

Прикладная статистика в машинном обучении

Домашнее задание #1

Часть 1

Дедлайн: 12 ноября, 23:59 МСК

Правила игры

1. Домашнее задание состоит из двух частей. Часть 1 содержит 13 обязательных и две бонусные задачи и предполагает решение «от руки». Часть 2 содержит 3 обязательных задачи и предполагает программное решение.
2. Домашнее задание оценивается в 80 баллов. При этом часть 1 оценивается в 65 баллов, а часть 2 – в 15 баллов. По умолчанию за каждый пункт каждой задачи можно получить 1 балл. Однако за некоторые пункты некоторых задач можно получить другое количество баллов, которое явно указано в скобках рядом с меткой пункта.
3. Каждый пункт оценивается с промежутком 0.5. Например, если за пункт можно получить максимум 1 балл, то за полностью корректное решение ставится 1 балл, за решение с небольшими ошибками – 0.5 балла, за решение с серьёзными ошибками или неправильное решение – 0 баллов. Для пунктов, за которые можно получить максимум 2 балла, в зависимости от решения можно получить 2, 1.5, 1 и т.д. баллов. При этом пункты проверяются независимо друг от друга: если пункт $t + 1$ зависит от численных результатов пункта t , и в пункте t допускается ошибка, из-за которой в пункт $t + 1$ приходят неверные входные данные, то при корректном решении пункт $t + 1$ оценивается в максимальное количество баллов, которое можно за него получить.
4. Бонусные задачи X и Y приведены в конце части 1 и обозначены значком †. Эти задачи необязательны к решению и учитываются сверх установленных 80 баллов. Баллы за корректно решённые бонусные задачи прибавляются к набранным баллам, даже если в сумме получается больше 80 баллов (оценка за домашнюю работу в этом случае будет больше 10, и так и будет внесена в таблицу с оценками).
5. Весь код должен быть написан на Python, R, C или C++.
6. Решения принимаются до 12 ноября 2021 года, 23:59 МСК включительно. Работы, отправленные после дедлайна, проверяются, но не оцениваются.
7. Все решения нужно загрузить в личный репозиторий на [GitHub Classroom](#).
8. Репозиторий должен содержать: PDF-файл с решениями задач части 1 и файл с кодом с решениями задач части 2. Решение задач части 1 можно набрать в любом электронном редакторе или написать от руки, а затем сделать качественный скан. Все решения должны быть расположены в правильном порядке в одном файле. Файлы должны быть названы по типу «name_surname_group_hw1_part1.pdf» и «name_surname_group_hw1_part2.ext», где вместо ext может быть .py, .ipynb, .R, .c, .cpp. Если решение части 2 разбивается на несколько файлов кода, то в репозиторий нужно загрузить все файлы, а в README.md подробно указать, что содержит каждый файл.
9. Разрешается использовать без доказательства любые результаты, встречавшиеся на лекциях или семинарах по курсу, если получение этих результатов не является вопросом задания. Разрешается использовать любые свободные источники с указанием ссылки на них.
10. Плагиат не допускается. При обнаружении случаев списывания, 0 за работу выставляется всем участникам нарушения, даже если можно установить, кто у кого списал.

Задача 1. Просто компания

Компания «Напиши-ка» производит три вида ручек: синие, красные и зелёные. Глава аналитического отдела компании Даниил хочет понять, какая из ручек скорее всего «выстрелит», а какая не будет пользоваться успехом у покупателей. Для этого он анализирует выборку в 300 проданных ручек. Оказалось, что из них 150 синих, 100 красных и 50 зелёных ручек. Даниил уверен, что ручки продаются независимо друг от друга, и вероятность того, что будет продана синяя ручка, равна p_1 , а что красная p_2 .

- [a] Обозначим $p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$. Найдите \hat{p}_{ML} интуитивно, не выписывая правдоподобие, и поясните, как вы это сделали.

Решение: Очевидная оценка может быть построена через однозначное определение вероятностей, основывающиеся на значениях выборки:

$$\hat{p}_{ML} = \begin{pmatrix} \frac{150}{300} \\ \frac{100}{300} \end{pmatrix} = \begin{pmatrix} 0.5 \\ \frac{1}{3} \end{pmatrix}$$

- [6] Выпишите функцию правдоподобия и найдите \hat{p}_{ML} как точку её глобального максимума.

Решение:

$$\begin{aligned} L &= p_1^{150} \cdot p_2^{100} \cdot (1 - p_1 - p_2)^{50} \rightarrow \ell = 150 \ln p_1 + 100 \ln p_2 + 50 \ln(1 - p_1 - p_2) \\ \begin{cases} \ell'_{p_1} = \frac{150}{p_1} - \frac{50}{1-p_1-p_2} \\ \ell'_{p_2} = \frac{100}{p_2} - \frac{50}{1-p_1-p_2} \end{cases} &= \begin{cases} \ell'_{p_1} = 150(1 - p_1 - p_2) - 50p_1 \\ \ell'_{p_2} = 100(1 - p_1 - p_2) - 50p_2 \end{cases} = \begin{cases} 3 - 3\hat{p}_2 = 4\hat{p}_1 \\ 2 - 2\hat{p}_1 = 3\hat{p}_2 \end{cases} = \\ &= \begin{cases} \hat{p}_1 = \frac{1}{2} \\ \hat{p}_1 = \frac{1}{3} \end{cases} \implies \hat{p}_{ML} = \begin{pmatrix} 0.5 \\ \frac{1}{3} \end{pmatrix} \end{aligned}$$

- [b] Проверьте гипотезу

$$\begin{cases} H_0 : p_1 = 0.2, \\ H_A : p_1 \neq 0.2 \end{cases}$$

на уровне значимости 5% при помощи тестов LR и LM .

Решение:

$$LR = 2 \cdot (\ell(\hat{\theta}) - \ell(\theta_0))$$

Определим значения каждого из элементов теста

$$\begin{aligned} \ell(\hat{\theta}) &= 150 \ln 0.5 + 100 \ln \frac{1}{3} + 50 \ln \frac{1}{6} \\ \ell(\theta_0) &= 150 \ln 0.2 + 100 \ln p_{2R} + 50 \ln(0.8 - p_{2R}) \longrightarrow \\ \longrightarrow \ell'_{p_{2R}} &= \frac{100}{p_{2R}} - \frac{50}{0.8 - p_{2R}} \longrightarrow \hat{p}_{2R} = \frac{80}{150} = \frac{8}{15} \longrightarrow \ell(\theta_0) = 150 \ln 0.2 + 100 \ln \frac{8}{15} + 50 \ln \frac{4}{15} \end{aligned}$$

Получим итоговый ответ

$$LR = 2(150 \ln 0.5 + 100 \ln \frac{1}{3} + 50 \ln \frac{1}{6} - 150 \ln 0.2 - 100 \ln \frac{8}{15} - 50 \ln \frac{4}{15}) = 300 \ln 2.5 + 300 \ln \frac{5}{8} = 300 \ln \frac{25}{16} \approx 134$$

Следовательно, так как $134 > 3.84$, гипотеза H_0 отвергается

$$LM = \frac{\ell'(\theta_0)^2}{\text{Var}(\ell'(\theta_0))}$$

Определим значения каждого из элементов теста

$$\ell'(\theta_0) = \frac{150}{0.2} - \frac{50}{\frac{4}{15}} = 750 + \frac{750}{4} = 562.5$$

$$\text{Var}(\ell'(\theta_0)) = \mathbb{E}(-\ell''(\theta_0)) = -\frac{150}{0.04} + \frac{50}{\frac{16}{225}} = 150 \cdot 25 + \frac{225 \cdot 50}{16} = 4453.125$$

Получим итоговый ответ

$$LM = \frac{(562.5)^2}{4453.125} \approx 71$$

Следовательно, так как $71 > 3.84$, гипотеза H_0 отвергается

[г] Проверьте гипотезу

$$\begin{cases} H_0 : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix}, \\ H_A : \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \neq \begin{pmatrix} 0.3 \\ 0.2 \end{pmatrix} \end{cases}$$

на уровне значимости 5% при помощи тестов LR и W .

