ТЕОРИЯ ВЕРОЯТНОСТЕЙ и МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Часть III. Математическая статистика

Оценки параметров линейной регрессионной модели по методу наименьших квадратов

Построение модели

Контроль расчетов

Проверка адекватности построенной модели

Построение доверительных интервалов для параметров регрессии

ЛИТЕРАТУРА

1. Карасев В.А., Богданов С.Н., Лёвшина Г.Д. Теория вероятностей и математическая статистика: Разд. 2. Математическая статистика: Учеб.-метод. Пособие. – М. МИСиС, 2005. – 117 с. № 1855. [печ.]

2. Карасев В.А., Лёвшина Г.Д. Теория вероятностей и математическая статистика: математическая статистика: практикум». – М. Изд. Дом МИСиС, 2016. № 2770. [электрон.]

3. Данченков И.В., Карасев В.А. Математическая статистика: проверка гипотезы о виде закона распределения: практикум. – М. : Изд. Дом НИТУ «МИСиС», 2017. – 54 с. № 2976

3. Гмурман В.Е. Теория вероятностей и математическая статистика: учебное пособие для вузов. – М.: Изд-во Юрайт, 2015. Работаем по: М.: Высшее образование, 2006. – 479 с.

4. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учебное пособие для вузов. – М.: Изд-во Юрайт, 2015.

5. Кремер Н.Ш. Теория вероятностей и математическая статистика. – М.: ЮНИТИ-ДАНА, 2004. – 573 с.

6. Лебедев А.В., Фадеева Л.Н. Теория вероятностей и математическая статистика. – М., 2018. – 480 с. [электрон.], Фадеева Л. Н., Лебедев А.В. 2011 [печ.]

7. Севастьянов Б.А. Курс теории вероятностей и математической статистики. – М.: Наука, 1982. – 256 с.

8. Ефимов А.В., Поспелов А.С. и др. Сборник задач по математике для втузов. Специальные курсы (ТВ. МС. МО. УрЧП). – М., 1984. – 608 с. [печ.]; В 4 ч. – Ч. 4 (ТВ. МС). – М., 2003. – 432 с. [электрон., печ.]

9. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. – М., 1983. – 304 с.

10. Лагутин М.Б. Наглядная математическая статистика. – М., 2009. – 472 с.

11. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс: Учеб. – 6-е изд., перераб. и доп. –М.: Дело, 2004. – 576 с.

Регрессия

Понятия регрессии и корреляции непосредственно связаны между собой. В то время как в корреляционном анализе оценивается сила стохастической связи, в регрессионном анализе исследуется ее форма.

Э.Фёрстер, Б. Рёнц [9], с. 18

Парная (простая) линейная регрессия

Линейной регрессией называется сведение наблюдаемой на опыте зависимости некоторой переменной (**зависимой** или **объясняемой**) от одной или более других переменных (**независимых** или **объясняющих**) к линейной (в предположении, что строгая линейная зависимость между ними нарушается случайными ошибками). Для проведения линейной регрессии часто используется метод наименьших квадратов.

В простейшем случае речь идет о двух переменных. Пусть x — независимая переменная, y — зависимая, и между ними существует следующая связь:

$$y_i = a + bx_i + \varepsilon_i,$$

где a и b — числовые коэффициенты, ε_i — случайные ошибки, $M\varepsilon_i = 0$ и $D\varepsilon_i < \infty$. Задача состоит в том, чтобы по имеющимся наблюдениям (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) построить оценки для a и b.

Согласно методу наименьших квадратов необходимо решить следующую математическую задачу:

$$T = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min.$$

О термине «регрессия», работах Ф. Гальтона и Ч. Дарвина см. [9], с. 46.

У Гмурмана [3] (изд. 2006 г., с. 255-256, §4) все это звучит как

Отыскание параметров выборочного уравнения прямой линии среднеквадратичной регрессии по несгруппированным данным

Пусть изучается система количественных признаков (X,Y). В результате n независимых опытов получены n пар чисел (x_1, y_1), (x_2, y_2),…, (x_n, y_n).

Найдем по данным наблюдений выборочное уравнение прямой линии среднеквадратичной регрессии. Для определенности будем искать уравнение

$$\hat{y} = a + bx \qquad (*_0)$$

регрессии Y на X.

«По несгруппированным данным» означает, что значения x признака X и соответствующие им значения признака Y наблюдались по одному разу (поэтому и группировать данные нет необходимости). Угловой коэффициент прямой линии регрессии Y на X называют *выборочным коэффициентом регрессии* Y на X. Гмурман обозначает его через ρ_{yx}, у нас он будет фигурировать ниже как $\hat{b}$.

Итак, будем искать выборочное уравнение прямой линии регрессии Y на X вида $(*_0)$.

Подберем параметры a и b так, чтобы точки (x_1, y_1), (x_2, y_2),…, (x_n, y_n), построенные по данным наблюдений, на плоскости xOy лежали как можно ближе к прямой $(*_0)$. Уточним смысл этого требования. Назовем отклонением разность

$$\hat{y}_i - y_i \ (i = 1, 2,\ldots, n),$$

где $\hat{y}_i$ – вычисленная по уравнению (*$_0$) ордината, соответствующая наблюдаемому значению x_i; y_i – наблюдаемая координата, соответствующая x_i.

Подберем параметры a и b так, чтобы сумма квадратов отклонений была минимальной (в этом состоит сущность метода наименьших квадратов). Так как каждое отклонение зависит от отыскиваемых параметров, то и сумма квадратов отклонений есть функция T этих параметров:

$$T = \sum_{i=1}^{n} (y_i - a - bx_i)^2 ,$$

которую и будем минимизировать:

$$T = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min$$

Далее придерживаемся изложения Фёрстера – Рёнца [9].

