age_i)$ — логарифм возраста слоника

Выборка	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	ESS	RSS	N
1. Самые молодые слоники	43.1	1.4	3.7	243	345	40
2. Самые старые слоники	48.4	3.6	1.1	489	194	40
3. Зелёные слоники	39.6	2.1	2.4	311	268	45
4. Розовые слоники	53.1	2.9	3.1	369	307	55
5. Все слоники	45.7	2.6	2.9	615	741	100

(a) Для выборки розовых слоников проинтерпретируйте коэффициент $\hat{\beta}_2$

(b) Определите на 5%-ом уровне значимости, можно ли использовать одну модель для розовых и зелёных слоников

(c) Пятачок уверен, что дисперсия ошибок модели падает с увеличением возраста слоника. Проверьте, так ли это, на 1% уровне значимости на основании соответствующего теста

При проверке гипотез: выпишите H_0, H_a, найдите значение тестовой статистики, укажите её распределение, найдите критическое значение, сделайте выводы

3. Царевна Несмеяна по 146 дням наблюдений построила регрессию:

$$\widehat{tear}_t = 200 - 5sun_t - 10prince_t - 15chocolate_t,\ R^2 = 0.8$$

где:

- $tear_t$ — количество пролитых слёз в мл
- sun_t — дамми-переменная для погоды (1 — солнечная, 0 — пасмурная)
- $prince_t$ — дамми-переменная для посещений Прекрасным Принцем (1 — Принц пришёл, 0 — нет)
- $chocolate_t$ — количество съеденного шоколада в плитках

Известно, что $\sum_{t=2}^{146}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2 = 876$, $\sum_{t=1}^{146}\hat{\varepsilon}_t^2 = 538$, $\sum_{t=2}^{146}|\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}| = 100$, $\sum_{t=1}^{146}|\hat{\varepsilon}_t| = 150$.

(a) На 1%-ом уровне значимости проверьте гипотезу об адекватности регрессии

(b) Проведите тест Дарбина-Уотсона на 5% уровне значимости

(c) Кроме того, была оценена следующая модель:

$$\hat{\hat{\varepsilon}}_t = 5 - 0.05sun_t + 0.02prince_t + 0.09\hat{\varepsilon}_{t-1} + 0.01\hat{\varepsilon}_{t-2} + 0.004\hat{\varepsilon}_{t-3},\ R^2 = 0.06$$

Проведите тест Бройша-Годфри на 1% уровне значимости

При проверке гипотез: выпишите H_0, H_a, найдите значение тестовой статистики, укажите её распределение, найдите критическое значение, сделайте выводы

4. Ослик Иа-Иа горюет и считает количество мёда в горшочках. Сейчас у него 50 горшочков. Горшочки отличаются друг от друга цветом: есть более розовые и менее розовые. Иа-Иа считает, что степень розовости влияет на количество мёда. Он смог оценить следующую регрессию:

$$\widehat{honey}_i = 15 + 3pinkness_i$$

Оценка ковариационной матрицы коэффициентов, $\widehat{\mathrm{Var}}(\hat{\beta}) = \begin{pmatrix} 3 & 1.5 \\ 1.5 & 9 \end{pmatrix}$

Оценка дисперсии ошибок равна $\hat{s}^2 = 137$.

Ослик нашёл новый горшочек с розовостью, равной 5.

(a) Постройте точечный прогноз для количества мёда

(b) Постройте 95%-ый доверительный интервал для $\mathbb{E}(honey_i|pinkness_i = 5)$, ожидаемого количества мёда в горшочке

(c) Постройте 95%-ый предиктивный интервал для фактического количества мёда в горшочке

5. Опишите F-тест для гипотезы о нескольких линейных ограничениях: сформулируйте нулевую и альтернативную гипотезы, способ получения тестовой статистики, её распределение при верной нулевой гипотезе, вид критической области.

6. В парной регрессии $y_t = \beta_1 + \beta_2 x_t + \varepsilon_t$ известно, что $\mathrm{Var}(\varepsilon_t) = \sigma^2 x_t^4$. Опишите процедуру получения эффективных оценок коэффициентов.

7. Опишите двухшаговый МНК: сформулируйте условия для его применения и опишите процедуру построения оценок.

8. 2016-2017

8.1. Праздник 1. Вспомнить всё! 12.09.2016

Сегодня 256-ой день года, всех с днём программиста! :) А ещё 12 сентября в 490 году до нашей эры Фидиппид добежал из Марафона в Афины с криком «Νενικήκαμεν»[2]!

