К.К. Фурманов, И.К. Чернышева

Задачи по эконометрике

без ответов и решений,

зато со вступительным словом от авторов.

Вот оно, это вступительное слово.

Здравствуйте, дорогие читатели!

Перед вами, конечно, не книга в полном смысле этого слова, а всего лишь задачник по эконометрике. Хочется верить, что даже если он не сильно поможет вам освоить эту дисциплину, то будет, по крайней мере, прочитан (или хотя бы просмотрен) с интересом. А уж если поможет и в освоении курса, и в подготовке к контрольной, мы будем счастливы вдвойне.

Мы постарались устроить задачник так, чтобы наши читатели[1] могли освежить в памяти элементы теории вероятностей и математической статистики, а затем приступить к собственно эконометрике. В курсе же собственно эконометрики предлагаются разноплановые задачи, различные по уровню сложности и тематике. Центральное место здесь, как и в большинстве учебных пособий по этому предмету, отведено классической нормальной регрессионной модели, однако мы надеемся познакомить читателей и с такими не очень широко освещенными разделами, как оценивание методом максимального правдоподобия, работа с моделями бинарного и множественного выбора, а также элементы анализа времени жизни.

Нашего задачника ни за что бы не случилось, если бы нас в своё время не обучали такие чудесные люди, как С.А. Айвазян, И.Б. Воскобойников, Б.Б. Демешев, О.А. Демидова, Т.А. Дуброва, Э.Б. Ершов, Г.Г. Канторович, Е.В. Коссова, А.А. Макаров, Г.Ю. Малышева, В.Ф. Матвеев, Е.Ю. Назруллаева, Т.А. Ратникова, А.А. Пересецкий, М.Ю. Турунцева, А.С. Шведов. Они вели у нас разные курсы, связанные с эконометрикой, математической и прикладной статистикой, и мы решили, что здесь самое место высказать им благодарность.

Успехов вам творческих и научных!

И. Чернышева, К. Фурманов,
кафедра математической экономики и эконометрики НИУ ВШЭ

По всем вопросам можно обращаться по адресам: ichernysheva@hse.ru и furmach@menja.net. Туда же мы просим сообщать о найденных в сборнике ошибках, нелепостях и недоразумениях.

Сборник, наверное, будет время от времени обновляться. Настоящая версия была подготовлена 14 января 2018 года.

[1] это, в частности, вы.

Часть 1. Элементы теории вероятностей и математической статистики

§1.1. Случайные события и случайные величины

№1.1.1. Известно, что события A и B несовместны, а события A и C независимы, причём $P(A)=0.4$, $P(A\cup B)=0.8$, $P(A\cap C)=0.2$, $P(B\cap C)=0.1$. Найти $P(B\cup C)$.

№1.1.2. Петя пригласил к себе друзей: Васю, Лену и Надю. Петя предполагает, что Лена придёт с вероятностью 0.6, а Надя – с вероятностью 0.3. При этом он знает, что Лена с Надей в ссоре, так что вместе они точно не придут. Вася, согласно Петиным ожиданиям, придёт с вероятностью 0.8 независимо от Лены и Нади. С какой вероятностью к Пете придут два гостя?

№1.1.3. Два стрелка стреляют по мишени. Для первого стрелка вероятность промаха составляет 0.3, для второго – 0.5. Результаты выстрелов независимы.
а) Какова вероятность того, что мишень будет поражена хотя бы одним из них?
б) При выстреле двух стрелков мишень была поражена (хотя бы одним выстрелом). Какова вероятность того, что второй стрелок промахнулся?

№1.1.4. Подбрасываются две игральные кости. Какова вероятность того, что на первой кости выпала четвёрка, если в сумме на двух костях выпало десять очков?

№1.1.5. Случайная величина X принимает значения -1 и 1, а случайная величина Y принимает значения 0, 1 и 2. В приведённой ниже таблице записаны вероятности событий $P(\{X=x_i\}\cap\{Y=y_j\})$ для различных значений x_i и y_j. Найдите числа a, b и c, если известно, что величины X и Y независимы.

X \ Y	0	1	2
-1	0.4	0.2	a
1	b	0.05	c

№1.1.6. Ряд распределения случайной величины X выглядит следующим образом:

$$X\sim\begin{pmatrix} x_i: & -4 & 0 & 4 \\ P(X=x_i): & a & 0.3 & b \end{pmatrix}$$

Также известно, что $E(X)=1.2$. Найдите:
а) числа a и b, б) $D(X)$, в) $E(7-3X)$, $D(7-3X)$.

№1.1.7. Средняя стоимость покупки в магазине, выраженная в рублях, равна 400, а дисперсия этой стоимости – 2500. Чему равны среднее, дисперсия и стандартное отклонение стоимости покупки, выраженной в тысячах рублей?

№1.1.8. Романтичная девушка Рая любит гадать на ромашках: «любит»-«не любит». Если получается, что любит, то Рая радуется, а если нет, то берёт новую ромашку – и так до тех пор, пока не получит желанный результат или не переберёт все ромашки вокруг. Будем считать, что вероятности сорвать ромашку с чётным и с нечётным количеством лепестков равны и что количества лепестков на разных ромашках – величины независимые.
Рая увидела на обочине дороги четыре ромашки.
а) С какой вероятностью эти ромашки доставят ей радость?
б) Каково математическое ожидание числа ромашек, которые переберёт Рая?
в) Каким будет ответ на пункт (б), если Рая забредёт на ромашковый луг, где число ромашек можно считать бесконечным?

№1.1.9. Про случайные величины X и Y известно, что $E(X+Y)=300$, $E(X-Y)=200$, $D(X)=1$, $D(X+Y)=1$, $Cov(X,Y)=-2$. Найти:
а) $E(X+2Y-50)$, б) $D(X+2Y-50)$, в) $Cov(2X+3,4-5Y)$, г) $Corr(5-X,7-2Y)$.

№1.1.10. Дана функция плотности:

$$f(x)=\begin{cases}0, & \text{если } x<a,\\ c, & \text{если } a\le x\le b,\\ 0, & \text{если } x>b.\end{cases}$$

Найдите значение параметра c.

№1.1.11. Когда гражданин Фёдор отправляется по воскресеньям в магазин, жена наказывает ему купить продукты, общая стоимость которых в тысячах рублей – случайная величина с функцией распределения:

$$F(x)=\begin{cases}1-\exp\left(-x^2\right), & \text{если } x\ge 0,\\ 0, & \text{если } x<0.\end{cases}$$

а) Сколько денег должен взять Фёдор, чтобы с вероятностью 0.9 он смог оплатить покупку?
б) Каким будет ответ в пункте (а), если известно, что покупка обойдётся Фёдору не меньше, чем в тысячу?

№1.1.12. Найдите математическое ожидание, дисперсию и функцию распределения случайной величины с заданной функцией плотности.

а) $f(x)=\begin{cases}4x^3, & \text{если } x\in[0;1],\\ 0, & \text{иначе.}\end{cases}$ б) $f(x)=\begin{cases}0, & x\le 0,\\ \dfrac{x}{2}, & 0<x\le 1,\\ \dfrac{2}{3}-\dfrac{x}{6}, & 1<x\le 4,\\ 0, & x>4.\end{cases}$

в) $f(x)=\begin{cases}\dfrac{1}{x^2}, & \text{если } x>1,\\ 0, & \text{если } x\le 1.\end{cases}$

№1.1.13. Случайная величина X распределена равномерно на интервале (0; 1). Найдите функции распределения случайных величин $Y = X^2$ и $Z = -\ln(X)$. Найдите функции квантилей величин X, Y и Z.

№1.1.14. Функция распределения с.в. X имеет следующий вид:

$$F(x)=\begin{cases}0, & x<0,\\ 1-\dfrac{1}{1+x^2}, & x\ge 0.\end{cases}$$

Найдите:
а) P(X>3), б) соответствующую функцию квантилей,
в) функцию распределения с.в. $Y = X^2$.

№1.1.15. Случайная величина X принимает значения -1, 2 и 3 с вероятностями 0.2, 0.4 и 0.4 соответственно. Начертите графики функции распределения и функции квантилей величины X.

№1.1.16. Приведите пример случайной величины, у которой нет математического ожидания, и пример величины, у которой есть математическое ожидание, но нет дисперсии.

№1.1.17. Распределение с.в. X известно с точностью до параметров *a* и *b*:

значение:	-1	0	1	2
вероятность:	*a*	2*a*	*b*	0.2

а) В каких пределах лежат значения *a* и *b*?
б) Допустим, *a* = 0.1. Начертите график соответствующей функции квантилей.
в) Величина Y имеет то же распределение, X и Y независимы. Найдите $E(X+XY)$, $Cov(X-7, X+Y)$. Предполагается, что *a* = 0.1.

§1.2. Многомерные случайные величины

№1.2.1. Задано совместное распределение случайной величины X, принимающей значения 0 и 1, и случайной величины Y, принимающей значения -1, 0 и 1:

$X \setminus Y$	-1	0	1
0	0.1	0.1	0.2
1	0.2	*a*	3*a*

а) Найти a, б) проверить, являются ли X и Y независимыми,
в) выписать частную функцию распределения с.в. X
г) Найти $E(X)$, $E(Y)$, $D(X)$, $D(Y)$, $Cov(X,Y)$, $Corr(X,Y)$
д) Найти $E(Y \mid X=0)$, $D(Y \mid X=0)$, $E(X \mid Y^2=1)$.

№1.2.2. Аналитик изучает связь между месячными доходностями акций компаний A и B. На основании располагаемой статистики он делает вывод, что доходности акций обеих компаний лежат в пределах от -1% до 2%, а их совместное распределение может быть описано следующей таблицей:

		Доходность акций компании A			
		-1%	0%	1%	2%
Доходность акций компании B	-1%	0.03	0.03	0.02	0.02
	0%	0.03	0.05	0.04	0.03
	1%	0.05	0.1	0.2	0.15
	2%	0.01	0.03	0.06	0.15

а) Найдите коэффициент корреляции между доходностями акций.
б) Согласно надёжному экспертному мнению, в следующем месяце доходность акций компании A будет положительной. Найдите математическое ожидание и стандартное отклонение дисперсии доходности акций компании B с учётом имеющейся информации.

№1.2.3. Задано совместное распределение трёх дискретных случайных величин X, Y, Z, принимающих значения 0 или 1:

Z = 0				**Z = 1**		
X \ Y	*0*	*1*		**X \ Y**	*0*	*1*
0	0.1	0.1		*0*	0.2	0.1
1	0.1	0.1		*1*	0.1	0.2

Рассчитайте вектор математических ожиданий и ковариационную матрицу для векторов $(X,Y,Z)'$ и $(X+Y, 2Z)'$.

№1.2.4. Определите, какие из приведённых ниже матриц могут быть ковариационными:
а) $\begin{pmatrix} 5 & 1 \\ 1 & -2 \end{pmatrix}$, б) $\begin{pmatrix} 9 & 2 \\ 2 & 4 \end{pmatrix}$, в) $\begin{pmatrix} 6 & 4 \\ 2 & 1 \end{pmatrix}$, г) $\begin{pmatrix} 7 & -2 \\ -2 & 3 \end{pmatrix}$, д) $\begin{pmatrix} 2 & 4 \\ 4 & 1 \end{pmatrix}$

№1.2.5. Совместная функция плотности случайных величин X и Y имеет вид:

1) $f(x,y)=\begin{cases} cx^2y^2, 0\le x\le 1, 0\le y\le 1 \\ 0, \quad \textit{иначе} \end{cases}$, 2) $f(x,y)=\begin{cases} c(x+y), \quad 0\le x\le 1, 0\le y\le 1 \\ 0, \quad \textit{иначе} \end{cases}$

Найдите: а) число c, б) частные функции плотности величин X и Y,
в) условные функции плотности величин X и Y, г) E(X) и D(X),
д) E(X|Y=0.5), е) Cov(X, Y) и Corr(X, Y)

№1.2.6. Случайные величины X и Y независимы, каждая из них имеет функцию распределения

$$F(x)=\begin{cases} 1-e^{-x}, \quad x\ge 0, \\ 0, \quad \textit{иначе.} \end{cases}$$

Выпишите совместные функции распределения и плотности для вектора $(X,Y)'$.

№1.2.7. Дана совместная функция плотности величин X и Y:

$$f(x,y)=\begin{cases} 2x, \quad 0\le x,y\le 1; \\ 0, \quad \textit{иначе.} \end{cases}$$

Найдите частные функции плотности X и Y и вероятность события {X<Y}.

№1.2.8. Вася различает два съедобных продукта: винегрет и маргарин. Совместное распределение массы винегрета X и маргарина Y в холодильнике задано следующей функцией (все величины — в килограммах):

$$F(x,y)=\begin{cases} 1-\dfrac{1+5x+3y^2}{(1+5x)(1+3y^2)} & \text{при } x\ge 0 \text{ и } y\ge 0, \\ 0 & \text{при } x<0 \text{ или } y<0. \end{cases}$$

а) Какова вероятность обнаружить в Васином холодильнике не менее 1 кг винегрета и от 200 до 500 г маргарина?
б) Найдите частные (маргинальные) функции распределения для массы винегрета и маргарина.
в) Проверьте независимость X и Y.

§1.3. Нормальное и логарифмически нормальное распределения. Многомерное нормальное распределение.

№1.3.1. Для случайной величины Z, имеющей стандартное нормальное распределение, найдите вероятности:

а) $P(-1.47<Z<1.47)$, б) $P(|Z|>1)$,
в) $P(-1<Z<2)$, г) $P(1<Z<2)$,
д) $P(Z>0.65)$, е) $P(Z<1.5)$.

№1.3.2. Продавец пирожков Анфиса платит за аренду торговой площади 400 рублей в день. Дневная выручка от продажи пирожков подчинена нормальному закону с математическим ожиданием 700 рублей и стандартным отклонением 200 рублей. Как часто Анфисе не хватает дневной выручки, чтобы заплатить за аренду?

№1.3.3. Известно, что $X\sim N(2,1), Y\sim N(3,4)$, X и Y независимы. Найти:

а) $P(X+Y<5)$, б) $P(Y<-9)$, в) такое число a, что $P(|3X-2Y|<a)=0.95$,
г) такое число a, что $P(X<a)=0.9$.

№1.3.4. Кондитер Иннокентий готовит эклеры. 20% приготовленных эклеров весят менее 130 грамм, и 20% эклеров весят более 150 грамм. Иннокентий убеждён, что масса эклера – нормально распределённая величина.

а) Найдите её математическое ожидание и стандартное отклонение.
б) Иннокентий продаёт только эклеры массой более 130 грамм, а остальные относит домой детям. Какую долю из числа эклеров, достающихся детям, составляют эклеры массой меньше 120 грамм?

№1.3.5. Для измерения уровня доходов населения часто вместо среднего значения используется медиана, потому что она меньше отклоняется в сторону нетипично больших доходов, которые получает небольшая часть населения. Известно, что в некотором регионе медианный доход семьи составляет 30 тысяч рублей в месяц, а стандартное отклонение – 8 тысяч рублей. Найдите средний доход в регионе и долю населения с доходом больше 40 тысяч, если распределение доходов в регионе подчинено логарифмически нормальному закону.

№1.3.6. Один из показателей неравенства доходов населения – коэффициент фондов (он же – децильный коэффициент), равный отношению квантили порядка 0.9 к квантили порядка 0.1 для дохода. В некотором регионе доходы (выраженные в тысячах рублей) имеют логарифмически нормальное распределение с параметрами $\mu = 3$ и $\sigma^2 = 0.5$. Найдите коэффициент фондов.

№1.3.7. Математическое ожидание логнормальной случайной величины равно 100. В каких пределах может находиться её медиана?

№1.3.8. Случайный вектор $\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ имеет двумерное нормальное распределение с математическим ожиданием $\mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ и ковариационной матрицей $\Sigma = \begin{pmatrix} 1 & -1 \\ -1 & 4 \end{pmatrix}$.
Найдите: а) $P(2X_1 + X_2 > 11)$, б) математическое ожидание и ковариационную матрицу вектора $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ 2X_1 - X_2 \end{pmatrix}$

№1.3.9. Ермолай Лопахин решил приступить к вырубке вишнёвого сада. Однако выяснилось, что растут в нём не только вишни, но и яблони. Причем, по словам Любови Андреевны Раневской, среднее количество деревьев (а они периодически погибают от холода или жары, либо из семян вырастают новые) в саду распределено в соответствии с нормальным законом (X – число яблонь, Y – число вишен) со следующими параметрами:
$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 20 \\ 130 \end{pmatrix}, \begin{pmatrix} 5 & 4 \\ 4 & 10 \end{pmatrix}\right)$$
Найдите вероятность того, что Ермолаю Лопахину придется вырубить менее 150 деревьев.
Каково ожидаемое число подлежащих вырубке вишен, если известно, что предприимчивый и последовательный Лопахин, не затронув ни одного вишнёвого дерева, начал очистку сада с яблонь и все 25 яблонь уже вырубил?

№1.3.10. Сэр Фрэнсис Гальтон – учёный XIX-XX веков, один из основоположников как генетики, так и статистики – изучал, среди всего прочего, связь между ростом детей и родителей. Он исследовал данные о росте 928 индивидов[2] (пусть их рост в дюймах описывается случайной величиной C) и о среднем арифметическом роста отца и матери каждого (эту величину, также измеренную в дюймах, обозначим за P). Согласно оценкам, полученным на основании тех данных, вектор математических ожиданий выглядит так:
$$\begin{pmatrix} E(C) \\ E(P) \end{pmatrix} = \begin{pmatrix} 68.09 \\ 68.31 \end{pmatrix}$$

[2] Рост женщин был помножен на 1.08 для сопоставимости с ростом мужчин.

А вот и ковариационная матрица:

$$\begin{pmatrix} D(C) & Cov(C,P) \\ Cov(P,C) & D(P) \end{pmatrix} = \begin{pmatrix} 6.34 & 2.06 \\ 2.06 & 3.19 \end{pmatrix}$$

1) Обратите внимание на то, что дисперсия роста детей выше дисперсии среднего роста родителей. С чем это может быть связано? Учтите, что рост детей измерялся уже по достижении зрелости, так что разброс не должен быть связан с возрастными различиями.

2) Рассчитайте корреляционную матрицу вектора $(C,P)'$.

3) Пусть величины $\tilde{C}$ и $\tilde{P}$ также отражают рост детей и средний рост родителей, но измерены не в дюймах, а в сантиметрах (1 дюйм равен 2.54 см). Выпишите вектор математических ожиданий, ковариационную и корреляционную матрицы вектора $(\tilde{C},\tilde{P})'$.

4) Предположим, что вектор $(C,P)'$ имеет совместное нормальное распределение. Определите, каков ожидаемый рост мужчины, средний рост родителей которого составляет 72 дюйма? Какова вероятность того, что рост такого мужчины превысит 68 дюймов?

№1.3.11. Случайный вектор (X,Y,Z)’ имеет совместное нормальное распределение:

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}, \begin{pmatrix} 9 & 0 & 4 \\ 0 & 4 & -1 \\ 4 & -1 & 1 \end{pmatrix}\right)$$

Рассчитайте вероятности: а) P(0<X+Y-2Z<10), б) P(2X-Y<Z)

№1.3.12. Одним из распространённых инструментов анализа финансовых рынков является модель оценки фондовых (или капитальных) активов CAPM (Capital Asset Pricing Model). Эта модель основана на предположении о наличии безрискового актива (гарантированно обеспечивающего некоторую доходность R_f) и на понятии рыночного портфеля – набора различных активов (включая безрисковый), имеющего максимально возможную ожидаемую доходность при заданной дисперсии доходности. Предполагается, что связь доходности некоторого актива R с доходностью рыночного портфеля R_m описывается следующим уравнением: $E(R)-R_f = \beta(E(R_m)-R_f)$, где $\beta = \dfrac{Cov(R,R_m)}{D(R_m)}$. То есть, превосходство ожидаемой доходности актива над безрисковой доходностью пропорционально превосходству ожидаемой доходности рыночного портфеля над безрисковой доходностью. Коэффициент пропорциональности β отражает чувствительность доходности выбранного актива к динамике рынка. Предположим, что доходность безрискового актива составляет 4%, ожидаемая доходность рыночного портфеля равна 6%, а случайный вектор (R, R_m) имеет совместное нормальное распределение с ковариационной матрицей $\Sigma = \begin{pmatrix} 4 & 1 \\ 1 & 1 \end{pmatrix}$.

а) Рассчитайте коэффициент β для рассматриваемого актива

б) Какова ожидаемая доходность рассматриваемого актива?

в) Каковы ожидаемая доходность и стандартное отклонение доходности рассматриваемого актива, если доходность рыночного портфеля равна 8%?

№1.3.13. Случайный вектор $\begin{pmatrix} X \\ Y \end{pmatrix}$ имеет нормальное распределение с математическим ожиданием $\begin{pmatrix} 1 \\ -4 \end{pmatrix}$ и ковариационной матрицей $\begin{pmatrix} 16 & -4 \\ -4 & 36 \end{pmatrix}$. Найдите: $P(X+2Y>-1)$, $E[(X-Y)^2]$.

№1.3.14. Затраты на отопление предприятия в январе Y (тыс. руб.) связаны со средней температурой января T соотношением $Y = 12 - 3T + \varepsilon$. Величины (T, ε) нормально распределены, известны их математические ожидания и ковариационная матрица:

$$E\begin{pmatrix} T \\ \varepsilon \end{pmatrix} = \begin{pmatrix} -18 \\ 0 \end{pmatrix}, \quad V\begin{pmatrix} T \\ \varepsilon \end{pmatrix} = \begin{pmatrix} 4 & 0 \\ 0 & 13 \end{pmatrix}.$$

а) Найдите математическое ожидание и дисперсию затрат на отопление.
б) Выпишите функцию плотности величины Y.
в) С какой вероятностью затраты на отопление превысят 73 тыс. руб?

§1.4. Точечные оценки

№1.4.1. В течение двух дней исследовательская группа проводила опрос населения некоторого города с целью выяснить уровень поддержки действующей администрации среди местного населения, для чего респондентам предлагалось оценить своё отношение к действиям администрации по 10-балльной шкале. В течение первого дня было опрошено 200 респондентов, среди которых средний балл составил $\bar{X}_1$. В течение второго дня было опрошено 300 респондентов и средний балл составил $\bar{X}_2$.

К сожалению, все результаты опроса, кроме значений $\bar{X}_1$ и $\bar{X}_2$, были утеряны, и исследователи не могли вспомнить, сколько респондентов было опрошено за эти два дня. В результате, для среднего по генеральной совокупности (математического ожидания) было решено использовать оценку $\hat{X} = \dfrac{\bar{X}_1 + \bar{X}_2}{2}$.

Покажите, что оценка $\hat{X}$ является несмещённой для среднего по генеральной совокупности. Рассчитайте, насколько её дисперсия больше дисперсии выборочного среднего $\bar{X}$, рассчитанного по всем 500 наблюдениям. Считайте, что все наблюдаемые случайные величины независимы и имеют одинаковые математическое ожидание и дисперсию.

№1.4.2. Имеется случайная выборка $X_1, \ldots, X_n$, где все X_i независимы и принимают значения 1, 3 и 5 со следующими вероятностями:

значения:	1	3	5
вероятности:	a	0.2	0.8- a

а) Каковы допустимые значения параметра a?

б) При каком значении m оценка $\hat{a} = mX_1 + 1.15 + \dfrac{1}{n-1}\sum_{i=2}^{n} X_i$ для параметра a будет несмещённой?

№1.4.3. Случайные величины $X_1, \ldots, X_n$ независимы, ряд распределения каждой из величин X_i известен с точностью до параметра μ:

Возможные значения величины X_i:	$\mu - 2\sqrt{i}$	μ	$\mu + 3\sqrt{i}$
Вероятности этих значений:	0.3	0.5	0.2

Пусть $\bar{X}_n$ - среднее значение величин $X_1, \ldots, X_n$: $\bar{X}_n = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$.

а) Найдите математическое ожидание и дисперсию $\bar{X}_n$.

б) Достаточно ли полученных в пункте (а) результатов, чтобы утверждать, что $\bar{X}_n$ является состоятельной оценкой для параметра μ?

№1.4.4. Имеются две независимые несмещённые оценки параметра θ, обозначим их $\hat{\theta}_1$ и $\hat{\theta}_2$, а их дисперсии соответственно σ_1^2 и σ_2^2. Найдите значение α, при котором оценка $\tilde{\theta} = \alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2$ будет иметь наименьшую дисперсию. Будет ли $\tilde{\theta}$ несмещённой оценкой для θ?

№1.4.5. Пусть X_i независимы и одинаково распределены с математическим ожиданием μ и дисперсией σ^2. Предлагаются две оценки параметра μ:

$\hat{\mu}_1 = X_1 + \alpha X_7 - 3\beta X_{20}$;

$\hat{\mu}_2 = \alpha X_2 - \beta X_7$.

При каких значениях α, β обе оценки для μ будут несмещёнными?
Выберите из двух более эффективную оценку.

№1.4.6. Имеются n наблюдений ($n \geq 3$), представляемых независимыми одинаково распределёнными случайными величинами $X_1, \ldots, X_n$ с математическим ожиданием μ и дисперсией σ^2. Определите, какие из приведённых ниже оценок $\hat{\mu}_1$, $\hat{\mu}_2$ и $\hat{\mu}_3$ являются несмещёнными? Какая из несмещённых оценок является относительно более эффективной?