Запишем необходимые условия экстремума:

$$\frac{\partial T}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0, \qquad \frac{\partial T}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0$$

или

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0, \qquad \sum_{i=1}^{n} (y_i - a - bx_i)x_i = 0 .$$

Раскроем скобки и получим (как это названо в [11], с. 35; см. также [9], с. 61) *стандартную форму нормальных уравнений* (для краткости уберем индексы суммирования у знака суммы Σ):

$$an + b\sum x_i = \sum y_i\,, \qquad a\sum x_i + b\sum x_i^2 = \sum x_i y_i\,. \qquad (*)$$

Для применения правила Крамера удобнее расположить уравнения (*) друг под другом:

$$\begin{cases} an + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases}$$

Тогда решения $a = \hat{a}$, $b = \hat{b}$ системы будут иметь вид ([9], с. 62; начнем с $\hat{b}$)

$$\hat{b} = \frac{\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \sum x_i \sum x_i} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{\mathrm{Cov}(x,y)_{\text{неиспр}}}{\mathrm{Var}(x)_{\text{неиспр}}},$$

$$\hat{a} = \frac{\begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}} = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - \sum x_i \sum x_i} = \frac{\overline{y}\,\overline{x^2} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2} \overset{\pm \overline{y}\overline{x}^2}{=} \overline{y} - \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}\overline{x} = \overline{y} - \hat{b}\overline{x}\,.$$

С использованием тождеств

$$\sum (y_i - \bar{y})(x_i - \bar{x}) = \sum x_i y_i - n\bar{x}\bar{y} = n(\overline{xy} - \bar{x}\bar{y})$$

и

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x} = n(\overline{x^2} - \bar{x}^2)$$

получаем следующие выражения для $\hat{b}$:

$$\hat{b} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\mathrm{Cov}(x,y)_{\text{неиспр}}}{\mathrm{Var}(x)_{\text{неиспр}}} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{s_{xy}}{s_x^2},$$

где

$\sigma_x^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2 = \mathrm{Var}(x)_{\text{неиспр}}$ – неисправленная выборочная дисперсия,

$\sigma_{xy} = \frac{1}{n}\sum (y_i - \bar{y})(x_i - \bar{x}) = \overline{xy} - \bar{x}\bar{y} = \mathrm{Cov}(x,y)_{\text{неиспр}}$ – неисправленная выборочная ковариация,

$s_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{n}{n-1}(\overline{x^2} - \bar{x}^2) = \frac{n}{n-1}\sigma_x^2$ – исправленная выборочная дисперсия,

$s_{xy} = \frac{1}{n-1}\sum (y_i - \bar{y})(x_i - \bar{x}) = \frac{n}{n-1}(\overline{xy} - \bar{x}\bar{y}) = \frac{n}{n-1}\sigma_{xy}$ – исправленная выборочная ковариация.

И это действительно минимум

Еще раз

$$T = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min$$

$$\left.\begin{aligned} T_a &= -2\sum (y_i - a - bx_i) \quad = 0 \\ T_b &= -2\sum (y_i - a - bx_i)x_i = 0 \end{aligned}\right\} \Rightarrow \begin{cases} a = \hat{a} \\ b = \hat{b} \end{cases}$$

Берем вторые производные:

$$A = T_{aa} = -2\sum(-1) = 2n > 0$$

$$C = T_{bb} = 2\sum x_i^2 = 2n\overline{x^2} \ > 0$$

$$B = \begin{cases} T_{ab} = 2\sum x_i = 2n\bar{x} \\ T_{ba} = 2\sum x_i = 2n\bar{x} \end{cases}$$

A, *B*, *C* и ниже *D* – это ностальгия по Б.П. Демидовичу (отд. VI, §7);

$$D = T_{aa}T_{bb} - (T_{ab})^2 = 4n^2(\overline{x^2} - \bar{x}^2) = 4n\sum (x_i - \bar{x})^2 > 0,$$

так как «значения *X* и соответствующие им значения *Y* наблюдались по одному разу» (Гмурман, с. 259), из чего следует, что существует *i*, такое что $x_i \neq \bar{x}$. Значит, согласно Демидовичу (*loc. cit.*),

$$D > 0,\ A > 0\ (C > 0)\ \Rightarrow\ T \text{ имеет минимум в } (a,b) = (\hat{a},\hat{b}).$$

Уравнение линейной регрессии

$$\hat{y} = \hat{a} + \hat{b}x =$$

$$= \bar{y} - \hat{b}\bar{x} + \hat{b}x = \bar{y} + \hat{b}(x - \bar{x})$$

Важный пример

На следующем примере из КБЛ [1] 2005 и КЛ [2] 2016 рассмотрим 1) (еще раз) построение линейной модели регрессии, 2) контроль расчетов, 3) проверку адекватности построенной модели и 4) построение доверительных интервалов для параметров регрессии.

(Постановку «по сгруппированным данным», множественную линейную регрессию, квадратичную регрессию, общую линейную регрессию постараемся здесь пока не рассматривать.)

Задача 1.25. Результаты экспериментов представлены в первых двух столбцах табл. 1.11. Экспериментальные значения Y являются независимыми и равноточными. Построить линейную и квадратичную регрессионные модели.

Таблица 1.11

Условие задачи 1.25 и результаты расчета

	x	Y	$X=\frac{x-0,6}{0,2}$	$X\cdot Y$	X^2	$Y_{лин}$	$\Delta Y_{лин}$	$\Delta Y^2_{лин}$
	0,2	4,5	–2	–9,0	4	5,3	–0,8	0,64
	0,4	7,0	–1	–7,0	1	6,25	0,75	0,56
	0,6	8,0	0	0,0	0	7,2	0,8	0,64
	0,8	7,5	1	7,5	1	8,15	–0,65	0,42
	1,0	9,0	2	18,0	4	9,1	–0,1	0,04
Σ	3,0	36,0	0	9,5	10		0	2,30

Эта постановка заимствована из [2]. А вот следующая – из [1]:

Задача 1.25 Результаты экспериментов представлены в первых двух столбцах табл. 1.11. Экспериментальные значения Y являются независимыми и равноточными. Построить линейную и квадратичную регрессионные модели.

Таблица 1.11

Исходные данные и результаты расчета (к задаче 1.25)

	x	Y	$X = \frac{x-0,6}{0,2}$	$X \cdot Y$	X^2	$Y_{лин}$	$\Delta Y_{лин}$	$X \cdot \Delta Y_{лин}$	$\Delta Y^2_{лин}$
	0,2	4,5	– 2	– 9,0	4	5,3	– 0,8	1,6	0,64
	0,4	7,0	– 1	– 7,0	1	6,25	0,75	– 0,75	0,56
	0,6	8,0	0	0,0	0	7,2	0,8	0,0	0,64
	0,8	7,5	1	7,5	1	8,15	– 0,65	– 0,65	0,42
	1,0	9,0	2	18,0	4	9,1	– 0,1	– 0,2	0,04
Σ	3,0	36,0	0	9,5	10	–	0	0	2,30

Отличие – в столбце $X \cdot \Delta Y_{лин}$ (кстати, почему?). Будем следовать в основном изложению **[2]**, так как оно было упрощено по сравнению с пособием **[1]**, которое, тем не менее, также будет участником дискуссии.