Решение:

• $LR = 2 \cdot (\ell(\hat{\theta}) - \ell(\theta_0))$

Определим значения каждого из элементов теста

$$\ell(\hat{\theta}) = 150 \ln 0.5 + 100 \ln \frac{1}{3} + 50 \ln \frac{1}{6}$$

$$\ell(\theta_0) = 150 \ln 0.3 + 100 \ln 0.2 + 50 \ln 0.5$$

Получим итоговый ответ

$$\begin{aligned} LR &= 2(150 \ln 0.5 + 100 \ln \frac{1}{3} + 50 \ln \frac{1}{6} - 150 \ln 0.3 - 100 \ln 0.2 - 50 \ln 0.5) = \\ &= 2(150 \ln \frac{5}{3} + 100 \ln \frac{5}{3} + 50 \ln \frac{1}{3}) = 500 \ln \frac{5}{3} + 100 \ln \frac{1}{3} \approx 146 \end{aligned}$$

Следовательно, так как $146 > 6.0$, гипотеза H_0 отвергается

• $W = (\hat{\theta} - \theta_0)^T \text{Var}(\hat{\theta})(\hat{\theta} - \theta_0)$

Определим значения каждого из элементов теста

$$\hat{\theta} - \theta_0 = \begin{pmatrix} 0.2 \\ \frac{2}{15} \end{pmatrix}$$

$$\text{Var}^{-1}(\hat{\theta}) = I_F(\hat{\theta}) = \mathbb{E}(-H) \longrightarrow$$

$$\ell''_{p_1} = -\frac{150}{0.25} - \frac{50}{\frac{1}{36}}$$

$$\ell''_{p_2} = -\frac{100}{\frac{1}{9}} - \frac{50}{\frac{1}{36}}$$

$$\ell''_{p_1 p_2} = -\frac{50}{\frac{1}{36}} \longrightarrow \mathbb{E}(-H) = \begin{pmatrix} 600 + 1800 & 1800 \\ 1800 & 900 + 1800 \end{pmatrix}$$

Получим итоговый ответ

$$W = \begin{pmatrix} 0.2 & \frac{2}{15} \\ 1800 & 2700 \end{pmatrix} \begin{pmatrix} 2400 & 1800 \\ 1800 & 2700 \end{pmatrix} \begin{pmatrix} 0.2 \\ \frac{2}{15} \end{pmatrix} = 240$$

Следовательно, так как $240 > 6.0$, гипотеза H_0 отвергается

- [д] Постройте график логарифма правдоподобия в трёхмерной плоскости. Покажите на графике \hat{p}_{ML} визуальную интерпретацию тестов LR и W для гипотезы из предыдущего пункта.

Решение: <https://www.math3d.org/mztOg9AZk>

- [е] Постройте 95%-ый доверительный интервал для p_3 .
[ж] Постройте 99%-ый доверительный интервал для $p_1 + p_2$.
[з] Постройте 90%-ый доверительный интервал для \hat{p}_1 .
Подсказка: помните, что мы работаем в рамках частотного подхода.
[и] Приведите разумное интерпретируемое определение того, что ручка «выстрелила».
[к] Пользуясь определением из предыдущего пункта, сформулируйте гипотезу о том, что «выстрелит» ручка синего цвета и проверьте её при помощи любого из тестов LR , LM или W на уровне значимости 5%.

Задача 2. Анекдоточная

Станислав знает, что хороший анекдот должен быть не очень коротким, но и не слишком длинным. Время, за которое Станислав произносит один анекдот, – это непрерывная случайная величина с плотностью

$$f(x|b) = \begin{cases} \frac{2x}{b} e^{-\frac{x^2}{b}}, & \text{если } x > 0, \\ 0, & \text{иначе,} \end{cases}$$

где b – некоторый параметр. Станислав собрал случайную выборку по продолжительности рассказанных им анекдотов: X_1, X_2, \dots, X_n , где $n = 10^6$. Оказалось, что $\sum X_i^2/n = 20$, $\sum X_i/n = 2$.

- [а] Найдите \hat{b}_{ML} .

Решение:

$$\begin{aligned} L &= \prod \frac{2x}{b} e^{-\frac{x^2}{b}} \longrightarrow \ell = \sum_{x=1}^n -\frac{x^2}{b} + \ln\left(\frac{2x}{b}\right) = -\frac{20n}{b} + \sum_{x=1}^n \ln\left(\frac{2x}{b}\right) \\ \ell'_b &= \frac{20n}{b^2} - \sum_{x=1}^n \frac{1}{b} = \frac{20n - nb}{b^2} \longrightarrow 20n - n\hat{b} = 0 \longrightarrow \hat{b} = 20 \end{aligned}$$

- [б] Проверьте гипотезу

$$\begin{cases} H_0 : b = 3, \\ H_A : b \neq 3 \end{cases}$$

на уровне значимости 5% при помощи теста LR .

Решение:

$$LR = 2 \cdot (\ell(\hat{\theta}) - \ell(\theta_0))$$

Определим значения каждого из элементов теста

$$\begin{aligned} \ell(\hat{\theta}) &= -\frac{20n}{20} + \sum_{x=1}^n \ln\left(\frac{2x}{20}\right) = -n + \sum_{x=1}^n \ln(0.1x) = -n - n \ln(10) + \sum_{x=1}^n \ln(x) \\ \ell(\hat{\theta}_0) &= -\frac{20n}{3} + \sum_{x=1}^n \ln\left(\frac{2x}{3}\right) = -\frac{20}{3}n + n \ln\left(\frac{2}{3}\right) + \sum_{x=1}^n \ln(x) \end{aligned}$$

Получим итоговый ответ

$$LR = 2(-n - n \ln(10) + \frac{20}{3}n - n \ln\left(\frac{2}{3}\right)) = 2\left(\frac{17}{3}n - n \ln\left(\frac{20}{3}\right)\right)$$

Следовательно, так как $7.54 \cdot 10^6 > 3.84$, гипотеза H_0 отвергается.

[в] Рассчитайте LM -статистику для проверки гипотезы

$$\begin{cases} H_0 : b = 1, \\ H_A : b \neq 1 \end{cases}$$

Чему приблизительно равно соответствующее p -значение?

Решение:

$$LM = \frac{\ell'(\theta_0)^2}{\hat{I}(\theta_0)}$$

Определим значения каждого из элементов теста

$$\ell'(\theta_0) = \frac{20n - n}{1} = 19n; \quad \hat{I}(\theta_0) = \mathbb{E}(-\ell''(\theta_0)_b) = 39n$$

Получим итоговый ответ

$$LM = \frac{361n}{39}$$

Следовательно, так как $\frac{361}{39} \cdot 10^6 > 3.84$, гипотеза H_0 отвергается.

[г] Проверьте гипотезу из предыдущего пункта, построив соответствующий доверительный интервал для b .

Задача 3. «Я не дерево. Я энт».

Исследователь Матвей подбрасывает монетку с вероятностью орла p до тех пор, пока не выпадет два орла (всего, не обязательно подряд). Он сыграл четыре игры, и оказалось, что первая завершилась за 3 хода, вторая – за 3 хода, третья – за 2 хода, четвёртая – за 4 хода. Будем считать, что подбрасывания в течение одной игры независимы. Также предположим, что игры происходили независимо друг от друга.