1. Найдите длины векторов $a = (1, 1, 1)$ и $b = (1, 2, 3)$ и косинус угла между ними. Найдите один любой вектор, перпенидкулярный вектору b.

2. Сформулируйте теорему о трёх перпендикулярах и обратную к ней

3. На плоскости α лежит прямая ℓ. Вне плоскости α лежит точка C. Ромео проецирует точку C на прямую ℓ и получает точку R. Джульетта проецирует точку C сначала на плоскость α, а затем проецирует полученную точку A на прямую ℓ. После двух действий Джульетта получает точку D. Обязательно ли R и D совпадают?

4. Для матрицы
$$A = \begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix}$$
   (a) Найдите собственные числа и собственные векторы матрицы
   (b) Найдите $\det(A)$, $\operatorname{tr}(A)$
   (c) Найдите собственные числа матрицы A^{2016}, $\det(A^{2016})$ и $\operatorname{tr}(A^{2016})$

5. Известно, что X — матрица размера $n \times k$ и $n > k$, известно, что $X'X$ обратима. Рассмотрим матрицу $H = X(X'X)^{-1}X'$. Укажите размер матрицы H, найдите H^{2016}, $\operatorname{tr}(H)$, $\det(H)$, собственные числа матрицы H. Штрих означает транспонирование.

6. Занудная халява: известно, что $\operatorname{Cov}(X, Y) = 5, \operatorname{Var}(X) = 10, \operatorname{Var}(Y) = 20, \mathbb{E}(X) = 10, \mathbb{E}(Y) = -10$. Найдите $\operatorname{Cov}(X + 2Y, Y - X)$, $\operatorname{Var}(X + 2Y)$, $\mathbb{E}(X + 2Y)$.

7. За 100 дней Ромео посчитал все глубокие вздохи Джульетты. Настроение Джульетты столь спонтанно, что глубокие вздохи за разные дни можно считать независимыми. В сумме оказалось 890 вздохов. Сумма квадратов оказалась равна 8000. Постройте 95%-ый доверительный интервал для математического ожидания ежедневного количества глубоких вздохов Джульетты. На уровне значимости 5%-ов проверьте гипотезу, что математическое ожидание равно 9.

8. Ромео подкидывает монетку два раза. Если монетка выпадает орлом, то Ромео кладет в мешок черный шар, если решкой —– белый. Джульетта не знает, как выпадала монетка, и достает шары из мешка наугад по очереди. Первый шар оказался черного цвета. Какова вероятность того, что второй шар Джульетты будет белым?

8.2. Контрольная работа-1, демо-версия

1. Эконометресса Ефросинья исследует, как зависит надой молока, $milk_i$, (в литрах) от возраста коровы, age_i, (в годах):
$$milk_i = \beta_1 + \beta_2 age_i + u_i$$

[2] Ликуйте! Мы победили!

Показатель	Значение
RSS	**B1**
ESS	**B2**
TSS	**1240**
R^2	**B3**
Стандартная ошибка регрессии	**1.45**
Количество наблюдений	340

Коэффициент	Оценка	$se(\hat{\beta})$	t-статистика	P-значение	Левая (95%)	Правая (95%)
Константа	4.565	0.207	**B4**	**B9**	**B5**	**B6**
age	**B7**	**B8**	3.670	0.000	0.036	0.119

Найдите пропущенные числа **B1**–**B9**.

Ответ округляйте до 2-х знаков после запятой. Кратко поясняйте формулой, как были получены результаты.

2. Гарри Поттер и Рон Уизли активно готовятся к чемпионату мира по квиддичу. В течение 30 дней они сначала посещают Хогсмид и выпивают некоторое количество сливочного пива в пинтах, $beer_t$, после забивают определённое количество квоффлов в штуках, $quaffle_t$. Гермиона Грейнджер оценила следующую регрессию:

$$\widehat{quaffle}_t = \underset{(2.83)}{80} - \underset{(1)}{3}\, beer_t$$

В скобках приведены стандартные ошибки. Оценка дисперсии ошибок равна $\hat{s}^2 = 238$.

Сегодня Гарри и Рон выпили 4 пинты сливочного пива.

(a) Проверьте гипотезы о значимости каждого коэффициента на уровне значимости 5%.

