МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
Национальный исследовательский ядерный университет «МИФИ»

Институт интеллектуальных кибернетических систем

КАФЕДРА КИБЕРНЕТИКИ

БДЗ

по курсу "Математическая статистика"

студента группы Б22-514 Юдиной Дианы Сергеевны

Вариант № 13

Оценка: _____

Подпись: _____

2024 г.

ОТЧЕТ № 1

по теме «Проверка статистических гипотез»

Вариант №_13_

ФИО студента Юдина Диана Сергеевна группа Б22-514

Оценка: _____ Подпись:_____

Результаты статистических тестов:

№ задания	Проверяемая гипотеза H_0	Критерий	Статистическое решение ($\alpha = 0.1$)	Вывод
4.1	$F(x) \sim N$	Хи-квадрат	Отклоняем H0	Распределение не является нормальным
4.2	$F(x) \sim N$	Харке-Бера	Отклоняем H0	Распределение не является нормальным
5.1	$F_1(x) = F_2(x)$	знаков	Отклоняем H0	Выборки не являются однородными
5.2	$F_1(x) = F_2(x)$	Хи-квадрат	Отклоняем H0	Выборки не являются однородными

Выводы:

В результате проведённого в п.4 статистического анализа обнаружено, что средняя заработная плата всех должностей не является нормально распределенной случайной величиной.
В результате проведённого в п.5 статистического анализа обнаружено, что выборки средних заработных плат всех профессоров и средних заработных плат всех должностей не являются однородными.

ОТЧЕТ № 2

по теме «Анализ статистических взаимосвязей»

Вариант №____13____

ФИО студента Юдина Диана Сергеевна группа Б22-514

Оценка: _____ Подпись:_____

Результаты статистических тестов:

№ задания	Проверяемая гипотеза H_0	Критерий	Статистическое решение ($\alpha = 0.1$)	Вывод
6	$F_Y(y \text{ при } X = No) = F_Y(y \text{ при } X = Yes)$	Хи-квадрат	Отклоняем H0	Присутствует стат. связь
7	$m_1 = \cdots = m_k$	ANOVA	Отклоняем H0	Присутствует стат. связь

Выводы:

В результате проведённого в п.6 статистического анализа обнаружено, что между «Средняя зарплата для всех категорий > средняя зарплата по колледжу» и «Средняя компенсация для всех категорий > средняя компенсация по колледжу» есть статистическая связь.

В результате проведённого в п.7 статистического анализа обнаружено, что средняя заработная плата всех должностей зависит от типа учебного заведения.

ОТЧЕТ № 3

по теме «Основы регрессионного анализа»

Вариант №___13_____

ФИО студента Юдина Диана Сергеевна группа Б22-514

Оценка: _____ Подпись:_____

Сводная таблица свойств различных регрессионных моделей:

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	32,7%	41,7%	91,5%
Значимость	да	да	да
Адекватность	неадекватная	неадекватная	адекватная
Степень тесноты связи	заметная	заметная	сильная

Выводы:

В результате проведённого в п.8 статистического анализа обнаружено, что между средней заработной платой всех должностей и средней заработной платой всех профессоров есть зависимость, также присутствует статистическая связь между средними заработными платами всех профессоров, всех должностей и всех доцентов.
В результате проведённого в п.9 статистического анализа обнаружено, что количество профессоров заметно влияет на среднюю компенсацию для всех должностей. Однако количество профессоров в учебном заведении и средняя заработная плата всех профессоров сильно влияет на среднюю компенсацию для всех должностей.

1. Описательные статистики

1.1. Выборочные характеристики

Анализируемый признак 1 – A5

Анализируемый признак 2 – A6

Анализируемый признак 3 – A8

а) Привести формулы расчёта выборочных характеристик

Выборочная хар-ка	Формула расчета
Объём выборки	$n = \sum_{i=1}^{k} n_i$
Среднее	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Выборочная дисперсия	$d_X^* = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$
Выборочное среднеквадратическое отклонение	$\sigma_X^* = \sqrt{d_X^*}$
Выборочный коэффициент асимметрии	$\gamma_X^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^2}\right)^3}$
Выборочный эксцесс	$\varepsilon_X^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x-\overline{x})^2\right)^2}$