[а] (2 балла) Найдите \hat{p}_{ML} .

$$\begin{aligned} L &= \prod_{i=1}^4 \mathbb{P}(X = s_i) = \prod_{i=1}^4 C_{s_i-1}^1 p^2 (1-p)^{s_i-2} \longrightarrow \ell = \sum \ln C_{s_i-1}^1 + 2 \ln(p) + (s_i - 2) \ln(1-p) = \\ &= 8 \ln(p) + \ln C_2^1 + 1 \ln(1-p) + \ln C_2^1 + 1 \ln(1-p) + \ln C_1^1 + \ln C_3^1 + 2 \ln(1-p) = 8 \ln(p) + 4 \ln(1-p) + \ln(12) \longrightarrow \\ &\longrightarrow \ell' = \frac{8}{p} - \frac{4}{1-p} \longrightarrow \hat{p}_{ML} = \frac{2}{3} \end{aligned}$$

[б] Найдите \hat{a}_{ML} для нового параметра $a = (p^2 + 3p^3 - 1)$.

$$\hat{a} = \hat{a}(p) = a(\hat{p}) = \frac{4}{9} + 3 \frac{8}{27} - 1 = \frac{1}{3}$$

[в] (2 балла) Покажите, что \hat{p} является состоятельной оценкой p .

Подсказка: для решения Задачи X потребуется доказать, что если M – число ходов, за которое завершится игра, то $\mathbb{E}(M) = \frac{2}{p}$. В этой задаче можно пользоваться этим утверждением без доказательства. Итак: "Оценка называется состоятельной, если она сходится по вероятности к оцениваемому параметру θ при $n \rightarrow \infty$ ". Следовательно,

$$\begin{aligned} \ell &= \dots = 2 \ln(p) \cdot n + \sum \ln C_{s_i-1}^1 + (s_i - 2) \ln(1-p) \rightarrow \ell'_p = \frac{2}{p}n - \frac{\sum s_i - 2}{1-p} \longrightarrow \hat{p} = \frac{2n}{\sum s_i - 2} \\ \text{ЗБЧ} \implies \hat{p} &\xrightarrow{p} \mathbb{E}(\hat{p}) = \frac{2n}{\mathbb{E}(s_i)n} \stackrel{\text{из условия}}{=} \frac{2n}{n \cdot \frac{2}{p}} = p \end{aligned}$$

$$\hat{p} \xrightarrow{p} p$$

Задача 4. Полезное утверждение

Гарри никак не может понять, почему при большой информации Фишера оценки максимального правдоподобия лежат к истинному параметру ближе, чем при малой информации Фишера. Гермиона решает продемонстрировать аналитическую интуицию, стоящую за этим утверждением:

«Если взять выборку независимых одинаково распределённых случайных величин Y_1, \dots, Y_N , каждая из которых имеет функцию плотности или функцию вероятности $f(y|\theta)$, и предположить, что выполнены все необходимые условия регулярности, то при $\phi \rightarrow \theta$:

$$D_{KL}[f(y|\theta) \| f(y|\phi)] = \frac{1}{2} I_f(\theta)(\phi - \theta)^2 + O((\phi - \theta)^3).$$

- [a] (2 балла) Докажите утверждение Гермионы либо для случая функций плотности, либо для случая функций вероятности.

$$\begin{aligned} D_{KL}[f(y|\theta) \| f(y|\phi)] &= CE(f(y|\theta) \| f(y|\phi)) - CE(f(y|\theta) \| f(y|\theta)) = \mathbb{E} \ln(f(y|\theta)) - \mathbb{E} \ln(f(y|\phi)) = \\ &= \mathbb{E}(\ln(f(y|\theta)) - \ln(f(y|\phi))) \stackrel{\text{Тейлор}}{=} \mathbb{E}(\ln(f(y|\theta)) - \ln(f(y|\theta)) - \frac{d}{d\phi} \ln(f(y|\theta))(\phi - \theta) - \\ &\quad - 0.5 \frac{d^2}{d\phi^2} \ln(f(y|\theta))(\phi - \theta)^2 + O((\phi - \theta)^3)) \stackrel{\ell(y|\theta) = \ln(f(y|\theta))}{=} \\ &= \mathbb{E}(-\ell'_\phi((y|\theta))(\phi - \theta) - 0.5\ell''_\phi(y|\theta))(\phi - \theta)^2 + O((\phi - \theta)^3)) \stackrel{\mathbb{E} \ell'(y|\theta) = 0}{=} \\ &= -0.5 \mathbb{E}(\ell''_\phi(y|\theta))(\phi - \theta)^2 + O((\phi - \theta)^3)) = 0.5I_F(\theta)(\phi - \theta)^2 + O((\phi - \theta)^3)) \end{aligned}$$

- [б] (2 балла) Поясните Гарри, почему при большей информации Фишера ML-оценки лежат ближе к истинному параметру.

Очевидно, что информация Фишера, равная дисперсии производной логистической функции, тем самым отражает степень изменения значений производной при колебании y . Следовательно, задавая большую информацию Фишера, мы указываем на большие изменения производной и близость несмешенной оценки к истинному значению параметра (график 1). Напротив, если график нельзя назвать крутым, то есть касательные сигнализируют малые изменения в производной и небольшие значения I_f , это указывает на большее расстояние до искомого (график 2).

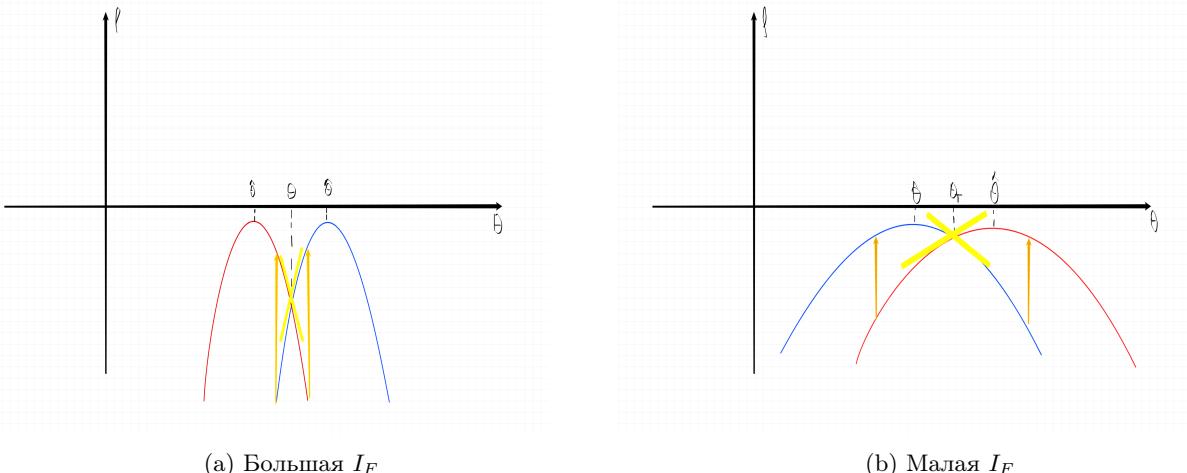


Рис. 1: Кейсы при разных информациониях Фишера

Подсказка: $H(f) = -\mathbb{E}(\ln f)$, аналогично для кросс-энтропии.

(По мотивам: Williams, Weighing the Odds)

Задача 5. Модель для зелий

Полумна хочет построить предсказательную модель, которая бы описывала зависимость популярности зелья y_i от силы его положительного влияния x_i . Обе величины являются количественными непрерывными переменными на \mathbb{R} . Предположим, что Полумна знает, как измерить популярность и силу влияния и верит, что искомая зависимость имеет следующий вид:

$$y_i = (\beta_1)^2 e^{-\beta_2 x_i} u_i,$$

где β_1 и β_2 – неизвестные положительные коэффициенты, u_i – случайная ошибка, причём $\ln u_i \sim \mathcal{N}(0, 2)$.

- [a] (2 балла) Введите любые разумные ограничения на переменные. Найдите $\hat{\beta}_1$ и $\hat{\beta}_2$ методом максимального правдоподобия.