(b) Постройте точечный прогноз количества квоффлов, забитых сегодня Гарри Поттером и Роном Уизли

(c) Постройте 90%-ый доверительный интервал для коэффициента наклона регрессии

3. Для модели $Y_i = \beta_1 + \beta_2 X_i + u_i$ выполнены все предпосылки теоремы Гаусса-Маркова.
   (a) Докажите, что МНК-оценка коэффициента β_2 является случайной величиной
   (b) Докажите, что эта оценка является несмещённой
   (c) Найдите дисперсию этой оценки
4. Для модели $Y_i = \beta_1 + \beta_2 X_i + u_i$ выполнены все предпосылки теоремы Гаусса-Маркова. Для МНК-оценок коэффициентов найдите $\widehat{\mathrm{Cov}}(\hat\beta_1, \hat\beta_2)$.
5. Дайте определения следующих понятий
   (a) Несмещённая оценка
   (b) Эффективная оценка
   (c) Состоятельная последовательность оценок

8.3. Демо-1. Решение

1.

$$B1 = 1.45^2 \cdot (340 - 2)$$

$$B2 = ESS = TSS - RSS$$

$$B3 = R^2 = ESS/TSS$$

$$B4 = t_c = 4.565/0.207 = 22$$

По таблицам (t-распределение с 338 степенями свободы или примерно нормальное) $t_{crit} = 1.96$

$$B5 = CI_{left} = 4.565 - 1.96 \cdot 0.207$$

$$B6 = CI_{right} = 4.565 + 1.96 \cdot 0.207$$

$$B7 = \hat\beta_{milk} = (0.036 + 0.119)/2 = 0.0775$$

$$B8 = se(\hat\beta_{milk}) = \hat\beta_{milk}/t_{milk} = 0.0775/3.670$$

$$B9 = P - value(22) \approx 0.000$$

2. (a) Находим t-статистики: $t_c = 80/2.83 = 28.3$, $t_{beer} = -3/1 = -3$. Если предположить нормальность ошибок, то $t_{crit} = 2.05$. Следовательно, в обоих случаях H_0: $\beta = 0$ отвергается и оба коэффициента значимо отличны от нуля.
   (b)

$$\hat Y_i = 80 - 3 \cdot 4 = 80 - 12 = 68$$

   (c) Для уровня доверия 90% получаем критическое значение $t_{crit} = 1.7$. Отсюда доверительный интервал равен

$$[-3 - 1 \cdot 1.7; -3 + 1 \cdot 1.7]$$

3. Решение изложено в лекциях

4. Решение изложено в лекциях

5. (a) Несмещённая оценка

   Оценка $\hat{\theta}$ называется несмещённой, если $\mathbb{E}(\hat{\theta}) = \theta$

   (b) Эффективная оценка

   Оценка $\hat{\theta}$ называется эффективной среди множества оценок Θ, если для любой оценки $\hat{\theta}'$ из множества Θ выполнено неравенство $\mathrm{Var}(\hat{\theta}) \leqslant \mathrm{Var}(\hat{\theta}')$

   (c) Состоятельная последовательность оценок

   Последовательность оценок $\hat{\theta}_1$, $\hat{\theta}_2$, ..., называется состоятельной, если

$$\lim_{n\to\infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1$$

   для любого числа $\varepsilon > 0$.

8.4. Контрольная работа-1, 24.10.2016

1. В течение 10 дней Василий записывал количество пойманных им покемонов, Y_i, и количество решённых задач по эконометрике, X_i. Оказалось, что $\sum X_i^2 = 44$, $\sum Y_i^2 = 197$, $\sum X_i = 15$, $\sum Y_i = 15$ и $\sum X_i Y_i = 44$. Василий предполагает корректность линейной модели $Y_i = \beta_1 + \beta_2 X_i + u_i$.

   (a) Найдите МНК-оценки коэффициентов регресси

   (b) Найдите RSS, ESS, TSS и R^2

2. Для модели $Y_i = \beta_1 + \beta_2 X_i + u_i$ выполнены все предпосылки теоремы Гаусса-Маркова, а случайные ошибки нормально распределены. Известны все значенения Y_i, все значения $\hat{Y}_i$ и часть значений X_i