б) Рассчитать выборочные характеристики

Выборочная хар-ка	Признак 1	Признак 2	Признак 3
Среднее	526.48	420.04	428.03
Выборочная дисперсия	13868.86	4957.76	8217.62
Выборочное среднеквадратическое отклонение	117.77	70.41	90.65
Выборочный коэффициент асимметрии	0.679	0.348	0.819
Выборочный эксцесс	0.530	0.176	0.976

1.2. Группировка и гистограммы частот

Анализируемый признак –A8

Объём выборки –1073

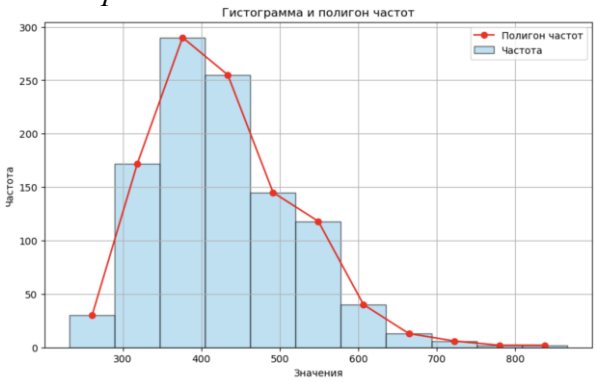
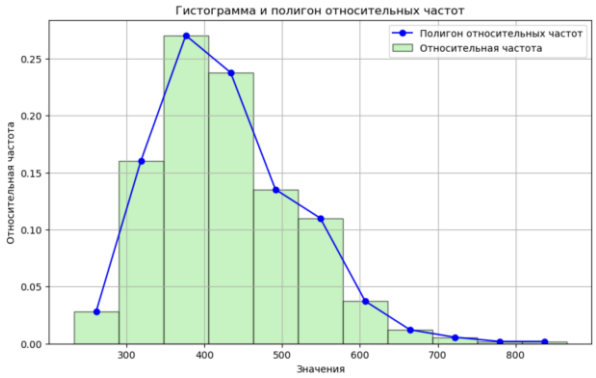
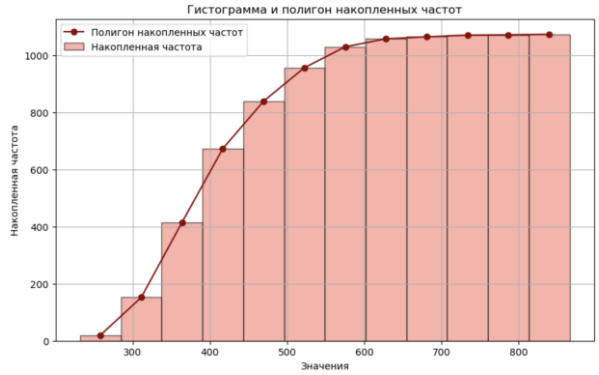
а) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
11	Формула Стерджесса $k \approx [1 + \log_2 \quad n]$	57.64