Ограничения на переменные? Интересно, конечно. Позже этим займемся. Сейчас пока займемся оценкой:

$$\begin{aligned} \ln(u_i) &= \ln\left(\frac{y_i}{\beta_1^2 \cdot \exp^{-\beta_2 x_i}}\right) = \ln(y_i) - 2 \ln(\beta_1) + \beta_2 x_i \sim \mathcal{N}(0, 2) \longrightarrow \\ &\longrightarrow \arg \max \ell(\ln u_i) = \arg \max \sum \ln f(\ln u_i) = \arg \max_{b_1, b_2} \sum \ln f(y_i) \\ \ell &= \sum \ln\left(\frac{1}{\sigma \sqrt{2\pi}} \cdot \exp^{-\frac{(\ln(u_i) - \mu)^2}{2\sigma^2}}\right) = \sum \ln\left(\frac{1}{2\sqrt{\pi}} - \frac{(\ln(u_i))^2}{4}\right) = \sum -\ln(2\sqrt{\pi}) - \frac{(\ln(u_i))^2}{4} \\ \begin{cases} \ell'_{\beta_1} = \sum \frac{\ln(y_i) - 2 \ln(\beta_1) + \beta_2 x_i}{\beta_1} \\ \ell'_{\beta_2} = -\sum \ln(y_i) - 2 \ln(\beta_1) + \beta_2 x_i \end{cases} &\implies \begin{cases} \sum \ln(y_i) - 2n \ln(\hat{\beta}_1) + \sum \hat{\beta}_2 x_i = 0 \\ \sum \ln(y_i) x_i - \sum 2 \ln(\hat{\beta}_1) x_i + \sum \hat{\beta}_2 x_i^2 = 0 \end{cases} \implies \\ \implies \begin{cases} \ln(\hat{\beta}_1) = \frac{\sum \ln(y_i) + \sum \hat{\beta}_2 x_i}{2n} \\ \sum \ln(y_i) x_i - \sum 2 \ln(\hat{\beta}_1) x_i + \sum \hat{\beta}_2 x_i^2 = 0 \end{cases} &\longrightarrow \text{sum } \ln(y_i) x_i - \frac{\sum \ln(y_j) + \sum \hat{\beta}_2 x_j}{n} \sum x_i + \sum \hat{\beta}_2 x_i^2 = 0 \longrightarrow \\ &\longrightarrow \begin{cases} \hat{\beta}_1 = \exp^{\frac{\sum \ln(y_i) + \sum \hat{\beta}_2 x_i}{2n}} \\ \hat{\beta}_2 = \frac{\sum_j \ln(y_j) \sum_i x_i - \sum \ln(y_i) x_i}{\sum x_i^2 - \frac{\sum_i x_i x_j}{n}} \end{cases} \end{aligned}$$

- [б] (2 балла) Полумна собрала выборку, для которой оказалось, что

$$\begin{aligned} \sum_{i=1}^n \ln y_i &= 100, \quad \sum_{i=1}^n x_i = 50, \quad \sum_{i=1}^n x_i \ln y_i = 200, \\ \sum_{i=1}^n x_i^2 &= 2500, \quad \sum_{i=1}^n (\ln y_i)^2 = 10000, \quad \sum_{i=1}^n e^{-\hat{\beta}_2 x_i} = 1, \\ n &= 500. \end{aligned}$$

Проверьте гипотезу

$$\begin{cases} H_0 : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \\ H_A : \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \neq \begin{pmatrix} 1 \\ 2 \end{pmatrix} \end{cases}$$

на уровне значимости 5% при помощи теста W .

Подсказка: $\ln Y \sim N(\mu, \sigma^2) \Rightarrow \mathbb{E}(Y) = e^{\mu + \frac{\sigma^2}{2}}$.

$$\begin{aligned} \begin{cases} \hat{\beta}_1 = \exp^{\frac{100+50 \cdot \hat{\beta}_2}{2 \cdot 500}} \\ \hat{\beta}_2 = \frac{\frac{100 \cdot 50}{500} - 200}{2500 - \frac{50 \cdot 50}{500}} \end{cases} & \implies \begin{cases} \hat{\beta}_2 \approx 0.076 \\ \hat{\beta}_1 \approx 1.1 \end{cases} \\ \frac{d^2 \ell}{db_1^2} = -\frac{1}{\hat{\beta}_1^2} (\sum \ln(y_i) - 2 \ln(\beta_1) + \beta_2 x_i) - \frac{2n}{\hat{\beta}_1^2} = -\frac{1}{1.21} (100 - 1000 \ln(1.1) - 50 \cdot 0.076) - \frac{1000}{1.21} \approx -827.18 \\ \frac{d^2 \ell}{db_2^2} = -0.5 \sum x_i^2 = -1250 \\ \frac{d^2 \ell}{db_1 b_2} = \sum \frac{x_i}{\hat{\beta}_1} = 45.45 \end{aligned}$$

$$W = \begin{pmatrix} 0.1 \\ -2.076 \end{pmatrix}^T \cdot \begin{pmatrix} 827.18 & -45.45 \\ -45.45 & 1250 \end{pmatrix} \cdot \begin{pmatrix} 0.1 \\ -2.076 \end{pmatrix} = \frac{67679533}{12500} \approx 5414.36 > \chi^2_2 \implies \text{Отвергаем } H_0$$

Задача 6. Функции правдоподобия

Пусть X_1, \dots, X_n – выборка независимых одинаково распределённых величин из распределения с параметром $p \in [0, 1]$. Известно, что $n = 100$, $\bar{X} = 20$, $\sum X_i^2/n = 400$. Найдите \hat{p}_{ML} для следующих функций (можно либо вывести в явном виде, либо использовать математический анализ):

[a]

$$\begin{aligned} \ell(p) &= \frac{\sqrt{X_1 + \dots + X_n}}{50 - p} + \frac{\ln p}{X_1^2 + \dots + X_n^2}. \\ \ell'(p) &= \frac{\sqrt{X_1 + \dots + X_n}}{(50 - p^2)} + \frac{1}{p} \frac{1}{X_1^2 + \dots + X_n^2} \longrightarrow \frac{\sqrt{n \bar{X}}}{(50 - \hat{p}^2)} + \frac{1}{p} \frac{1}{400n} \longrightarrow \hat{p}^2 + (4 \cdot 10^5 \cdot \sqrt{20} - 100) \hat{p} + 2500 = 0 \longrightarrow \\ &\longrightarrow D = (8 \cdot 10^5 \cdot \sqrt{5} - 100)^2 - 10000 \approx 1788754.4 \longrightarrow \begin{cases} \hat{p}_1 = 0.001 \\ \hat{p}_2 = 1.79 \cdot 10^5 \end{cases} \end{aligned}$$

Таким образом, учитывая отрезок принадлежности p , ни одна из критических точек не подходит. Учитывая рост функции на отрезке значений (основаваясь на положительной производной), максимум достигается в правой границе:

$$\hat{p}_{ML} = 1$$

[б]

$$\ell(p) = \frac{(p^2 - \ln p) \sum X_i^2/n}{\bar{X}}.$$

$$\ell'(p) = \frac{(2p - \frac{1}{p}) \sum \frac{x_i^2}{n}}{\bar{X}} = 20(2p - p^{-1}) \longrightarrow p = \pm \sqrt{0.5}$$

Учитывая принадлежность p к отрезку значений, подходит $\sqrt{0.5}$ – минимум. В таком случае будем искать максимум:

$$\begin{cases} \lim_{p \rightarrow 0} \ell(p) = \infty \\ \ell(\sqrt{0.5}) \approx 17 \\ \ell(1) = 20 \end{cases} \implies \text{Максимум в левой границе}$$

$$\hat{p}_{ML} = 0$$

Задача 7. Дивергент

Рассмотрим распределения $p = \mathcal{N}(1, 2)$, $q = \text{Exp}(1)$, $r = \text{Bin}(3, 0.5)$. Для каждого пункта приведите математическое обоснование ответа.