1. Задача регрессии для линейной функции (теория)

Рассмотрим случай, когда уравнение регрессии (1.61) является линейной функцией

$$y = \beta_1 + \beta_2 x, \tag{1.69}$$

Уравнение (1.61) – это

$$y = f(x,\beta_1,\beta_2,...,\beta_m) = \beta_1\varphi_1(x) + \beta_2\varphi_2(x) + ... + \beta_m\varphi_m(x). \tag{1.61}$$

Упр. Как это увязывается с исходной постановкой на с. 5–8 выше?

т.е. базисные функции $\varphi_1(x) = 1$, $\varphi_2(x) = x$. В этом случае система (1.63) имеет вид

$$\begin{cases} \beta_1 n + \beta_2 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} Y_i, \\ \beta_1 \sum_{i=1}^{n} x_i + \beta_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i. \end{cases} \tag{1.70}$$

Расчет упростится, если ввести замену $X = \dfrac{x - \overline{x}}{h}$ и рассматривать уравнение

$$y = B_1 + B_2 X = B_1 + B_2 \frac{x - \overline{x}}{h}, \tag{1.71}$$

где $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ – среднее арифметическое аргументов x; h выбирается из условия, чтобы значения X были целыми не имеющими общего множителя.

Уравнение (1.71) будем называть уравнением с *кодированным переменным*, в отличие от уравнения (1.69) с *реальным переменным*. В этом случае $\sum_{i=1}^{n} X_i = 0$ и система (1.70) будет иметь вид

$$\begin{cases} B_1 n = \sum_{i=1}^{n} Y_i, \\ B_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} Y_i x_i. \end{cases}$$

Откуда имеем формулы для оценок коэффициентов регрессии уравнения с кодированным переменным:

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n}, \quad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}. \qquad (1.72)$$

Упр. Что будет, если перейти к реальным переменным?

1а. Задача регрессии для линейной функции (решение задачи 1.25)

Сначала найдем решение задачи регрессии в кодированных значениях переменной x (1.71). Введем новую переменную по формуле $X = \frac{x - \overline{x}}{h}$, где $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5} \cdot 3 = 0,6$. Если значения величины $x - 0,6$ поделить на число $h = 0,2$, то получатся целые значения, не имеющие общего множителя. Поэтому $X = \frac{x - 0,6}{0,2}$.

По формулам (1.72) находим оценки коэффициентов линейной регрессии

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{36}{5} = 7,2; \quad \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} = \frac{9,5}{10} = 0,95.$$

Получили линейную модель регрессии

$$\hat{Y}_{\text{лин}} = 7,2 + 0,95X.$$

Уравнение линейной регрессии Y от реального переменного x найдем, сделав преобразование

$$Y_{\text{лин}} = \beta_1 + \beta_2 x = 7,2 + 0,95\frac{x - 0,6}{0,2} = 4,35 + 4,75x.$$

2. Контроль расчетов (теория)

Напомним:

$$\tilde{B}_1 = \frac{\sum_{i=1}^{n} Y_i}{n},\ \tilde{B}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}. \qquad (1.72)$$

Для контроля расчетов удобно воспользоваться свойством отклонений $\Delta Y_i = Y_i - \tilde{Y}(x_i)$ экспериментальных результатов Y_i от рассчитанных по оценкам (1.72) значений функции регрессии $\tilde{Y}(x_i) = \tilde{B}_1 + \tilde{B}_2 x_i$:

$$\sum_{i=1}^{n} \Delta Y_i = 0\,. \qquad (1.73)$$

Пособие **[1]** добавляет к (1.73) еще одно условие контроля:

$$\sum_{i=1}^{n} X_i \Delta Y_i = 0$$

Теперь понятно, почему **[1]**-вариант таблицы 1.11 содержит столбец $X \cdot \Delta Y_{\text{лин}}$.

2а. Контроль расчетов (решение задачи 1.25)

Вернемся к полученной выше (с. 13) модели линейной регрессии в кодированных переменных:

Получили линейную модель регрессии

$$\hat{Y}_{\text{лин}} = 7{,}2 + 0{,}95X.$$

По полученной формуле вычисляем значения линейной функции регрессии $\hat{Y}_{\text{лин}}$ при всех значениях аргумента X, а затем рассчитываем $\Delta Y_i = \hat{Y}_i - \hat{Y}_{i\,\text{лин}}$ отклонения экспериментальных значений Y_i от значений $\hat{Y}_{i\,\text{лин}}$, полученных по функции регрессии. Контроль, согласно формуле $\sum_{i=1}^{n} \Delta Y_i = 0$, выполнен. Все расчеты приведены в табл. 1.11.

Это же касается и равенства $\sum_{i=1}^{n} X_i \Delta Y_i = 0$.

3. Проверка адекватности линейной модели (теория)

Регрессионная модель называется *адекватной*, если предсказанные по ней значения переменной Y согласуются с результатами эксперимента.

Для проверки адекватности регрессионной модели вычисляют остаточную дисперсию (так называемую дисперсию адекватности) по формуле

$$S^2_{\text{ад}} = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{\text{ад}}}; \quad k_{\text{ад}} = n - m, \tag{1.66}$$

где ΔY_i – отклонения результатов эксперимента Y_i от проверяемой модели регрессии; $k_{\text{ад}}$ – число степеней свободы дисперсии адекватности; n – число точек, в которых проводился эксперимент; m – число оцениваемых параметров β_j в проверяемой модели.

Далее идет не с первого раза понимаемый текст:

Если истинная функция регрессии имеет тот же вид, что и рассматриваемая модель (например, так же, как и модель, представляет собой квадратичную функцию), то дисперсия адекватности служит несмещенной оценкой истинной дисперсии эксперимента и ее можно сравнивать с другими подобными оценками. В частности, может быть проведена независимая серия измерений для получения оценки дисперсии эксперимента $S^2_{\text{эксп}}$. В этом случае $S^2_{\text{эксп}}$ оценивает дисперсию эксперимента $D_{\text{эксп}}$, $S^2_{\text{ад}}$ характеризует степень отклонения экспериментальных точек от регрессионной модели, т.е. оценивает некую дисперсию адекватности $D_{\text{ад}}$.

Проверка адекватности модели заключается в проверке гипотезы H_0: $D_{ад} = D_{эксп}$ при альтернативной гипотезе H_1: $D_{ад} > D_{эксп}$ (если модель неадекватна, отклонения экспериментальных точек от модели будут больше погрешностей эксперимента). Таким образом, задача сводится к проверке гипотезы о равенстве дисперсий, которая решается с помощью критерия Фишера. Вычисляем отношение

$$F = S^2_{ад} / S^2_{эксп}. \qquad (1.67)$$

Если при заданном уровне значимости α отношение F окажется меньше квантили $F_{1-\alpha}(k_1, k_2)$, где $k_1 = k_{ад}$, $k_2 = k_{эксп}$, то рассматриваемая модель не противоречит результатам эксперимента и принимается; в противоположном случае модель отвергается с уровнем значимости α, как противоречащая результатам эксперимента.