а) $\hat{\mu}_1 = X_1$,

б) $\hat{\mu}_2 = X_1 - X_2 + \frac{1}{n-2}\sum_{i=3}^{n} X_i$,

в) $\hat{\mu}_3 = \frac{X_1 + 3X_2}{2}$.

№1.4.7. Имеются n наблюдений ($n \geq 2$), представляемых независимыми одинаково распределёнными случайными величинами $X_1, \ldots, X_n$ с математическим ожиданием μ и дисперсией σ^2. Докажите, что приведённые в пунктах (а) и (б) оценки для μ являются состоятельными, а приведённая в пункте (в) оценка состоятельной не является.

а) $\hat{\mu}_1 = \frac{2X_1 + \sum_{i=2}^{n} X_i}{n+1}$,

б) $\hat{\mu}_2 = \frac{1}{n+2}\sum_{i=1}^{n} X_i$,

в) $\hat{\mu}_3 = X_1$.

№1.4.8. Имеется независимая случайная выборка $X_1, \ldots, X_n$ из геометрического распределения: $P(X_i = x) = (1-p)^x p$, $x = 0, 1, 2, \ldots$ Предложите несмещённую и состоятельную оценку для параметра p.

№1.4.9. Имеется выборка из n наблюдений, описываемых независимыми случайными величинами $X_1, \ldots, X_n$, имеющими распределение Бернулли ($X_i = 1$ с вероятностью p и $X_i = 0$ с вероятностью $1-p$). Найдите математическое ожидание и дисперсию предложенных оценок параметра p:

а) $\hat{p} = \frac{X_1 + 3X_2}{4}$,

б) $\hat{p} = \frac{1}{n} X_1 + \frac{1}{n-1} \sum_{i=2}^{n} X_i$,

в) $\hat{p} = X_1 - X_2$,

г) $\hat{p} = 2X_1 - \frac{1}{n-1} \sum_{i=2}^{n} X_i$.

Для каждой из оценок проверьте, выполняется ли достаточное условие состоятельности.

№1.4.10. Дан ряд распределения случайной величины X:

$$X \sim \begin{pmatrix} \textit{значения}: & 0 & 1 & 2 \\ \textit{вероятности}: & \theta & 2\theta & 1-3\theta \end{pmatrix}.$$

а) Определите множество допустимых значений параметра θ.

б) При каких значениях a и b величина $Y = a + bX$ будет несмещённой оценкой для θ?

№1.4.11. Случайные величины X_i независимы и одинаково распределены с математическим ожиданием μ и дисперсией σ^2. Покажите, что выборочное среднее $\bar{X} = \frac{X_1 + ... + X_n}{n}$ является эффективной оценкой для μ в классе линейных несмещённых оценок, т.е. обладает наименьшей дисперсией среди всех несмещённых оценок вида $\hat{\mu} = \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_n X_n$, где α_i - некоторые коэффициенты.

№1.4.12. Пусть случайные величины $X_1, ..., X_n$ независимы, а их распределение известно с точностью до параметра θ: $X_i \sim \begin{pmatrix} \textit{значения}: & \theta & i \\ \textit{вероятности}: & 1 - \frac{1}{i} & \frac{1}{i} \end{pmatrix}$.

Покажите, что X_n является состоятельной оценкой для θ, но достаточное условие состоятельности для неё не выполняется.

№1.4.13. Случайные величины $X_1, ..., X_n$ независимы и одинаково распределены с функцией плотности, известной с точностью до параметра $a \in [0; 2]$:

$$f(x) = \begin{cases} a + 2x(1-a), & 0 \le x \le 1, \\ 0, & \text{иначе}. \end{cases}$$

а) При каком значении m оценка $\hat{a} = m - 3(X_1 + X_2)$ для параметра a будет несмещённой?

б) Пусть a=0. Найдите P(0.5<X<0.7).

№1.4.14. Тимофей и Надежда независимо друг от друга проводят выборочное обследование населения, пытаясь определить долю людей, дающих взятки. Тимофей решает, что вопрос достаточно пикантный, чтобы не ожидать честного ответа, поэтому поступает так: каждому респонденту он даёт тайно подбросить монетку, оговорив, что если выпадет «орёл», респонденту следует отвечать «да» на последующий вопрос, а если выпадет «решка», то от респондента ожидается честный ответ. После этого Тимофей спрашивает: «Доводилось ли вам за последний год давать кому-либо взятку?» и записывает полученный ответ. Предположим, что опрашиваемые ведут себя именно так, как ожидает Тимофей.

У Нади такие глаза, что ей невозможно врать, поэтому она просто спрашивает, доводилось ли респонденту давать взятки, и ей честно отвечают.

Пусть $X_1, ..., X_m$ — ответы, получаемые Тимофеем, а $Y_1, ..., Y_n$ — ответы, получаемые Надеждой (ответы кодируются так: 1 — «да», 0 — «нет»). В качестве оценки для доли дающих взятки

Тимофей использует величину $\tilde{p} = 2 \times \left(\frac{X_1 + ... + X_m}{m} - 0.5 \right)$, а Надя — обычную выборочную долю $\hat{p} = \frac{Y_1 + ... + Y_n}{n}$.

а) Покажите, что используемая Тимофеем оценка несмещённая. Проверьте, выполняется ли для неё достаточное условие состоятельности.

б) Тимофей опросил 100 человек. Сколько человек достаточно опросить Надежде, чтобы её оценка не уступала по точности оценке Тимофея?

№1.4.15. Случайные величины $X_1, ..., X_n$ независимы, их распределение известно с точностью до параметра θ:

Значения:	–1	0	1
Вероятности:	θ	1–2θ	θ

а) Найдите $MSE(\tilde{\theta})$, где $\tilde{\theta} = X_1^2$.

б) При каком значении c оценка $\hat{\theta} = c\sum_{i=1}^{n} X_i^2$ будет несмещённой?

в) Найдите $MSE(\hat{\theta})$.

№1.4.16. Докажите формулу $MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + D(\hat{\theta})$, где $Bias(\hat{\theta}) = E(\hat{\theta} - \theta)$ — смещение оценки.

§1.5. Выборочные характеристики. Основные распределения математической статистики.

№1.5.1. Курьерская служба изучает среднее время доставки посылок по городу. Случайная выборка из времён доставки для 10 посылок дала следующие результаты (в часах):
7, 3, 4, 6, 10, 5, 6, 4, 3, 8.

а) Рассчитайте выборочные характеристики: среднее, медиану, дисперсию и стандартное отклонение.

б) Постройте график выборочной функции распределения времени доставки.

№1.5.2. Романтичная девушка Рая перебрала 50 цветков сирени и обнаружила четыре цветка с пятью лепестками (в остальных было по четыре лепестка). Рассчитайте выборочное среднее и выборочную дисперсию для числа лепестков в одном цветке.

№1.5.3. Пусть $X \sim \chi^2_3, Y \sim t_7$. Найдите:

а) P(X>4.642), б) P(1.424<X<7.815), в) P(|Y|<1.895), г) P(-0.402<Y<0.130)

№1.5.4. Меткий стрелок Василиса стреляет по мишени так, что координаты (X, Y) точки, в которую она попадает, имеют совместное стандартное нормальное распределение. Начало координат совпадает с центром мишени.

Какова вероятность того, что при очередном выстреле Василиса отклонится от центра мишени более чем на 2 дюйма? Именно в дюймах измеряются координаты (X, Y).

№1.5.5. Менеджер кофейни «У Агафьи Матвеевны» знает, что объём стандартной порции эспрессо должен составлять 50 мл со стандартным отклонением в 2 мл. Желая понять, насколько точно официанты наливают посетителям положенный объём кофе, менеджер выборочно обследует несколько чашек. Сколько чашек он должен обследовать, чтобы с вероятностью 95%

отклонение среднего объёма порции в выборке от генерального среднего не превышало двух процентов?

№1.5.6. Случайные величины X и Y независимы и распределены по стандартному нормальному закону. Какова вероятность того, что одна из них превысит другую более чем в пять раз по модулю?

№1.5.7. Случайные величины $X_1,\dots,X_n$ независимы, $X_i \sim N(\mu,\sigma^2)$. Сравните по MSE две оценки для дисперсии: $S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$ и $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$. При каком значении c оценка $\tilde{\sigma}^2=c\sum_{i=1}^{n}(X_i-\overline{X})^2$ будет иметь наименьшую среднеквадратическую ошибку?

№1.5.8. Есть две независимые выборки из нормального распределения с одинаковой дисперсией, в каждой по 8 наблюдений. С какой вероятностью несмещённая оценка дисперсии, рассчитанная по первой выборке, окажется больше оценки дисперсии по второй выборке хотя бы в семь раз?

§1.6. Доверительные интервалы.

№1.6.1. С целью изучения трудовой мобильности проводится выборочный опрос населения. Рассчитанный по ответам 400 респондентов средний стаж работы на текущем рабочем месте составил 2.2 года, а оценка дисперсии составила $\hat{\sigma}^2=4$. Рассчитайте 99% доверительный интервал для среднего стажа по генеральной совокупности.

№1.6.2. На кондитерской фабрике отдел контроля качества отобрал 25 плиток горького шоколада для проверки состава продукции. На основании анализа отобранных образцов был построен 90% доверительный интервал для среднего веса какао в составе одной плитки в граммах: [59.5; 62.5]. При расчёте интервала предполагалось, что вес какао имеет нормальное распределение с неизвестной дисперсией.

Начальник отдела заявил, что хочет иметь более надёжные результаты и попросил рассчитать 95% доверительный интервал для среднего веса какао. Выполните просьбу начальника.

№1.6.3. В течение двух дней проводился опрос населения города с целью определения отношения жителей к действующей администрации. В первый день было опрошено 256 человек. По этим наблюдениям статистик построил доверительный интервал для доли тех, кто выразил положительное отношение к работе администрации города: (31,57%; 43,43%). Во второй день были опрошены ещё 144 человека, из которых 46 выразили положительное отношение к работе администрации.

а) Какой уровень доверия (доверительная вероятность) использовался при построении доверительного интервала по данным первого дня опроса?

б) По данным за оба дня рассчитайте 90% доверительный интервал для доли населения, положительно относящегося к работе администрации.

№1.6.4. Статистик Тимофей оценивает доверительный интервал для математического ожидания по большой выборке по формуле $\overline{X}-z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}}<\mu<\overline{X}+z_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{n}}$. Тимофей забыл таблицы нормального распределения и не может точно вспомнить значение $z_{\frac{\alpha}{2}}$ для уровня доверия 95%. Определите, каков будет уровень доверия, если

а) Тимофей подставит значение $z_{\frac{\alpha}{2}} = 2$;

б) Тимофей воспользуется следующим выражением для доверительного интервала:
$\bar{X} - 1.5\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{X} + 2.5\frac{\hat{\sigma}}{\sqrt{n}}$.

№1.6.5. По выборке из нормально распределённой генеральной совокупности объёмом в 16 наблюдений рассчитаны выборочное среднее $\bar{X} = 12.6$ и оценка для дисперсии $\hat{\sigma}^2 = 6.25$. Рассчитайте 90% доверительный интервал для математического ожидания и 80% доверительный интервал для дисперсии.

№1.6.6. Пусть известно, что дисперсия по нормальной генеральной совокупности равна 4. Каким должен быть объём выборки, чтобы ширина 95% доверительного интервала для математического ожидания (разность верхней и нижней границы интервала) была не более 1?

№1.6.7. По 25 наблюдениям из нормальной генеральной совокупности с неизвестной дисперсией был оценен 95% доверительный интервал для математического ожидания: [37.3; 40.7]. Найдите выборочное среднее и оценку дисперсии $\hat{\sigma}^2$.

№1.6.8. Статистику Тимофею надоели обычные, симметричные по вероятности доверительные интервалы. Он хочет построить 90% доверительный интервал для математического ожидания так, чтобы математическое ожидание недооценивалось с вероятностью 2%, а переоценивалось с вероятностью 8%. Помогите Тимофею вывести выражение для такого интервала в случае нормальной генеральной совокупности и известной дисперсии.

№1.6.9. Производитель одежды хочет узнать, какой цвет футболок предпочитает целевая группа: малиновый или салатовый. В выборке из 225 человек 90 высказались в пользу малинового цвета, а 135 – в пользу салатового.
а) Рассчитайте 95% доверительный интервал для доли предпочитающих салатовый цвет.
б) Строгий начальник хочет, чтобы ширина доверительного интервала (разница между его верхней и нижней границей) была не больше 0.1. Какой доверительной вероятности можно добиться в таком случае?

№1.6.10. Перед разработкой собственного сайта Василий планирует опрос населения с целью выяснить, какой из двух проектов сайта (обозначим их А и Б) больше нравится потенциальным посетителям. Василий намерен опросить 400 человек и узнать, какой проект предпочитает каждый из них.
а) В каких пределах лежать математическое ожидания и дисперсия *доли* опрошенных, предпочитающих проект А?
б) Помогите Василию понять, сколько человек нужно опросить, чтобы доля предпочитающих проект А в выборке отличалась от доли в генеральной совокупности не более чем на 0.08 с вероятностью 95%.
в) Василий всё равно опросил 400 человек. Из них 280 высказались в пользу проекта А. Найдите 95% доверительный интервал для доли предпочитающих этот проект в генеральной совокупности.

№1.6.11. По выборке из 14 наблюдений статистик Тимофей и почтенный старец Феодосий рассчитывают 95% доверительный интервал для математического ожидания. Оба уверены, что выборка взята из нормальной генеральной совокупности, но Тимофей оценивает дисперсию по имеющимся наблюдениям, а Феодосий по опыту знает, что дисперсия равна 25.
а) С какой вероятностью интервал Тимофея окажется шире интервала Феодосия?
б) Феодосий рассчитал выборочное среднее — 10.5. Помогите ему рассчитать интервал до конца.

§1.7. Проверка гипотез.

№1.7.1. Известно, что $X \sim R[0;a]$. Исследователь проверяет гипотезу $H_0 : a = 10$ против $H_A : a > 10$ с помощью следующего критерия: отвергнуть H_0 в пользу H_A, если $X > c$. Каким должно быть число c, чтобы обеспечить уровень значимости 10%? При найденном c выразите мощность критерия как функцию от a.

№1.7.2. Гражданин Фёдор решает проверить, не жульничает ли напёрсточник Афанасий, для чего предлагает Афанасию сыграть 5 партий в напёрстки. Фёдор решает, что в каждой партии будет выбирать один из трёх напёрстков наугад, не смотря на движения рук ведущего. Основная гипотеза: Афанасий честен, и вероятность правильно угадать напёрсток, под которым спрятан шарик, равна $^1/_3$. Альтернативная гипотеза: Афанасий каким-то образом жульничает (например, незаметно прячет шарик), так что вероятность угадать нужный напёрсток меньше, чем $^1/_3$. Статистический критерий: основная гипотеза отвергается, если Фёдор ни разу не угадает, где шарик.
а) Найдите уровень значимости критерия.
б) Найдите мощность критерия в том случае, когда Афанасий жульничает, так что вероятность угадать нужный напёрсток равна $^1/_5$.

№1.7.3. Перед выборами проводится опрос с целью выяснения уровня поддержки кандидатов А и В (других кандидатов нет). Основная гипотеза: голоса избирателей распределяются между кандидатами поровну. Альтернативная гипотеза: доля избирателей, поддерживающих кандидата А, больше 50%. Для проверки выбран уровень значимости $\alpha = 0.01$. Из 500 опрошенных 280 ответили, что поддерживают кандидата А.
а) Рассчитайте статистику, с помощью которой проверяется указанная гипотеза.
б) Рассчитайте критическое значение (или значения) для этой статистики.
в) Рассчитайте *p*-значение.
г) Определите, есть ли основания отвергнуть основную гипотезу на заданном уровне значимости.

№1.7.4. Производитель шампуня следит за тем, чтобы в среднем в упаковке было 400 мл шампуня. При обследовании был измерен объём шампуня в 12 упаковках:
451 454 444 454 447 450 445 447 442 446 448 451
На уровне значимости 10% определите, есть ли основания считать, что производственный процесс требует переналадки.

№1.7.5. При отлаженном процессе упаковки чая в одну упаковку в среднем помещается 125 граммов чая, при этом дисперсия массы чая в упаковке не должна превышать 9 (граммов в квадрате). Отдел контроля качества отобрал 25 упаковок и рассчитал несмещённую оценку дисперсии $\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 13.5$.
Предполагается, что масса чая в упаковке имеет нормальное распределение.
а) Есть ли основания считать, что дисперсия массы чая превышает допустимый предел? Используйте уровень значимости 1%.
б) Гипотезу о равенстве средней массы 125 граммам решено проверять против двусторонней альтернативы с помощью следующего критерия: гипотеза о равенстве не отвергается, если средняя масса чая в выборке лежит в пределах $[123.452;126.548]$. В противном случае основная гипотеза отвергается, процесс упаковки останавливается для переналадки.
Предположим, что настоящая дисперсия массы чая в упаковке равна 9. С какой вероятностью хорошо отлаженный процесс упаковки будет остановлен для переналадки (иначе говоря, произойдёт ошибка первого рода)?

№1.7.6. Сёстры Аня и Таня по очереди моют посуду и время от времени эту посуду бьют. Аня считает сестру неумехой и утверждает, что Таня бьёт посуду чаще. Таня оправдывается и предлагает устроить следующий эксперимент: мыть тарелки по очереди (начиная с Ани) до тех пор, пока сёстры не разобьют две тарелки. Если оба раза виновата будет Таня, то она признаётся неумехой. Основная гипотеза Тани заключается в том, что она бьёт посуду так же часто, как и Аня. В каких пределах находится вероятность ошибки первого рода? В каком случае критерий является более мощным: если первой мыть посуду начинает Аня или Таня?

№1.7.7. В случайной выборке $X_1,...,X_n$ величины X_i независимы и принимают значения 0 и 1, вероятность события $\{X_i = 1\}$ обозначим p. Для проверки гипотезы $H_0 : p = 1$ против альтернативы $H_A : p < 1$ используется критерий: отвергать H_0, если $\sum_{i=1}^{n} X_i < n$.

а) Предположим, что на самом деле $p = 1$. С какой вероятностью основная гипотеза будет отвергнута?
б) Пусть $0 \le p < 1$. Выпишите мощность критерия как функцию от p.

№1.7.8. Случайные величины X_1, X_2 независимы и равномерно распределены на отрезке [0; a]. Исследователь проверяет гипотезу H_0: a=0 против альтернативы H_A: a>1 с помощью критерия: отвергнуть основную гипотезу, если $\max(X_1, X_2) > 0.9$. Какой уровень значимости соответствует этому критерию? Выпишите мощность критерия как функцию от a.

Часть 2. Классическая линейная нормальная регрессионная модель: парная регрессия

§2.1. Метод наименьших квадратов

№2.1.1. По парам наблюдений за длиной линии судьбы на ладони в см (X) и временем, прошедшим до момента вступления человека в брак в годах (Y) методом наименьших квадратов оценена зависимость $Y_i = \alpha + \beta X_i + \varepsilon_i$. Найдите оценки коэффициентов α и β, если $X_1 = 4$; $X_2 = 6$; $X_3 = 8$, $Y_1 = 20$; $Y_2 = 30$; $Y_3 = 25$. Рассчитайте коэффициент детерминации.

№2.1.2. Постройте оценку метода наименьших квадратов для коэффициента θ по набору наблюдений $(X_1, Y_1), ..., (X_n, Y_n)$ в следующих моделях:

а) $Y_i = \theta - \theta X_i + \varepsilon_i$, б) $Y_i = \theta + 2X_i + \varepsilon_i$.

№2.1.3. По наблюдениям за средней температурой зимы T (ºC) и затратами на отопление зимой C (тыс. руб.) на некотором предприятии была оценена линейная зависимость с помощью МНК: $\hat{C}_i = 18 - 0.2T_i$. Скажите, какими были бы МНК-оценки коэффициентов, если бы:

а) затраты измерялись в рублях,

б) температура измерялась в градусах по Фаренгейту ($T_F = 32 + \frac{9}{5}T$).

№2.1.4. По четырём наблюдениям оценивалась парная регрессия $Y_i = \alpha + \beta X_i + \varepsilon_i$ с помощью метода наименьших квадратов. Ниже приведены остатки регрессии e_i и значения регрессора X_i:

e_i:	6	0	a	$2a$
X_i:	10	13	12	b

Найдите значения a и b.

№2.1.5. Имеется оценённое с помощью МНК уравнение регрессии: $\hat{Y}_i = 17 - 0.3X_i$, RSS=200, ESS=300. Найдите выборочный коэффициент корреляции между X и Y, между Y и $\hat{Y}$.

№2.1.6. Найдите МНК-оценку $\hat{\beta}$ в регрессии $Y_i = \alpha + \beta X_i + \varepsilon_i$, если $\bar{X} = 10, \bar{Y} = 10, \hat{\alpha} = 2$.

№2.1.7. В Городе Всепобеждающего Оптимизма доля людей, ожидающих наступления полного счастья в ближайшие сроки, с каждым годом растёт. Эконометрист Надежда пытается построить модель динамики этой доли (обозначим её за Y) и сталкивается с проблемой: доля не может выходить за пределы от 0 до 1, так что линейная модель тренда не годится. Поэтому Надежда решает оценить зависимость вида $Y_t = \frac{\exp(\alpha + \beta t + \varepsilon_t)}{1 + \exp(\alpha + \beta t + \varepsilon_t)}$, где t - время (число лет, прошедших с начала наблюдения), а величина ε отражает отклонение моделируемой доли от тренда. Помогите Наде свести эту зависимость к линейной (по коэффициентам α и β), чтобы её можно было оценить с помощью обычного МНК.

№2.1.8. Составьте две выборки наблюдений за признаками X и Y так, чтобы в каждой из них уравнение регрессии $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ имело нулевой коэффициент детерминации, а при оценивании того уравнения по объединённой выборке R^2 был близок к единице (например, больше 0.9).

№2.1.9. Оценивание уравнения $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ с помощью метода наименьших квадратов дало оценки $\hat{\beta}_1$ и $\hat{\beta}_2$. Какими будут МНК-оценки коэффициентов γ_1 и γ_2 уравнения:

а) $\frac{Y_i}{10} = \gamma_1 + \gamma_2 X_i + \upsilon_i$, б) $Y_i = \gamma_1 + \gamma_2 \frac{X_i}{10} + \upsilon_i$, в) $(Y_i - \bar{Y}) = \gamma_1 + \gamma_2 (X_i - \bar{X}) + \upsilon_i$.

№2.1.10. Не проводя вычислений, догадайтесь, какими будут коэффициенты уравнения $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, подогнанного МНК по четырём наблюдениям:

Y_i:	0	0	1	1
X_i:	0	1	0	1

Каким будет коэффициент детерминации? Проверьте свои догадки.
Как добавить одно наблюдение, чтобы коэффициент детерминации стал приблизительно равен единице?

№2.1.11. Статистик Тимофей как-то обмолвился, что если при оценивании зависимости $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ коэффициент наклона оказывается нулевым, то коэффициент детерминации также равен нулю. Старец Феодосий привёл пример данных, для которых это не выполняется. Догадайтесь, что это был за пример.

№2.1.12. Придумайте два таких набора наблюдений за признаками X и Y, чтобы в каждом наборе метод наименьших квадратов давал положительный наклон при оценивании зависимости $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, а при оценивании по объединённому набору данных наклон подгоняемой прямой был бы отрицательным.

№2.1.13. В восьми торговых точках в центре города средняя стоимость пирожка с капустой составила 30 рублей, а выборочная дисперсия (смещённая оценка) — 9 руб^2. В десяти торговых точках на окраине средняя стоимость пирожка составила 25, а выборочная дисперсия — 12. Методом наименьших квадратов оцените коэффициенты уравнения $P_i = \beta_1 + \beta_2 C_i + \varepsilon_i$, где P_i — цена пирожка в точке i, C_i=1, если точка i находится в центре города, и C_i=0, если точка i находится на окраине. Рассчитайте коэффициент детерминации.

§2.2. Оценивание параметров регрессионной модели

№2.2.1. Оценивается регрессия случайной величины на номер наблюдения в выборке: $Y_i = \beta_1 + \beta_2 i + \varepsilon_i$, $i = 1, \ldots, n$. Предполагается, что случайные составляющие ε_i независимы, $E(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma_\varepsilon^2$.

а) Правда ли, что величина $\hat{\beta}_2 = \frac{Y_n - Y_1}{n-1}$ - несмещённая оценка параметра β_2? Найдите $D\left(\hat{\beta}_2\right)$.

б) При каком значении α величина $\hat{\beta}_1 = \alpha Y_2 - Y_1$ будет несмещённой оценкой для β_1? Найдите $D\left(\hat{\beta}_1\right)$.

№2.2.2. Эконометрист Надежда убеждена, что оценивание регрессии методом наименьших квадратов – излишняя морока. По мнению Нади, коэффициент β в уравнении $Y_i = \beta X_i + \varepsilon_i$ можно оценить, опираясь на одно-единственное наблюдение (обозначим номер этого наблюдения за j): $\hat{\beta} = \frac{Y_j}{X_j}$. Покажите, что такая оценка несмещённая, и посоветуйте Наде, как выбрать наблюдение j, чтобы результат был наиболее точным.

№2.2.3. По n наблюдениям оценивается модель $Y_i = \alpha + \beta X_i + \varepsilon_i$. Известно, что $E(\varepsilon_i) = 2$, остальные же предпосылки теоремы Гаусса-Маркова выполнены. Покажите, что оценка метода наименьших квадратов для параметра β является несмещённой, а для параметра α - смещённой. Предложите несмещённую оценку для α.