[а] Найдите $D_{KL}(p\|q)$.

$$D_{KL}(p\|q) = \infty \Leftarrow \text{supp}(q) \subset \text{supp}(p)$$

[б] Найдите $D_{KL}(q\|p)$.

$$D_{KL}(q\|p) = CE(q\|p) - CE(q\|q)$$

$$CE(q\|q) = -\mathbb{E}_q \ln(q) = \int_0^\infty -e^{-x} \ln(e^{-x}) dx = (-e^{-x}x - e^{-x}) \Big|_0^\infty = 1$$

$$\begin{aligned} CE(q\|p) &= -\mathbb{E}_q \ln(p) = \int_0^\infty -e^{-x} \ln\left(\frac{1}{2\sqrt{\pi}} \cdot e^{-\frac{(x-1)^2}{4}}\right) dx = \ln(2\sqrt{\pi}) \int_0^\infty e^{-x} dx + 0.25 \int_0^\infty (x-1)^2 e^{-x} dx \\ &\quad 0.25 \int_0^\infty (x-1)^2 e^{-x} dx = 0.25 \left(\int_0^\infty x^2 e^{-x} dx - \int_0^\infty 2xe^{-x} dx + \int_0^\infty e^{-x} dx \right) = \\ &= \left[\begin{array}{l} \int_0^\infty x^2 e^{-x} dx = -e^{-x}x^2 \Big|_0^\infty + 2 \int_0^\infty e^{-x} dx = 2 \\ \int_0^\infty 2xe^{-x} dx = -2e^{-x}x \Big|_0^\infty + 2 \int_0^\infty e^{-x} dx = 2 \\ \int_0^\infty e^{-x} dx = 1 \end{array} \right] = 0.25 + \ln(2\sqrt{\pi}) \end{aligned}$$

$$D_{KL}(q\|p) = 0.25 + \ln(2\sqrt{\pi}) - 1 = \ln(2\sqrt{\pi}) - 0.75$$

[в] Найдите $D_{KL}(p\|r)$.

$$D_{KL}(p\|r) = \infty \Leftarrow \text{supp}(r) \subset \text{supp}(p)$$

[г] Найдите $D_{KL}(q\|r)$.

$$D_{KL}(q\|r) = \infty \Leftarrow \text{supp}(r) \subset \text{supp}(q)$$

[д] Возможно ли применить линейное преобразование к p , q или r так, чтобы ответ на хотя бы один из пунктов выше изменился?

Задача 8. Между молотом и наковальней

Одной из симметричных альтернатив KL -дивергенции является взаимная информация: для случайных величин X и Y она определяется как

$$I(X, Y) = H(X) - H(X|Y),$$

где $H(X|Y) = -\int p(x, y) \ln \frac{p(x, y)}{p(y)}$.

[а] Покажите, что $I(X, Y) = I(Y, X)$.

$$\begin{aligned} H(X) &= -\int p(x, x) \ln \frac{p(x, x)}{p(x)} dx; \quad H(Y) = -\int p(y, y) \ln \frac{p(y, y)}{p(y)} dy \\ H(X, Y) &= -\int p(x, y) \ln p(x, y) dx dy = -\left(\int p(x, y) \ln \frac{p(x, y)}{p(y)} dx dy + \int p(x, y) \ln p(y) dx dy \right) = \\ &= H(X\|Y) - \int \ln p(y) p(y) dy = H(X\|Y) + H(Y) \end{aligned}$$

Аналогичные преобразования позволяют получить симметричную формулу:

$$\begin{cases} H(X, Y) = H(X\|Y) + H(Y) \\ H(X, Y) = H(Y\|X) + H(X) \end{cases} \rightarrow I(X, Y) = H(X) - H(X\|Y) = \underbrace{H(X) - H(X, Y)}_{H(Y\|X)} + H(Y) = I(Y, X)$$

- [б] (2 балла) Покажите, что $I(X, Y) = D_{KL}(p(x, y) \| p(x) \times p(y))$.

$$\begin{aligned}
 D_{KL}(p(x, y) \| p(x) \times p(y)) &= CE(p(x, y) \| p(x) \times p(y)) - (p(x) \times p(y)) = \\
 &= \int p(x, y) \ln p(x, y) dx dy - \int p(x, y) (\ln(p(x)) + \ln(p(y))) dx dy = \\
 &= \int p(x, y) \ln p(x, y) dx dy - \int p(x) \ln p(x) dx - \int p(y) \ln p(y) dy = \underbrace{-H(X, Y) + H(X)}_{H(Y \| X)} + H(Y) = I(Y, X)
 \end{aligned}$$

- [в] Поясните интуитивную интерпретацию $I(X, Y)$.

Честно говоря, тяжеловато мне дается все эта статистика в абстрактных ее формах. Вот про данетки это да, это мне близко и понятно. В терминах этой игры $I(X, Y)$ подразумевает общее число вопросов, которое в среднем понадобится задать для получения итогового ответа. Другими словами, мы знаем Y и ищем X , а $I(X, Y)$ отражает кол-во информации, содержащиеся о первом во втором.

Задача 9. Хорошая задача на экзамен

Случайная величина X принимает значение 0 с вероятностью p , значение 1 с вероятностью $1/3$ и значение 2 с вероятностью $2/3 - p$.

- [а] Постройте график зависимости $H(X)$ как функцию от p .

Заметим, что имеем дело с задачей бинарной энтропии, так как параметр p влияет на два возможных сценария. Так же мы доопределяем функцию в граничных значениях.



Рис. 2: A simple caption

- [б] При каком p энтропия будет максимальна? Поясните полученный результат.

Что есть энтропия? Хм. Мера неопределенности! Бинго. Значит, надо выбрать p , которое бы обеспечивало наиболее равновероятные сценарии развития событий. У нас есть три возможных пути. p принадлежит от 0 до $\frac{2}{3}$. В таком случае, рационально будет выбрать $p = \frac{1}{3}$, тем самым создав "идеальный баланс".

Задача 10. Порисуем!

Рассмотрим модель множественной регрессии $y = X\beta + u$, которая оценивается при помощи МНК. Число наблюдений равно $n = 400$, число регрессоров равно $k = 10$, включая константный. Все регрессоры ортогональны друг другу.

- [а] Долорес Амбридж строит регрессию по константному и следующим за ним четырём регрессорам. Корнелиус Фадж строит регрессию по константному и оставшимся пятью регрессорам. Покажите на единой картинке МНК \hat{y} , TSS , ESS , RSS и R^2 в их регрессиях.
- [б] Альбус Дамблдор строит регрессию по всем 10 регрессорам. Покажите на той же картинке МНК \hat{y} , TSS , ESS , RSS и R^2 в его регрессии.
- [в] (2 балла) Гарри Поттер хочет сравнить регрессии Амбридж и Дамблдора при помощи F -теста. Напомним, что

$$F = \frac{(RSS_R - RSS_{UR})/(k_{UR} - k_R)}{RSS_{UR}/(n - k_{UR})}.$$

Покажите на картинке МНК RSS_R , RSS_{UR} и угол, квадрату тангенса которого пропорциональна F -статистика.