X_i	5	3	.	.
Y_i	4	7	7	2
$\hat{Y}_i$	5	7	4	4

   (a) Найдите МНК-оценки коэффициентов регрессии

   (b) Найдите стандартную ошибку коэффициента $\hat{\beta}_2$

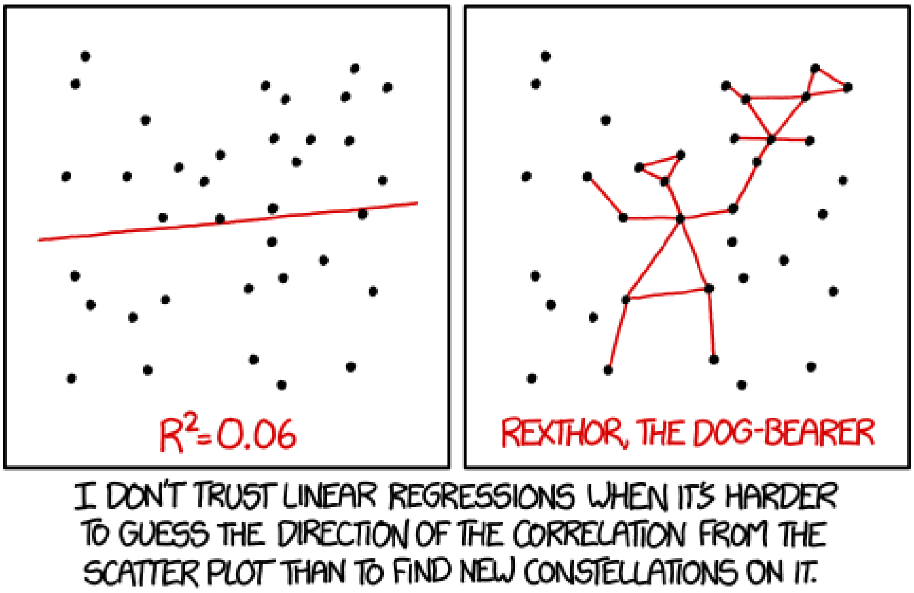
   (c) Постройте 95%-ый доверительный интервал для коэффициента $\hat{\beta}_2$

   (d) Проверьте гипотезу о незначимости коэффициента β_2 на уровне значимости 5%

3. Для модели $Y_i = \beta_1 + \beta_2 X_i + u_i$ выполнены все предпосылки теоремы Гаусса-Маркова. Докажите несмещённость МНК-оценки коэффициента β_1.

4. Для модели $Y_i = \beta_1 + \beta_2 X_i + u_i$ выполнены все предпосылки теоремы Гаусса-Маркова. Выведите формулу для дисперсии МНК-оценки, $\mathrm{Var}(\hat{\beta}_1)$.

5. Рассмотрим модель $Y_i = \beta_1 + \beta_2 X_i + u_i$ с неслучайным регрессором. Аккуратно сформулируйте теорему Гаусса-Маркова, пояснив смысл используемых понятий



Randall Munroe, xkcd

8.5. Демо брутальной части

1. Убедитесь, что в рамках классической модели $y = X\beta + u$ и предпосылок $\mathbb{E}(u) = 0$, $\mathrm{Var}(u) = \sigma^2 I$ умеете находить любые $\mathbb{E}()$, $\mathrm{Var}()$ и $\mathrm{Cov}(,)$ для векторов β, $\hat{\beta}$, y, $\hat{y}$, u, $\hat{u}$.

2. Вектор u размера 4×1 имеет стандартное нормальное распределение, $u \sim \mathcal{N}(0; I)$. Известен вектор d, $d = (1, -1, 2, -2)$.

   (a) Найдите такую матрицу H, что её умножение на произвольный вектор y означает проецирование вектора y на прямую порождённую вектором d

   (b) Как распределена случайная величина $u'Hu$? Чему равно её математическое ожидание и дисперсия?

3. Вектор u имеет стандартное нормальное распределение, $u \sim \mathcal{N}(0; I)$. Матрица A такова, что Au также имеет стандартное нормальное распределение, $Au \sim \mathcal{N}(0; I)$.

   (a) Выпишите уравнение, которому подчиняется матрица A

   (b) Чему может равняться $\det A$?