№2.2.4. Методом наименьших квадратов оценивается уравнение $Y_i = \alpha + \beta X_i + \varepsilon_i$. Покажите, что в случае $\beta = 0$ получаемая оценка параметра α оказывается менее эффективной, чем оценка $\hat{\alpha} = \bar{Y}$.

№2.2.5. Оценивание регрессии $Y_i = \alpha + \beta X_i + \varepsilon_i$ по 16 наблюдениям дало результаты: $TSS = 63$, $\hat{\sigma}_\varepsilon^2 = 2.5$. Найдите R^2.

§2.3. Прогнозирование, доверительные интервалы и проверка гипотез

№2.3.1. По ежегодным данным с 2000 по 2010 год (всего 11 наблюдений) оценивается тренд туристического потока из России в Финляндию с помощью уравнения регрессии $Trips_t = \beta_1 + \beta_2 t + \varepsilon_t$. Здесь t – год, которому соответствует наблюдение (t=0 для 2000 года, t=10 для 2010 года), а $Trips_t$ – число туристических поездок (в тысячах) российских граждан в Финляндию в году t. Вот результаты оценивания:

$$\hat{Trips}_t = \underset{(51.1)}{278.8} + \underset{(8.6)}{40.4}t, \quad R^2 = 0.7, \; TSS = 253000.$$

В скобках под оценками коэффициентов приведены их стандартные ошибки.

а) Согласно оценённой модели, насколько в среднем увеличивается поток туристов из России в Финляндию за два года?

б) Постройте 90% доверительный интервал для коэффициента β_2.

в) Оцените дисперсию случайной составляющей ε_t.

№2.3.2. По ежегодным данным[3] за 1975-1988 гг. (14 наблюдений) оценивалась зависимость цены на бензин (Petrol, центы за галлон) от цены на сырую нефть (Oil, доллары за баррель). Результаты оценивания приведены ниже:

$$\hat{Petrol}_i = \underset{(4.6)}{41.9} + \underset{(0.2)}{3.0} Oil_i$$

В скобках под оценками коэффициентов приведены их стандартные ошибки.

Также известно, что RSS=631.1, а TSS=12622. Предполагается, что все предпосылки классической линейной нормальной регрессионной модели выполнены.

а) Рассчитайте коэффициент детерминации R^2.

б) Проверьте гипотезу о том, что рост цены на нефть на 1 долл. за баррель соответствует росту цены на бензин на 2 цента за галлон, используя уровень значимости 10%.

в) В 1988 году цена на нефть составила 12.57 доллара за баррель. Какова была ожидаемая (прогнозируемая согласно имеющейся модели регрессии) цена на бензин? Какой должна быть цена на нефть, чтобы ожидаемая цена на бензин составила 100 центов за галлон?

№2.3.3. На рынке пирожков основным заменителем пирожков с капустой являются пирожки с картошкой. Исследователь, стараясь разобраться в механизмах ценообразования, оценивает регрессию $P_i = \beta_1 + \beta_2 C_i + \varepsilon_i$, где P_i - цена пирожка с картошкой у i-го продавца, а C_i - цена пирожка с капустой у того же продавца. Вот результат оценивания по 12 наблюдениям:

[3] эти данные мы взяли из книжки Л.И. Ниворожкиной и др. «Основы статистики с элементами теории вероятностей для экономистов».

$\hat{P}_i = \underset{(1.8)}{5.2} + \underset{(0.1)}{0.8} C_i, \quad RSS = 23.$

а) Найдите несмещённую оценку дисперсии случайной составляющей ε_i.

б) Постройте 95% доверительный интервал для коэффициента β_2.

в) Исследователь предполагает, что ожидаемая цена пирожка с картошкой у продавца, торгующего пирожками с капустой по 20 рублей, равна 22 рублям. Сформулируйте гипотезу исследователя в терминах коэффициентов регрессии.

№2.3.4. По ежегодным данным с 2002 по 2009 год оценивался тренд в динамике общей стоимости экспорта из РФ: $Ex_t = \alpha + \beta t + \varepsilon_t$, где t – год (t=0 для 2002 г., t=1 для 2003 г., …, t=7 для 2009 г.), Ex_t – стоимость экспорта из РФ во все страны в млрд. долл. Оценённое уравнение выглядит так: $\hat{Ex}_t = 111.9 + 43.2t$. Получены также оценки дисперсии случайной ошибки $\hat{\sigma}_\varepsilon^2 = 4009$ и ковариационной матрицы оценок коэффициентов:

$$\hat{V}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = \begin{pmatrix}1671 & -334\\ -334 & 95\end{pmatrix}$$

Спрогнозируйте стоимость экспорта на 2010 год и постройте 90% доверительный интервал для прогноза.

№2.3.5. Рассчитанный по 20 наблюдениям за признаками X и Y выборочный коэффициент корреляции составил 0.5. Есть ли основания считать, что между признаками существует регрессионная зависимость? Сформулируйте основную и альтернативную гипотезы в терминах коэффициентов регрессии, проверку проведите на уровне значимости 5%.

№2.3.6. По наблюдениям за 25 дней продавец Анфиса оценила зависимость числа проданных пирожков от их цены: $\hat{Q}_i = 83.7 - 1.5P_i, \quad R^2 = 0.7$.

а) Помогите Анфисе определить наилучшую цену с точки зрения максимизации выручки.
б) Постройте 90% доверительный интервал для коэффициента при цене.

№2.3.7. Считается, что Александр Дюма-отец намеренно удлинял романы, вводя множество второстепенных героев, поскольку оплачивалось его творчество в соответствии с количеством напечатанных строк.
Ленивый российский лингвист считает, что вовсе не обязательно считать строки оригинальных изданий и проверять, действительно ли Дюма получал по 3 франка за строку. Поскольку лингвисту не хочется идти в библиотеку, он берет 17 произведений Дюма, которые уже давно пылятся на полке дома у лингвиста и сопоставляет гонорар (сведения о гонорарах за каждую из этих 17 книг лингвист достал где-то в Интернете) с количеством страниц в русскоязычном издании романа.
Согласно регрессии, оценённой МНК, гонорар Дюма (в тыс. франков) зависит от количества страниц следующим образом:

$\hat{H}_i = 18 + 0,9P_i;\ RSS = 8,97;\ TSS = 25,75\ ;\ \mathrm{Cov}(\hat{\alpha}, \hat{\beta}) = 1,11$.

а) Проверьте, есть ли основания считать, что гонорар действительно зависел от числа страниц.
б) Постройте точечный и интервальный прогноз для гонорара, полученного Дюма за роман «Три мушкетёра», если в русскоязычном издании романа, доступном вышеупомянутому лингвисту, 542 страницы.

№2.3.8. На основании 10 наблюдений[4] за пациентами, страдающими эмфиземой лёгких, врач-исследователь выясняет зависимость площади поражённой части лёгких (переменная *Emph*, измеряется в процентах от общей площади лёгких) от числа лет курения (переменная *Years*, измеряется в годах). Вот оценённое уравнение регрессии:

[4] Эти наблюдения мы тоже взяли из книжки Л.И. Ниворожкиной и др.

$$\hat{Emph}_i = \underset{(5.6)}{11.2} + \underset{(0.4)}{1.3}\,Years_i$$

В скобках под оценками коэффициентов приведены их стандартные ошибки.
а) Постройте 90% доверительный интервал для свободного члена.
б) Определите стаж курения, который приводит к поражению лёгких в среднем на 40%.
в) Проверьте гипотезу о том, что каждый дополнительный год курения приводит к увеличению поражённой площади в среднем на два процента, используя уровень значимости 5%.
г) Рассчитайте R^2.

№2.3.9. Рассмотрим классическую линейную нормальную регрессионную модель $y_i = \alpha + \varepsilon_i$, которая оценивается по 4 наблюдениям y_1, y_2, y_3, y_4.
а) Выведите оценку метода наименьших квадратов для коэффициента α.
б) Найдите дисперсию оценки из пункта (а), считая дисперсию случайной ошибки равной 1.
в) При оценивании были получены результаты: $\hat{\alpha} = 8, \sum_{i=1}^{4}(y_i - \hat{\alpha})^2 = 6.$ Рассчитайте 90% доверительный интервал для нового наблюдения y_5.
г) Каким был бы доверительный интервал для прогноза при неизвестной дисперсии случайной составляющей?

Часть 3. Классическая линейная нормальная регрессионная модель: множественная регрессия

§3.1. Метод наименьших квадратов, оценивание и интерпретация параметров регрессионной модели

№3.1.1. Оцените коэффициенты модели $y = X\beta + \varepsilon$ по данным: $y = \begin{pmatrix} 6 \\ 8 \\ 6 \\ 4 \end{pmatrix}$, $X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \\ 1 & 3 & 1 \end{pmatrix}$.

Рассчитайте прогнозные значения, остатки и коэффициент детерминации.

№3.1.2. Оценена регрессия: $\hat{Y} = 34 + 2X_2 - 4X_3$. Какими будут оценки коэффициентов модели $Y = \beta_1 + \beta_2(X_2 + X_3) + \beta_3 X_3 + \varepsilon$?

№3.1.3. При оценивании по 40 наблюдениям зависимости $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i$ были получены значения $R^2 = 0.6, TSS = 200$. Рассчитайте несмещённую оценку дисперсии случайной составляющей и скорректированный коэффициент детерминации.

№3.1.4. Оценивалась регрессия величины заработной платы (*w*) на образование, измеряемое количеством лет обучения (*sch*), стаж работы на последнем рабочем месте (*ten*), возраст (*age*) и квадрат возраста. Также в модель была включена дамми-переменная *sex*, равная единице для мужчин и нулю для женщин. Ниже приведено оценённое уравнение регрессии:

$$\widehat{\ln w_i} = 2.13 + 0.07 \cdot sch_i + 0.01 \cdot ten_i + 0.05 \cdot age_i - 0.0005 \cdot age_i^2 + 0.1 \cdot sex_i$$

а) Оцените, на сколько процентов зарплата мужчин выше зарплаты женщин при прочих равных условиях.

б) Определите возраст, в котором для индивида с заданными характеристиками образования, стажа и пола ожидается наибольший уровень зарплаты.

№3.1.5. Исследуется зависимость расходов семей на продукты питания от их доходов. По данным некоторой выборки оценены две регрессионные модели:

(1) $\widehat{\ln E_i} = -0.77 + 0.72 \ln I_i + 0.22 \ln(1 + N_i)$,

(2) $\hat{E}_i = 0.72 + 0.13 I_i + 1.04 N_i$,

где E – расходы семьи на продукты питания в расчёте на человека, тыс. руб., I – суммарный доход всей семьи в тыс. руб., поделенный на число членов семьи, N - количество детей в семье. В уравнении (1) к числу детей прибавлена единица, т.к. для бездетных семей логарифм числа детей не существует.

Рассмотрим семью с двумя детьми, средний доход члена которой равен 30 тыс. руб. По каждой из моделей (1) и (2) рассчитайте эластичность расходов на питание по доходу.

№3.1.6. Придумайте такой набор наблюдений за величинами (Y, X_2, X_3), чтобы в регрессии $Y_i = \alpha_1 + \alpha_2 X_{2,i} + \upsilon_i$ оценка коэффициента при X_2 была положительной, а в регрессии $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$ оценка коэффициента при X_2 была отрицательной.

№3.1.7. Оценивается модель $Y_i = \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$. Эконометрист Надежда допускает, что коэффициент β_1 может оценить, применив МНК для парной связи (также без свободного члена): $Y_i = \beta_1 X_{1,i} + \nu_i$. Покажите, что получаемая таким способом оценка будет несмещённой в двух случаях:

1) если $\beta_2 = 0$,

2) если $\sum_{i=1}^{n} X_{1,i} X_{2,i} = 0$.

№3.1.8. Модель $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$ удовлетворяет всем предпосылкам классической линейной нормальной регрессионной модели. Известно, что $D(\varepsilon_2) = 6$. Заполните пропуски:

$D(\varepsilon_4) =$_____ $Cov(\varepsilon_2, \varepsilon_4) =$_____ $E(\varepsilon_2) =$_____

Множество возможных значений коэффициента R^2 в регрессии $X_{2,i} = \alpha_1 + \alpha_2 X_{3,i} + \upsilon_i$: _____

Функция плотности ε_2:

№3.1.9. Рассмотрим классическую линейную нормальную регрессионную модель $y = X\beta + \varepsilon$, $D(\varepsilon_i) = \sigma_\varepsilon^2$.

а) Найдите $MSE(\hat{\sigma}_\varepsilon^2)$, где $\hat{\sigma}_\varepsilon^2 = \dfrac{RSS}{n-k}$.

б) Рассмотрим оценку $\tilde{\sigma}_\varepsilon^2 = c \cdot RSS$. При каком *c* она будет иметь наименьший средний квадрат ошибки?

№3.1.10. Для уравнения $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ рассчитаны оценки $\hat{\beta}_1$ и $\hat{\beta}_2$ и коэффициент R^2. Какими будут оценки уравнения $Y_i = \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$ по тем же данным, если $Z_i = 1 - X_i$? Каким будет коэффициент детерминации?

§3.2. Прогнозирование, доверительные интервалы и проверка гипотез

№3.2.1. Для оценивания уравнения спроса на импортируемое мыло и моющие средства была выбрана следующая спецификация: $\ln M_i = \beta_1 + \beta_2 \ln GDP_i + \beta_3 \ln Pf_i + \beta_4 \ln Pd_i + \varepsilon_i$, где:

M – объём импорта мыла и моющих средств в килограммах,
GDP – ВВП России в месяце, которому соответствует наблюдение,
Pf – индекс цен на импортное мыло и моющие средства
Pd – индекс цен на отечественное мыло и моющие средства.

Следующая таблица содержит результаты оценивания:

Коэфф-т	**Оценка коэфф-та**	**Станд. ошибка**	***t*-статистика**	***P*-value**	**95% доверит. интервал**
β_1	3.188	2.067	1.54	0.132	(-1.003; 7.379)
β_2	0.596	0.346	1.73	0.093	(-0.104; 1.297)
β_3	-0.781	0.167	-4.68	0.000	(-1.120; -0.443)
β_4	0.375	0.854	0.44	0.664	(-1.357; 2.107)

Количество наблюдений:	40	**Коэффициент детерминации R^2:**	0.8176
***F*-статистика:**	53.79	***P*-значение для *F*-статистики:**	0.0000

Проверьте значимость регрессии в целом и определите, какие коэффициенты оказываются значимыми по отдельности на уровне 10%.

№3.2.2 По 24 наблюдениям, соответствующим 24 представительствам национальной компании по торговле недвижимостью, была оценена регрессия объёма годовых продаж (*Sales*, млн. долл.) на число агентов в представительстве (*Agents*) и объём затрат на рекламу (*AdvCosts*, тыс. долл.). Ниже приведены результаты оценивания:

$$\hat{Sales}_i = -7.7 + 0.3 \cdot AdvCosts_i + 0.8 \cdot Agents_i, \quad RSS = 42$$

$$(X'X)^{-1} = \begin{pmatrix} 3.2 & -0.003 & -0.3 \\ -0.003 & 0.005 & 0.002 \\ -0.3 & 0.002 & 0.4 \end{pmatrix}$$

а) Найдите оценку дисперсии случайной составляющей.
б) Найдите оценку дисперсии коэффициента перед *AdvCosts*.
в) Проверьте значимость коэффициента при затратах на рекламу на уровне значимости 1%.

№3.2.3. По 27 наблюдениям оценена регрессия $\hat{Y}_i = 5.2 - 0.5X_{2,i} + 0.6X_{3,i}$, RSS = 12.

$$(X'X)^{-1} = \begin{pmatrix} 0.5 & -0.2 & -0.2 \\ -0.2 & 0.1 & -0.1 \\ -0.2 & -0.1 & 0.2 \end{pmatrix}$$

Рассчитайте прогнозное значение $\hat{Y}_0$ для случая $X_{2,0} = 0$, $X_{3,0} = 1$ и найдите 95% доверительный интервал для прогноза.

№3.2.4. Исследователь изучал, как формируется стоимость подержанного автомобиля, для чего решил оценить регрессию:

$$Цена_i = \beta_1 + \beta_2 Пробег_i + \beta_3 Число\_поломок_i + \beta_4 Иномарка_i + \beta_5 Литраж_i + \beta_6 Длина_i + \varepsilon_i.$$

Перед вами результаты оценивания:

```
      Source |       SS       df       MS              Number of obs =      50
-------------+------------------------------           F(  5,    44) =    4.86
       Model |   176083058     5  35216611.6           Prob > F      =  0.0013
    Residual |   318894258    44  7247596.78           R-squared     =  0.3557
-------------+------------------------------           Adj R-squared =  XXXX
       Total |   494977316    49  10101577.9           Root MSE      =  2692.1

------------------------------------------------------------------------------
       цена  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      пробег |  -321.6181   172.2067    -1.87   0.068     -668.678     25.4418
число_поломок|   -466.653   603.6094    -0.77   0.444    -1683.148    749.8418
     иномарка|   5102.066   1749.387     2.92   0.006       XXXX        XXXX
       литраж|  -196.3938    173.674    -1.13   0.264    -546.4106    153.6231
       длина |   121.7116   54.30542     XXXX   0.030     12.26624     231.157
       _cons |  -7961.162   11653.58    -0.68   0.498    -31447.42    15525.09
------------------------------------------------------------------------------
```

а) Заполните пропуски в таблице (обозначены «**XXXX**»).
б) Определите, при каких уровнях значимости регрессия в целом будет значима. Сформулируйте основную и альтернативную гипотезы в терминах коэффициентов регрессии.
в) Проверьте значимость коэффициента при переменной «литраж» на уровне значимости 5%. Сформулируйте основную и альтернативную гипотезы в терминах коэффициентов регрессии.
г) Исследователь предполагает, что ожидаемая цена иномарки, претерпевшей четыре поломки, совпадает с ценой ни разу не ломавшегося отечественного автомобиля с аналогичными характеристиками. Сформулируйте эту гипотезу в терминах коэффициентов регрессии.

№3.2.5. По 25 наблюдениям оценено уравнение регрессии количества проданных в некоторой торговой точке бутылок газированной воды Q_i от цены газировки P_i, средней температуры дня T_i. В модель также была включена дамми-переменная W_i, равная единице для наблюдений, соответствующих выходных дням, и нулю для будних дней. Оценивание дало следующие результаты: $\hat{Q}_i = \underset{(5.7)}{67.5} - \underset{(1.0)}{2.4} P_i + \underset{(1.5)}{2.1} T_i + \underset{(3.2)}{13.7} W_i, \quad R^2 = 0.7$ (в скобках под оценками коэффициентов указаны их стандартные ошибки).

а) Проверьте адекватность регрессии на уровне значимости 1%.

б) Определите, какие из включённых в модель регрессоров оказывают значимое влияние на объём продаж? Используйте уровень значимости 5%.

в) Постройте 95% доверительный интервал для свободного члена.

г) Представим, что в данное уравнение регрессии вместо переменной W_i включена переменная N_i, равная нулю для выходных и единице для будних дней. Как изменится оценённое уравнение регрессии?

№3.2.6. Склонный заниматься ерундой хозяин кафе разрабатывает регрессионную модель спроса на кофе, связывающую количество проданных за день чашек кофе с ценами кофе и чая. Кроме того, он учитывает, что спрос в будни, в выходные дни и в праздники различен. Перед вами распечатка таблицы, в которую он вбивал результаты наблюдений за 8 дней.

№	Чашек кофе продано	Цена чашки кофе, руб.	Цена чашки чая, руб.	День
1	87	50	40	будний
2	75	50	40	будний
3	101	50	45	выходной
4	109	60	50	выходной
5	48	70	50	будний
6	139	60	45	праздничный
7	75	60	45	будний
8	77	60	50	будний

а) Сначала хозяин оценивает регрессию числа проданных чашек кофе только на цены кофе и чая. Выпишите матрицу регрессоров (X в матричной записи уравнения регрессии: $y = X\beta + \varepsilon$).

б) Теперь хозяин кафе решает учесть различия между будними, выходными и праздничными днями. Выпишите матрицу регрессоров для уравнения с учётом этих различий.

в) Наконец, хозяин собирается проверить, действительно ли различия между буднями, выходными и праздниками имеют место. Выпишите основную и альтернативную гипотезы в терминах коэффициентов уравнения регрессии.

№3.2.7. По 30 наблюдениям была оценена некая регрессия (в скобках под оценками коэффициентов приведены их стандартные ошибки):

$$\hat{Y}_i = \underset{(2.4)}{23.5} - \underset{(0.2)}{0.6} X_{2,i} + \underset{(0.7)}{1.2} X_{3,i} + \underset{(0.1)}{0.9} X_{4,i} - \underset{(1.2)}{0.2} X_{5,i}, \; RSS = 150, \; R^2 = 0.7$$

а) Рассчитайте $\hat{\sigma}_\varepsilon^2$ - несмещённую оценку дисперсии случайной составляющей.

б) Рассчитайте скорректированный коэффициент детерминации R^2_{adj}.

в) Укажите переменные, при исключении которых R^2_{adj} увеличится?

г) Постройте 95% доверительный интервал для коэффициента при X_2.

д) На уровне значимости 5% проверьте гипотезу о том, что увеличение переменной X_4 на единицу при прочих равных условиях соответствует увеличению Y в среднем на 0.5 (укажите основную и альтернативную гипотезы).
е) В целях проверки некоторой гипотезы исследователь оценил дополнительную регрессию: $\hat{Y}_i = \underset{(3.2)}{28.4} - \underset{(0.2)}{0.5}(X_{2,i} - X_{3,i}) + \underset{(0.2)}{1.1} X_{4,i}$. Выпишите в терминах коэффициентов β_j основную гипотезу, которую хотел проверить исследователь.

№.3.2.8. Побывав у пяти помещиков, Чичиков выяснил, что цена «мёртвой души» Y (в рублях), по которой её согласен продать *i*-ый помещик, следующим образом связана с количеством «мертвых душ» (Q) у этого помещика и с расстоянием (D) от города N до деревни, где живет помещик:

$$\hat{Y}_i = \underset{(0.31)}{2.08} - \underset{(0.22)}{0.73} Q_i - \underset{(0.37)}{0.36} D_i, \; R^2 = 0{,}81.$$

Рассчитайте скорректированный коэффициент детерминации. Как он изменится, если исключить из регрессии «расстояние от города до деревни»? Почему вы так считаете?

№3.2.9. Исследовательская группа в составе статистика Тимофея и эконометриста Надежды изучала различия в оплате труда между мужчинами и женщинами на основании выборки из 200 наблюдений за сотрудниками и сотрудницами концерна имени Международного женского дня 8 марта.
Тимофей проверил гипотезу о совпадении средних зарплат для мужчин и женщин и получил такие результаты:

Двухвыборочный t-тест с одинаковыми дисперсиями

	Мужчины	*Женщины*
Среднее	18.30197	16.52504
Дисперсия	16.26656	10.89242
Наблюдения	100	100
Объединенная дисперсия	13.57949	
Гипотетическая разность средних	0	
Df	198	
t-статистика	3.409681	
P(T<=t) одностороннее	0.000394	
t критическое одностороннее	1.652586	
P(T<=t) двухстороннее	0.000788	
t критическое двухстороннее	1.972017	

Отсюда он сделал вывод, что средняя заработная плата мужчин выше, чем у женщин.
Надежда оценила уравнение регрессии заработной платы на переменные пола (Female=1 для женщин, 0 иначе) и наличия высшего образования (High=1 для рабочих с высшим образованием, 0 иначе):

Регрессионная статистика	
Множественный R	0.925987
R-квадрат	0.857452
Нормированный R-квадрат	0.856004
Стандартная ошибка	1.4352
Наблюдения	200

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	2440.833039	1220.41652	592.49339	4.6269E-84
Остаток	197	405.7801463	2.0597977		

Итого	199	2846.613186				

	Коэфф-ты	*Станд. ош.*	*t-стат.*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	14.37657	0.1857431	77.4003128	1.994E-149	14.01027398	14.74287411
Female	1.049355	0.220006867	4.7696455	3.5847E-06	0.615483842	1.483225681
High	7.850794	0.235817915	33.2917603	7.8595E-83	7.385741949	8.315845059

Посмотрев эту табличку, она заявила, что принадлежность к мужскому полу отрицательно связана с заработной платой.
Чем объясняется разница между результатами, полученными Тимофеем и Надеждой? Как эти результаты согласуются?