Дисперсия адекватности (1.66) для проверки адекватности линейной регрессионной модели вычисляется по формуле

$$S^2_{ад} = \frac{\sum_{i=1}^{n} (\Delta Y_i)^2}{k_{ад}}; \quad k_{ад} = n - 2. \qquad (1.74)$$

3а. Проверка адекватности линейной модели (решение задачи 1.26)

Задача 1.26 – продолжение задачи 1.25, как раз и посвященное адекватности модели:

Задача 1.26. В задаче 1.25 рассчитаны линейная и квадратичная модели регрессии. По отдельной независимой серии измерений получена несмещенная оценка дисперсии $S^2 = 0{,}32$ с числом степеней свободы k, равным 20. Проверить адекватность линейной и квадратичной моделей регрессии с уровнем значимости $\alpha = 0{,}05$.

Решение

Для нахождения дисперсии адекватности (1.66) необходимо вычислить сумму квадратов отклонений результатов эксперимента от функции регрессии $\sum \Delta Y^2$ для каждой модели регрессии. Для линейной модели эта сумма равна 2,3 (см. последний столбец табл. 1.11). Число точек, в которых проводился эксперимент, $n = 5$; число оцениваемых параметров $m = 2$, тогда $k_{ад} = 5 - 2 = 3$. Дисперсия адекватности (1.74) равна $S^2_{ад\ лин} = 2{,}3/3 = 0{,}767$. Для проверки гипотезы об адекватности линейной модели вычисляем критерий Фишера (1.67):

$F = S^2_{ад} / S^2_{эксп} = 0{,}767/0{,}32 = 2{,}40$. Квантиль распределения Фишера $F_{0{,}95}(3; 20) = 3{,}10$. Так как $2{,}4 < 3{,}1$, то гипотеза об адекватности линейной модели принимается с уровнем значимости $\alpha = 0{,}05$.

Отметим, что адекватность квадратичной модели проверяется аналогично (соответствующие выкладки приводятся как в **[1]**, так и в **[2]**). Но:

Квадратичную модель в данной задаче строить нецелесообразно. Так как линейная модель оказалась адекватной, ее уточнять не было необходимости.

Изложение было бы неполным без следующей таблицы, где пришлось поискать квантиль 3.10:

Квантили распределения Фишера, $F_p(k_1, k_2)$, $p = 0{,}95$

k_2 \ k_1	1	2	3	4	5	6	7	8	9	10	12	15	20
1	161,4	199,5	199,5	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0
2	18,51	19,00	19,0	19,25	19,30	19,33	19,35	19,37	19,37	19,40	19,41	19,43	19,45
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,85	8,79	8,74	8,70	8,66
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,04	5,96	5,91	5,86	5,80
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,82	4,74	4,68	4,62	4,56
6	5,99	5,14	4,75	4,53	4,39	4,28	4,21	4,15	4,15	4,06	4,00	3,94	3,87
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,73	3,64	3,57	3,51	3,44
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,44	3,35	3,28	3,22	3,15
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,23	3,14	3,07	3,01	2,94
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,07	2,98	2,91	2,85	2,77
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,95	2,85	2,79	2,72	2,65
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,85	2,75	2,69	2,62	2,54
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,77	2,67	2,60	2,53	2,46
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,70	2,60	2,53	2,46	2,39
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,64	2,54	2,43	2,40	2,33
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,59	2,49	2,42	2,35	2,28
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,55	2,45	2,38	2,31	2,23
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,51	2,41	2,34	2,27	2,19
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,48	2,38	2,31	2,23	2,16
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,45	2,35	2,28	2,20	2,12
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,42	2,32	2,25	2,18	2,10
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,40	2,30	2,23	2,15	2,07
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,37	2,27	2,20	2,13	2,05
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,36	2,25	2,18	2,11	2,03

4. Построение доверительных интервалов для параметров регрессии (теория)

Если результаты экспериментов независимы и подчиняются нормальному закону распределения с дисперсией σ^2, то доверительный интервал с доверительной вероятностью $P=1-\alpha$ для каждого параметра $\tilde{\beta}_j$ можно определить неравенством

$$\left|\beta_j-\tilde{\beta}_j\right|<\varepsilon_j \text{ или } \beta_j=\tilde{\beta}_j\pm\varepsilon_j,$$

где
$$\varepsilon_j=t_{1-\alpha/2}(k)S\sqrt{a_{jj}}\ \ (j=1,2,...,m), \qquad (1.68)$$

здесь S^2 – несмещенная оценка дисперсии σ^2 с числом степеней свободы k; $t_{1-\alpha/2}(k)$ – квантиль распределения Стьюдента; a_{jj} – диагональный элемент матрицы A^{-1}, $\alpha=1-P$, P – доверительная вероятность.

Как было указано выше, если рассматриваемая модель регрессии адекватна, то дисперсия адекватности служит несмещенной оценкой истинной дисперсии эксперимента. Следовательно, в случае адекватности модели в формулу (1.68) в качестве оценки среднего квадратического отклонения S можно подставить корень из дисперсии адекватности $S_{ад}=\sqrt{S_{ад}^2}$.

В построенной регрессионной модели (1.61) некоторые коэффициенты могут быть незначимы, т.е. может выполняться гипотеза H_0: $\beta_j = 0$. Для проверки этой гипотезы можно найти доверительный интервал для коэффициента β_j с уровнем значимости α. Если этот интервал «накрывает» значение $\beta_j = 0$, гипотеза H_0 принимается и коэффициент β_j признается незначимым, в противоположном случае коэффициент β_j значим.

Границы доверительных интервалов для параметров линейной функции регрессии с кодированным переменным (1.71) имеют вид

$$\tilde{B}_1 \pm \varepsilon_1; \quad \varepsilon_1 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{n}}; \quad \tilde{B}_2 \pm \varepsilon_2; \quad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2}}. \quad (1.75)$$

4а. Построение доверительных интервалов для параметров регрессии (решение задачи 1.27)

Задача 1.27 – продолжение задачи 1.25, как раз и посвященное построению доверительных интервалов для параметров регрессии (тем более, что, как выяснится ниже, переход от линейной к квадратичной регрессии необоснован и следовало бы остановиться на линейной регрессии):

Задача 1.27. В задаче 1.25 получено уравнение линейной и квадратичной модели регрессии. По отдельной независимой серии измерений получена несмещенная оценка дисперсии $S^2 = 0{,}32$ с числом степеней свободы k равным 20. Считая результаты экспериментов независимыми, равноточными и подчиняющимися нормальному закону распределения, построить доверительные интервалы для истинных значений коэффициентов обеих моделей с надежностью $P = 0{,}95$.