б) Построить таблицу частот

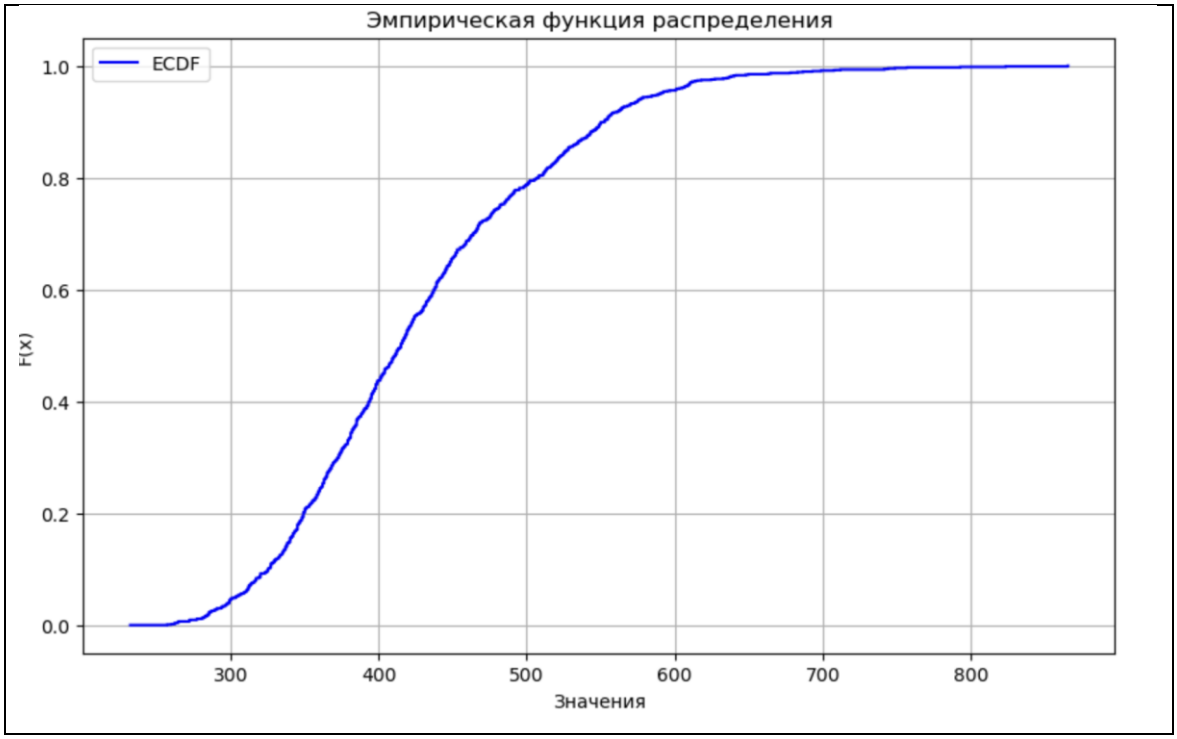
Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Накопл. частота	Относит. накопл. частота
1	232.00	289.63	30	0.027	30	0.027
2	289.63	347.27	172	0.16	202	0.1883
3	347.27	404.91	290	0.27	492	0.4585
4	404.91	462.55	255	0.24	747	0.6963
5	462.55	520.18	145	0.14	892	0.8313
6	520.18	577.82	118	0.11	1010	0.9413
7	577.82	635.45	40	0.037	1050	0.9786
8	635.45	693.09	13	0.012	1063	0.9907
9	693.09	750.73	6	0.0056	1069	0.9963
10	750.73	808.36	2	0.0019	1071	0.9981
11	808.36	866.00	2	0.0018	1073	1.00

в) Построить гистограммы частот и полигоны частот

Гистограмма и полигон частот 	*Гистограмма и полигон относительных частот* 
Гистограмма и полигон накопленных частот 	*Гистограмма и полигон накопленных относительных частот* 

г) Построить график эмпирической функции распределения

Эмпирическая функция распределения


Эмпирическая функция распределения
ECDF
1.0
0.8
0.6
0.4
0.2
0.0
F(x)
300
400
500
600
700
800
Значения

2. Интервальные оценки

2.1. Доверительные интервалы для мат. ожидания

Анализируемый признак – A8

Объём выборки – 1073

Оцениваемый параметр – m

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\overline{X} - \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$
Верхняя граница	$\overline{X} + \frac{S}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1)$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	420.90425	422.60860	423.48064
Верхняя граница	435.16099	433.45664	432.58460

2.2. Доверительные интервалы для дисперсии

Анализируемый признак –A8

Объём выборки – 1073

Оцениваемый параметр – σ^2

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}$
Верхняя граница	$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	7371.6480	7564.0110	7665.0364
Верхняя граница	9210.0315	8960.2160	8835.8714

2.3. Доверительные интервалы для разности мат. ожиданий

Анализируемый признак 1 – A5

Анализируемый признак 2 –A8

Объёмы выборок –1073

Оцениваемый параметр – ($m_1 - m_2$)

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$(\overline{x_1} - \overline{x_2}) - t_{\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Верхняя граница	$(\overline{x_1} - \overline{x_2}) + t_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	86.7647	89.5588	90.9885
Верхняя граница	110.1375	107.3433	105.9137

2.4. Доверительные интервалы для отношения дисперсий

Анализируемый признак 1 – A5

Анализируемый признак 2 –A8

Объёмы выборок –1073

Оцениваемый параметр – $\frac{\sigma_1^2}{\sigma_2^2}$

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{S_1^2}{S_2^2} \cdot f_{\frac{\alpha}{2}}(n_2 - 1,\, n_1 - 1)$
Верхняя граница	$\frac{S_1^2}{S_2^2} \cdot f_{1-\frac{\alpha}{2}}(n_2 - 1,\, n_1 - 1)$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	1.4418	1.4972	1.5263
Верхняя граница	1.9755	1.9025	1.8662