№3.2.10. По квартальным данным оценивается зависимость объёма импорта косметики (M) от ВВП страны-импортёра (Y), индекса внутренних цен на косметику (Pd) и индекса импортных цен на косметику (Pf). При оценивании используется логарифмическая спецификация.
Для учёта возможных сезонных колебаний в модель введены дамми-переменные D2, D3 и D4. D2 = 1, если наблюдение соответствует 2-му кварталу, в остальных случаях D2=0. D3 и D4 равны единице для 3-го и 4-го кварталов соответственно.
Ниже приведены результаты оценивания пяти регрессий, каждая из которых оценивалась по 40 наблюдениям.
Оценка уравнения регрессии без ограничений:

$$\ln \hat{M}_i = 7.6 + 0.8\ln Y_i + 0.1\ln Pd_i - 0.3\ln Pf_i - 0.4D2_i - 0.2D3_i - 0.2D4_i, \quad RSS = 0.37$$

Оценка уравнений регрессии с ограничениями:

$$\ln \hat{M}_i = 9.3 + 0.6\ln Y_i - 0.1(\ln Pf_i + \ln Pd_i) - 0.3D2_i - 0.1D3_i - 0.3D4_i, \quad RSS = 0.48$$

$$\ln \hat{M}_i = 9.0 + 0.6\ln Y_i - 0.2(\ln Pf_i - \ln Pd_i) - 0.4D2_i - 0.0D3_i - 0.3D4_i, \quad RSS = 0.40$$

$$\ln \hat{M}_i = 12.9 + 0.3\ln Pd_i + 0.1\ln Pf_i - 0.2D2_i - 0.1D3_i - 0.3D4_i, \quad RSS = 0.45$$

$$\ln \hat{M}_i = 9.4 + 0.5\ln Y_i - 0.3D2_i - 0.2D3_i - 0.3D4_i, \quad RSS = 0.42$$

$$\ln \hat{M}_i = 10.1 + 0.4\ln Y_i - 0.1\ln Pd_i + 0.1\ln Pf_i, \quad RSS = 1.48$$

а) Проверьте наличие сезонных эффектов на уровне значимости 1%.
б) На уровне значимости 5% проверьте гипотезу о том, что коэффициент при логарифме индекса внутренних цен имеет то же абсолютное значение, но другой знак, что и коэффициент при логарифме индекса импортных цен. Иначе говоря, проверьте гипотезу о том, что на импорт влияют не внутренняя и внешняя цены по отдельности, но их отношение.

№3.2.11. По 55 наблюдениям за 30 мужчинами и 25 женщинами экономист Маня Манкина методом наименьших квадратов оценила регрессию:

$\hat{EARNINGS} = 3744,44 - 35,54 \cdot S + 3,59 \cdot TEST$, $TSS = 100000000$, $ESS = 50000000$, где $EARNINGS$ – почасовая зарплата (\$), S – длительность обучения (годы), $TEST$ – результаты ЕГЭ по математике (баллы).
Регрессия, оценённая для женщин, имеет вид:

$\hat{EARNINGS} = 5664,05 - 114,22 \cdot S - 9,05 \cdot TEST$, $TSS = 51000000$, $ESS = 550000$

Регрессия, оценённая для мужчин, имеет вид:

$\hat{EARNINGS} = 2783,67 + 112,11 \cdot S + 13,50 \cdot TEST$, $TSS = 48000000$, $ESS = 870000$

а) Найдите коэффициенты такого уравнения регрессии с фиктивными переменными для пола, оценённого по всей выборке:

$$\hat{EARNINGS} = \hat{\beta}_1 + \hat{\beta}_2 \cdot S \cdot MALE + \hat{\beta}_3 \cdot S \cdot FEMALE + \hat{\beta}_4 \cdot FEMALE + \hat{\beta}_5 \cdot TEST \cdot MALE + \hat{\beta}_6 \cdot TEST$$

Здесь *MALE* – фиктивная переменная, равная 1, если индивид мужчина, и равная 0, если индивид женщина, *FEMALE* – фиктивная переменная, равная 1, если индивид женщина, и равная 0, если индивид мужчина.

б) Есть ли основания считать, что регрессионные зависимости для мужчин и для женщин различаются?

№3.2.12. По 30 наблюдениям были оценены два уравнения регрессии:

(1) $\hat{Y}_i = 27.3 + 1.6X_{2,i} + 3.5X_{3,i} - 0.5X_{4,i}, \quad RSS = 29.5$

(2) $\hat{Y}_i = 39.6 + 1.0X_{2,i}, \quad RSS = 38.1.$

а) На уровне значимости 1% проверьте гипотезу о том, что коэффициенты при переменных X_3 и X_4 равны нулю.

б) Предположим, что вам нужно проверить гипотезу о том, что коэффициенты при переменных X_2 и X_3 совпадают. Выпишите в общем виде уравнение регрессии, оценив которое вы могли бы проверить эту гипотезу.

№3.2.13. Для 50 штатов США оценена зависимость участия женщин в рабочей силе (переменная PLF, измерена в процентах) от медианного дохода домохозяйства (MHPI, тыс. долл.), уровня образования женщин (EDUC, количество лет обучения) и уровня безработицы среди женщин (UR, %). Ниже приведены описательные статистики изучаемых переменных[5]:

Переменная	Выборочное среднее	Выб. станд. отклонение
PLF	58.72	4.37
MHPI	18.19	2.76
EDUC	12.48	0.19
UR	5.96	1.43

А вот результаты оценивания регрессии $PLF_i = \beta_1 + \beta_2 MHPI_i + \beta_3 EDUC_i + \beta_4 UR_i + \varepsilon_i$:

Коэффициент	**Оценка коэфф-та**	**Станд. ошибка**	***t*-статистика**	***p*-значение**
β_1 (константа)	0.158	34.91	0.00	0.996
β_2 (доход)	0.406	0.17	2.34	0.024
β_3 (образование)	4.842	2.81	1.72	0.092
β_4 (безработица)	-1.554	0.34	-4.57	0.000
Число наблюдений:		50	R^2**:**	0.54
***F*-статистика:**		18.24	***p*-значение:** (для *F*-стат.)	0.000

а) Является ли оценённая регрессия значимой в целом?

б) Какие из оцениваемых коэффициентов значимо отличаются от нуля на уровне значимости 5%?

в) Какая из объясняющих переменных модели вносит наибольший вклад в различия в уровне участия женщин в рабочей силе между штатами? Рассчитайте стандартизированные коэффициенты регрессии.

г) Допустим, что для одного из штатов США все объясняющие переменные равны их средним значениям в нашей выборке. Каков ожидаемый уровень участия женщин в рабочей силе для этого штата?

[5] Данные взяты из книжки P.Newbold “Statistics for Business and Economics”

№3.2.14. На рынке подержанных электрических чайников в селе Самоварово представлены товары, отличающиеся один от другого, в основном, по трём характеристикам: сроку эксплуатации *Age* (годы), объёму вмещаемой жидкости *Volume* (литры) и потреблению электроэнергии *PC* (киловатт в час). По выборке из 26 образцов товара была оценена зависимость цены *Price* (рубли) от этих характеристик:

$\hat{Price} = 2150 - 350 \cdot Age + 400 \cdot Volume - 330 \cdot PC, \quad R^2 = 0.75.$

Описательная статистика выборки (приведены несмещённые оценки):

Признак:	*Age*	*Volume*	*PC*	*Price*
Среднее:	1.4	1.6	1.9	???
Ст. отклонение:	1.0	0.4	0.4	330

а) Какой из факторов вносит наибольший вклад в разброс цен на чайники?
б) Какова средняя цена чайника в выборке?
в) Какова оценка дисперсии случайной составляющей в регрессионной модели?
г) При проверке значимости потребления электроэнергии оказалось, что *p*-значение равно 0.02. Найдите стандартную ошибку коэффициента при этой переменной.

№3.2.15. Эконометрист Надежда разрабатывает модель удоя. Надя считает, что средний удой молока от коровы году *t* в её местности ($Yield_t$, кг) определяется расходом кормов на корову ($Fodder_t$, центнеры) и тенденцией к улучшению качества коровьей жизни, для учёта которой в модель включена переменная времени *t* (номер года в выборке). Ниже приведены результаты оценивания уравнения $Yield_t = \beta_1 + \beta_2 Fodder_t + \beta_3 t + \varepsilon_t$ в программе Stata:

```
      Source |       SS       df       MS              Number of obs =      12
-------------+------------------------------           F(  2,     9) =    7.06
       Model |  801273.721     2   400636.86           Prob > F      =  0.0143
    Residual |  510826.279     9  56758.4755           R-squared     =  0.6107
-------------+------------------------------           Adj R-squared =  0.5242
       Total |     1312100    11  119281.818           Root MSE      =  238.24

------------------------------------------------------------------------------
       yield |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      fodder |   32.64206   11.00422     2.97   0.016     7.748778    57.53535
           t |   29.90836   20.64307     1.45   0.181    -16.78951    76.60623
       _cons |   1523.787   459.7586     3.31   0.009     483.7411    2563.834
------------------------------------------------------------------------------
```

а) Выпишите основную и альтернативную гипотезы для проверки значимости регрессии в целом:
H_0: _____ H_A: _____
б) Сформулируйте вывод о значимости/незначимости регрессии в целом на уровне 5%:
_____
в) Рассчитайте прогнозное значение удоя на следующий год ($t = 13$) в предположении, что расход кормов составит 50 центнеров кормовых единиц на корову:
_____
г) Выпишите основную и альтернативную гипотезы для проверки значимости коэффициента при переменной расхода кормов:
H_0: _____ H_A: _____
д) Сформулируйте вывод о значимости/незначимости этого коэффициента на уровне 5%:
_____

№3.2.16. Демограф Дмитрий, изучая репродуктивное поведение жительниц Ивановской области, оценивает модель $\ln EM_i = \beta_1 + \beta_2 Rubia_i + \beta_3 Pelirroja_i + \beta_4 Ciudad_i + \varepsilon_i$, где EM_i - возраст вступления в брак, $Rubia_i = 1$ для светловолосых женщин, 0 иначе, $Pelirroja_i = 1$ для рыжеволосых женщин, 0 иначе, $Ciudad_i = 1$ для жительниц городской местности, 0 иначе.

Дмитрий делит всех женщин по типу местности на горожанок и сельчанок, по естественному цвету волос – на темноволосых, рыжих и светловолосых.
Дмитрий хочет проверить следующие гипотезы:
I. Средний возраст при вступлении в брак не связан с типом местности.
II. Средний возраст при вступлении в брак не связан с цветом волос.
III. Средний возраст при вступлении в брак для рыжих и светловолосых женщин одинаков.
а) В каждом случае выразите основную и альтернативную гипотезы через коэффициенты модели.
б) Какое уравнение регрессии вы бы посоветовали оценить Дмитрию, чтобы учесть ограничение, соответствующее гипотезе III?
в) Коллега Ирина убеждает Дмитрия, что роль цвета волос в определении возраста вступления в брак зависит от того, проживает женщина в селе или городе (например, рыжие сельчанки могут выходить замуж в том же возрасте, что и светловолосые, но среди рыжих горожанок возраст вступления в брак выше, чем среди горожанок со светлыми волосами, и т.п.) Выпишите уравнение регрессии, которое стоит оценить, чтобы учесть предполагаемый Ириной эффект.

№3.2.17. Старший экономист тётя Циля исследует связь между изменениями цен на недвижимость в Москве ΔPR_RUR и динамикой курса доллара ΔUSD. Используя месячные данные, для периода с января 2000 года по февраль 2011 года (n=133) она получает следующий результат:

$$\underset{(350)}{\Delta PR\_RUR} = \underset{(1450)}{784 + 5058\,\Delta USD\_RATE}; \quad RSS = 573\,000\,000$$

Однако, хорошенько поразмыслив, тётя Циля решила, что стоит включить в уравнение модели динамику среднедушевого дохода ΔINC и процентной ставки ΔIR, и оценила следующую модель:

$$\underset{(921)}{\Delta PR\_RUR} = \underset{(1054)}{1876 + 2012\,\Delta USD\_RATE} - \underset{(4055)}{6027\,\Delta IR} - \underset{(0.540)}{0.132\,\Delta INC}; \quad RSS = 280\,000\,000$$

а) На 1% уровне значимости проверьте гипотезу о значимости процентной ставки и среднедушевого дохода.
б) Проверьте гипотезу о равенстве коэффициента при среднедушевом доходе нулю. Найдите *p*-значение.
в) Какое уравнение вы бы оценили, чтобы учесть возможную сезонность?

№3.2.18. По ежегодным данным за 23 года была оценена зависимость индекса цен на индийский чай $PrInd_t$ от индексов цен на цейлонский чай $PrCey_t$ и на бразильский кофе $PrCoffee_t$:

$$\hat{PrInd}_t = \underset{(3.4)}{33.6} + \underset{(0.2)}{0.9\,PrCey_t} - \underset{(0.1)}{0.2\,PrCoffee_t}, \quad R^2 = 0.6.$$

В скобках под оценками коэффициентов приведены их стандартные ошибки. Все переменные измерялись в процентах.
а) Рассчитайте ожидаемое значение индекса цен на индийский чай для случая, когда индекс цен на цейлонский чай составит 110%, а на кофе – 100%.
б) Проверьте значимость регрессии на уровне 1%.
в) Используя уровень значимости 5%, выясните, есть ли основания считать, что цена индийского чая связана с ценой бразильского кофе.

№3.2.19. По 45 наблюдениям методом наименьших квадратов экономист Вася Пробкин оценил регрессию $EARNINGS = -15 + 2.5 \cdot S + 8 \cdot MALE,\ TSS = 300,\ ESS = 150$,
где *EARNINGS* – почасовая зарплата (\$), *S* – уровень образования (годы обучения),
MALE – фиктивная переменная, равная 1, если индивидуум мужчина, и равная 0, если индивидуум женщина.
Однако, хорошенько подумав, Вася Пробкин решил включить в число объясняющих переменных результаты ЕГЭ (из 100 баллов). Вот какая регрессия у него получилась:

$$EARNINGS = -18 + 2,7 \cdot S + 6,1 \cdot MALE + 0,07 \cdot EGE, \quad R^2 = 0,79$$

Вычислите, как изменился скорректированный коэффициент детерминации. Что вы можете сказать о стандартной ошибке коэффициента при результатах ЕГЭ?

№3.2.20. Рассмотрим пример применения регрессионного анализа при разрешении конфликтов в ВТО. В ответ на жалобу Евросоюза о дискриминации, проводимой Чили в отношении импортируемого виски в пользу чилийских производителей писко (крепкий спиртной напиток из винограда – национальный напиток чилийцев), со стороны Чили были предъявлены результаты регрессионного анализа спроса на писко. Оценивалось следующее уравнение регрессии:

$$Объём\ спроса\ на\ писко_i = \beta_0 + \beta_1 \times Доход_i + \beta_2 \times Цена\ писко_i + \\ + \beta_3 \times Цена\ виски_i + \beta_4 \times Цена\ вина_i + \beta_5 \times Цена\ пива_i + \varepsilon_i$$

Вот результаты оценивания:

Коэффициент	**Оценка коэфф.**	**Станд. ошибка**	***t*-статистика**	***P*-значение**	**95% доверит. интервал**
β_0	3,5771	3,6554	0,9786	0,3534	(-4.6920; 11.8461)
β_1 (доход)	-0,0072	1,2109	-0,0059	0,9954	(-2.7465; 2.7321)
β_2 (цена писко)	-1,3109	0,4574	-2,8661	0,0186	(-2.3456; -0.2762)
β_3 (цена виски)	0,1248	0,5158	0,2419	0,8143	(-1.0421; 1.2917)
β_4 (цена вина)	0,5963	0,4030	1,4796	0,1731	(-0.3154; 1.5079)
β_5 (цена пива)	0,3622	1,2132	0,2985	0,7721	(-2.3823; 3.1067)
Количество наблюдений:		15	**Коэффициент детерминации R^2:**		0,9758
***F*-статистика:**		72,6767	***P*-значение для *F*-статистики:**		0,0000

Как представители Чили использовали эти результаты для укрепления своей позиции в споре?

Часть 4. Трудности регрессионного анализа и отклонения от классической модели

§4.1. Регрессионная диагностика.

№4.1.1. Дана матрица $X = \begin{pmatrix} 1 & 10 & 3 \\ 1 & 6 & 1 \\ 1 & 12 & a \\ 1 & 8 & 2 \end{pmatrix}$.

Каким должно быть число a, чтобы между столбцами матрицы была строгая мультиколлинеарность?

Допустим, $a = 2$. Рассчитайте значения *VIF* для оценок коэффициентов регрессии $y = X\beta + \varepsilon$, где y - любая величина.

№4.1.2. Оценив зависимость логарифма заработной платы от возраста (age), квадрата возраста (age_sq), стажа работы (ten) и наличия высшего образования (higheduc), Надя представила результаты своего исследования:

```
. reg lnwage ten higheduc age age_sq

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  4,   495) =   69.50
       Model |  56.6465102     4  14.1616276           Prob > F      =  0.0000
    Residual |  100.865932   495  .203769559           R-squared     =  0.3596
-------------+------------------------------           Adj R-squared =  0.3545
       Total |  157.512442   499  .315656196           Root MSE      =  .45141

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ten |    -.04216   .0180433    -2.34   0.020    -.0776108   -.0067092
    higheduc |   .1032464   .0445535     2.32   0.021     .0157091    .1907838
         age |   .0562991   .0270907     2.08   0.038     .0030722     .109526
      age_sq |   -.001414   .0003692    -3.83   0.000    -.0021394   -.0006886
       _cons |  -.9322643   .4941175    -1.89   0.060    -1.903091     .038562
------------------------------------------------------------------------------
```

Обратив внимание на отрицательный коэффициент при стаже, Надежда заметила, что этот результат плохо согласуется с экономической интуицией и, видимо, вызван наличием мультиколлинеарности – тесной связи между возрастом и стажем. Слушавший доклад статистик Тимофей сказал, что мультиколлинеарность может присутствовать, но не должна приводить к таким результатам. Почему Тимофей решил, что причина Надиного результата не связана с мультиколлинеарностью?

№4.1.3. По набору наблюдений за тремя величинами *Y*, X_2 и X_3 была оценена корреляционная матрица:

```
             |       x2        x3         y
-------------+---------------------------
          x2 |      1.0      -0.9      -0.6
          x3 |     -0.9       1.0       0.7
           y |     -0.6       0.7       1.0
```

Найдите значение $VIF(\hat{\beta}_3)$ для регрессии $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$.

№4.1.4. Исследователь оценил регрессионную модель $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i$ и провёл диагностику различных проблемных явлений. Результаты его стараний приведены ниже:

График зависимости остатков регрессии от прогнозных значений

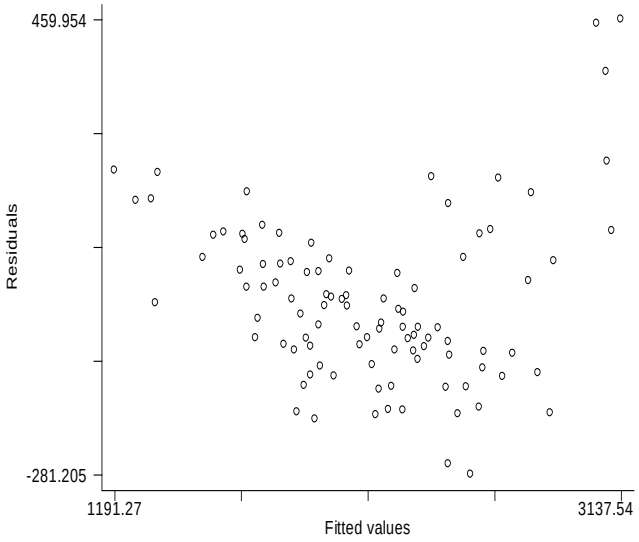
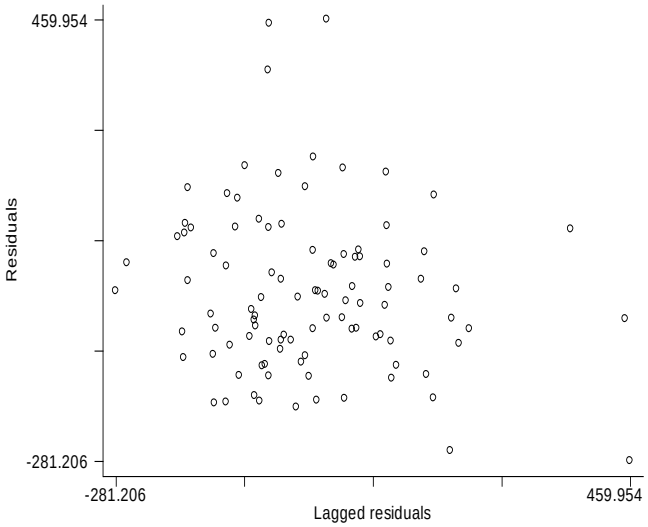


График зависимости остатков e_i *от остатков* e_{i-1}

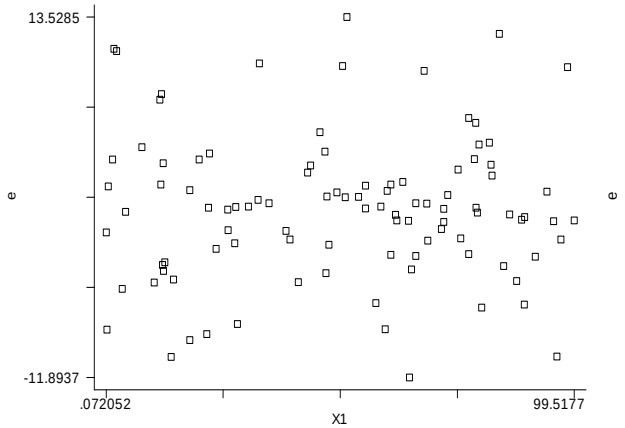


$VIF(\hat{\beta}_2) = 1.02$, $VIF(\hat{\beta}_3) = 2.17$, $VIF(\hat{\beta}_4) = 1.75$

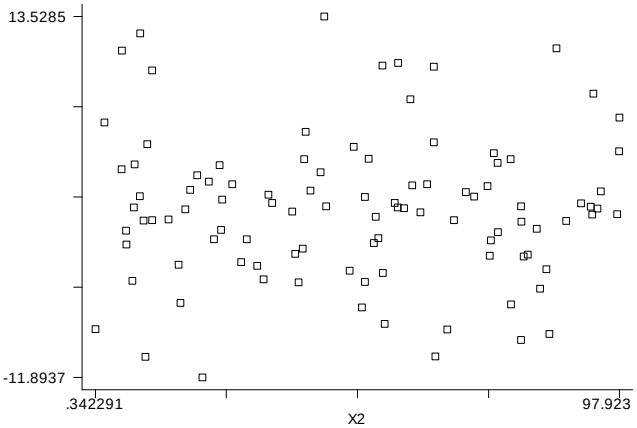
Определите, какие проблемные явления обнаружил исследователь. Обоснуйте свой ответ.

№4.1.5. Оценивалась некоторое уравнение регрессии: $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$. С помощью метода наименьших квадратов были получены оценки коэффициентов и рассчитаны остатки регрессии e_i. Ниже приведены графики зависимости e от X_2 и X_3, а также значения VIF.

e от X_2



e от X_3



$VIF(\hat{\beta}_2) = VIF(\hat{\beta}_3) = 100$

а) С какими проблемами, судя по приведённым данным, столкнулся исследователь? Объясните свой ответ.
б) Найдите коэффициент детерминации в регрессии $X_{2,i} = \alpha_1 + \alpha_2 X_{3,i} + \upsilon_i$.

№4.1.6 Оценив уравнение $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i$, исследователь занялся выявлением возможных проблем в модели. Определите, с какими проблемами он столкнулся, по результатам его стараний:

График в осях e_i *и* $\hat{Y}_i$

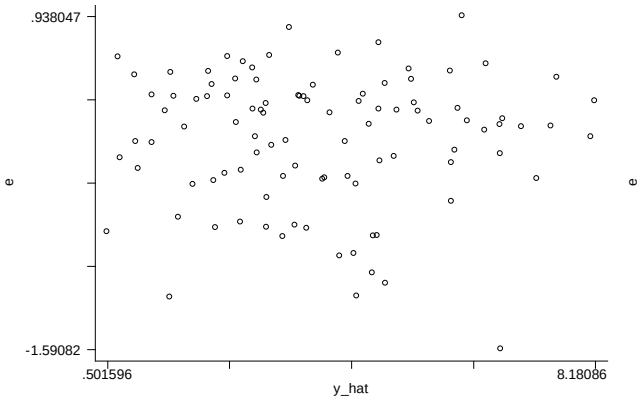
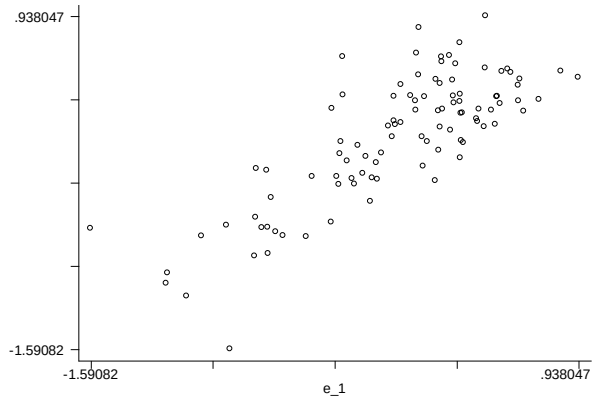


График в осях e_i *и* e_{i-1}



$VIF(\hat{\beta}_2) = 1.2$, $VIF(\hat{\beta}_3) = 21.1$, $VIF(\hat{\beta}_4) = 20.6$

№4.1.7. Мартовский Заяц и Безумный Шляпник почти всё время пьют чай. Известно, что количество выпитого за день чая (в чашках) зависит от количества пирожных (в штуках) и печенья (в штуках).

Алиса, гостившая у героев в течение 25 дней, заметила, что если оценить зависимость выпитого чая от закуски для Мартовского Зайца и Шляпника, то получится:

$\hat{Tea}_i = 6 + 0,5Biscuit_i + 1,5Cake_i$; RSS=11,5 .

Чтобы понять, удачную ли модель она построила, Алиса оценила ещё одну регрессию:

$\tilde{Tea}_i = 12,7 + 0,65Biscuit_i - 0,8Cake_i - 0,59\hat{Tea}_i^2 + 0,03\hat{Tea}_i^3$; RSS=9,5

а) Помогите героине понять, верную ли спецификацию модели она выбрала: проведите соответствующий тест, не забудьте сформулировать основную и альтернативную гипотезы.