Решение

В задаче 1.25 получена линейная модель регрессии от кодированной переменной

$$y = \bar{B}_1 + \bar{B}_2 X = 7{,}2 + 0{,}95X.$$

Границы доверительных интервалов (1.75) для коэффициентов B_1 и B_2 при этом будут:

$$B_1 = 7{,}2 \pm \varepsilon_1;$$

$$\varepsilon_1 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{n}} = t_{0{,}975}(20)\frac{\sqrt{0{,}32}}{\sqrt{5}} = 2{,}086\frac{\sqrt{0{,}32}}{\sqrt{5}} = 0{,}52;$$

$$B_2 = 0{,}95 \pm \varepsilon_2; \qquad \varepsilon_2 = t_{1-\frac{\alpha}{2}}(k)\frac{S}{\sqrt{\sum_{i=1}^{n} X_i^2}} = 2{,}086\frac{\sqrt{0{,}32}}{\sqrt{10}} = 0{,}38.$$

Уравнение линейной регрессии от реального переменного x

$$y = \beta_1 + \beta_2 x = 4,35 + 4,75x.$$

Границы доверительных интервалов для коэффициентов β_1 и β_2 находим по формулам (1.77):

$$\beta_1 = 4,35 \pm \hat{\varepsilon}_1;$$

$$\hat{\varepsilon}_1 = t_{1-\frac{\alpha}{2}}(k)\, S \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{h^2 \sum_{i=1}^{n} X_i^2}} = 2,086\sqrt{0,32} \cdot \sqrt{\frac{1}{5} + \frac{0,36}{0,04 \cdot 10}} = 1,24;$$

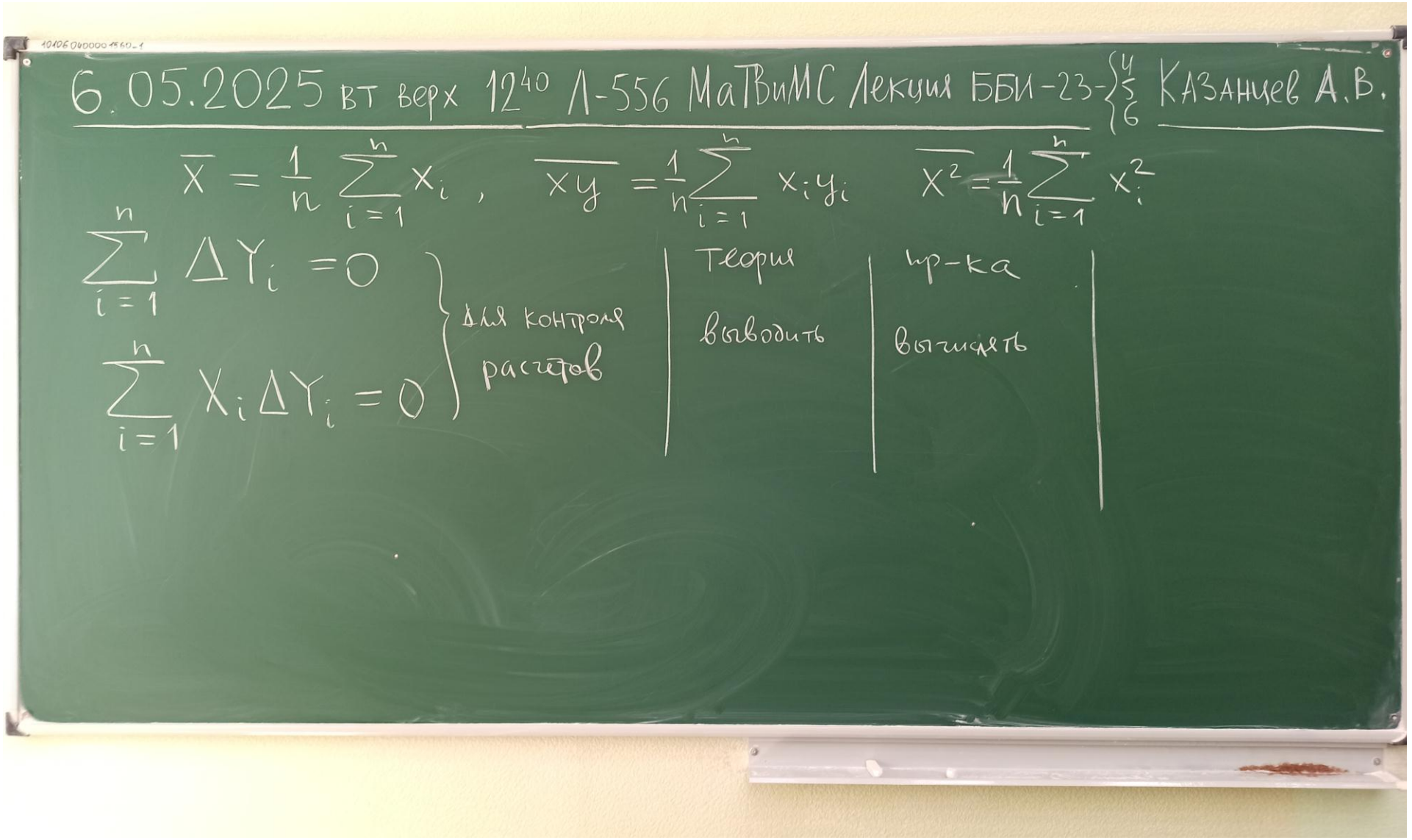
$$\beta_2 = 4,75 \pm \hat{\varepsilon}_2; \qquad \hat{\varepsilon}_2 = t_{1-\frac{\alpha}{2}}(k) \frac{S}{h\sqrt{\sum_{i=1}^{n} X_i^2}} = 2,086 \frac{\sqrt{0,32}}{0,2\sqrt{10}} = 1,9.$$