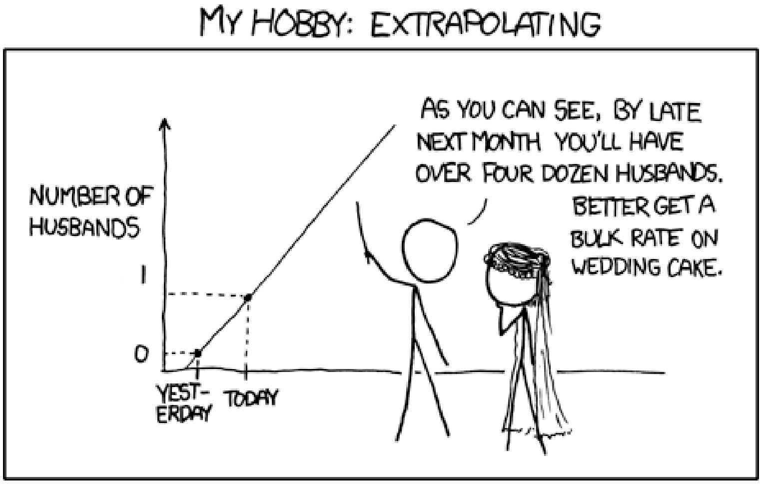
   (c) Рассмотрим c_1 и c_2 — первый и второй столбец матрицы A. Найдите $c_1'c_1$ и $c_1'c_2$

4. Предположим, что функция $f(x) = x'Ax + Bx + c$ имеет минимум. Найдите его, используя технику дифференциирования по вектору

5. Рассмотрим систему уравнений $X\beta = y$. Здесь y — известный вектор размера $n \times 1$, β — неизвестный вектор размера $k \times 1$, и X — известная матрица размера $n \times k$ полного ранга. Мы хотим решить эту систему относительно β. Если $n = k$, то решать эту систему скучно, и, конечно, $\beta = X^{-1}y$. Гораздо интереснее решать систему, когда решений нет или когда их бесконечно много :)

   (a) Если решений нет, то найдите наилучшее приближение к решению, то есть такое β при котором длина $(y - X\beta)$ минимальна.

   (b) Если решений, β, бесконечно много, то найдите решение с наименьшей длиной.



Randall Munroe, xkcd

8.6. Битва под Малоярославцем, 24.10.2016/24.10.1812

1. Цель этой кампании — доказать теорему Гаусса-Маркова для множественной регрессии. Итак, $y = X\beta + u$, про случайные ошибки u известно, что $\mathbb{E}(u) = 0$, $\mathrm{Var}(u) = \sigma^2 \cdot I$. Поехали! :)

   Тульский дворянин Дмитрий Сергеевич Дохтуров использует классическую МНК-оценку, $\hat{\beta}_{OLS}$:

   (a) Вспомните формулу для МНК-оценки $\hat{\beta}_{OLS}$
   (b) Является ли оценка $\hat{\beta}_{OLS}$ линейной по y?
   (c) Докажите, что оценка $\hat{\beta}_{OLS}$ является несмещённой

   Корсиканец Наполеон Бонапарт предлагает альтернативную несмещённую оценку $\hat{\beta}_{alt} = A \cdot y$:

   (d) Является ли оценка $\hat{\beta}_{alt}$ линейной по y?
   (e) Чему равняется матрица AX? Да поможет здесь условие несмещённости!
   (f) Найдите $\mathrm{Cov}(\hat{\beta}_{alt}, \hat{\beta}_{OLS})$ и $\mathrm{Cov}(\hat{\beta}_{OLS}, \hat{\beta}_{alt})$.
   (g) Полученное в предыдущем пункте выражение должно вызывать ностальгию, так как очень похоже на... На что?

   Теперь пора посмотреть на разницу $\hat{\beta}_{alt} - \hat{\beta}_{OLS}$:

   (h) Вспомните или выведите формулу для $\mathrm{Var}(r + s)$, где r и s — случайные векторы одинаковой длины
   (i) Докажите, что $\mathrm{Var}(\hat{\beta}_{alt} - \hat{\beta}_{OLS}) = \mathrm{Var}(\hat{\beta}_{alt}) - \mathrm{Var}(\hat{\beta}_{OLS})$
   (j) Рассмотрим матрицу $C = \mathrm{Var}(\hat{\beta}_{alt}) - \mathrm{Var}(\hat{\beta}_{OLS})$. Что находится на её диагонали?
   (k) Является ли матрица C симметричной?
   (l) Докажите, что матрица C является положительно полуопределённой. Если кто забыл, то это означает, что для любого вектора a выполнено неравенство $a'Ca \geqslant 0$
   (m) Докажите, что диагональные элементы матрицы C не меньше нуля

2. Вектор u размера 3×1 имеет стандартное нормальное распределение, $u \sim \mathcal{N}(0; I)$. Дана матрица D:

$$D = \frac{1}{14}\begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{pmatrix}$$

(a) Является ли матрица D симметричной? Идемпотентной?

(b) Найдите собственные числа матрицы D с учётом кратности

(c) Как распределена случайная величина $u'Du$?