б) Алиса решила проверить первоначальную (короткую) модель на наличие гетероскедастичности с помощью теста Уайта. Выпишите уравнение регрессии, которое она должна оценить.

№4.1.8. Продавец мороженого оценил динамическую модель объёмов продаж:

$\hat{\ln Q}_t = 26.7 + 0.2\ln Q_{t-1} - 0.6\ln P_t$

Здесь Q_t - число проданных в день t вафельных стаканчиков, а P_t - цена одного стаканчика в рублях. Продавец также рассчитал остатки e_t.

а) Чему, согласно полученным оценкам, равна долгосрочная эластичность объёма продаж по цене?

б) Предположим, что продавец решил проверить наличие автокорреляции первого порядка с помощью теста Бройша-Годфри. Выпишите уравнение регрессии, которое он должен оценить.

№4.1.9. По 22 наблюдениям сотрудники НИИ оценили уравнение регрессии:

$\hat{Y} = 10.9 - 1.3X_2 + 0.8X_3, \quad R^2 = 0.6$

- и рассчитали прогнозные значения $\hat{Y}_i$ и остатки e_i. После того они приступили к диагностике возможных недостатков модели.

а) Самый младший научный сотрудник решил проверить наличие гетероскедастичности с помощью теста Уайта. Выпишите уравнение регрессии, которое он должен оценить.

б) Профессор для проверки на гетероскедастичность по тем же данным оценил другое уравнение:

$|e_i| = 10.9 - \frac{1.3}{X_3}$, $R^2 = 0.06$, – и сообщил, что никакой гетероскедастичности нет. Поясните, прав ли профессор и стоит ли самому младшему научному сотруднику с ним спорить, если последний получил для вспомогательной регрессии $R^2 = 0.86$? Все сотрудники НИИ используют уровень значимости 5%.

№4.1.10. Многие знают песню «Во саду ли, в огороде», в которой есть такие строки:

Кумачу я не хочу,
Китайки не надо;
Принеси, моя надёжа,
Алого гризета,
Чтоб не стыдно было девке
На улицу выйти.

Однако далеко не все ценители фольклора осведомлены, что цены на гризет, кумач и китайку можно связать оценённым уравнением регрессии:

$$\Delta\hat{P}_{\text{гризет}} = \underset{(0.02)}{0.03} + \underset{(0.12)}{0.44}\Delta P_{\text{кумач}} + \underset{(0.18)}{0.63}\Delta P_{\text{китайка}},\ R^2 = 0.94.$$

Все переменные здесь – приросты реальных цен на соответствующие товары, в скобках указаны стандартные ошибки. Оценивание проводилось по квартальным данным за 6 полных лет.

а) Проверьте гипотезу о том, что в случае неизменности реальных цен на кумач и китайку, ожидаемое изменение цены на гризет будет также нулевым.

б) Купец Пафнутий намерен оценить зависимость объёма продаж гризета от цен на все три конкурирующих товара по данным за тот же период времени. Какие трудности поджидают Пафнутия?

№4.1.11. Оценивая уравнение $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \beta_4 X_{4,i} + \varepsilon_i$, исследовательская группа проверяла данные на наличие мультиколлинеарности. Статистик Тимофей оценил корреляционную матрицу регрессоров:

	x2	x3	x4
x2	1.0000		
x3	-0.4934	1.0000	
x4	-0.6144	0.9018	1.0000

А эконометрист Надежда рассчитала значения VIF:

Variable	VIF	1/VIF
x4	2.53	0.394967
x3	2.08	0.480034
x2	1.62	0.616889
Mean VIF	2.08	

Изучив эти результаты, почтенный старец Феодосий сказал, что они не согласуются – кто-то допустил ошибку в расчётах. Выясните, почему приведённые VIF и корреляции не могут относиться к одним и тем же данным.

№4.1.12. В былые времена для диагностики автокорреляции использовалась статистика Дарбина-Уотсона, рассчитываемая по формуле $DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$, где e_i - остатки, полученные

при МНК-оценивании диагностируемого уравнения регрессии. Известно, что эта статистика принимает значения от 0 до 4. Подумайте: какие значения статистики говорят о наличии положительной корреляции между случайными составляющими в соседних наблюдениях, а какие – об отрицательной?

№4.1.13. Оценивается зависимость $Y_i = \beta_1 + \beta_{2,i} X_{2,i} + \varepsilon_i$, в которой коэффициент $\beta_{2,i}$ зависит от детерминированной величины Z_i, в остальном все предпосылки классической модели выполнены. Предположим, что

а) $\beta_{2,i} = \alpha_1 + \alpha_2 Z_i$;

б) $\beta_{2,i} = \alpha_1 + \alpha_2 Z_i + \upsilon_i$, где $\upsilon_i \sim N(0, \sigma_\upsilon^2)$, причём $\upsilon_1, ..., \upsilon_n, \varepsilon_1, ..., \varepsilon_n$ независимы.

Чем различаются случаи (а) и (б) с точки зрения оценивания параметров моделей и возникающих при оценивании проблем? Как бы вы оценили параметры $\beta_1, \alpha_1, \alpha_2$? Какими свойствами обладали бы получаемые оценки в каждом из случаев?

№4.1.14. По выборке $Y_1, ..., Y_n$ оценивается зависимость $Y_i = \beta + \varepsilon_i$, где случайные ошибки образуют случайное блуждание: $\varepsilon_i = \varepsilon_{i-1} + \upsilon_i$. Все υ_i имеют нулевое математическое ожидание и дисперсию σ_υ^2, такова же случайная ошибка в первом наблюдении: $E(\varepsilon_1) = 0, D(\varepsilon_1) = \sigma_\upsilon^2$. Величины $(\varepsilon_1, \upsilon_1, ..., \upsilon_n)$ независимы.

а) Найдите математическое ожидание и ковариационную матрицу вектора $(\varepsilon_1, ..., \varepsilon_n)'$.

б) Выясните, несмещённая ли оценка МНК для коэффициента β.

в) Найдите дисперсию этой оценки и её предел при $n \to \infty$.

г) Сравните точность оценки МНК и альтернативной оценки $\hat{\beta} = Y_1$.

д) Предложите несмещённую оценку для σ_υ^2.

№4.1.15. Девочка Алиса пишет диктанты по сольфеджио не очень хорошо, потому что никогда не выполняет домашних заданий. Если она делает меньше двух ошибок, то получает «5», если 2-3 ошибки, то «4», если 4-5 ошибок, то «3». Когда ошибок оказывается больше, Алиса получает двойку. Алисин брат первокурсник Тимофей считает, что успехи Алисы зависят от опыта, измеряемого количеством пройденных уроков, и от продолжительности этих занятий, т.е. от количества минут, проведённых на уроках. Зная, что урок сольфеджио длится 110 минут, статистику Алисиного пребывания в музыкальной школе можно представить так:

Оценка	Количество уроков	Количество минут	Количество ошибок
3	3	330	5
4	4	440	3
3	5	550	4
3	6	660	4

Тимофей пытается оценить регрессию вида: $Y_i = \alpha + \beta \cdot Минуты_i + \gamma \cdot Уроки_i + \varepsilon_i$, где в качестве регрессанта выступает оценка.

Подскажите, с какой проблемой (связанной с нарушениями предпосылок КЛНР) может столкнуться Тимофей. Почему вы так считаете? Помогите Тимофею избавиться от этой проблемы. Предложите свою модель, оцените её по доступным данным. Спрогнозируйте оценку, которую получит Алиса на 10-ом уроке.

№4.1.16. По наблюдениям за 54 месяца была оценена модель $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$ и рассчитаны остатки e_i. Для диагностики автокорреляции была оценена вспомогательная регрессия $e_i = 0.05 - 0.01 X_{2,i} - 0.09 X_{3,i} - 0.14 e_{i-1} - 0.08 e_{i-2}, R^2 = 0.05$.

Проверьте наличие автокорреляции. Что бы вы дополнительно проверили и как бы усовершенствовали модель с учётом того, что данные ежемесячные?

№4.1.17. По наблюдениям за 31 год[6] была оценена зависимость импорта в Германии M от частного потребления C и обменного курса E:

$\ln \hat{M} = -4.1 + 1.4 \ln C + 0.1 \ln E, \; RSS = 0.081.$

Для проверки спецификации было оценено дополнительное уравнение:

$\ln \tilde{M} = 57.0 - 11.3 \ln C - 0.8 \ln E + 1.2 \left(\ln \hat{M}\right)^2 - 0.1 \left(\ln \hat{M}\right)^3, \; RSS = 0.063.$

Проверьте правильность выбранной функциональной формы на уровне 5%.

§4.2. Оценивание регрессии при гетероскедастичности и автокорреляции.

№4.2.1. Рассмотрим модель $Y_i = \beta X_i + \varepsilon_i$, где $D(\varepsilon_i) = \alpha X_i^2$, α - неизвестный параметр модели. Все остальные предпосылки КЛНРМ выполнены. Предлагается оценка для коэффициента регрессии: $\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i}{X_i}$. Покажите, что она несмещённая, найдите её дисперсию. Докажите, что предложенная оценка эффективна в классе линейных несмещённых оценок.

№4.2.2 Обследовав выборку из 27 домохозяйств, исследователь оценил уравнение регрессии:

$$\frac{\hat{Exp}_i}{Size_i} = 926 + 235 \times \frac{1}{Size_i} + 0.3 \times \frac{Income_i}{Size_i}, \quad R^2 = 0.3,$$

где Exp_i - месячные затраты i-го домохозяйства на питание в рублях, $Income_i$ - месячный доход домохозяйства (также в рублях), $Size_i$ - число членов домохозяйства.

а) Каково, согласно оценённой модели, ожидаемое различие в затратах на питание между двумя домохозяйствами с одинаковым доходом, первое из которых больше второго на одного человека?

б) Доходы и затраты включены в модель в расчёте на одного члена домохозяйства. Известно, что это было сделано с целью избавления от гетероскедастичности, которая обнаружилась в предыдущей модели: $Exp_i = \beta_1 + \beta_2 Size_i + \beta_3 Income_i + \varepsilon_i$. На какое предположение о случайной составляющей ε_i опирался исследователь?

в) Для проверки правильности выбранной спецификации было оценено ещё одно уравнение:

$$\frac{\hat{Exp}_i}{Size_i} = 513 + 1499 \times \frac{1}{Size_i} + 0.5 \times \frac{Income_i}{Size_i} - 0.001 \times \left(\frac{\hat{Exp}_i}{Size_i} \right)^2, \quad R^2 = 0.4$$

Даёт ли эта проверка основание считать модель исследователя неверно специфицированной? Используйте уровень значимости 1%.

№4.2.3. По n наблюдениям оценивается регрессия величины Y_i на свободный член α: $Y_i = \alpha + \varepsilon_i$, где $E(\varepsilon_i) = 0$, $Cov(\varepsilon_i, \varepsilon_j)=0$ для $i \neq j$, а $D(\varepsilon_i) = \sigma^2$, для $i = 1, ..., n_1$ и $D(\varepsilon_i) = 4\sigma^2$ для $i = n_1+1, ..., n$ (т.е. дисперсия в последних $n - n_1$ наблюдениях в четыре раза больше дисперсии в первых n_1 наблюдениях).

а) Постройте оценку для параметра α с помощью МНК и найдите её дисперсию.

б) Предложите несмещённую оценку параметра α, более эффективную, чем оценка МНК. Найдите её дисперсию.

---

[6] Данные из книги P. Newbold, “Statistics for Business and Economics”.

№4.2.4. Сотрудники НИИ столкнулись в своём исследовании с гетероскедастичностью и решают, что с ней делать. Младший научный сотрудник предлагает использовать МНК, но оценить ковариационную матрицу с учётом гетероскедастичности (оценка Уайта). Старший предлагает взвешенный МНК: перед оцениванием поделить исходное уравнение на один из регрессоров (возможно, на корень из регрессора). Обсудите достоинства и недостатки этих способов. Как бы поступали вы?

№4.2.5. Пусть поведение случайной величины ε_i описывается процессом скользящего среднего второго порядка (ε_i ~ MA(2)): $\varepsilon_i = \alpha_1 \upsilon_{i-1} + \alpha_2 \upsilon_{i-2} + \upsilon_i$, где α_1 и α_2 – некоторые числа (параметры процесса), все величины υ_i независимы, имеют нулевое математическое ожидание и дисперсию σ_υ^2. Найдите Cov(ε_i, ε_{i-1}), Cov(ε_i, ε_{i-2}).

№4.2.6. В модели $Y_i = \beta_1 + \beta_2 X_{2,i} + ... + \beta_k X_{k,i} + \varepsilon_i$, $i = 1,...,n$ случайная составляющая описывается процессом случайного блуждания: $\varepsilon_i = \varepsilon_{i-1} + \upsilon_i, \varepsilon_0 = 0$, где υ_i независимы. Как вы посоветуете преобразовать модель, чтобы избежать автокорреляции?

№4.2.7. Однажды любознательные барышни Оля и Маша оценили зависимость $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$. Оля провела тест, выявивший наличие гетероскедастичности. Маша решила, что результаты теста могут быть следствием ошибки спецификации. Оля, чтобы одолеть гетероскедастичность, оценила модель $\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \upsilon_i$. Маша, чтобы устранить ошибку спецификации, перешла к логарифмам: $\ln Y_i = \alpha_1 + \alpha_2 \ln X_i + u_i$. И тот, и другой подход дали хорошие результаты: в новых моделях тесты не выявили ни гетероскедастичности, ни ошибки спецификации. Представьте себя на месте любознательных барышень: какими соображениями вы бы руководствовались при выборе одной из этих двух моделей?

№4.2.8. Исследователь оценил модель $y = X\beta + \varepsilon$ и обнаружил автокорреляцию. Он решил описать автокорреляцию моделью $\varepsilon_i = \rho \varepsilon_{i-1} + \upsilon_i$ и оценил коэффициент авторегрессии: $\hat{\rho} = 0.5$. Вот первые три наблюдения в его данных:

$$y_{(1-3)} = \begin{pmatrix} 10 \\ 8 \\ 14 \end{pmatrix}, \quad X_{(1-3)} = \begin{pmatrix} 1 & 20 & 0 \\ 1 & 22 & 0 \\ 1 & 20 & 1 \end{pmatrix}.$$

Как будут выглядеть первые три строки матрицы регрессоров и вектора объясняемой переменной при оценивании модели с помощью процедуры Кохрейна-Оркатта с поправкой Прайса-Винстена?

№4.2.9. Продавец Анфиса оценивает модель объёма продаж мороженого: $Q_i = \beta_1 + \beta_2 P_i + \beta_3 T_i + \varepsilon_i$, где Q_i - количество проданных стаканчиков в день i P_i - цена стаканчика, T_i - средняя температура дня. Обнаружив автокорреляцию, Анфиса оценила модель $\hat{Q}_i = 11.1 + 0.25 Q_{i-1} - 3.38 P_i + 0.96 P_{i-1} + 3.92 T_i - 0.85 T_{i-1}$.

а) Найдите оценку коэффициента авторегрессии в модели $\varepsilon_i = \rho \varepsilon_{i-1} + \upsilon_i$.

б) После поправки на автокорреляцию Анфиса получила такие оценки:

$\hat{Q}_i = 16.50 - 3.18 P_i + 3.94 T_i$.

Вчера Анфиса продала 28 стаканчиков по цене 20 рублей при температуре 18 градусов. Рассчитайте прогноз числа проданных стакаников на сегодня, если Анфиса решила не менять цену, а температура ожидается такая же.

Часть 5. Метод максимального правдоподобия

§5.1. Оценивание параметров регрессионных моделей.

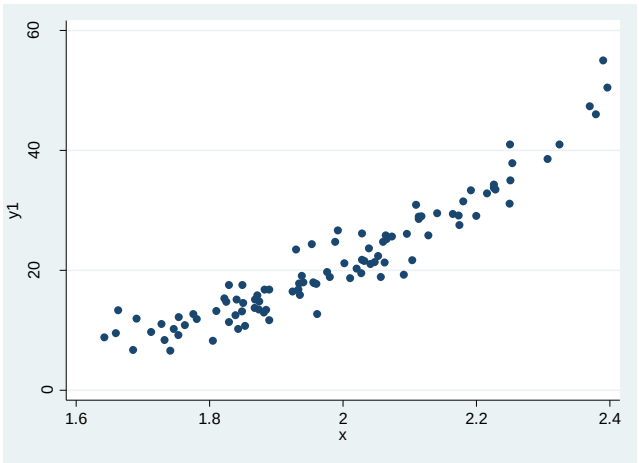
№5.1.1 По n наблюдениям оценивается модель $Y_i = \beta X_i + \varepsilon_i$ (регрессия без свободного члена). Предполагается, что все предпосылки классической линейной нормальной регрессионной модели выполнены. С помощью метода максимального правдоподобия постройте оценки для параметра β и для дисперсии случайной составляющей.

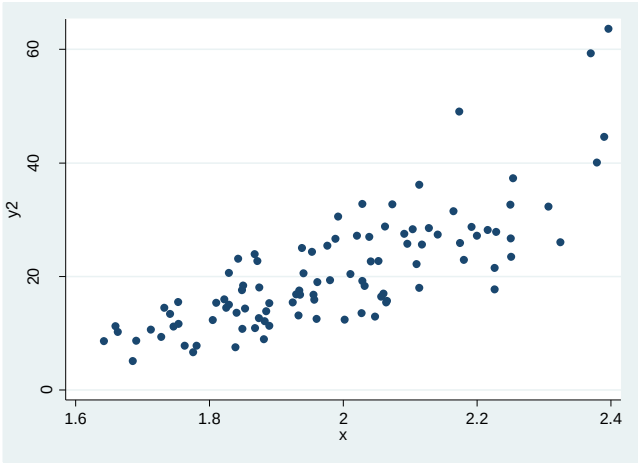
№5.1.2. Рассмотрим так называемую модель экспоненциальной регрессии, в которой случайная величина Y_i имеет показательное (экспоненциальное) распределение с параметром $\lambda_i = \exp(\alpha + \beta X_i)$, где X_i – детерминированный регрессор. Предположим, что все Y_i независимы. Параметры α и β предполагается оценивать методом максимального правдоподобия. Выпишите функцию правдоподобия и условия экстремума первого порядка.

№5.1.3. Оценивается линейная нормальная регрессионная модель с гетероскедастичностью: предполагается, что независимые величины $Y_1,...,Y_n$ имеют нормальное распределение с математическим ожиданием $\mu_i = x_i'\beta$ и дисперсией $\sigma_i^2 = \exp(z_i'\gamma)$, где x_i и z_i - векторы объясняющих переменных, β и γ - векторы оцениваемых коэффициентов. Выпишите логарифмическую функцию правдоподобия, условия максимума первого и второго порядков.

№5.1.4. Выпишите логарифмическую функцию правдоподобия для модели пуассоновской регрессии: $Y_i \sim \Pi(\lambda_i)$, $\lambda_i = \exp(x_i'\beta)$ по выборке из независимых величин $Y_1,...,Y_n$.

№5.1.5. Сотрудники НИИ размышляют над тем, как оценивать экспоненциальную зависимость Y от X. Старший научный сотрудник предлагает свести зависимость к линейной: $\ln(Y_i) = \beta_1 + \beta_2 X_i + \varepsilon_i$ - и оценить её с помощью МНК. Его молодой коллега предлагает применить метод максимального правдоподобия для оценивания нелинейной модели $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \exp(\beta_1 + \beta_2 X_i)$. Посмотрите на две возможные диаграммы рассеяния признаков Y и X. Скажите, в каком случае вы бы поддержали старшего, а в каком – младшего научного сотрудника.





№5.1.6. Меткий стрелок Василиса регулярно тренируется, а на её тренировках так же регулярно присутствует статистик Тимофей, который записывает число выстрелов, сделанных Василисой до первого промаха включительно. Так Тимофей собрал уже продолжительный ряд наблюдений $Y_1,...,Y_n$. Теперь он хочет понять, какие факторы определяют меткость стрелка. К счастью,

помимо результатов стрельбы, Тимофей отмечал и потенциальные детерминанты меткости $x_i, i = 1,...,n$: погодные условия, время тренировки и т.п.

Статистик предполагает, что результаты выстрелов независимы, а вероятность попадания в отдельном выстреле на тренировке i связана с объясняющими переменными так: $p_i = \dfrac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$.

Помогите Тимофею составить логарифмическую функцию правдоподобия.

№5.1.7. Модель $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ удовлетворяет всем предпосылкам КЛНРМ за тем исключением, что параметр β_2 — случайная величина, зависящая от переменной Z_i: $\beta_2 = \alpha_1 + \alpha_2 Z_i + \upsilon_i$ (для этого уравнения все предпосылки КЛНРМ выполнены без исключения). Предполагается, что величины $\varepsilon_1,...,\varepsilon_n, \upsilon_1,...,\upsilon_n$ независимы.

а) Выпишите функцию правдоподобия для оценивания неизвестных коэффициентов β_1, α_1, α_2 и дисперсий случайных составляющих σ_ε^2 и σ_υ^2.

б) Как бы вы оценили эту модель с помощью МНК? Каковы недостатки этого метода в данном случае? Какими будут свойства оценок МНК?

№5.1.8. Для модели $Y_i = x_i'\beta + \varepsilon_i$ выполнены почти все предпосылки КЛНРМ, вот только случайная составляющая имеет функцию распределения $F_\varepsilon(x) = \exp(-\exp(-x))$. Для оценивания Тимофей предлагает использовать метод максимального правдоподобия.

а) Выпишите логарифмическую функцию правдоподобия.

б) Надя обращает внимание на то, что модель удовлетворяет всем предпосылкам теоремы Гаусса-Маркова, так что можно применять МНК — правда, случайная составляющая имеет ненулевое математическое ожидание, но и Тимофей, и Надежда сходятся во мнении, что это не важно[7]. Какие преимущества даёт метод максимального правдоподобия? Как они согласуются с теоремой Гаусса-Маркова?

§5.2. Доверительные интервалы и проверка гипотез. Информационный критерий Акаике.

№5.2.1. Найдите оценку $\{-H(\hat\beta, \hat\sigma_\varepsilon^2)\}^{-1}$ для ковариационной матрицы оценок параметров β и σ_ε^2 из задачи №5.1.1. Здесь $H(\hat\beta, \hat\sigma_\varepsilon^2)$ - матрица вторых производных логарифмической функции правдоподобия для значений параметров, равных их оценкам.

№5.2.2. Плотность обобщённого гамма-распределения задаётся функцией:

$$f(x) = \begin{cases} \dfrac{\gamma^\gamma}{x\sqrt{\gamma\sigma^2\Gamma(\gamma)}}\exp\left(z\sqrt{\gamma} - u\right), & \text{если } \kappa \neq 0, \\ \dfrac{1}{x\sqrt{2\pi\sigma^2}}\exp\left(-\dfrac{z^2}{2}\right), & \text{если } \kappa = 0. \end{cases}$$

Пусть она вас не пугает, хотя бы потому что приведена здесь только для справки. Распределение имеет три параметра: μ, $\sigma^2 > 0$ и κ. Величины $u = |\kappa|^{-2}$, $z = \operatorname{sgn}(\kappa)(\ln x - \mu)/\sigma$ и $u = \gamma\exp(|\kappa| z)$ введены в запись, чтобы сделать её менее громоздкой. Для удобства численной максимизации

[7] Кстати, почему они так считают? Каковы последствия нецентральности случайной составляющей для оценок МНК?

функции правдоподобия вместо σ^2 подбирается параметр $\ln\sigma$ - он может принимать любые вещественные значения, так что не требуется учитывать ограничение $\sigma^2 > 0$.
При $\kappa = 0$ распределение совпадает с логарифмически нормальным, при $\kappa = 1, \sigma^2 = 1$ - с показательным.
Вот результаты подгонки обобщённого гамма-распределения под некую выборку:

$$\begin{pmatrix} \hat{\mu} \\ \ln\hat{\sigma} \\ \hat{\kappa} \end{pmatrix} = \begin{pmatrix} 1.72 \\ -0.03 \\ 0.88 \end{pmatrix}, \quad \hat{V}\begin{pmatrix} \hat{\mu} \\ \ln\hat{\sigma} \\ \hat{\kappa} \end{pmatrix} = \begin{pmatrix} 0.006 & -0.002 & 0.007 \\ -0.002 & 0.002 & -0.003 \\ 0.007 & -0.003 & 0.015 \end{pmatrix}$$

а) Есть ли основания считать, что выборка была взята не из логнормального распределения? Не из показательного?
б) Вот значения логарифмической функции правдоподобия для трёх распределений:

Обобщённое гамма	-605.09
Логнормальное	-634.34
Показательное	-606.49

Проверьте те же гипотезы критерием отношения правдоподобия.
в) Какое из трёх распределений предпочтительнее по критерию Акаике?

№5.2.3. При оценивании классической линейной нормальной модели $Y_i \sim N(\mu_i, \sigma^2)$, $\mu_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i}$ по некоторой выборке было получено значение логарифмической функции правдоподобия $\ln L = -328.2$. Оценивание аналогичной модели, учитывающей гетероскедастичность: $\ln\sigma_i^2 = \gamma_1 + \gamma_2 X_{2,i} + \gamma_3 X_{3,i}$ - дало значение $\ln L^* = -321.5$. Проверьте гипотезу о гомоскедастичности. Какая из двух моделей предпочтительнее по критерию Акаике?