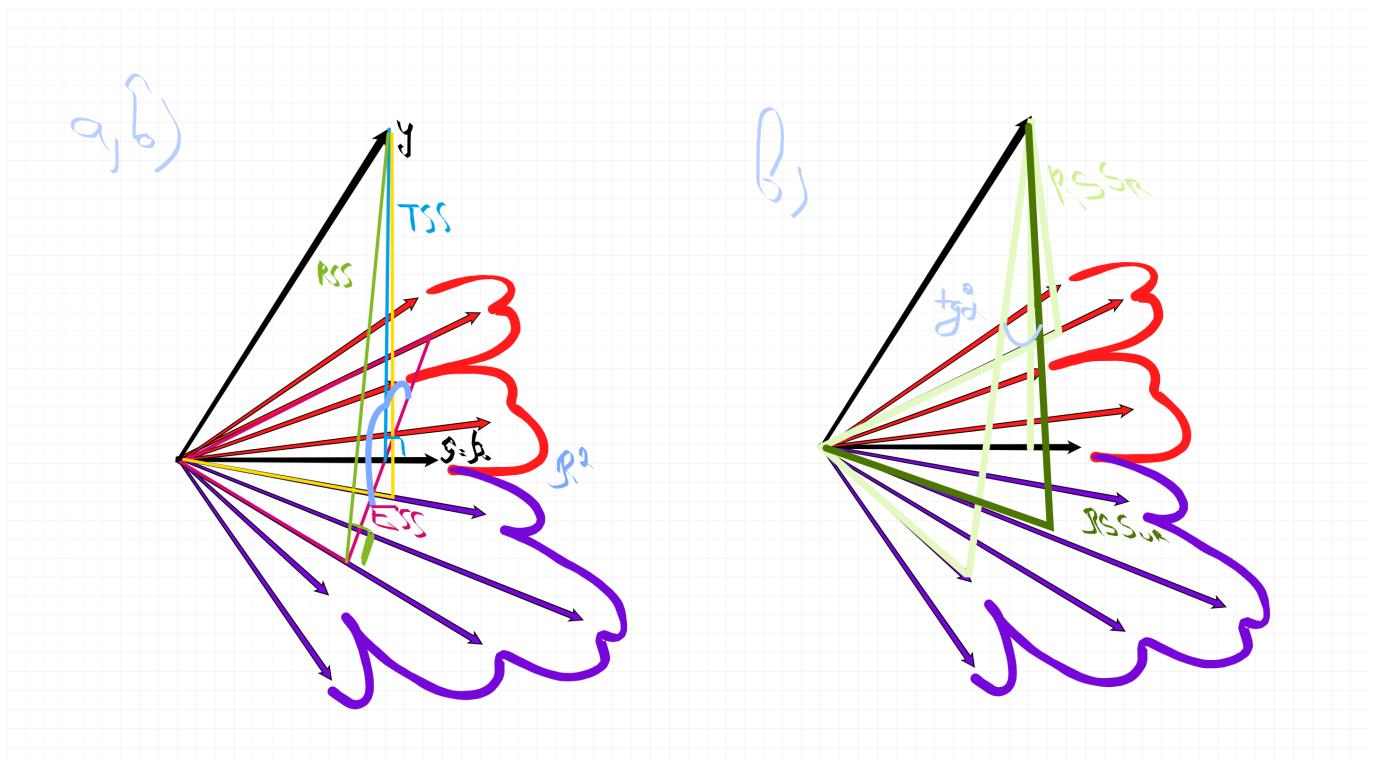


Рис. 3: Решение 10

- [г] Приведите геометрическую интерпретацию F -теста.

В целом, F тест указывает на разницу между построенной регрессией и моделью, не использующей независимые переменные. То есть насколько линия регрессии соответствует всем имеющимся данным.

Задача 11. Подпространства

Рассмотрим пространство \mathbb{R}^3 и два подпространства в нём

$$W = \{(x_1, x_2, x_3) | 3x_1 + 2x_2 - x_3 = 0\}$$

и

$$V = \text{Lin}[(1, 1, 1)^T].$$

[a] Найдите $\dim V$, $\dim W$, $\dim(V \cap W)$, $\dim V^\perp$, $\dim W^\perp$.

$\dim V$ vs $\dim V^\perp$

$$V = \text{Lin}[(1, 1, 1)^T] \longrightarrow \dim V = 1 \longrightarrow \dim V^\perp = 3 - 1 = 2$$

$\dim W$ vs $\dim W^\perp$

$$W = \{(x_1, x_2, x_3) | 3x_1 + 2x_2 - x_3 = 0\} \longrightarrow W^\perp = \text{Lin}(3, 2, -1)^T \longrightarrow \dim W = 2 \longrightarrow \dim W^\perp = 3 - 2 = 1$$

$W \cap V$

$$W \cap V = \{x | x \perp \text{базисы}\}$$

$$V : (1, 1, 1) \longrightarrow \begin{cases} v_1 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} \\ v_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \end{cases}$$

$$\begin{pmatrix} 3 & 2 & -1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \longrightarrow \dim W \cap V = 0$$

[6] Найдите проекцию произвольного вектора u на V , W , $V \cap W$, V^\perp , W^\perp . Найдите квадрат длины каждой проекции.

$$\begin{cases} v_0 - \text{базис } V \\ w_1, w_2 - \text{базис } W \end{cases} \Leftrightarrow \begin{cases} v_1, v_2 - \text{базис } V^\perp \\ w_0 - \text{базис } W^\perp \end{cases}$$

$pr_v u$ vs $pr_{v^\perp} u$

$$\begin{cases} pr_v u = \frac{\langle u, v_0 \rangle}{\langle v_0, v_0 \rangle} v_0 = \frac{u_1 + u_2 + u_3}{3} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ pr_{v^\perp} u = u - pr_v u = \begin{pmatrix} u_1 - \frac{u_1 + u_2 + u_3}{3} \\ u_2 - x \\ u_3 - x \end{pmatrix} = \begin{pmatrix} \frac{2u_1 - u_2 - u_3}{3} \\ \frac{2u_2 - u_1 - u_3}{3} \\ \frac{2u_3 - u_1 - u_2}{3} \end{pmatrix} \end{cases}$$

$$\begin{cases} \|pr_v u\|^2 = \frac{(u_1 + u_2 + u_3)^2}{3} \\ \|pr_{v^\perp} u\|^2 = \frac{(2u_1 - u_2 - u_3)^2 + (2u_2 - u_1 - u_3)^2 + (2u_3 - u_1 - u_2)^2}{9} \end{cases}$$

$pr_{w^\perp} u$ vs $pr_w u$

$$\begin{cases} pr_{w^\perp} u = \frac{\langle u, w_1 \rangle}{\langle w_1, w_1 \rangle} w_1 = \frac{3u_1 + 2u_2 - u_3}{14} \begin{pmatrix} 3 \\ 2 \\ -1 \end{pmatrix} \\ pr_w u = u - pr_{w^\perp} u = \begin{pmatrix} u_1 - \frac{9u_1 + 6u_2 - 3u_3}{14} \\ u_2 - \frac{6u_1 + 4u_2 - 2u_3}{14} \\ u_3 - \frac{-3u_1 + 2u_2 - u_3}{14} \end{pmatrix} = \begin{pmatrix} \frac{5u_1 - 6u_2 + 3u_3}{14} \\ \frac{5u_2 - 3u_1 + u_3}{14} \\ \frac{13u_3 + 3u_1 + 2u_2}{14} \end{pmatrix} \end{cases}$$

$$\begin{cases} \|pr_w u\|^2 = \frac{(9u_1 + 6u_2 - 3u_3)^2 + (6u_1 + 4u_2 - 2u_3)^2 + (13u_3 + 3u_1 + 2u_2)^2}{196} \\ \|pr_{w^\perp} u\|^2 = \frac{(5u_1 - 6u_2 + 3u_3)^2 + (5u_2 - 3u_1 + u_3)^2 + (13u_3 + 3u_1 + 2u_2)^2}{196} \end{cases}$$

$pr_{v \cap w} u$

$$pr_{v \cap w} u = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\|pr_{v \cap w} u\|^2 = 0$$

- [в] Как распределён квадрат длины проекции в каждом случае, если дополнительно известно, что вектор u имеет многомерное стандартное нормальное распределение?

$$V \quad \|pr_v u\|^2 \sim \chi_1^2$$

$$V^\perp \quad \|pr_{v^\perp} u\|^2 \sim \chi_2^2$$

$$W \quad \|pr_w u\|^2 \sim \chi_2^2$$

$$W^\perp \quad \|pr_{w^\perp} u\|^2 \sim \chi_1^2$$

Задача 12. Парная регрессия

Исследователь Борис работает с обычной парной регрессией

$$y_i = \beta_0 + \beta_1 X_i + u_i,$$

которую он оценивает при помощи МНК:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

- [а] Просто для удобства выпишите RSS в этой регрессии и условия первого порядка в задаче минимизации.
- [б] Докажите, что $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$.
- [в] Докажите, что $\bar{y} = \bar{\hat{y}}$.
- [г] Докажите, что точка (\bar{x}, \bar{y}) лежит на линии оценённой регрессии. К сожалению, не успел выполнить данное задание в виде теса. Скрины решенных пунктов я разместил в конце работы.
- [д] Докажите, что $\sum_{i=1}^n x_i(y_i - \hat{y}_i) = 0$.