3. Проверка статистических гипотез о математических ожиданиях и дисперсиях

3.1. Проверка статистических гипотез о математических ожиданиях

Анализируемый признак – A8

Объём выборки –1073

Статистическая гипотеза – $\begin{matrix} H_0 : m = m_0 \\ H' : m \neq m_0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z = \frac{\overline{x} - m_0}{\frac{s}{\sqrt{n}}}$
Закон распределения статистики критерия при условии истинности основной гипотезы	T(n – 1)
Формулы расчета критических точек	$+t_{1-\frac{\alpha}{2}}(n-1)$ $-t_{1-\frac{\alpha}{2}}(n-1)$
Формула расчета *p-value*	2*min($F_Z(z), 1 - F_Z(z)$) (в условиях истинности H0)

б) Выбрать произвольные значения m_0 и проверить статистические гипотезы

m_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
500	0.1	-26.005	0.0	Отклоняем H0	Отсутствует ошибка принятия стат. решения $m \neq 500$
428	0.1	0.012	0.99	Принимаем H0	Отсутствует ошибка принятия стат. решения $m = 428$
400	0.1	10.13	0.0	Отклоняем H0	Отсутствует ошибка принятия стат. решения $m \neq 400$

3.2. Проверка статистических гипотез о дисперсиях

Анализируемый признак – A8

Объём выборки –1073

Статистическая гипотеза – $\begin{matrix} H_0 : \sigma = \sigma_0 \\ H' : \sigma \neq \sigma_0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z=\frac{(n-1)S^2}{\sigma_0^2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(n-1)$
Формулы расчета критических точек	$\chi^2_{\frac{\alpha}{2}}(n-1), \chi^2_{1-\frac{\alpha}{2}}(n-1)$
Формула расчета *p-value*	2*min($F_Z(z), 1-F_Z(z)$) (в условиях истинности H0)

б) Выбрать произвольные значения σ_0 и проверить статистические гипотезы

σ_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
100	0.1	880.93	0.99	Принимаем H0	Ошибка 2 рода
90	0.1	1087.57	0.364	Принимаем H0	Отсутствует ошибка принятия стат. решения $\sigma = 90$
70	0.1	1797.81	0.0	Отклоняем H0	Отсутствует ошибка принятия стат. решения $\sigma \neq 70$

3.3. Проверка статистических гипотез о равенстве математических ожиданий

Анализируемый признак 1 – A5

Анализируемый признак 2 – A8

Объёмы выборок –1073

Статистическая гипотеза – $\begin{matrix} H_0 : m_1 = m_2 \\ H' : m_1 \neq m_2 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Закон распределения статистики критерия при условии истинности основной гипотезы	T(2012), число степеней свободы 1/k, k=

	$\frac{\left(\frac{\frac{s_1^2}{n1}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)^2}{n_1-1}+\frac{\left(\frac{\frac{s_2^2}{n_2}}{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right)^2}{n_2-1}$
Формулы расчета критических точек	$+t_{1-\frac{\alpha}{2}}(2012), -t_{1-\frac{\alpha}{2}}(2012)$
Формула расчета *p-value*	2*min($F_Z(z), 1-F_Z(z)$) (в условиях истинности H0)

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	21.6999	0.0	Отклоняем H0	Отсутствует ошибка принятия стат. решения $m_1 \neq m_2$
0.05			Отклоняем H0	Отсутствует ошибка принятия стат. решения $m_1 \neq m_2$
0.1			Отклоняем H0	Отсутствует ошибка принятия стат. решения $m_1 \neq m_2$

3.4. Проверка статистических гипотез о равенстве дисперсий

Анализируемый признак 1 – A5

Анализируемый признак 2 – A8

Объёмы выборок – 1073

Статистическая гипотеза – $\begin{matrix} H_0 : \sigma_1 = \sigma_2 \\ H' : \sigma_1 \neq \sigma_2 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z = \frac{S_1^2}{S_2^2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(n_1-1,\, n_2-1)$
Формулы расчета критических точек	$f_{\frac{\alpha}{2}}(n_1-1,\, n_2-1), f_{1-\frac{\alpha}{2}}(n_1-1,\, n_2-1)$
Формула расчета *p-value*	2*min($F_Z(z), 1-F_Z(z)$) (в условиях истинности H0)