№5.2.4. Под данные пяти наблюдений подгоняются показательное и нормальное распределения. Определите, какое распределение лучше соответствует данным согласно критерию Акаике.

0.5 1.0 0.5 3.0 1.5

№5.2.5. По данным о 80 забастовках было рассчитана средняя длительность забастовки – 39 дней. Считая, что длительность распределена по показательному закону, оцените параметр λ, рассчитайте 90% доверительные интервалы для λ, математического ожидания и дисперсии длительности забастовки.

№5.2.6. В XVIII веке естествоиспытатель Жорж Луи де Бюфон увлёкся подбрасыванием монетки и, подкинув её 4040 раз, отметил, что решка выпала 1992 раза. С помощью критерия отношения правдоподобия установите, не противоречит ли этот результат гипотезе о правильности монетки.

Часть 6. Модели дискретного выбора

§6.1. Шансы и отношение шансов.

№6.1.1. Установите, верны ли следующие утверждения, ответ обоснуйте.
а) Если шансы события A выше шансов события B, то и вероятность события A больше вероятности события B.
б) Если $\frac{P(A)}{P(B)} > \frac{P(C)}{P(D)}$, то и отношение шансов событий A и B больше, чем отношение шансов C и D.

№6.1.2. Известно, что шансы события равны 3:2. Какова вероятность этого события?

№6.1.3. Вася Сидоров участвовал в двух соревнованиях по стрельбе: на первых присутствовала его подруга Аня Иванова, а на вторых её не было[8]. В первом состязании нужно было сделать 20 выстрелов по мишени, а во втором – 10. Вот таблица Васиных результатов:

	Аня есть	Ани нет
Попадания	16	6
Промахи	4	4

Каково отношение шансов попадания в присутствии Ани к шансам попадания в её отсутствие? А отношение шансов промаха?
Каково отношение доли попаданий в присутствии Ани к доле попадания в её отсутствие? Отношение доли промахов?

№6.1.4. Шансы события *A* в три раза больше, чем шансы *B*. Во сколько раз вероятность события *A* может быть больше вероятности *B*?

№6.1.5. Постройте график зависимости шансов события от его вероятности. В какой области шансы и вероятность близки друг к другу?

№6.1.6. Романтичная девушка Рая увлеклась и перебрала 100 цветков сирени, кушая цветки с пятью лепестками. На основании этого опыта она построила 95% доверительный интервал для логарифма шансов наткнуться на такой цветок: $-4.625 < \ln\frac{p}{1-p} < -2.327$.
а) Исходя из результатов Раи, рассчитайте доверительный интервал для вероятности того, что в случайно отобранном цветке сирени будет 5 лепестков.
б) Как вы думаете, почему Рая использовала такой необычный доверительный интервал, а не воспользовалась традиционным интервалом для доли/вероятности?
в) Известно, что Рая скушала три цветка. Рассчитайте обычный 95% доверительный интервал для вероятности встретить цветок с пятью лепестками.

§6.2. Модели бинарного выбора

№6.2.1. Исследователь считает, что основным фактором, определяющим решение индивида о том, добираться ли ему до работы на автобусе или на автомобиле, является время, затрачиваемое на дорогу. Он оценивает три модели:
Линейная модель вероятности: $\hat{A}_i = 0.485 + 0.007(BT_i - AT_i)$

[8] Другие истории из жизни Василия Сидорова и Анны Ивановой вы можете найти в задачнике Б.Б. Демешева по теории вероятностей и математической статистике.

Логит: $\hat{P}(A_i = 1) = \dfrac{\exp(-0.238 + 0.053(BT_i - AT_i))}{1 + \exp(-0.238 + 0.053(BT_i - AT_i))}$

Пробит: $\hat{P}(A_i = 1) = \Phi(-0.065 + 0.03(BT_i - AT_i))$

Здесь $A_i = 1$, если *i*-й индивид склонен добираться до работы на автомобиле, и $A_i = 0$, если он предпочитает автобус, BT_i – среднее время, затрачиваемое на путь до работы на автобусе в минутах, AT_i – среднее время, затрачиваемое на путь до работы на автомобиле.

На основании каждой из оценённых моделей, определите:

а) вероятность того, что индивид предпочтёт пользоваться автомобилем, если дорога до работы на автобусе и на машине занимает одинаковое время;

б) разницу между *BT* и *AT*, при которой индивиду безразлично, какой транспорт выбрать;

в) предельный эффект от увеличения разности $BT_i - AT_i$ для индивида, которому на работу добираться 40 минут на автомобиле и 50 минут на автобусе.

№6.2.2. Случайная величина Y_i^* линейно связана с вектором-строкой детерминированных объясняющих детерминированных величин x_i': $Y_i^* = x_i'\beta + \varepsilon_i$, где β – вектор оцениваемых коэффициентов, а ε_i – случайная величина. Все ε_i независимы и имеют функцию распределения $F_\varepsilon(x) = \exp(-(З^{-x}))$. Однако величина Y_i^* является ненаблюдаемой, и вместо неё исследователь фиксирует значения Y_i, равные единице, если $Y_i^* > 0$, и нулю иначе. Выпишите функцию правдоподобия для оценивания коэффициентов β.

№6.2.3. По данным из задачи 6.1.3 оцените коэффициенты модели $\dfrac{P(Y_i = 1)}{P(Y_i = 0)} = \exp(\beta_1 + \beta_2 X_i)$, где Y_i - индикатор Васиного попадания в мишень в *i*-м выстреле, а X_i - индикатор Аниного присутствия. Как изменятся оценки, если объясняемой величиной будут шансы промаха?

№6.2.4. Изучается зависимость вероятности рождения ребёнка с малым весом (<2500 г) от возраста матери и того, курила ли она во время беременности. Вот результаты оценивания логит-модели:

$$\frac{\hat{P}(LBW_i = 1)}{\hat{P}(LBW_i = 0)} = \exp(-1.405 - 0.014 Age_i + 0.001 Age_i^2 + 0.827 Smoke_i),\ \ln L = -291.443$$

Здесь $LBW_i = 1$, если вес *i*-го новорождённого меньше 2500 г, и $LBW_i = 0$ иначе,

$Smoke_i = 1$, если мать курила во время беременности, и $Smoke_i = 0$ иначе,

Age_i - возраст матери.

Также известна оценка ковариационной матрицы оценок коэффициентов:

$$\hat{V}(\hat{\beta}) = \begin{array}{c} \\ \text{3.401} \\ -0.252 \\ 0.004 \\ -0.025 \end{array} \begin{array}{c} \begin{array}{cccc} \textit{Константа} & \textit{Age} & \textit{Age}^2 & \textit{Smoke} \end{array} \\ \begin{pmatrix} 3.401 & -0.252 & 0.004 & -0.025 \\ -0.252 & 0.019 & -0.001 & 0 \\ 0.004 & -0.001 & 6.35 \times 10^{-6} & 0 \\ -0.025 & 0 & 0 & 0.04 \end{pmatrix} \end{array}$$

а) На уровне значимости 5% проверьте, есть ли основания считать, что у курящих матерей больше вероятность рождения ребёнка с малым весом, чем у некурящих матерей того же возраста.

б) Требуется проверить, значима ли связь изучаемой вероятности с возрастом матери. Выпишите основную и альтернативную гипотезы в терминах коэффициентов модели. Какую модель вы бы дополнительно оценили, если проверку нужно провести с помощью критерия отношения правдоподобия?

№6.2.5. По набору наблюдений $(X_1,Y_1),...,(X_n,Y_n)$, где Y_i - независимые случайные величины, принимающие значения 0 и 1, а X_i - значения детерминированного регрессора, оценивается модель $P(Y_i=1)=\left(1+\alpha^{\beta+X_i}\right)^{-1}$. Выпишите логарифмическую функцию правдоподобия и задачу максимизации для оценивания параметров α и β.

№6.2.6. По наблюдениям за 450 семьями была оценена логит-модель:

$$\frac{\hat{P}(Dacha_i=1)}{\hat{P}(Dacha_i=0)}=\exp(-3.5+0.3\ln Income_i+0.4Auto_i),\quad \ln L=-185.45,$$

где $Dacha_i$ - переменная, равная единице, если у семьи есть загородный дом, и нулю иначе,

$Income_i$ - месячный доход семьи в тысячах рублей,

$Auto_i$ - переменная, равная единице, если семья владеет автомобилем, и нулю иначе,

$\ln L$ - логарифм функции правдоподобия.

При оценивании модели без объясняющих переменных (регрессии только на константу) логарифм функции правдоподобия оказался равен -187.6.

а) Рассмотрите случайно отобранную семью с автомобилем и доходом в 110 тыс. руб. в месяц. Что, согласно оценённой модели, более вероятно: владеет такая семья загородным домом или нет?

б) Проверьте значимость модели в целом на уровне 5%.

№6.2.7. По одному набору данных исследователь оценил коэффициенты логит и пробит моделей зависимости *Y* от *X*:

Коэффициент	*Оценка в модели I*	*Оценка в модели II*
Константа	0.25	0.16
Коэфф. при *X*	-0.12	-0.07

В каком столбце приведены оценки модели логит, а в каком – пробит?

№6.2.8. При проверке спецификации моделей бинарного выбора нередко используется тест, предложенный в диссертации Дарила Прегибона “Goodness of link tests for generalized linear models”. После оценивания проверяемой модели $P(Y_i=1)=F(x_i'\beta)$ рассчитывается линейная комбинация регрессоров и оценённых коэффициентов – она выступает как объясняющая переменная в новой модели с той же связующей функцией $F(\cdot)$:

$$P(Y_i=1)=F(\alpha_0+\alpha_1(x_i'\hat{\beta})+\alpha_2(x_i'\hat{\beta})^2)$$

Какими должны быть коэффициенты $\alpha_0,\alpha_1,\alpha_2$, если первоначальная модель верно специфицирована?

Вспомните похожий способ проверки спецификации линейного уравнения регрессии. В чём его отличие от теста Прегибона?

№6.2.9. Имеется *n* объектов, которые наблюдаются в течение некоторого промежутка времени (условимся считать, что длина этого промежутка равна единице). Объекты отличаются друг от друга вектором характеристик $x_i, i=1,...,n$. За время наблюдения с ними может произойти или не произойти некоторое событие (допустим, объект – автомобиль, ему может потребоваться обслуживание в автосервисе или нет). Величина Y_i принимает значение 1, если событие с объектом *i* произошло, и 0 иначе.

Исследователь предполагает, что наступление событий описывается пуассоновским потоком с интенсивностью $\lambda_i=\exp(x_i'\beta)$, где β - вектор неизвестных коэффициентов при характеристиках объектов.

а) Найдите вероятности значений Y_i как функции от x_i и β.

б) Выпишите логарифмическую функцию правдоподобия для оценивания коэффициентов β.

в) Что произойдёт с оценками коэффициентов β, если сменить единицы измерения времени, так что продолжительность периода наблюдения будет равна не единице, а, например, двенадцати?

№6.2.10. Оценена модель пробит:

$\hat{P}(Y=1) = \Phi(-0.23 + 0.6X_2 - 0.4X_3 + 0.1X_2X_3)$.

а) Пусть $X_2 = 2,\ X_3 = 1$. Какое значение величины Y более вероятно?

б) Найдите предельный эффект от приращения переменной X_2 в точке $X_2 = 2,\ X_3 = 1$.

в) При каком значении X_2 моделируемая вероятность не зависит от X_3?

г) Что ещё нужно знать, чтобы проверить гипотезу об отсутствии связи между моделируемой вероятностью и переменной X_3 с помощью критерия Вальда? Критерия отношения правдоподобия?

№6.2.11. Надежда изучает вероятность прекращения трудовой деятельности среди пенсионеров. Она обследовала выборку из 400 работающих пенсионеров мужского пола, а через год провела повторное обследование, чтобы узнать, кто из опрошенных ушёл с работы. После этого Надя оценила модель логит, где объясняемая величина *Retired* равна 1 для прекративших работать и 0 для продолжающих, а объясняющие переменные – возраст (*Age*, годы), стаж работы на последнем рабочем месте (*Tenure*, годы), состояние здоровья (*Health*, измеряется по 10-балльной шкале). Вот результаты оценивания:

```
Logistic regression                               Number of obs   =        400
                                                  LR chi2(3)      =      17.43
                                                  Prob > chi2     =     0.0006
Log likelihood = -255.91205                       Pseudo R2       =     0.0329

------------------------------------------------------------------------------
     retired |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0698429   .0223703     3.12   0.002     .0259978    .1136879
      tenure |  -.0188485   .0207149    -0.91   0.363     -.059449     .021752
      health |   -.296244   .1124497    -2.63   0.008    -.5166412   -.0758467
       _cons |  -3.671451   1.607947    -2.28   0.022     -6.82297   -.5199327
------------------------------------------------------------------------------
```

А вот описательная статистика:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      health |       400     5.12875    .9547394          2        7.6
      tenure |       400      3.4775    5.369037          0         29
         age |       400       67.61    4.774924         60         77
```

а) Определите, какой объясняющий признак вносит наибольший вклад в разброс вероятностей прекращения работы.

б) Дайте по возможности полную интерпретацию результатам оценивания.

в) Рассчитайте 95% доверительный интервал для отношения шансов ухода с работы между двумя пенсионерами с одинаковым возрастом и стажем, у одного из которых индекс здоровья ниже на единицу, чем у второго.

№6.2.12. Оценивание пробит модели $P(Y=1) = \Phi(x'\beta)$ дало такой результат:

$\hat{P}(Y=1) = \Phi(0.6 - 0.5X_2 + 1.2X_3)$.

а) Какие бы вы ожидали оценки коэффициентов в логит модели?

б) Для проверки спецификации была оценена модель $\hat{P}(Y=1)=\Phi(\underset{(0.08)}{-0.04}+\underset{(0.14)}{1.11}x'\beta+\underset{(0.15)}{0.12}(x'\beta)^2)$. В скобках под оценками коэффициентов указаны стандартные ошибки.
Есть ли основания считать первоначальную модель неверно специфицированной?

№6.2.13. Психолог, изучая половую активность подростков, оценил модель логит, где объясняемая величина *SI* отражала наличие (1) или отсутствие (0) опыта полового сношения. Наличие этого опыта увязывалось с полом подростка и занятостью матери (полный рабочий день, неполный рабочий день, не работает). Посмотрите на таблицу с результатами оценивания[9]:

Переменная	**Коэффициент**	**Потенцированный коэффициент**	***p*-значение**
Male	0.102	1.107	0.652
Full	0.292	1.339	0.187
Part	-0.854	0.426	0.001
Male×Full	0.077	1.080	0.803
Male×Part	0.951	2.589	0.005
Своб. член	-1.046	0.351	

Дайте интерпретацию каждому из коэффициентов, включая свободный член. Дайте интерпретацию результатам проверки значимости.
Какими были бы оценки, если занятость полный рабочий день была бы выбрана базовой категорией?
Какова стандартная ошибка коэффициента при переменной *Male*?

№6.2.14. Связь между бинарными величинами Y и D может быть описана моделью логит: $\frac{P(Y=1)}{P(Y=0)}=\exp(0.3-0.2D)$. Найдите коэффициенты соответствующей линейной модели и пробит-регрессии.

№6.2.15. Перед вами результаты оценивания логит-модели, которая исследует факторы, способные повлиять на вероятность смерти индивида. Зависимая переменная бинарная, принимающая значение 1, если индивид умер, и 0, если остался жив в *i*-ом периоде. Среди регрессоров – самооценка здоровья (sah), курение (nosmoke), перенесенные инфаркты и/или инсульты (infmins), квадрат возраста (agesq).

```
Logistic regression                               Number of obs   =      10883
                                                  LR chi2(4)      =     245.39
                                                  Prob > chi2     =     0.0000
Log likelihood = -501.28999                       Pseudo R2       =     0.1966