Задача 13. Гипотезы в линейной регрессии

Линейная регрессионная модель задаётся в следующем виде:

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + u_i.$$

Предположим, что $u \sim \mathcal{N}(0, \sigma^2 I)$. Известно, что

$$X = \begin{pmatrix} 1 & 1 & 3.1 \\ 1 & 12 & 2.2 \\ 1 & -3 & 0.1 \\ 1 & 2 & 0.5 \\ 1 & 0 & 11.3 \end{pmatrix}, \quad y = \begin{pmatrix} 1.1 \\ 2.5 \\ 2.2 \\ 4 \\ 1 \end{pmatrix}$$

В процессе решения используйте калькулятор, все числа округляйте до сотых.

- [а] Найдите $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.

$$RSS = (y - \hat{y})^T (y - \hat{y}) \rightarrow \min_{\hat{\beta}} \implies \hat{\beta} = (X^T X)^{-1} X^T y = ?$$

$$1) \quad (X^T X)^{-1} = \begin{pmatrix} \frac{539679}{1328851} & \frac{-29734}{1328851} & \frac{-58880}{1328851} \\ \frac{-29734}{1328851} & \frac{10404}{1328851} & \frac{1385}{1328851} \\ \frac{-58880}{1328851} & \frac{1385}{1328851} & \frac{16150}{1328851} \end{pmatrix}$$

$$2) (X^T X)^{-1} X^T = \begin{pmatrix} \frac{327417}{1328851} & \frac{53335}{1328851} & \frac{622993}{1328851} & \frac{450771}{1328851} & \frac{-3065}{32411} \\ \frac{-30073}{2657702} & \frac{98161}{1328851} & \frac{-121615}{1328851} & \frac{-16467}{61420} & \frac{-687}{48035} \\ \frac{-7430}{1328851} & \frac{1328851}{1328851} & \frac{2657702}{1328851} & \frac{2657702}{1328851} & \frac{64822}{3015} \\ \frac{17707499}{6644255} & \frac{961367}{26577020} & \frac{-6730}{1328851} & \frac{-48035}{1328851} & \frac{32411}{1328851} \end{pmatrix}$$

$$3) (X^T X)^{-1} X^T y = \begin{pmatrix} 2.183 \\ 2.776 \\ 2.533 \\ 2.665 \\ 0.749 \end{pmatrix}$$

Таким образом,

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 2.67 \\ 0.04 \\ -0.17 \end{pmatrix}$$

[б] Найдите \hat{y} .

$$\hat{y} = X\hat{\beta} = \begin{pmatrix} 2.183 \\ 2.776 \\ 2.533 \\ 2.665 \\ 0.749 \end{pmatrix}$$

[в] Найдите TSS, ESS, RSS и R^2 .

TSS

$$\sum_i (y_i - \bar{y})^2 = (-1.06)^2 + 0.34^2 + 0.04^2 + 1.84^2 + (-1.16)^2 \approx 5.97$$

ESS

$$\sum_i (\hat{y}_i - \bar{y})^2 = 0.004^2 + 0.604^2 + 0.324^2 + 0.494^2 + (-1.426)^2 \approx 2.76$$

RSS

$$\sum_i (y_i - \hat{y}_i)^2 = 27/25^2 + 7/25^2 + 3/10^2 + (-133/100)^2 + (-1/4)^2 \approx 3.19$$

R^2

$$R^2 = \frac{ESS}{TSS} = 0.46$$

[г] Найдите $\hat{\sigma}$.

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-k}} = \sqrt{\frac{3.19}{5-3}} \approx \sqrt{1.6}$$

[д] Найдите $\widehat{\text{Var}}(\hat{\beta})$.

$$\text{Var}(\hat{\beta}) = \hat{\sigma}^2 (X^T X)^{-1} = 1.6 \cdot \begin{pmatrix} 0.41 & -0.02 & -0.04 \\ -0.02 & 0.01 & 0.001 \\ -0.04 & 0.001 & 0.01 \end{pmatrix} \approx \begin{pmatrix} 0.66 & -0.03 & -0.06 \\ -0.03 & 0.02 & 0.002 \\ -0.06 & 0.002 & 0.02 \end{pmatrix}$$

[е] На уровне значимости 5% проверьте гипотезу

$$\begin{cases} H_0 : \beta_1 = 1, \\ H_1 : \beta_1 \neq 1. \end{cases}$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\text{Var}}(\hat{\beta}_1)}} \sim t_2$$

$$t_{obs} = \frac{0.04 - 1}{\sqrt{0.02}} = -6.79 \stackrel{H_0}{\sim} t_{2,0.975} = -4.3 \implies H_0 \text{ отвергается}$$

[ж] На уровне значимости 10% проверьте гипотезу

$$\begin{cases} H_0 : \beta_2 = 1, \\ H_1 : \beta_2 < 1. \end{cases}$$

$$\begin{cases} t_{2,0.05} = -2.9 \\ t_{obs} = \frac{-0.17-1}{\sqrt{0.02}} \approx -8.27 \end{cases} \implies H_0 \text{ отвергается}$$

[з] Проверьте регрессию на значимость в целом.

$$\begin{aligned} F &= \frac{(RSS_R - RSS_{UR})/(K_{UR} - K_R))}{RSS_{UR}/(n - K_{UR})} \stackrel{H_0}{\sim} F_{(K_{UR} - K_R, n - K_{UR})} \longrightarrow \begin{cases} UR : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \\ R : y_i = \beta_0 + u_i \end{cases} \\ beta_{0,R} &= \bar{y} \longrightarrow RSS_R = \sum_i (y_i - \bar{y}_{R,i})^2 \longrightarrow \begin{cases} RSS_{UR} = 3.19 \\ RSS_R = \sum (y_i - 2.18)^2 \approx 5.97 \end{cases} \implies \\ &\implies \frac{(5.97 - 3.19)/(3-1)}{3.19/(5-3)} \approx 0.87 \stackrel{H_0}{\sim} F_{2,2,0.975} = 19 \implies H_0 \text{ не отвергается, регрессия не значима в целом.} \end{aligned}$$

[и] На уровне значимости 5% проверьте гипотезу

$$\begin{cases} H_0 : \beta_1 = \beta_2, \\ H_1 : \beta_1 \neq \beta_2. \end{cases}$$

$$\begin{cases} UR : y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \\ R : y_i = \beta_0 + \beta_1(x_{1i} + x_{2i}) + u_i \end{cases} \implies X^R = \begin{pmatrix} 1 & 4.1 \\ 1 & 14.2 \\ 1 & -2.9 \\ 1 & 2.5 \\ 1 & 11.3 \end{pmatrix}$$

$$\hat{\beta}_R \approx \begin{pmatrix} 2.41 \\ -0.04 \end{pmatrix} \longrightarrow \hat{y}_R = X^R \cdot \hat{\beta}_R = \begin{pmatrix} 2.23 \\ 1.8 \\ 2.53 \\ 2.3 \\ 1.92 \end{pmatrix}$$

$$F = \frac{RSS_R - 3.19}{3.19/2} \approx \frac{5.62 - 3.19}{3.19/2} \approx 1.52 \stackrel{H_0}{\sim} F_{1,2,0.95} = 18.5 \implies H_0 \text{ не отвергается}$$

[к] Постройте 95%-ый доверительный интервал для β_1 .

$$\beta_1 \in [\hat{\beta}_1 - t_{2,0.975} \cdot \sqrt{\hat{\text{Var}}(\hat{\beta}_1)}; \hat{\beta}_1 + t_{2,0.975} \cdot \sqrt{\hat{\text{Var}}(\hat{\beta}_1)}]$$

$$\beta_1 \in [0.04 - 4.3\sqrt{0.02}; 0.04 + 4.3\sqrt{0.02}]$$

[л] Пусть $x_{1,6} = 10$, $x_{2,6} = 7$. Найдите \hat{y}_6 .