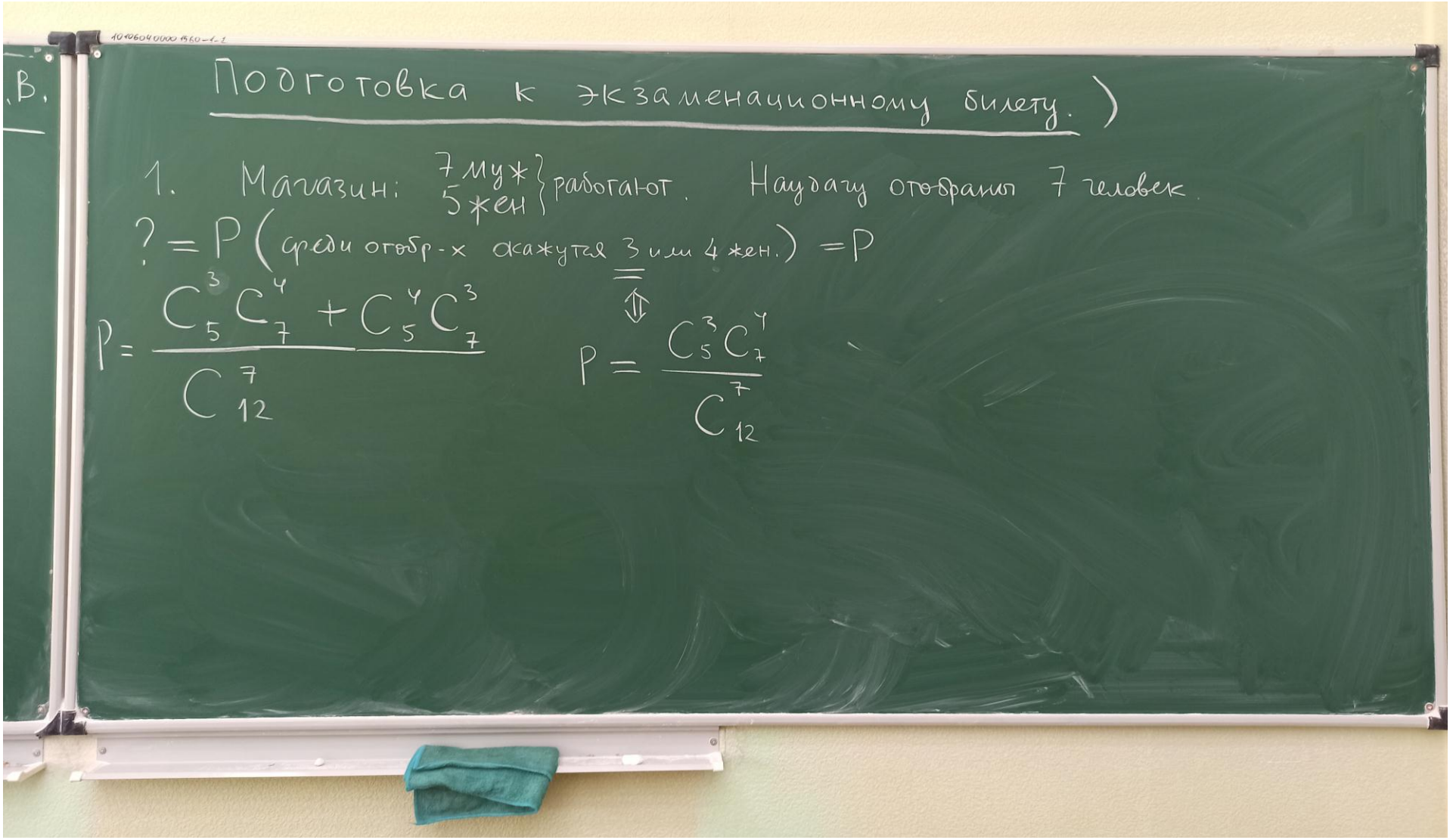
Все то же самое разобрано для квадратичной модели. Пропуская соответствующие выкладки (они есть и в **[1]**, и в **[2]**), приведем завершающий вывод:

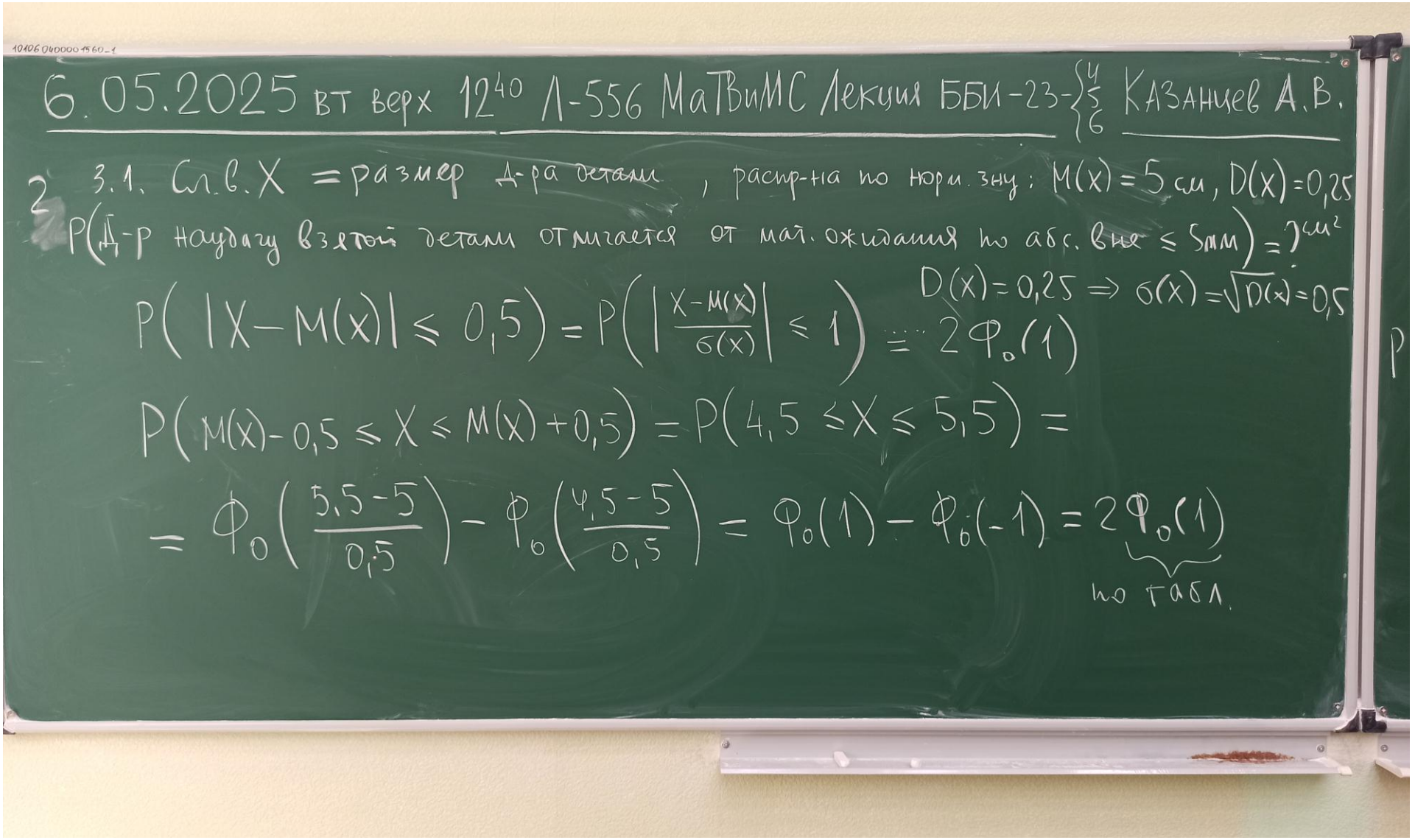
Полуширина доверительного интервала для коэффициента B_3 оказалась больше абсолютной величины этого коэффициента, т.е. доверительный интервал для этого коэффициента накрывает значение $B_3 = 0$. В этом случае говорят, что коэффициент $\tilde{B}_3$ незначим, т.е. переход от линейной модели к квадратичной необоснован и следовало остановиться на линейной модели. Аналогичный вывод был получен при проверке адекватности полученных моделей регрессии.