------------------------------------------------------------------------------
     outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sah |   .3711193   .0908177     4.09   0.000     .1931199    .5491187
     nosmoke |  -1.302109   .2324934    -5.60   0.000    -1.757788   -.8464305
     infmins |   .9779169   .2278314     4.29   0.000     .5313756    1.424458
       agesq |   .0006243   .0000582    10.72   0.000     .0005102    .0007384
       _cons |  -7.237883   .3546244   -20.41   0.000    -7.932934   -6.542832
------------------------------------------------------------------------------
```

Запишите уравнение логит-модели в общем виде. Как можно оценить качество приведённой выше модели? Объясните, какова связь каждого из регрессоров с риском смерти в рамках конкретной модели.

[9] Данные этой таблицы взяты из книги J. Jaccard “Interaction effects in logistic regression”

№6.2.16. По данным о результатах трудоустройства 671 человека в далёкой Америке[10] была оценена модель:

$\frac{\hat{P}(H=1)}{\hat{P}(H=0)} = \exp(-1.29 + 0.68Male - 0.59Felony - 0.82NatAmer)$, где

H =1, если человек был принят на работу, 0 иначе,
Male - индикатор принадлежности к мужскому полу,
Felony - индикатор наличия судимости,
NatAmer - индикатор принадлежности к коренным американцам (индейцам).

А вот оценка ковариационной матрицы оценок коэффициентов и статистика отношения правдоподобия для проверки значимости модели в целом:

$$\hat{V}(\hat{\beta}) = \begin{array}{c} \begin{matrix} \textit{Константа} & \textit{Male} & \textit{Felony} & \textit{NatAmer} \end{matrix} \\ \begin{pmatrix} 0.028 & -0.027 & -0.006 & -0.007 \\ -0.027 & 0.038 & -0.003 & -0.002 \\ -0.006 & -0.003 & 0.098 & -0.004 \\ -0.007 & -0.002 & -0.004 & 0.04 \end{pmatrix} \end{array}, \quad LR = 23.3$$

а) Как вы думаете, успехом или неудачей чаще заканчивались попытки устроиться на эту работу?
б) Есть ли основания считать, что хотя бы одна из объясняющих переменных связана с вероятностью трудоустройства?
в) Дайте интерпретацию полученным оценкам.
г) Постройте 95% доверительный интервал для коэффициента при переменной *Felony* и для соответствующего отношения шансов. Найдите *p*-значение для проверки значимости этого коэффициента.
д) Проверьте гипотезу о том, что при одинаковом статусе судимости мужчина-индеец и женщина не из коренных американцев имеют одинаковые шансы устроиться на эту работу.

№6.2.17. Надюша изучает розовоносых единорогов. Каждый розовоносый единорог рождается с белым носом, который с возрастом приобретает розовый окрас. Время, за которое нос окончательно меняет окраску, Надюша описывает случайной величиной с функцией распределения $F(t) = 1 - \frac{1}{1+(\lambda t)^2}$, $t \geq 0$. Она отобрала n только что родившихся розовоносых единорогов и через месяц после их рождения (t = 1) отметила, приобрёл ли нос каждого розовую окраску. Так получился ряд наблюдений $y_1, \ldots, y_n$, где $y_i = 1$, если нос у единорога i приобрёл розовый цвет, и $y_i = 0$, если окрашивание не завершилось. Кроме того, Надюша собрала данные о характеристиках каждого единорога x_i', которые предположительно определяют неизвестный параметр распределения: $\lambda_i = \exp(x_i'\beta)$. Вектор коэффициентов β предполагается оценивать методом максимального правдоподобия. Выпишите функцию правдоподобия (выразите её через y_i, x_i', β).

№6.2.18. Винни-Пух обследовал 60 мест обитания пчёл и оценил зависимость:

$\frac{\hat{P}(RightHoney=1)}{\hat{P}(RightHoney=0)} = \exp(0.73 + 0.37RightBees), \quad \ln L = -36.47$.

Здесь RightBees=1, если в месте обитают правильные пчёлы, 0 иначе.
RightHoney=1, если пчёлы делают правильный мёд, 0 иначе.

а) Винни-Пух находит новое место, где обитают правильные пчёлы. С какой вероятностью (по его оценкам), они делают правильный мёд?
б) Проверьте значимость оценённой Винни-Пухом модели на уровне 5%. Известно, что без объясняющих переменных $\ln L = -36.65$.

[10] Это данные к книге J. Hilbe “Logistic regression models”.

§6.3. Модели множественного выбора

№6.3.1. По результатам опроса 300 жителей некоторого города изучалась зависимость уровня поддержки главы города от пола и семейных расходов. Используемые переменные:

Att_i - отношение к действиям главы города по пятибалльной шкале (1 - «полностью отрицательное», 5 – «полностью положительное»),

$Male_i = 1$ для мужчин и 0 для женщин,

$Expend_i$ - месячные затраты семьи, тыс. руб. (используется как измеритель дохода).

Оценивалась упорядоченная логит-модель:

$$\frac{P(Y_i \geq s)}{P(Y_i < s)} = \exp(\beta_1 Male_i + \beta_2 \ln Expend_i - \alpha^s), \quad s = 2,3,4,5.$$

Результаты оценивания:

Коэффициент	β_1	β_2	α^2	α^3	α^4	α^5
Оценка	-0.70	1.73	2.68	4.50	7.16	10.00
Стандартная ошибка	0.23	0.40	1.39	1.39	1.43	1.51

а) Оцените вероятность того, что мужчина из семьи с месячными расходами в 30 тыс. руб относится к действиям главы города полностью отрицательно. Каково наиболее вероятное значение Att_i в этом случае?

б) Рассчитайте для этого случая предельный эффект от увеличения затрат.

в) Проверьте значимость коэффициента β_1, рассчитайте *p*-значение.

г) Скажите, какими были бы (приблизительно) оценки коэффициентов логит модели с теми же регрессорами, в которой объясняемая переменная Z_i принимала бы значение 1 при $Y_i \geq 2$ и 0 иначе.

д) Оценив модель из предыдущего пункта, вы обнаружили, что оценки коэффициентов существенно расходятся с вашими ожиданиями. О чём это может говорить?

№6.3.2. Величина Y принимает значения целые значения от 0 до *p*, связь распределения Y с вектором переменных x' задаётся множественной логит-моделью:

$\frac{P(Y = s)}{P(Y = 0)} = \exp(x'\beta^s)$, $s = 1,..., p$, где β^s - векторы коэффициентов.

Покажите, что а) условная вероятность $P(Y = 1 | Y \leq 1)$ задаётся бинарной логит-моделью,

б) условное распределение Y при условии $\{Y \leq 2\}$ задаётся множественной логит-моделью.

№6.3.3. В таблице приведены результаты исследования связи между занятостью и полом по ответам 290 респондентов:

Сфера занятости \ Пол	Мужской	Женский
Производство	75	50
Сфера услуг	55	60
Вне рабочей силы	25	25

Представьте эти результаты в виде оценённой множественной логит-модели, в которой объясняемый признак – сфера занятости.

№6.3.4. Любознательная барышня Ольга изучает человеческое представление о блаженстве. В своём исследовании она опросила выборку из 263 человек в возрасте от 15 до 35 лет, которым предлагалось ответить на вопрос:

На мой взгляд, чувство блаженства наиболее близко тем ощущениям, что возникают, когда

1) гладишь мурчащего кота,

2) гуляешь под тёплым летним ливнем,
3) просыпаешься утром и понимаешь, что наступил первый день каникул или отпуска.
Выбор респондента увязывался с его полом и возрастом с помощью множественной логит-модели, оценки которой приведены ниже (базовая категория – прогулка под ливнем):

Переменные	**Коэффициенты**	
	Поглаживание кота	*Начало каникул*
Пол (1 - мужчина)	-0.15	0.04
Возраст	-0.01	0.02
Константа	0.1	0.05

а) Случайно отобранный человек оказывается женщиной 20 лет. С какой вероятностью она свяжет своё представление о блаженстве с поглаживанием мурчащего кота? Каково различие в этой вероятности между мужчиной и женщиной 20 лет?
б) Какими будут оценки той же модели по тем же данным, если выбрать базовой категорией третий вариант ответа?
в) Предположим, Ольга удалила из выборки всех, кто выбрал вариант с поглаживанием кота, и оценила логит-модель $\frac{P(\text{начало каникул})}{P(\text{прогулка под ливнем})} = \exp(\beta_1 + \beta_2 Пол + \beta_3 Возраст)$. Будь вы на месте Ольги, какие значения оценок коэффициентов вы бы ожидали получить?
г) Подумайте, как графически представить оценённую зависимость. Попробуйте осуществить свои идеи (наверное, это удобнее делать на компьютере).

№6.3.5. Какая из четырёх моделей в наибольшей степени подходит для описания перечисленных статистических признаков?
Модели:
1) линейная регрессия,
2) бинарная логит,
3) упорядоченная логит,
4) номинальная логит.
Признаки:
а) наличие задолженности по заработной плате;
б) специальность, по которой респондент обучается в институте;
в) наступление страхового события за определённый период времени;
г) затраты семьи на покупку продуктов питания;
д) удовлетворённость работой, измеренная по пятибалльной шкале от 1 («совсем не удовлетворён») до 5 («полностью удовлетворён»);
е) интегральный индикатор качества жизни в регионе, измеренный по непрерывной шкале от 0 до 10;
ё) ответ на вопрос «как часто вы проходите профилактический медицинский осмотр?» с предложенными вариантами: «1 – реже, чем раз в два года, 2 – раз в год или два года, 3 – чаще раза в год»;
ж) предпочитаемый утренний напиток: сок, вода, чай, кофе или другой;
з) наличие какого-либо домашнего животного у респондента.

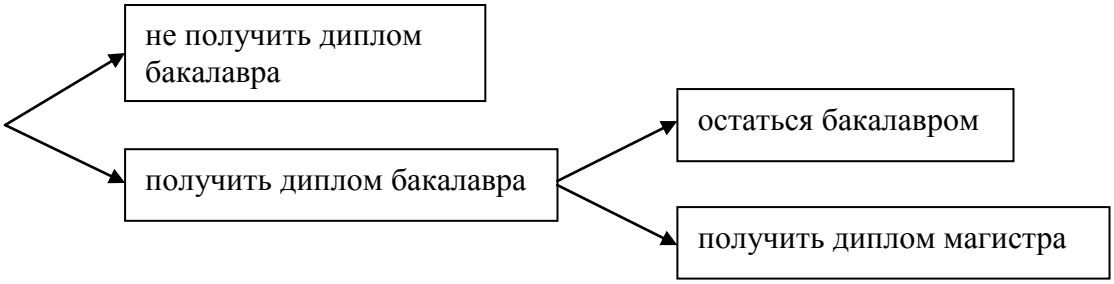
№6.3.6. Среди нефритовых и золотых рощ горы Пэньлай проводит свои дни поэт Ли Бо, уйдя от мирской суеты. Каждое утро приходит к нему госпожа Ли Цинчжао и приносит несколько кубков вина, перед сном же она приходить вновь и отмечает уровень счастья Ли Бо по трёхбалльной шкале (1 – «мне грустно так, что лучше помолчать», 2 – «забыли мы про старые печали», 3 – «пляшу, и пляшет тень моя»). Наблюдая поэта долгие годы, Ли Цинчжао поняла, что три вещи определяют его блаженство: число выпитых кубков вина, возможность любоваться луной и созвездием Винных Звёзд (чему мешают иногда облака) и редкие визиты его друга Ду Фу.

Госпожа оценила модель $\frac{P(Bliss_i \geq s)}{P(Bliss_i < s)} = \beta_1 Wine_i + \beta_2 Wine_i^2 + \beta_3 Clouds_i + \beta_4 Friend_i - \alpha^s,\ s = 2,3$:

Коэффициент	β_1	β_2	β_3	β_4	α^2	α^3
Оценка	1.20	-0.12	-1.26	1.42	0.02	10.93
Станд. ошибка	0.11	0.05	0.16	0.29	0.15	0.56

а) Помогите Ли Цинчжао понять, сколько кубков вина стоит приносить Ли Бо, чтобы доставить ему наивысшее блаженство.
б) Оцените наибольшую возможную для Ли Бо вероятность достигнуть высшего уровня блаженства по трёхбалльной шкале.
в) Ли Бо не вполне уверен, что вино может уменьшать счастье. Помогите Ли Цинчжао убедить его в этом, опираясь на оценённую модель: выясните, есть ли основания считать, что истинный коэффициент β_2 отрицателен.
г) Подумайте, как графически представить оценённую зависимость. Попробуйте осуществить свои идеи (наверное, это удобнее делать на компьютере).

№6.3.7. Опираясь на индивидуальные данные о студентах университета, наблюдавшихся с первого курса до прекращения обучения или получения магистерского диплома, исследователь пытается разработать модель, в которой объясняемая величина – полученный уровень образования. Эта величина может принимать три значения: «не получил диплом», «бакалавр», «магистр».
Исследователь выбирает между двумя вариантами. Первый – применить упорядоченную логит модель, второй – представить выбор уровня образования в виде такого дерева:



В каждом разветвлении выбор описывается с помощью бинарной модели логит.
Какой подход вы посоветовали бы использовать? Обсудите преимущества и недостатки каждого варианта.

№6.3.8. Американский исследователь Огден Фредерик Нэш утверждал, что обнаружил прямую зависимость между любвеобильностью собак и их мокростью[11]. Любознательная барышня Маша задалась целью проверить этот результат, для чего осмотрела 150 бродячих собак в своём регионе, отмечая их мокрость и любвеобильность:
$Wet_i = 1$, если *i*-й пёс мокрый; 0, если сухой.
$Lovingness_i = 2$, если пёс любвеобилен; 1, если осторожно ластится; 0, если чуждается.
По собранным данным Маша оценила модель:

$$\frac{\hat{P}(Lovingness_i \geq s)}{\hat{P}(Lovingness_i < s)} = \exp(0.05 Wet_i - \hat{\alpha}^s), \quad \ln L = -139.5.$$

$\hat{\alpha}_1 = 0.1, \quad \hat{\alpha}_2 = 2.2$.
а) Оцените долю любвеобильных собак среди генеральной совокупности мокрых.
б) Каково, согласно полученным оценкам, отношение шансов любвеобильности между мокрыми и сухими псами?

[11] Свои наблюдения он изложил в стихах (см. O.Nash, “The dog”)

в) Проверьте, есть ли основания считать, что существует связь между признаками, если в модели без объясняющей переменной логарифм функции правдоподобия равен -140.0? Рассчитайте псевдо- R^2 .
г) Маша обдумывала и другой способ – проверить независимость с помощью обычного критерия хи-квадрат для таблиц сопряжённости. Скажите, в чём существенные различия между двумя способами выявления связи? В чём их достоинства и недостатки?

№6.3.9. Вот выборка из 100 наблюдений за признаком, принимающим значения 1, 2 и 3:

Значение	Частота
1	40
2	30
3	30

Опишите эту выборку в форме порядковой модели логит без объясняющих переменных.

§6.4. Чувствительность и специфичность. Кривая ROC.

№6.4.1. Разработанная в банке система выявления неблагонадёжных заёмщиков показала на обучающей выборке такие результаты:

		классифицирован	
		надёжный	ненадёжный
своевременная выплата	да	800	40
	нет	100	60

Рассчитайте чувствительность и специфичность.

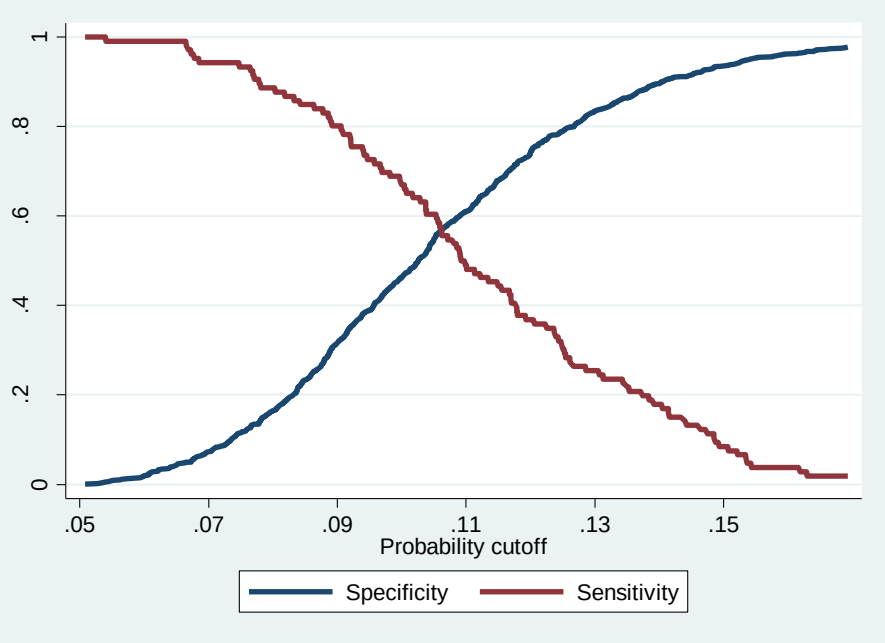
№6.4.2. Для бинарных случайных величин X, Y, Z выполняется соотношение: $P(Y=1) = -0.4 + 0.6X - 0.2Z$. Совместное распределение X и Z задаётся таблицей:

	$X = 0$	$X = 1$
$Z = 0$	0.3	0.2
$Z = 1$	0.1	0.4

а) Для прогнозного правила $\{\hat{Y}=1$ при $P(Y=1)>c\}$ Постройте график ROC, найдите площадь под кривой.
б) Постройте график зависимости чувствительности и специфичности от c .

№6.4.3. Рассмотрим прогнозное правило $\{\hat{Y}=1$ при $P(Y=1)>c\}$. При каких значениях c правило имеет: а) наибольшую чувствительность, б) наибольшую специфичность, в) наибольшую вероятность правильного прогноза?

№6.4.4. Разработанная в банке модель вероятности невыплат по кредиту характеризуется таким графиком чувствительности и специфичности:



По оценкам банка, плохие заёмщики (не выплачивающие долг) попадаются в 10 раз реже хороших, а ожидаемые потери от предоставления кредита плохому заёмщику в 20 раз выше, чем от непредоставления хорошему. При какой модельной вероятности невыплат стоит отказывать в предоставлении кредита, если поставлена цель минимизировать ожидаемый убыток?

Часть 7. Модели времени жизни

§7.1. Функции дожития и риска, интегральная функция риска.

№7.1.1. Распределение случайной величины T описывается функцией дожития $S(t)=\frac{1}{(1+t)^2}$. Выпишите функцию риска и интегральную функцию риска величины T, опишите характер временной зависимости. Выпишите условную функцию дожития при условии $\{T\geq 1\}$.

№7.1.2. Пусть время жизни распределено равномерно на отрезке [0;1]. Как вы думаете, каков в этом случае характер временной зависимости? Проверьте свою догадку, найдя функцию риска.

№7.1.3. Время работы телевизора до поломки описывается показательным распределением. Известно, что в среднем только что купленный телевизор работает 5 лет.
а) Какова вероятность того, что только что купленный телевизор проработает 10 лет?
б) Купленный 5 лет назад телевизор до сих пор не вышел из строя. Какова вероятность того, что он проработает ещё 10 лет?
в) Какими будут ответы на вопросы (а) и (б), если время работы телевизора распределено по закону Вейбулла с параметрами $\lambda=0.1$ и $p=1.3$?

№7.1.4. Прибор состоит из двух независимо функционирующих и последовательно соединённых блоков, так что для работы прибора необходима работа каждого из этих блоков. Время работы первого блока описывается показательным распределением с параметром λ_A, время работы второго – показательным распределением с параметром λ_B. Найдите функцию дожития для времени работы прибора.

№7.1.5. Распределение случайной величины T описывается функцией плотности $f(t)=\begin{cases}2t, & t\in[0;1],\\ 0, & \textit{иначе}\end{cases}$
Найдите: а) соответствующую функцию дожития,
б) функции риска и интегрального риска на интервале [0; 1),
в) медиану величины T, функцию квантилей.

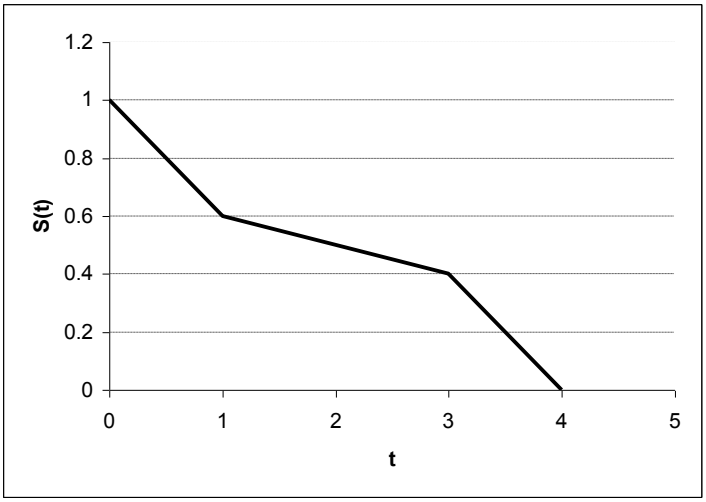
№7.1.6. Домохозяюшка Настя любит устраивать генеральные уборки. Число лет между уборками – случайная величина с функцией дожития $S(t)=\exp\left(-t^2\right)$.
а) Найдите функцию риска проведения уборки. Опишите характер временной зависимости, дайте интерпретацию этой зависимости: почему она именно такова? Какие соображения могут руководить Настей?
б) Настя уже полгода не проводила генеральную уборку. Какова вероятность того, что она проведёт её в течение следующей половины года?

№7.1.7. Известно, что из проданных холодильников 5% выходят из строя в течение первого года эксплуатации, а 10% - в течение второго года. Рассчитайте вероятность того, что недавно проданный холодильник будет работать не менее полутора лет в предположении, что а) момент поломки равномерно распределён в течение года, б) риск выхода из строя постоянен в течение года.

№7.1.8. Беспокойный медведь Михаил, лёжа в берлоге, ворочается с боку на бок. Время, которое он проводит на одном боку, имеет показательное распределение, в среднем равно 2 суткам и не

зависит от того, сколько он лежал на том или ином боку раньше. Какова вероятность того, что за 140 дней зимы Миша перевернётся с боку на бок не менее 64 раз?

№7.1.9. Функция дожития величины T изображена на графике:



Найдите: а) $P(1<T<3)$, б) медиану T, в) $E(T)$,
г) $P(T>3\,|\,T>1)$, д) $E(T\,|\,T>1)$.

№7.1.10. Пусть длительность T имеет показательное распределение. Покажите, что величина T^{0}, равная целой части T, распределена по геометрическому закону.

№7.1.11. Для освещения шпульно-катушечного цеха на ткацкой фабрике требуется 1000 лампочек. Причем на заводе имеется достаточный запас лампочек и, как только какая-то лампочка в цеху перегорает, её сразу же заменяют новой. Функция риска для каждой лампочки постоянна и равна 0.5.
1) Укажите среднее время жизни лампочки.
2) Определите, сколько (в среднем) лампочек будет сгорать ежемесячно.
3) Найдите интегральнкю функцию риска и функцию дожития.
4) Нарисуйте график функции квантилей.
5) Найдите математическое ожидание числа лампочек, которые придётся придется заменить, если шпульно-катушечный цех будет непрерывно работать целый год.

№7.1.12. Время дожития описывается функцией $S(t)=\dfrac{64}{(t+8)^2}$. Найдите медианное время дожития, выпишите функцию риска. Найдите условную функцию дожития при условии, что объект дожил до 8 лет.

№7.1.13. По данным о продолжительности забастовок в промышленности США медианная длительность забастовки составляет 27 дней. Предположим, что продолжительность забастовки имеет показательное распределение.
а) Рассчитайте вероятность того, что забастовка будет длиться более 40 дней.
б) Некоторая забастовка длится уже 20 дней. Какова вероятность того, что она продлится ещё не менее 40 дней?

№7.1.14. Докажите формулу $E(T)=\sum_{t=1}^{\infty}S(t)$ для величины T, принимающей неотрицательные целые значения (обоснуйте геометрическую интерпретацию математического ожидания).

№7.1.15. Пусть T_1, T_2 - независимые случайные величины, имеющие геометрическое распределение с одинаковым параметром p. Какое распределение будет у величины а) $\min(T_1, T_2)$, б) $\max(T_1, T_2)$, в) $T_1 + T_2$? Каков будет характер временной зависимости в каждом из случаев?

№7.1.16. Покажите, что функция плотности может не убывать только в области с положительной временной зависимостью.

§7.2.Статистический анализ данных о длительности состояний.

№7.2.1. Китайский поэт Ли Бо приобрёл четыре кувшина для вина и в течение трёх лет наблюдал за их жизнью. Результаты его наблюдений приведены в таблице:

№ кувшина	*Время жизни*	*Состояние*
1	1 год	разбился
2	1 год	разбился
3	3 года	разбился
4	3 года	цел

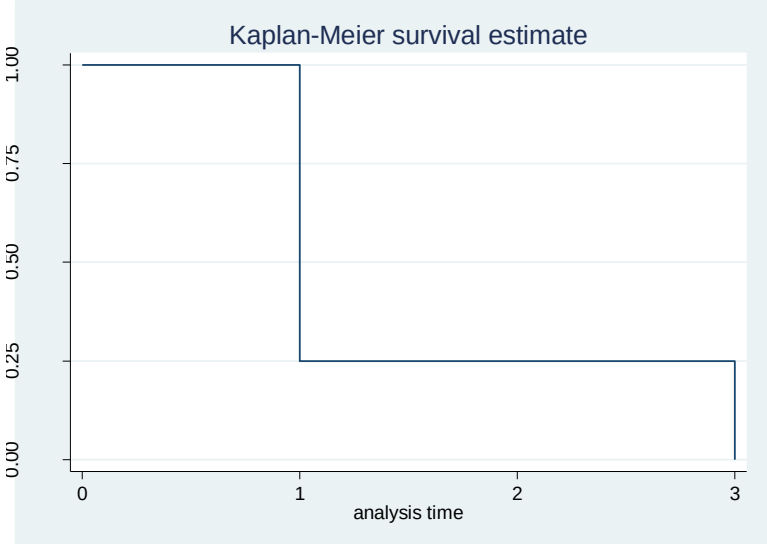
а) Нарисуйте график оценки Каплана-Майера для функции дожития и оценки Нельсона-Аалена для интегральной функции риска.
б) Предположим, что время жизни кувшина имеет показательное распределение: $S(t) = e^{-\lambda t}, \lambda > 0$. Выпишите функцию правдоподобия для данной выборки.

№7.2.2. Василий собрал данные о времени работы 100 холодильников до появления неисправностей:

Срок работы:	менее года	1-2 года	более 2 лет
Число холодильников:	20	15	65

Предполагая, что время работы холодильника имеет показательное распределение, выпишите функцию правдоподобия и оцените среднее время работы холодильника.

№7.2.3. По графику оценки Каплана-Майера оцените математическое ожидание и дисперсию и постройте график оценки Нельсона-Аалена (данные не подвержены усечению):



№7.2.4. Имеется цензурированная справа выборка $t_1, \ldots, t_n$, взятая из показательного распределения с параметром λ; d_i - индикатор цензурирования. Предполагается, что

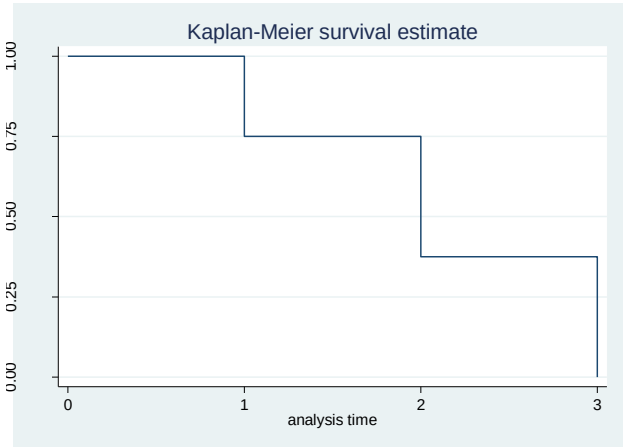
наблюдения есть реализации независимых случайных величин, а цензурирование детерминировано. Постройте оценку для λ с помощью метода максимального правдоподобия.

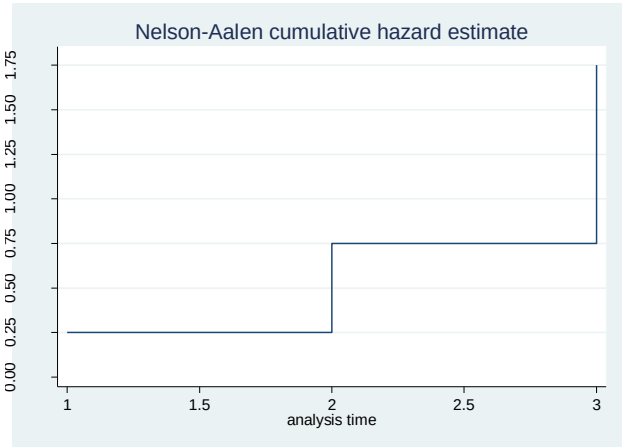
№7.2.5. По данным о длительности пребывания бывших заключённых на свободе до повторного ареста оцените соответствующие функции Каплана-Майера и Нельсона – Аалена. Постройте график для функции дожития. Оцените по графику среднее и медианное время до момента повторного ареста. Рассчитайте грубую оценку функции риска (incidence rate, среднее число завершений в единицу времени).

№ заключённого	Число недель, проведённых на свободе	Арестован (1 – да)
174	17	1
411	27	1
99	30	0
273	32	1
77	43	1
229	43	1
301	47	1
65	52	1
168	52	0
408	52	0

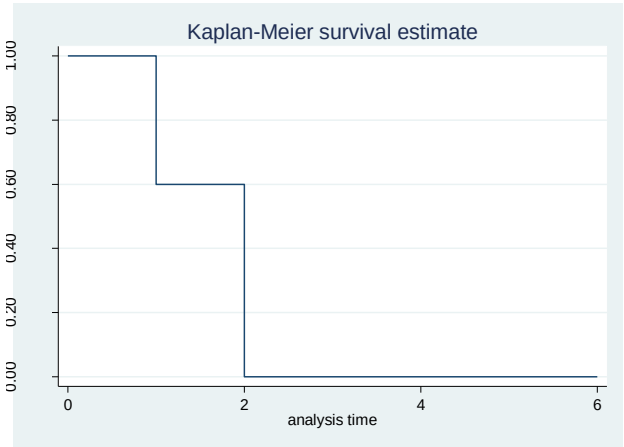
№7.2.6. Составьте выборку из 8 наблюдений, соответствующую графикам оценок Каплана-Майера и Нельсона-Аалена:

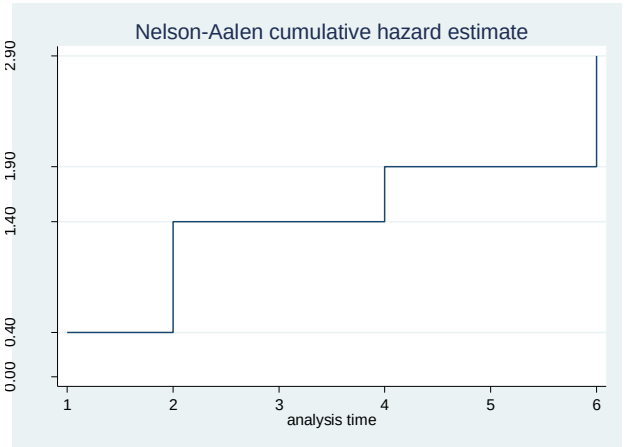
а)



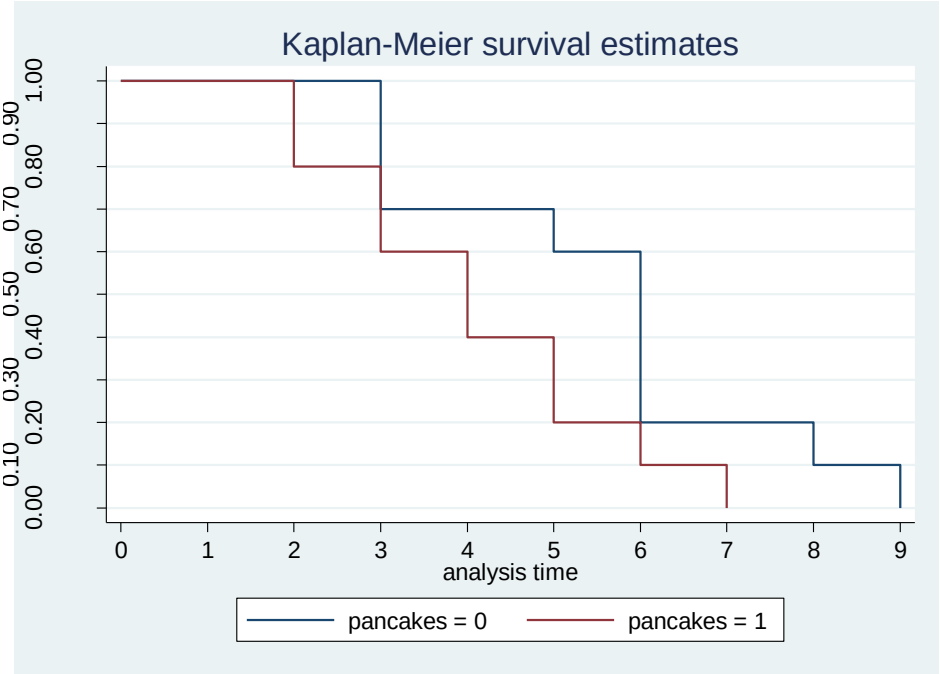


б)





№7.2.7. Две лисички и крокодил устраивают опыты: поджигают в ванной воду и пытаются её тушить либо блинами, либо сушёными грибами, замеряя время до погашения. Ниже приведены графики оценок Каплана-Майера для времени погашения, полученные из двух серий опытов: десяти попыток тушения огня сушёными грибами (синяя линяя) и десяти попыток тушения блинами (красная линия).



а) При каком способе тушения среднее время погашения было меньше? Попробуйте дать обоснованный ответ, не прибегая к вычислениям.
б) Каково среднее время тушения пожара блинами?
в) Начертите график оценки интегрального риска для времени тушения пожара блинами.
г) Лисички и крокодил сравнили две выборки с помощью критерия логарифмических рангов, получив такой результат:

	Events observed	Events expected
грибы	10	13.18
блины	10	6.82
Total	20	20.00
	chi2(1) =	3.31
	Pr>chi2 =	0.0687

Цветом выделены критическая статистика и *p*-значение.
Каковы основная и альтернативная гипотезы проведённого теста? Какой можно сделать вывод?

§7.3. Регрессионные модели длительности.

№7.3.1. Модель длительности забастовки T (измеряется в неделях) включает в качестве объясняющих переменных индекс промышленного производства *INDP* и переменную *FEB*, равную единице для забастовок, начавшихся в феврале, и 0 иначе. Ниже приведены результаты оценивания модели в двух параметрических спецификациях (экспоненциальная и Вейбулла) и полупараметрической (модель пропорциональных рисков Кокса)[12].

[12] Эти результаты частично взяты из текста G.R. Neumann “Search models and duration data”.

Коэффициент	Экспоненциальная модель	Модель Вейбулла (метрика проп. рисков)	Модель Вейбулла (метрика ускор. времени)	Модель Кокса
β_1 (свободный член)	-3.72	-3.67	3.71	-
β_2 (при *INDP*)	3.30	3.25	-3.28	3.25
β_3 (при *FEB*)	-0.57	-0.55	0.56	-0.57
p (параметр формы)	-	0.99	0.99	-
$\ln L$	-800.57	-800.55	-800.55	-2696.66

Для модели Кокса приведено значение функции частичного правдоподобия.

а) Согласно оценкам экспоненциальной регрессии, как выглядит функция риска для забастовки, начавшейся в июне, когда индекс промышленного производства был равен 0.01?
б) Согласно оценкам, полученным из модели Вейбулла, каковы различия в ожидаемой продолжительности и функции риска между забастовками, начавшимся в феврале, и теми, что начались в другие месяцы, при прочих равных условиях? Дайте количественную оценку этих различий.
в) Выпишите в линейной форме оценённую модель Вейбулла.
г) Предположим, что вы хотите выбрать модель, наилучшим образом описывающую данные. Какой вывод вы можете сделать, руководствуясь критерием Акаике?

№7.3.2. По данным о миротворческих миссиях ООН с 1948 по 2001 год оценена регрессия продолжительности миссии на переменные Interstate (равна 1 для межстранового конфликта и 0 иначе) и Civil (1 для гражданской войны, 0 иначе). В качестве базовой категории конфликтов выступала категория гражданских войн с иностранным вмешательством (internationalized civil war – гражданская война, в которой хотя бы одна из сторон получает поддержку со стороны другого государства). При оценивании использовалась модель Вейбулла: $h(t) = \lambda p t^{p-1}$, где параметр λ зависел от вектора объясняющих переменных $x = (1 \quad Interstate \quad Civil)'$ экспоненциально: $\lambda = \exp(x'\beta)$. Также была оценена показательная (экспоненциальная) модель, где параметр p полагался равным 1. Результаты оценивания приведены ниже[13]:

Переменные	Оценки коэффициентов (станд. ошибки)	
	Показательная модель	Модель Вейбулла
Константа	-4.35 (0.21)	-3.46 (0.50)
Civil	1.16 (0.36)	0.89 (0.38)
Interstate	-1.64 (0.50)	-1.40 (0.51)
p	1.00 (---)	0.81 (0.10)
Логарифм функции правдоподобия	-86.35	-84.66
Число наблюдений	54	

а) По результатам оценивания модели Вейбулла выясните, есть ли основания считать, что продолжительность миротворческой миссии при урегулировании гражданской войны связана с иностранным вмешательством. Используйте уровень значимости 5%.
б) Проверьте гипотезу об отсутствии временной зависимости.
в) Сравните качество подгонки двух моделей по критерию Акаике.
г) Нарисуйте графики функций риска для каждого типа конфликта (наверное, это удобнее делать на компьютере).

№7.3.3. По выборке из n независимых наблюдений T_1, …, T_n оценивается модель пропорциональных рисков. Предполагается, что функция риска величины T_i имеет вид

[13] А это мы взяли из книжки J.M. Box-Steffensmeier, B.S. Jones “Event history modeling: a guide for social scientists”

$h(t) = h_0(t)\exp(x_i'\beta)$, где $h_0(t) = 1 + \alpha t$. Здесь x_i – вектор объясняющих переменных для i-го наблюдения, β – вектор оцениваемых коэффициентов при объясняющих переменных, α – неотрицательный параметр опорной функции риска. Часть наблюдений цензурированы справа (тогда T_i отражает длительность i-го состояния на момент цензурирования), в остальных изучаемая длительность точно известна. Обозначим за δ_i индикатор цензурирования, равный единице для цензурированных наблюдений и нулю для остальных. Выпишите функцию правдоподобия для оценивания параметров α и β при известных T_i, x_i, δ_i.

№7.3.4. Исследователь предполагает, что величина T описывается функцией дожития $S(t) = (1+t)^{-\lambda}$, где параметр λ связан с вектором объясняющих переменных x: $\lambda = \exp(x'\beta)$. Покажите, что исследователь имеет дело с моделью пропорциональных рисков.

№7.3.5. В программе лечения от табакокурения участвовало 200 человек, по наблюдениям за которыми была оценена модель длительности воздержания от курения. Предполагалось, что время воздержания T имеет распределение Вейбулла с функцией риска $h(t) = \lambda p t^{p-1}$, где параметр λ зависит от объясняющих переменных *male* (1 для мужчин и 0 для женщин) и *exper* (стаж курения в годах): $\lambda = \exp(\beta_1 + \beta_2 male + \beta_3 exper)$. Вот результаты оценивания:

Параметр:	β_1	β_2	β_3	p
Оценка:	-1.95	-0.16	0.14	0.82
Стандартная ошибка:	0.15	0.10	0.04	0.05

а) Что можно сказать о характере временной зависимости?
б) Есть ли основание считать, что при одинаковом стаже курения риск заново начать курить для мужчин и женщин различен?
в) Сформулируйте в терминах параметров модели гипотезу: для женщины с двухлетним стажем курения риск заново начать курить не зависит от продолжительности воздержания и равен 0.2.

№7.3.6. Давным-давно в одной далёкой-далёкой области демограф Дмитрий добыл данные о возрасте вступления в брак среди 500 мужчин и оценил логлогистическую регрессию этой величины на переменные *Ciudad* (1 для горожан, 0 для сельчан) и *Rubio* (1 для светловолосых, 0 иначе).