(d) Аккуратно объясните, в чём состоит геометрический смысл умножения произвольного вектора y на матрицу D?

3. Найдите величины ESS, RSS, TSS и R^2 для регрессии $y_i = \mu + u_i$

4. В прошлом году, в курсе теории вероятностей и математической статистики, использовалось без доказательства следующее утверждение:

   Если случайные величины y_i независимы и нормально распределены $y_i \sim \mathcal{N}(\mu; \sigma^2)$, то $q = \sum_{i=1}^{n}(y_i - \bar{y})^2/\sigma^2$ имеет хи-квадрат распределение с $(n-1)$ степенью свободы.

   Настала пора вернуть долг чести и доказать это утверждение :)

   (a) Рассмотрим вектор центрированных y, то есть такой вектор $\tilde{y}$, что $\tilde{y}_i = y_i - \bar{y}$. Представьте вектор $\tilde{y}$ в виде $\tilde{y} = Ay$. Как выглядит матрица A?

   (b) Является ли матрица A симметричной? Идемпотентной? Найдите все её собственные числа с учётом кратности.

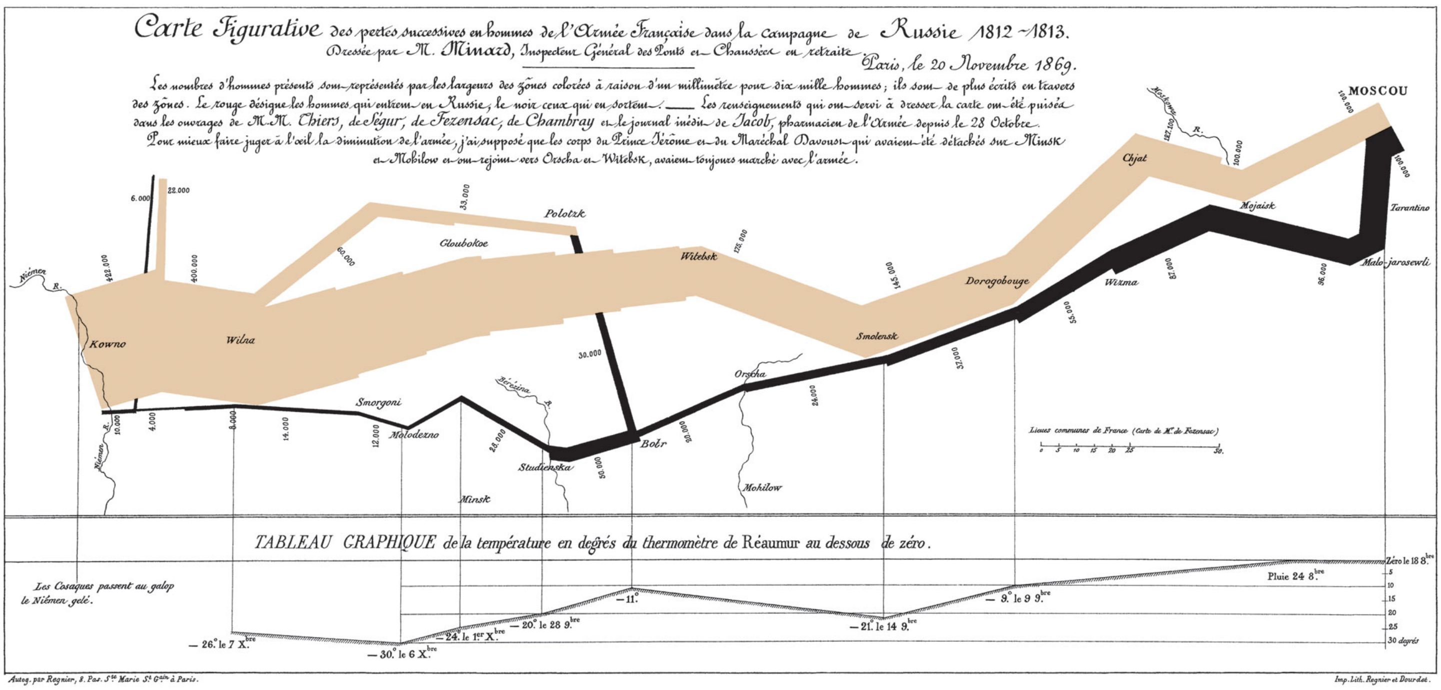
   (c) Представьте скаляр q в виде $q = \frac{1}{\sigma^2}\tilde{y}'B\tilde{y}$. Как выглядит матрица B?