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	1.688	0.0	Отклоняем H0	Отсутствует ошибка принятия стат. решения $\sigma_1 \neq \sigma_2$
0.05			Отклоняем H0	Отсутствует ошибка принятия стат. решения $\sigma_1 \neq \sigma_2$
0.1			Отклоняем H0	Отсутствует ошибка принятия стат. решения $\sigma_1 \neq \sigma_2$

4. Критерии согласия

Анализируемый признак – A8

Объём выборки – 1073

4.1. Критерий хи-квадрат

Теоретическое распределение – нормальное.

Статистическая гипотеза – $H_0 : F(x)$ □ N

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \sum_{i=0}^{k} \frac{(n_i - np_i)^2}{np_i}$	n_i – число элементов в i-ом интервале, n-объем выборки, p_i - вероятность попадания в i-ый интервал, k - число интервалов
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z \sim \chi^2(k-r-1)$	r – число неизвестных параметров распределения, k - количество интервалов
Формула расчета критической точки	$\chi^2_{1-\alpha}(k-r-1)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

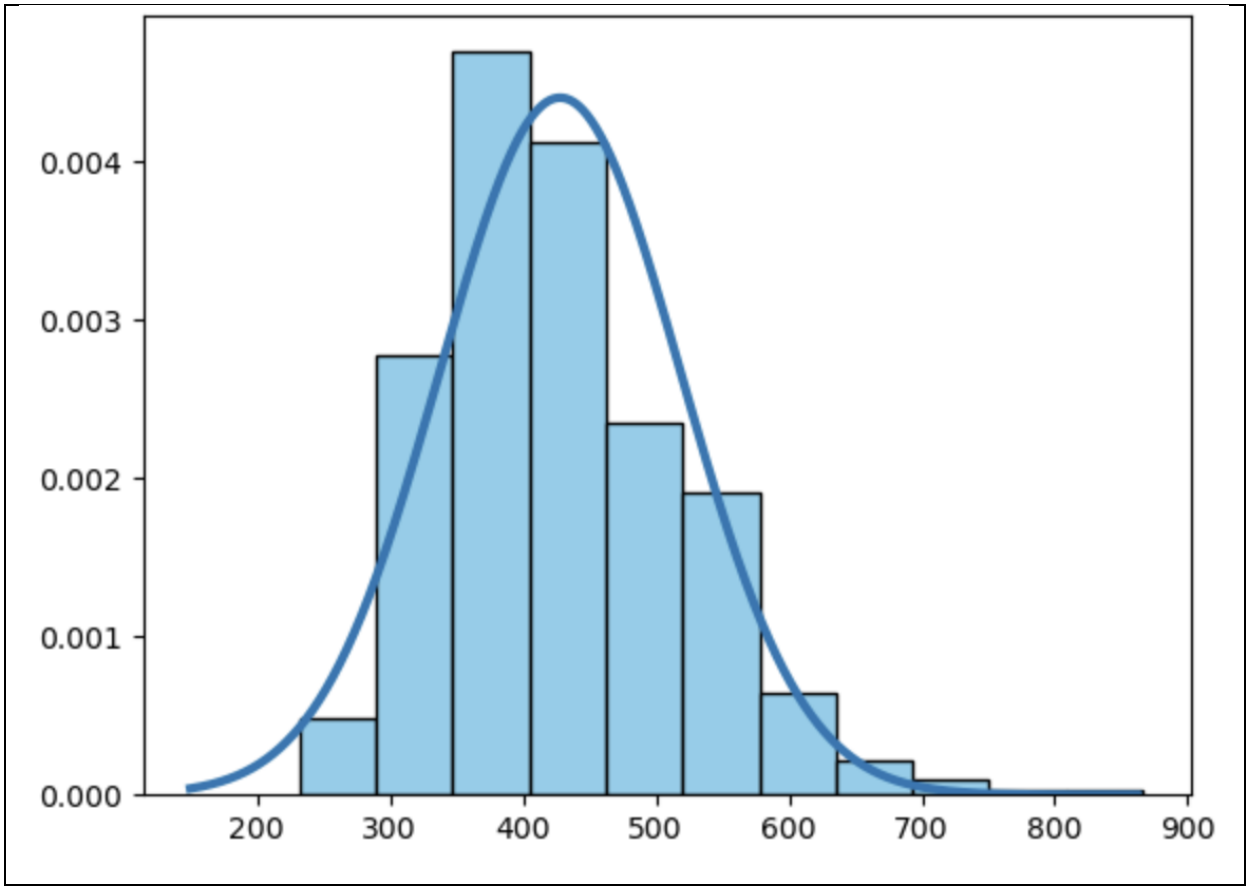
б) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
11	Формула Стерджесса $k \approx [1 + \log_2 \quad n]$	57.64