```
Loglogistic regression -- accelerated failure-time form

No. of subjects =          500                     Number of obs   =       500
No. of failures =          405
Time at risk    =  14796.55582
                                                   LR chi2(2)      =     12.00
Log likelihood  =    -214.4989                     Prob > chi2     =    0.0025
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ciudad |   .1327109   .0380371     3.49   0.000     .0581596    .2072621
       rubio |  -.0214258   .0315395    -0.68   0.497     -.083242    .0403904
       _cons |   3.279762   .0341681    95.99   0.000     3.212794    3.346731
-------------+----------------------------------------------------------------
     /ln_gam |  -1.666244   .0408503   -40.79   0.000    -1.746309   -1.586179
-------------+----------------------------------------------------------------
       gamma |   .1889554   .0077189                      .1744165    .2047063
------------------------------------------------------------------------------
```

а) Дайте по возможности полную интерпретацию результатов оценивания.
б) Выпишите модель в линейной форме.
в) Оцените медианный возраст вступления в брак для темноволосого сельчанина.
г) Начертите графики функций дожития и риска для четырёх групп мужчин по типу поселения и цвету волос (наверное, это удобнее делать на компьютере).

№7.3.7. Оценивается модель расторжения договоров страхования жизни, где объясняемая величина – время действия договора до его расторжения lifetime (в днях), а регрессоры – дамми-

переменные возраста (age_30=1 для клиентов моложе 30 лет, age50=1 для клиентов от 50 лет и старше), пола (male=1 для мужчин) и типа договора (prestige=1 для дорогих договоров типа «Престиж», 0 для более дешёвых договоров типа «Классика»). Вот результаты оценивания модели Вейбулла в метрике ускоренного времени:

```
                                                LR chi2(4)      =     21.37
Log likelihood  =   -174.16693                  Prob > chi2     =    0.0003
------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_30 |  -1.414679   .5591233    -2.53   0.011    -2.510541    -.318818
      age50_ |   .4295516   .5389189     0.80   0.425    -.6267101    1.485813
        male |   .3393536   .5893608     0.58   0.565    -.8157724     1.49448
    prestige |  -1.485285   .4947198    -3.00   0.003    -2.454918   -.5156518
       _cons |   8.589703   .5142198    16.70   0.000     7.581851    9.597555
-------------+----------------------------------------------------------------
           p |   .6174127   .0731074                      .4895369     .778692
         1/p |   1.619662    .191783                      1.284205    2.042747
------------------------------------------------------------------------------
```

а) Женщина 35 лет только что застрахована по договору типа «Классика». С какой вероятностью договор будет расторгнут в течение года?
б) Как различаются медианы времени расторжения для мужчины 58 лет и женщины 35 лет, если они заключили договор одинакового типа?
в) Выпишите модель в линейной форме.
г) Рассчитайте значение критерия Акаике.

№7.3.8. Исследователь предполагает, что продолжительность безработицы описывается моделью экспоненциальной регрессии $h(t \mid x) = \exp(x'\beta)$. Для определения коэффициентов β он собирает выборку из *n* безработных объёма и измеряет их характеристики $x_1, ..., x_n$. Через месяц он находит тех же людей и узнаёт, продолжают ли они быть быть безработными, дополняя свои наблюдения величинами $Y_1, ..., Y_n$, где $Y_i = 1$, если безработный под номером *i* прекратил поиск, и 0 иначе.
а) Выпишите логарифмическую функцию правдоподобия для оценивания коэффициентов β по данным $x_1, ..., x_n$, $Y_1, ..., Y_n$.
б) Каковы могут быть последствия «задержанного входа» - того, что опрашиваемые уже находились в состоянии безработицы какое-то время на момент их первого опроса исследователем?

§7.4.Ненаблюдаемая разнородность.

№7.4.1. Пусть совокупность безработных делится на две группы: активно ищущие работу и не прилагающие к поиску особых усилий. Риск выхода из безработицы для активно ищущих задаётся функцией $h_A(t) = 0.5$, а для малоактивных - $h_{NA}(t) = 0.2$ Доля активных безработных среди только что начавших поиск составляет 40%.
а) Каково среднее время поиска работы для активно ищущих работу? Для малоактивных?
б) Некий человек только что потерял работу. Неизвестно, насколько активно он будет её искать. Найдите функции дожития и риска для продолжительности поиска работы этим человеком. Постройте график функции риска.

№7.4.2. Легкомысленная девушка Таня ищет мужчину своей мечты. Встретив кандидата на эту роль, Таня немедленно выходит за него замуж. Будем считать, что с вероятностью 20% она не ошибается и живёт затем счастливой семейной жизнью, не разводясь. В противном случае Таня

начинает постепенно осознавать свою оплошность, и осознаёт её тем полнее, чем больше времени проводит в браке. Как следствие, риск развода растёт: $h(t) = 0.5t^2$.

а) Только что Таня вышла замуж. Выпишите функцию дожития и функцию риска для величины T – продолжительности пребывания в браке. Дайте интерпретацию характеру временной зависимости.

б) Прошло три года, а Таня так и не развелась. Рассчитайте вероятность того, что Таня нашла мужчину своей мечты.

№7.4.3. Завод выпускает телевизоры, доля брака в продукции составляет 5%. Риск поломки не зависит от срока эксплуатации телевизора и описывается функцией $h_1(t) = 0.2$ для качественных телевизоров и $h_2(t) = 2$ для бракованных (время измеряется в годах).

а) Только что покупатель приобрёл телевизор с этого завода. Выпишите функции дожития и риска для времени его жизни.

б) Телевизор проработал год, не сломавшись. Какова вероятность того, что он бракованный?

№7.4.4. Случайная величина T имеет показательное распределение со случайным параметром λ. Найдите функции дожития и риска, если λ имеет

а) равномерное распределение на отрезке [0; 1],

б) показательное распределение с единичным параметром.

№7.4.5. В городе НН все новорождённые девочки делятся на невест (их 80%) и недотрог (20%). Недотроги никогда не выходят замуж. Для невест возраст вступления в брак – логарифмически нормальная случайная величина с параметрами $\mu_M = 3.2$ и $\sigma^2 = 0.01$ для темноволосых девочек и $\mu_R = 3.4$ и $\sigma^2 = 0.01$ для светловолосых. Принадлежность к недотрогам не зависит от цвета волос новорождённой.

Красавец Митрофан собирается жениться и рассматривает две кандидатуры: темноволосую Прасковью 25 лет от роду и светловолосую Софью 32 лет (обе замужем не были). Митрофан уверен, что отказать ему может только недотрога. К кому стоит свататься, чтобы иметь бóльшие шансы на успех?

№7.4.6. Оценивалась экспоненциальная регрессия с ненаблюдаемой разнородностью: $h(t) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2) \cdot \alpha$, где случайная величина α имеет гамма-распределение с математическим ожиданием 1 и дисперсией θ. Вот результаты оценивания:

```
Exponential regression -- log relative-hazard form
                       Gamma frailty

No. of subjects =         1000                     Number of obs   =      1000
No. of failures =         1000
Time at risk    =  5883.972279
                                                   LR chi2(2)      =     19.62
Log likelihood  =   -2219.7535                     Prob > chi2     =    0.0001

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1563679   .0387083     4.04   0.000     .0805011    .2322347
          x2 |   -.235001   .1183053    -1.99   0.047    -.4668751   -.0031269
       _cons |  -.3089989   .0887406    -3.48   0.000    -.4829272   -.1350706
-------------+----------------------------------------------------------------
     /ln_the |   .0323916   .0729977     0.44   0.657    -.1106813    .1754645
-------------+----------------------------------------------------------------
       theta |   1.032922   .0754009                      .895224      1.1918
------------------------------------------------------------------------------
Likelihood-ratio test of theta=0: chibar2(01) =   637.76 Prob>=chibar2 = 0.000
```

а) Есть ли основания включать в модель ненаблюдаемую разнородность?

б) Выпишите функцию риска для случая $X_1 = 0, X_2 = 1$ при неизвестном значении α.

в) Рассчитайте псевдо-R^2.

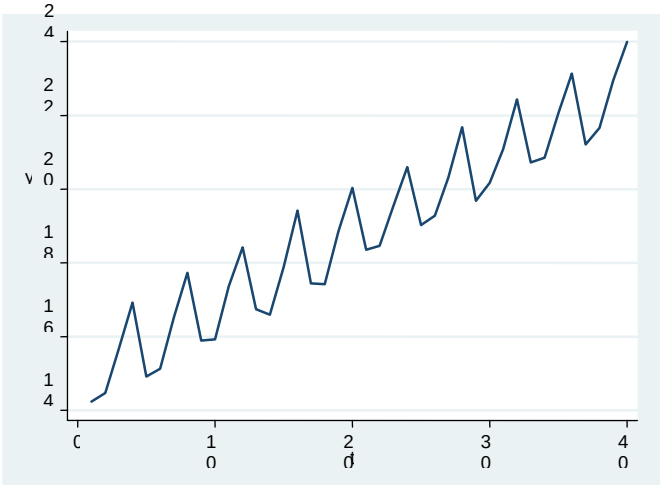
Часть 8. Стохастические регрессоры. Системы уравнений

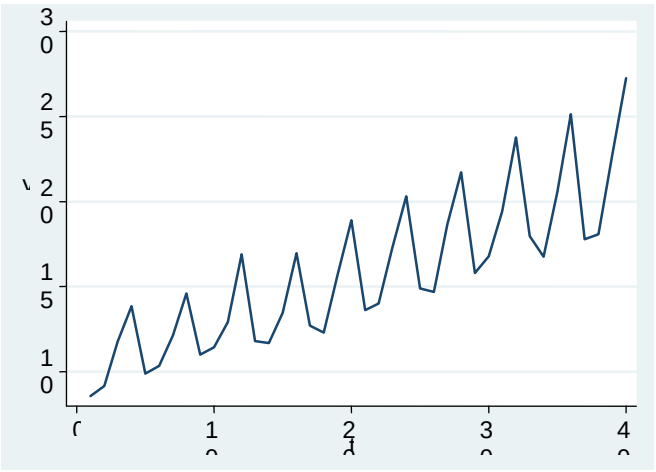
Часть 9. Анализ временных рядов

§9.1. Классическое разложение временного ряда. Сглаживание сезонности.

№9.1.1. Объясните, в чём различия между циклической и сезонной состоавляющей временного ряда?

№9.1.2. Посмотрите на графики ниже. В каком случае сезонная компонента входит в ряд аддитивно, а в каком — мультипликативно?





№9.1.3. В чём разница между двумя подходами к учёту мультипликативной сезонности:
1) сглаживание временного ряда скользящими средними и расчёт мультипликативных сезонных индексов;
2) логарифмирование ряда и выделение аддитивной сезонной компоненты?

№9.1.4. В чём разница между рядом скользящих средних и рядом с поправкой на сезонность?

№9.1.5. Статистик Тимофей размышляет так: для поправки на сезонность нужно знать, входит ли сезонная компонента аддитивно или мультипликативно в поправляемый ряд, а скользящие средние в обоих случаях лишены сезонных колебаний, поэтому для удаления сезонности лучше просто использовать ряд скользящих средних. Допустим ли такой подход, если цель сглаживания сезонности а) выявление исторических долгосрочных тенденций, б) получение оперативной информации об изучаемом показателе?

§9.2. Экспоненциальное сглаживание.

№9.2.1. Проведите простое экспоненциальное сглаживание ряда y_t с параметром сглаживания $\alpha = 0.2$. Рассчитайте прогноз для y_6.

t:	1	2	3	4	5
y_t:	5	5	2	4	4

№9.2.2. Проводится простое экспоненциальное сглаживание ряда y_t с параметром $\alpha = 0.5$. С каким весом входит первое наблюдение в сглаженное значение $\tilde{y}_4$, если $\tilde{y}_1 = y_1$?

№9.2.3. Верно ли, что сглаживание средним арифметическим $\bar{y}_t = \frac{1}{t}\sum_{i=1}^{t} y_i$ — частный случай экспоненциального сглаживания? Зависит ли ответ от выбора начального значения?

№9.2.4. Верно ли, что сглаживание средним арифметическим $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T} y_t$ по всему ряду из T наблюдений — частный случай экспоненциального сглаживания? Разберите случай $\tilde{y}_1 = y_1$. Зависит ли ответ от выбора начального значения?

§9.3. Стационарность случайных процессов. Модели ARMA.

№9.3.1. Пусть $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, $y_t = \varepsilon_t - \varepsilon_{t-1}$. Покажите, что процесс y_t стационарный и выпишите его автоковариационную и автокорреляционную функции.

№9.3.2. Пусть ε_t и υ_t — независимые белые шумы. Выясните, будут ли стационарными процессы: $x_t = \varepsilon_t + \upsilon_t$, $y_t = \varepsilon_t + \varepsilon_{t-1}$.

№9.3.3. Придумайте два стационарных процесса, сумма которых нестационарна.

№9.3.4. Придумайте два нестационарных процесса, сумма которых стационарна.

№9.3.5. Придумайте слабо стационарный процесс, который не был бы строго стационарным, и строго стационарный процесс, который не был бы слабо стационарным.

№9.3.6. Стационарен ли процесс авторегрессии $y_t = 0.5y_{t-1} + \varepsilon_t$? Покажите, что $\Delta y_t = y_t - y_{t-1}$ — процесс ARMA, и проверьте его стационарность.

№9.3.7. Найдите автоковариационные и частные автоковариационные функции процессов $y_t = \beta y_{t-1} + \varepsilon_t$ и $z_t = \varepsilon_t + \gamma\varepsilon_{t-1}$, где $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$.

№9.3.8. Пусть $y_t = 12.6 - 0.3y_{t-1} + 0.5y_{t-2} + \varepsilon_t$. Найдите *pacf*(3), *pacf*(2), *pacf*(1).

№9.3.9. Пусть $y_t = 3.3 + \varepsilon_t + 0.8\varepsilon_{t-1} - 0.2\varepsilon_{t-2}$. Выпишите автокорреляционную функцию.

№9.3.10. Проверьте стационарность процессов:
а) $y_t = 0.7y_{t-1} + \varepsilon_t$, б) $y_t = 2.3 + 1.4y_{t-1} - 0.4y_{t-2} + \varepsilon_t$, в) $y_t = 0.25y_{t-2} + \varepsilon_t$.

№9.3.11. Проверьте стационарность процессов:
а) $y_t = \varepsilon_t + 0.4\varepsilon_{e-1} - 0.6\varepsilon_{e-2}$, б) $y_t = 1.1y_{t-1} + \varepsilon_t - 0.8\varepsilon_{e-1}$,
в) $y_t = -0.8 + 0.7y_{t-1} - 0.1y_{t-2} + \varepsilon_t$, г) $y_t = 0.64y_{t-2} + \varepsilon_t + 2\varepsilon_{e-1}$.

№9.3.12. Пусть $y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \varepsilon_t$ — стационарный процесс. Покажите, что $|\beta_p| < 1$.

№9.3.13. Пусть $\varepsilon_t \sim WN(0, 1)$. Найдите математическое ожидание и дисперсию процесса y_t:
а) $y_t = 15 + \varepsilon_t - 2\varepsilon_{t-1}$, б) $y_t = 0.6y_{t-1} + \varepsilon_t$, в) $y_t = 10 + 0.5y_{t-1} + \varepsilon_t$.

№9.3.14. Имеется временной ряд y_t:

t:	1	2	3	4	5
y_t:	4	4	8	8	6

а) Проведите простое экспоненциальное сглаживание ряда с параметром сглаживания $\alpha = 0.5$, рассчитайте прогноз на два года вперёд.
б) Оценённая по этим данным модель авторегрессии имеет вид: $\hat{y}_t = 5.0 + 0.25 y_{t-1}$. Рассчитайте прогноз на два года вперёд по этой модели.

§9.4. Нестационарные случайные процессы. Модели ARIMA.

№.9.4.1. Покажите, что перечисленные процессы нестационарны, и проверьте, стационарны ли их разности:
а) $y_t = \alpha + \beta t + \varepsilon_t$, б) $y_t = y_{t-1} + \varepsilon_t$,
в) $y_t = \beta + y_{t-1} + \varepsilon_t$, г) $y_t = 2y_{t-1} + \varepsilon_t$.
Везде ε_t — белый шум.

№9.4.2. Придумайте такие процессы ε_t и υ_t, чтобы процесс y_t был стационарным в разностях (но не стационарным с трендом), а процесс z_t был стационарным с трендом:
$y_t = \alpha + \beta t + \varepsilon_t$, $z_t = \beta + z_{t-1} + \upsilon_t$.

№9.4.3. Придумайте такие процессы ε_t и υ_t, чтобы процессы x_t и y_t были стационарными:
$x_t = x_{t-1} + \varepsilon_t$, $y_t = \alpha + \beta t + \varepsilon_t$.

№9.4.4. Рассмотрим простое экспоненциальное сглаживание: $\tilde{y}_t = (1-\alpha)\tilde{y}_{t-1} + \alpha y_t$. Допустим, что $\varepsilon_t = y_t - \tilde{y}_t$ — белый шум. Покажите, что $y_t \sim ARIMA(0,1,1)$, если $0 < \alpha < 1$. Каким будет процесс y_t в случаях $\alpha = 0$ и $\alpha = 1$?

№9.4.5. По 30 наблюдениям оценена зависимость $\Delta y_t = \underset{(0.08)}{-0.2}\, y_{t-1} + e_t$. Есть ли основания считать, что процесс y_t стационарный? Используйте уровень значимости 5%.

№9.4.6. По 28 наблюдениям оценена зависимость $\Delta y_t = \underset{(0.18)}{0.62} - \underset{(0.06)}{0.17}\, y_{t-1} + e_t$. Согласуется ли этот результат с гипотезой, что y_t — случайное блуждание с дрейфом? Используйте уровень значимости 10%.

№9.4.7. По 50 наблюдениям оценена зависимость $y_t = -0.52 + 0.9 y_{t-1} + e_t$, $R^2 = 0.6$. Согласуется ли этот результат с гипотезой, что y_t — случайное блуждание с дрейфом? Используйте уровень значимости 5%.

№9.4.8. При проведении теста Дики-Фуллера основная гипотеза была отвергнута. Даёт ли этот результат основание считать, что изучаемый процесс стационарный? Что изучаемый процесс — случайное блуждание? Дайте ответ в двух случаях: а) проводился тест Дики-Фуллера без константы и тренда, б) проводился тест Дики-Фуллера с константой и без тренда.

№9.4.9. Статистик Тимофей ленится анализировать временные ряды. Если нужно делать прогноз, он всегда поступает так: строит график ряда (зависимость y_t от t), соединяет первую и последнюю точку по линейке и проводит линию дальше — значения, лежащие на линии, и будут прогнозом. Завтра Тимофею предстоит защищать результаты своей работы перед начальством. Помогите ему подготовиться: придумайте такие предположения о процессе y_t, при которых метод Тимофея будет разумно обоснованным.

№9.4.10. Надежда предлагает свою версию подхода, изложенного в предыдущей задаче: она соединяет две последние точки на графике (вместо первой и последней) и продолжает полученную линию в будущее. Придумайте модель прогнозируемого процесса, при которой такой подход будет разумным.

№9.4.11. По данным за 25 лет была оценена модель динамики индекса цен на чай по отношению к индексу потребительских цен: $\hat{P}_t = \underset{(0.16)}{0.37} + \underset{(0.18)}{0.62} P_{t-1}$.

а) Определите, каково среднее отношение индекса цен на чай к ИПЦ в долгосрочном периоде.
б) Есть ли основания считать, что отношение цены чая к ИПЦ не описывается процессом случайного блуждания с дрейфом? Используйте уровень значимости 5%

Эти милые ёлочки – напоминание о нормальном распределении каждому из прохожих. Такое фото могло бы украсить обложку книжки по теории вероятностей. Наша книжка не по теории вероятностей, и у неё нет обложки. Но ёлочки в ней всё равно есть.

Всего доброго!