   (d) Представьте скаляр q в виде $q = \frac{1}{\sigma^2}y'Cy$. Как выглядит матрица C?

   (e) Представьте скаляр q в виде $q = u'Du$, где вектор $u \sim \mathcal{N}(0; I)$. Как выглядит матрица D?

   (f) Является ли матрица D симметричной? Идемпотентной? Найдите все её собственные числа с учётом кратности.

   (g) Сформулируйте теорему о законе распределение квадратичной формы нормальных случайных величин и верните долг чести.



Charles Joseph Minard, Схема потерь наполеоновской армии в компании 1812-1813 годов

8.7. Контрольная работа-1. Решение брутальной части

1. (25 points)

   (a) (1)

   (b) (1)

   (c) (2)

   (d) (2) Является, каждый элемент вектора оценок, это линейная комбинация значений y.

   (e) (2)
   $$\mathbb{E}(\hat{\beta}_{alt}) = \mathbb{E}(Ay) = A\mathbb{E}(y) = AX\beta = \beta \Rightarrow AX = I$$

   (f) (2)
   $$\operatorname{Cov}(Ay, (X'X)^{-1}X'y) = A\operatorname{Cov}(y,y)X(X'X)^{-1} = \sigma^2 AX(X'X)^{-1} = \sigma^2(X'X)^{-1}$$
   $$\operatorname{Cov}((X'X)^{-1}X'y, Ay) = (X'X)^{-1}X'\operatorname{Cov}(y,y)A' = \sigma^2(X'X)^{-1}(AX)' = \sigma^2(X'X)^{-1}$$

   (g) (2) На ковариационную матрицу $\hat{\beta}_{OLS}$!

   (h) (2)
   $$\operatorname{Var}(r+s)$$
   Понятно, что на диагонали такой матрицы будут находится элементы следующего вида
   $$\operatorname{Var}(r_i + s_i) = \operatorname{Var}(r_i) + \operatorname{Var}(s_i) + 2\operatorname{Cov}(r_i, s_i)$$
   Все остальные элементы
   $$\operatorname{Cov}(r_i + s_i, r_j + s_j) = \operatorname{Cov}(r_i, r_j) + \operatorname{Cov}(r_i, s_j) + \operatorname{Cov}(s_i, r_j) + \operatorname{Cov}(s_i, s_j)$$
   Первое слагаемое принадлежит матрице $\operatorname{Var}(r)$, третье – $\operatorname{Var}(s)$, второе – $\operatorname{Cov}(r,s)$, четвертое – $\operatorname{Cov}(s,r)$. Получаем
   $$\operatorname{Var}(r+s) = \operatorname{Var}(r) + \operatorname{Var}(s) + \operatorname{Cov}(r,s) + \operatorname{Cov}(s,r)$$

   (i) (1)
   $$\operatorname{Var}(\hat{\beta}_{alt} - \hat{\beta}_{OLS}) = \operatorname{Var}(\hat{\beta}_{alt}) + \operatorname{Var}(\hat{\beta}_{OLS}) - 2\operatorname{Cov}(\hat{\beta}_{alt}, \hat{\beta}_{OLS}) =$$
   $$\operatorname{Var}(\hat{\beta}_{alt}) + \operatorname{Var}(\hat{\beta}_{OLS}) - 2\operatorname{Var}(\hat{\beta}_{OLS}) = \operatorname{Var}(\hat{\beta}_{alt}) - \operatorname{Var}(\hat{\beta}_{OLS})$$

   (j) (2) На диагонали находится разница в дисперсиях коэффициентов.

   (k) (2)
   $$\operatorname{Var}(Ay) = A\operatorname{Var}(y)A' = \sigma^2 AA'$$
   Поэтому C симметрична, как разница симметричных матриц.

   (l) (4) Константа не влияет на определенность матрицы, поэтому докажем этот факт без неё.
   $$a'Ca = a'(AA' - (X'X)^{-1})a = a'(AA' - AX(X'X)^{-1}X'A')a = a'A(I-H)A'a =$$
   $$a'A(I-H)(I-H)'A'a = ||(I-H)'A'a||_2^2 \geqslant 0$$

   (m) (2) В матрице C на диагонали находятся дисперсии разности оценок, а дисперсия всегда неотрицательна.