.....

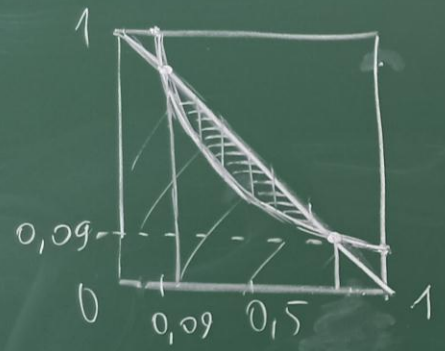
ФОТО НА ЛЕКЦИИ




Подготовка к экзаменационному билету.)
1. Магазин: 7 муж } 5 жен } работают. Наудачу отобраны 7 человек.
? = P (среди отобр-х окажутся 3 или 4 жен.) = P
$P = \frac{C_5^3 C_7^4 + C_5^4 C_7^3}{C_{12}^7}$
$P = \frac{C_5^3 C_7^4}{C_{12}^7}$


6.05.2025 вт верх 12^40 Л-556 МаТВиМС Лекция ББИ-23-{4 5 6 Казанцев А.В.
2 3.1. Сл.в. X = размер Д-ра детали, распр-на по норм. зну: M(X)=5 см, D(X)=0,25
P(Д-р наудачу взятой детали отличается от мат. ожидания по абс. вне ≤ 5мм) = ? см²
D(X)=0,25 ⇒ σ(X)=√D(X)=0,5
P(|X−M(X)| ≤ 0,5) = P(|(X−M(X))/σ(X)| ≤ 1) ≈ 2Φ0(1)
P(M(X)−0,5 ≤ X ≤ M(X)+0,5) = P(4,5 ≤ X ≤ 5,5) =
= Φ0((5,5−5)/0,5) − Φ0((4,5−5)/0,5) = Φ0(1) − Φ0(−1) = 2Φ0(1)
по табл.

Подготовка к экзаменационному билету.)