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Вероятность попадания в интервал при условии истинности основной гипотезы
1	232.00	289.64	30	0.027	0.048
2	289.64	347.27	172	0.16	0.123
3	347.27	404.91	290	0.27	0.212
4	404.91	462.55	255	0.24	0.249
5	462.55	520.18	145	0.14	0.197
6	520.18	577.82	118	0.11	0.105
7	577.82	635.45	40	0.037	0.038
8	635.45	693.09	13	0.012	0.009
9	693.09	750.73	6	0.0056	0.002
10	750.73	808.36	2	0.0019	0.000172
11	808.36	866.00	2	0.0018	0.000013

г) Построить гистограмму относительных частот и функцию плотности теоретического распределения на одном графике



д) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	373.83	0.0	Отклоняем H0	Распределение не является нормальным

0.05	373.83	0.0	Отклоняем H0	Распределение не является нормальным
0.1	373.83	0.0	Отклоняем H0	Распределение не является нормальным

4.2. Проверка гипотезы о нормальности на основе коэффициента асимметрии и эксцесса (критерий Харке-Бера)

Статистическая гипотеза – $H_0 : F(x) \sim N$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{n}{6}\left((\gamma^*)^2 + \frac{(\varepsilon^*)^2}{4}\right)$	n- объем выборки γ^* - выборочный коэф. ассиметрии ε^* - выборочный эксцесс
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(2)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(2)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	162.561	*0.0*	Отклоняем H0	Распределение не является нормальным
0.05			Отклоняем H0	Распределение не является нормальным
0.1			Отклоняем H0	Распределение не является нормальным

Вывод (в терминах предметной области)

В результате проведённого в п.4 статистического анализа обнаружено, что средняя заработная плата всех должностей не является нормально распределенной случайной величиной.

5. Проверка однородности выборок

Анализируемый признак 1 – A5

Анализируемый признак 2 –A8

Объёмы выборок – 1073

5.1 Критерий знаков

Статистическая гипотеза – $H_0 : F_1(x) = F_2(x)$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{(k_+) - \frac{n}{2}}{\frac{\sqrt{n}}{2}}$	k_+ - число знаков ‘+’ n- объем выборки
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z \sim N(0,1)$	N(0,1) - стандартное нормальное распределение
Формула расчета критической точки	$U_{1-\frac{\alpha}{2}}, U_{\frac{\alpha}{2}}$	α - уровень значимости
Формула расчета *p-value*	$2\min(F_Z(z), 1 - F_Z(z))$ (в условиях истинности H0)	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.0	0.0	Отклоняем H0	Выборки неоднородные
0.05			Отклоняем H0	Выборки неоднородные
0.1			Отклоняем H0	Выборки неоднородные

5.2. Критерий хи-квадрат

Статистическая гипотеза – $H_0 : F_1(x) = F_2(x)$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = n_X n_Y \sum_{i=1}^{k} \frac{1}{n_i + m_i} \left(\frac{n_i}{n_X} - \frac{m_i}{n_Y} \right)^2$	n_i количество наблюдений в i-ом интервале первой выборки, m_i - количество наблюдений в i-ом интервале второй выборки, n_X, n_Y - общее количество наблюдений в 1 или 2 выборке, k - число интервалов
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z \sim \chi^2(k-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(k-1)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