2. (10)

(a) (2) Является симметричной и идемпотентной

(b) (3) Можем посчитать по известной формуле, но лучше не надо :) Заметим, что ранг равен 1. Значит у нас 2 нулевых собственных значения. Матрица идемпотентна, значит третье собственное значение равно 1.

(c) (3) Матрица D представима в следующим виде:

$$\frac{1}{14}(1,2,3)'(1,2,3) \Rightarrow u'Du = \frac{1}{14}u'(1,2,3)'(1,2,3)u$$

$u'(1,2,3)'$ – это сумма нормально распределенных случайных величин с некоторыми коэффициентами, такая сумма тоже имеет нормальное распределение с матожиданием 0 и дисперсией равной 14. Нормируя на корень из 14, мы получаем стандартную нормальную случайную величину. Значит

$$\frac{1}{14}u'(1,2,3)'(1,2,3)u \sim \chi^2_1$$

(d) (2) Это проекция на пространство строк/столбцов матрицы D. В частности, на прямую, порожденную вектором (1,2,3).

3. (7)

Оценка МНК даст $\mu = \overline{y}$.

$$ESS = \sum(\hat{y}_i - \overline{y})^2 = \sum(\overline{y} - \overline{y})^2 = 0$$

$$RSS = \sum(\hat{y}_i - y_i)^2 = \sum(\overline{y} - y_i)^2 = TSS$$
$$R^2 = 0$$

4. (16)

(a) (2) Пусть G – матрица из одних единиц. Её смысл в том, что она будет суммировать элементы вектора, который был умножен на неё. Тогда

$$A = I - \frac{1}{n}G$$

(b) (3)
$$A' = A$$
$$(I - \frac{1}{n}G)(I - \frac{1}{n}G) = I - \frac{1}{n}G - \frac{1}{n}G + \frac{1}{n^2}GG$$
Легко видеть, что GG – это матрица, где каждый элемент равен n. Получаем

$$A^2 = A$$

Заметим, что матрица $\frac{1}{n}G$ также идемпотентна. Значит её собственные значения равны единицам и нулям. Легко также видеть, что ранг равен 1. Значит у этой матрицы одно собственное значение равное 1, и все остальные равны 0. Умножая матрицу на -1, мы меняем знак собственных значений. А прибавляя единичную матрицу, мы увеличиваем все собственные значения на 1. Значит мы имеем $n-1$ единичное собственное значение, и одно собственное значение равное 0.

(c) (1)
$$B = I$$

(d) (3) Чтобы перейти к равенству предыдущего пункта, нужно применить преобразование A на вектора y. Тогда

$$q = \frac{1}{\sigma^2}\tilde{y}'\tilde{y} = \frac{1}{\sigma^2}y'A'Ay \Rightarrow C = A'A$$

(e) (3) Заметим, что деление y на σ нормирует вектор y.

$$\frac{1}{\sigma^2}y'A'Ay = \frac{1}{\sigma^2}y'A'Ay = \left(\frac{\sigma u + \mu}{\sigma}\right)' A'A\left(\frac{\sigma u + \mu}{\sigma}\right) = (u + \frac{\mu}{\sigma})'A'A(u + \frac{\mu}{\sigma}) =$$

Так как A центрирует переменные, то $A\frac{\mu}{\sigma} = 0$, следовательно

$$(u + \mu)'A'A(u + \mu) = u'A'Au$$

$$D = A'A = AA = A$$

(f) (2) Аналогично пункту б)

(g) (3) Осталось понять, какое распределение имеет $u'Au = u'(I - \frac{1}{n}G)u$. Как известно, матрицу G можно представить в следующем виде

$$\frac{1}{n}G = S\Lambda S'$$

Где у матрицы лямбда все элементы 0 кроме элемента $S_{1,1}$. Пусть s – это собственный вектор, который соответствует собственному значению 1, тогда

$$u'\frac{1}{n}Gu = u'ss'u$$

Понятно, что собственный вектор с собственным значением 1, это вектор констант, но так как в данном разложении матрица S ортогональна, то длина собственного вектора s равна 1. Тогда все элементы вектора s равны $\frac{1}{\sqrt{n}}$.

$$u'ss'u = \left(\sum_{i=1}^{n}\frac{u_i}{\sqrt{n}}\right)^2$$

$$\sum_{i=1}^{n}\frac{u_i}{\sqrt{n}} \sim \mathbb{N}(0,1) \Rightarrow u'ss'u \sim \chi^2_1$$

$$u'u \sim \chi^2_n \Rightarrow q = u'Au \sim \chi^2_{n-1}$$