$y = 1 - x$

$y = \frac{0,09}{x}$

$$P = \int_{0,1}^{0,9} \left(1 - x - \frac{0,09}{x}\right) dx$$

$x(1-x) = 0,09$

$x_1 = 0,1$

$x_2 = 0,9$

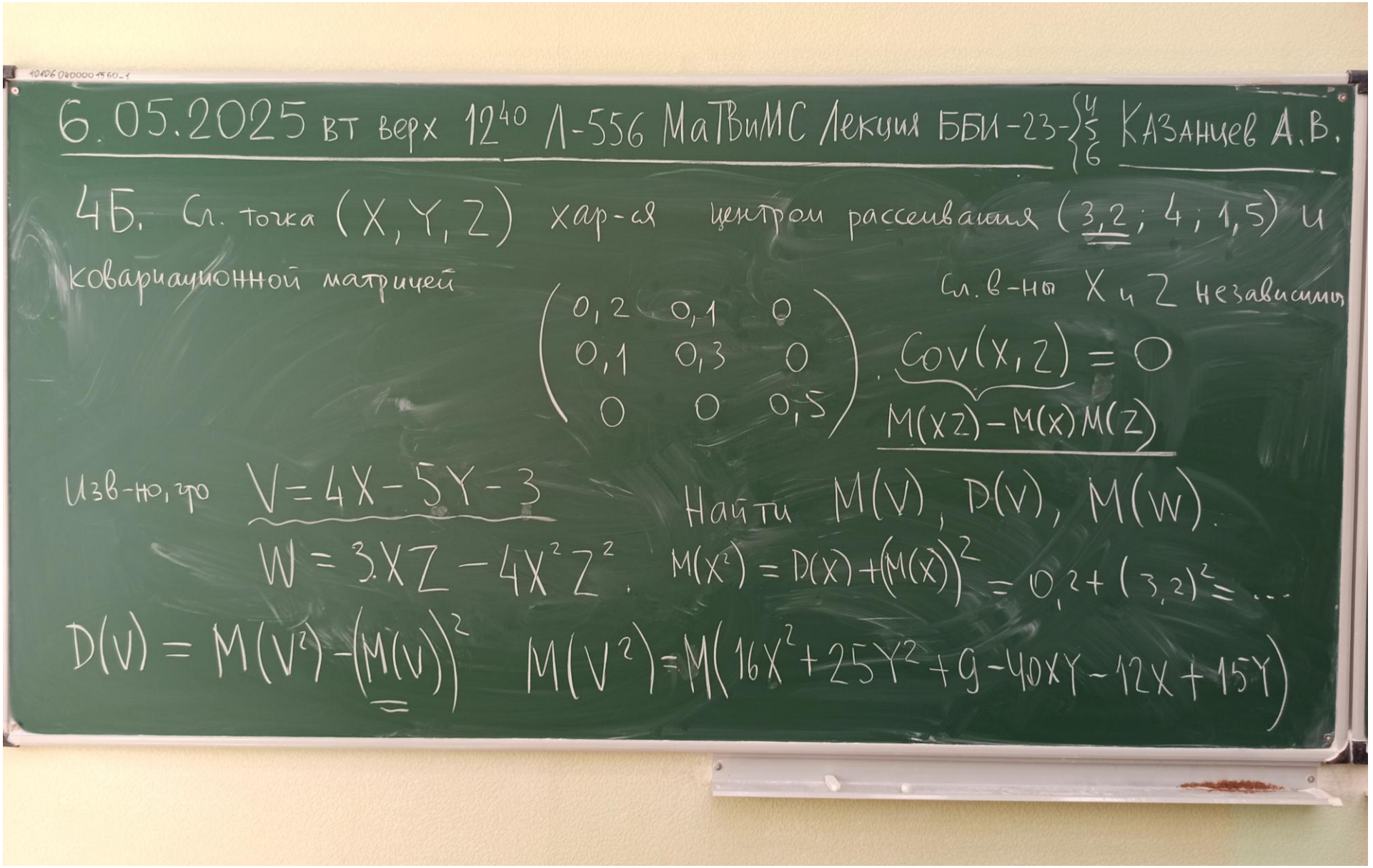
$x^2 - x + 0,09 = 0$

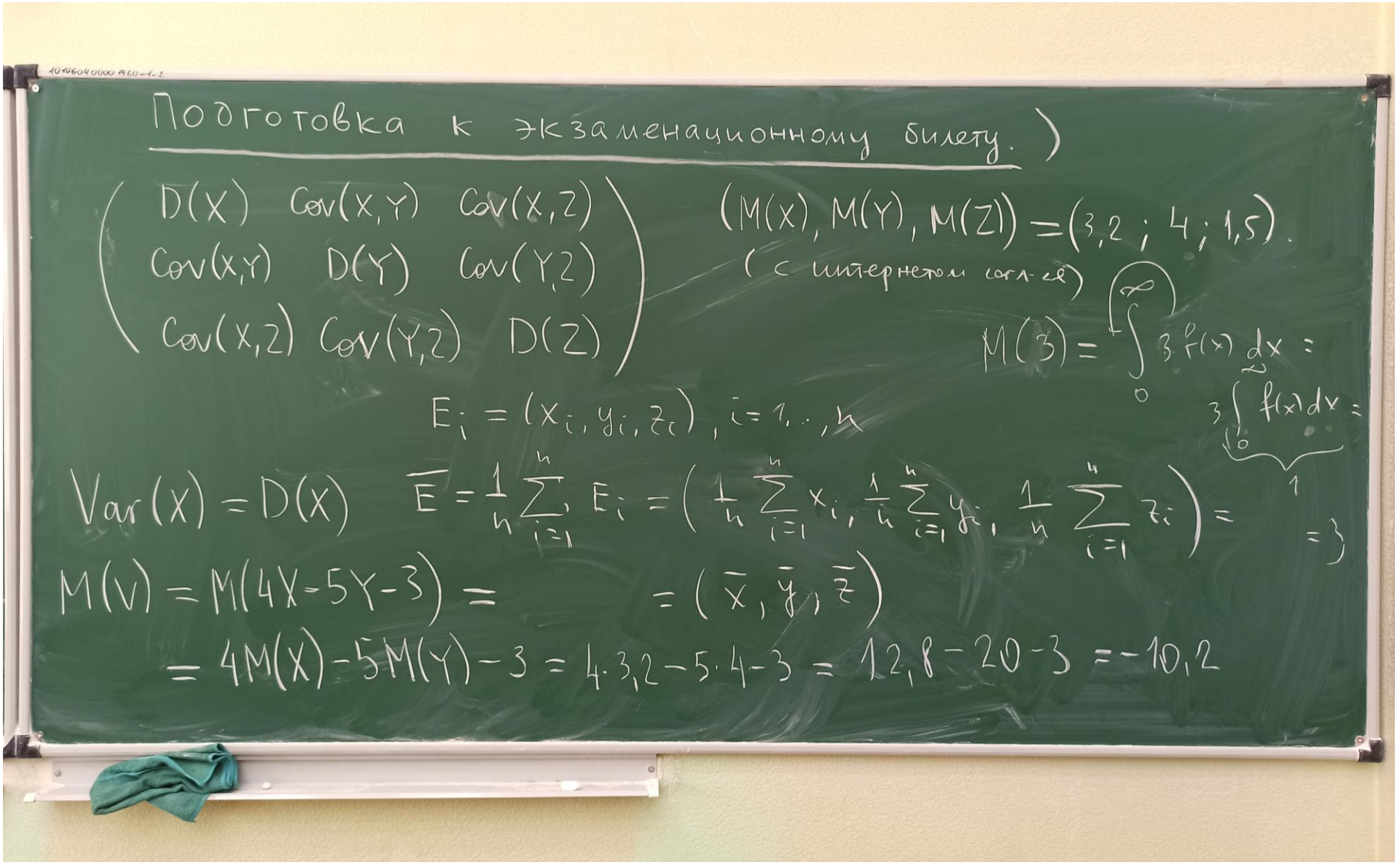
$D = 1 - 4 \cdot 0,09 = 1 - 0,36 = 0,64 = 0,8^2$

$xy \geq 0,09$

$y = \frac{0,09}{x}$

$y(0,5) = \frac{0,09}{0,5} = \frac{9}{50} = 0,18$


6.05.2025 вт верх 12^40 Л-556 МаТВиМС Лекция ББИ-23-{4 5 6 Казанцев А.В.
4Б. Сл. точка (X, Y, Z) хар-ся центром рассеивания (3,2; 4; 1,5) и ковариационной матрицей
0,2 0,1 0
0,1 0,3 0
0 0 0,5
Сл. в-ны X и Z независимы
Cov(X,Z) = 0
M(XZ) − M(X)M(Z)
Изв-но, что V = 4X − 5Y − 3
W = 3XZ − 4X²Z².
Найти M(V), D(V), M(W).
M(X²) = D(X) + (M(X))² = 0,2 + (3,2)² = ...
D(V) = M(V²) − (M(V))²
M(V²) = M(16X² + 25Y² + 9 − 40XY − 12X + 15Y)


Подготовка к экзаменационному билету.)
$\begin{pmatrix} D(X) & Cov(X,Y) & Cov(X,Z) \\ Cov(X,Y) & D(Y) & Cov(Y,Z) \\ Cov(X,Z) & Cov(Y,Z) & D(Z) \end{pmatrix}$
$(M(X), M(Y), M(Z)) = (3,2 ; 4 ; 1,5)$.
(с интернетом согл-ся)
$M(3) = \int_0^{+\infty} 3 \cdot f(x)\,dx =$
$E_i = (x_i, y_i, z_i),\ i = 1, \dots, n$
$3 \int_0^{!} f(x)\,dx = 1$
$Var(X) = D(X)$
$\overline{E} = \frac{1}{n}\sum_{i=1}^{n} E_i = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{n}\sum_{i=1}^{n} z_i\right) = =)$
$M(V) = M(4X - 5Y - 3) = = (\bar{x}, \bar{y}, \bar{z})$
$= 4M(X) - 5M(Y) - 3 = 4 \cdot 3,2 - 5 \cdot 4 - 3 = 12,8 - 20 - 3 = -10,2$