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$$

$$\hat{y}_6 = 2.67 + 0.04 \cdot 10 - 0.17 \cdot 7 = 2.67 + 0.4 - 1.19 = 1.88$$

[м] Постройте 95%-ый доверительный интервал для $\mathbb{E}(y_6 | x_{1,6}, x_{2,6})$.

$$\begin{aligned} \mathbb{E}(y_6 | X_6) &= \beta_0 + \beta_1 \cdot 10 + \beta_2 \cdot 7 \longrightarrow \hat{\mathbb{E}}(y_6 | X_6) = 1.88 \longrightarrow \\ \longrightarrow \hat{\text{Var}}(\hat{\mathbb{E}}(y_6 | X_6)) &= \hat{\text{Var}}(\hat{\beta}_0) + \hat{\text{Var}}(\hat{\beta}_1) \cdot 100 + \hat{\text{Var}}(\hat{\beta}_2) \cdot 49 + 2 \cdot 10 \cdot \hat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_0) + 2 \cdot 7 \cdot \hat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_2) + 2 \cdot 70 \cdot \hat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = \\ &= 1.04 + 0.02 \cdot 100 + 0.03 \cdot 49 + 2 \cdot 10 \cdot (-0.06) + 2 \cdot 7 \cdot (-0.11) + 2 \cdot 70 \cdot 0.003 = 2.19 \\ \mathbb{E}(y_6 | X_6) &\in [1.88 - 4.3\sqrt{2.19}; 1.88 + 4.3\sqrt{2.19}] \end{aligned}$$

Задача X[†]. Методы моментов и первого шага

Альтернативой методу максимального правдоподобия является метод моментов, суть которого заключается в том, чтобы приравнять теоретические моменты как функции от оцениваемых параметров к их выборочным аналогам, и из полученной системы найти оценки.

- [a] Пункт для тренировки. Рассмотрим выборку $X_1, X_2, X_3 \sim i.i.d. \mathcal{N}(\mu, 1)$. Оказалось, что $X_1 = 1$, $X_2 = 2$, $X_3 = 3$. Найдите $\hat{\mathbb{E}}(X_1)_{MM}$.
- [б] (6 баллов) Исследователь Матвей подбрасывает монетку с вероятностью орла p до тех пор, пока не выпадет два орла (всего, не обязательно подряд). Оказалось, что среднее число ходов, за которое завершится игра, равно 40. Найдите \hat{p}_{MM} .
Подсказка: докажите, что если M – число ходов, за которое завершится игра, то $\mathbb{E}(M) = \frac{2}{p}$.

Задача Y[†]. Известное неравенство

(6 баллов)

Рассмотрим линейную модель

$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + u_i,$$

где $u_i \sim \mathcal{N}(0, 1)$, а все предпосылки ТГМ выполнены. Исследователь Вадим тестирует гипотезу вида

$$\begin{cases} H_0 : \beta_1 = C \\ H_A : \beta_1 \neq C, \end{cases}$$

где C – некоторая константа, при помощи тестов LR , LM и W . Докажите, что в такой постановке всегда верно, что $LM \leq LR \leq W$.

X

a) $\bar{X} = 2$; $\hat{E}(X_1)_{\text{нн}} = \hat{P}_{\text{нн}}$ согласно $E(X_1)$

Значит, $\hat{E}(X_1)_{\text{нн}} = \hat{P}_{\text{нн}} = \bar{X} = 2$

$S \cdot P = P(M=n) = C_{n-1}^1 p^n q^{n-2}$

- $E(M) = \sum P \cdot n = p \sum q^{n-2} n(n-1)$
- $\left(\frac{1}{p}\right)^2 = \frac{2}{p^3}$; $\frac{1}{p} = \frac{1}{1-q} = ? \rightarrow [1/q = \sum_i (1-q)^i]$
на Тестоффу

Проверка:

$$\begin{cases} E(M) = p^2 \sum n(n-1) q^{(n-2)} \\ \frac{2}{p^3} = \left(\frac{1}{1-q}\right)^2 \left(1 + \sum q^n\right)^2 = \sum q^{n-2} \cdot n(n-1) \end{cases}$$

$$E(M) = p^2 \cdot \frac{2}{p^3} = \frac{2}{p} \quad \checkmark$$

Однако, не проверено?.. Все как на Презентации

$\bar{X} = 40 = \frac{2n}{P_{\text{нн}}} \Rightarrow \hat{P}_{\text{нн}} = 1/20$

пункт А

Рис. 4: Задача X

$\hat{\beta}_0 = 12$

$$\text{a). RSS} = \sum (y_i - \hat{y}_i)^2 \rightarrow \min_{\hat{\beta}_0, \hat{\beta}_1}$$

$$\begin{aligned} \cdot RSS'_{\hat{\beta}_0} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\ \cdot RSS'_{\hat{\beta}_1} &= -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \frac{x_i}{\sum (y_i - \bar{y})(x_i - \bar{x})} \\ \cdot \hat{\beta}_0 \text{ MME} &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

$\therefore \sum (y_i - \hat{y}_i) = \sum y_i - \bar{y} - \frac{\sum (y_i - \bar{y})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \cdot (x_i - \bar{x}) = 0$

$$\Rightarrow \sum_j ((x_j^2 - 2x_j \bar{x} + \bar{x}^2)(y_j - \bar{y}) - (y_j - \bar{y})(x_j x_j - x_j \bar{x} - x_j \bar{y} + \bar{y}^2)) = 0$$

$\therefore = ① + ② \equiv$

\Rightarrow норма
сток
норма
сток $\left\{ \begin{array}{l} ①: \sum x_j y_i - 2 \bar{x} \sum y_i + \bar{y} \bar{x}^2 + 2 \bar{x} \bar{y} \sum y_i - x_j^2 \bar{y} - \bar{x}^2 \bar{y} \\ ②: \sum x_i x_j y_i - x_i \bar{x} y_i - y_i x_j \bar{x} + \bar{x}^2 y_i - x_i \bar{y} \bar{y} + x_i \bar{x} \bar{y} + x_j \bar{x} \bar{y} - \bar{x}^2 \bar{y} \end{array} \right.$

$\equiv x_j \sum y_i - 2 \bar{x} \sum y_i + \bar{x}^2 n \sum y_i + 2 \bar{x} \bar{y} \sum x_j - \bar{y} n \sum x_j$

$$- \sum x_i y_j y_i + \bar{y} \sum y_i \sum y_i + \bar{x} n \sum y_j x_j - \bar{x}^2 n \sum y_i$$

$$+ \bar{y} \sum x_i x_j - \bar{y} \bar{y} n \sum x_i - \bar{y} \bar{y} n \sum x_i =$$

$$= \sum y_i \sum x_j - (\bar{x}) \sum x_j \sum y_i - (\bar{y} n) \sum x_j^2 - \sum x_i y_j x_j$$

$$+ \sum y_j x_j \cdot (\bar{x} n) + (\bar{y}) \sum x_i x_j = 0 \Rightarrow \sum (y_i - \hat{y}_i) = 0$$

\uparrow
 z.r.o.D.

Рис. 5: Задача 12.а и б

b) $\bar{\hat{y}} = \hat{y}$

$$\bar{y} = \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0 + \frac{\sum x_i \cdot \hat{\beta}_1}{n} = \frac{\sum \hat{\beta}_0 + x_i \hat{\beta}_1}{n} = \bar{y}$$

2) Аналогично предыдущему

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \quad \text{согласно, что на линии}$$

Рис. 6: Задача 12.в и г