б) Выбрать число групп

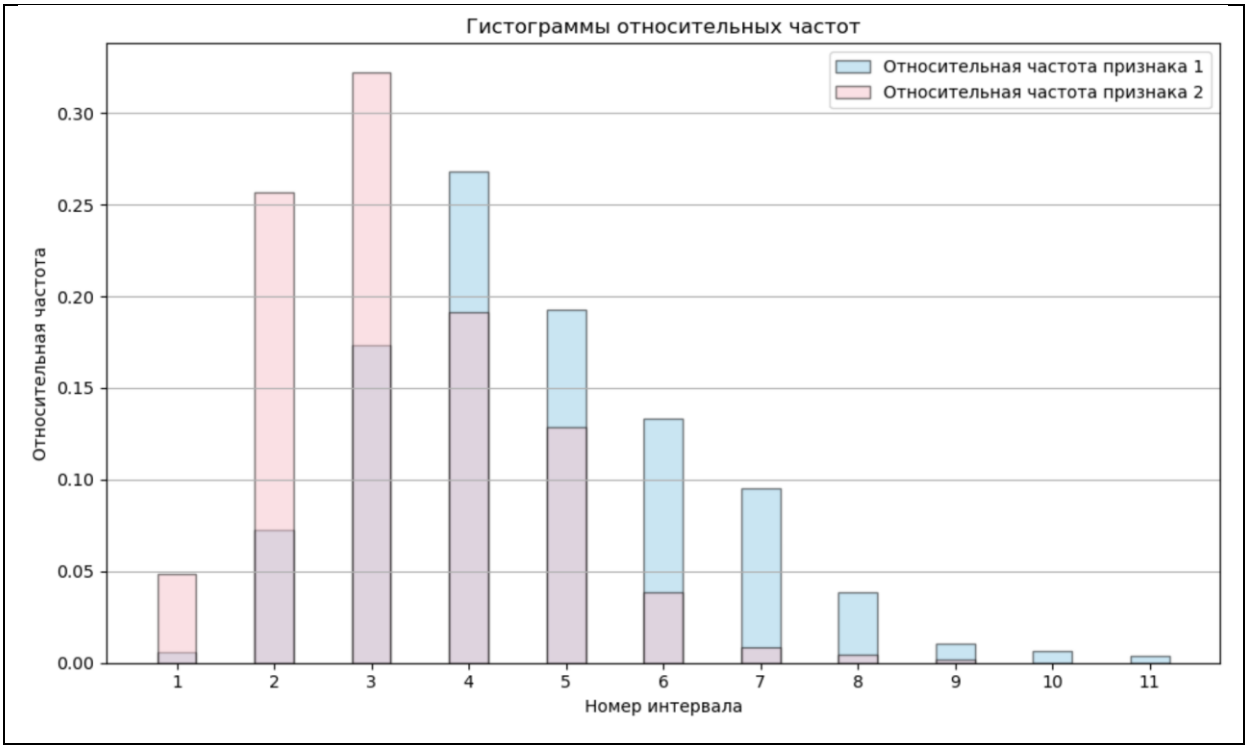
Число групп	Обоснование выбора числа групп	Ширина интервалов
11	Формула Стерджесса k ≈ $[1 + \log_2 \quad n]$	70,63

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота признака 1	Частота признака 2	Относит. частота признака 1	Относит. частота признака 2
1	232.00	302.63	6	52	0.006	0.049
2	302.63	373.27	78	275	0.072	0.257
3	373.27	443.91	186	345	0.173	0.322
4	443.91	514.55	288	205	0.268	0.191
5	514.55	585.18	207	138	0.193	0.129
6	585.18	655.82	143	41	0.133	0.038
7	655.82	726.45	102	9	0.095	0.008
8	726.45	797.09	41	5	0.038	0.005
9	797.09	867.73	11	2	0.010	0.002

10	867.73	938.36	7	0	0.007	0
11	938.36	1009.00	4	0	0.004	0

г) Построить гистограммы относительных частот на одном графике



д) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	153.688	*0.0*	Отклоняем H0	Выборки неоднородные
0.05			Отклоняем H0	Выборки неоднородные
0.1			Отклоняем H0	Выборки неоднородные

Вывод (в терминах предметной области)

В результате проведённого в п.5 статистического анализа обнаружено, что выборки средних заработных плат всех профессоров и средних заработных плат всех должностей не являются однородными.

6. Таблицы сопряжённости

Факторный признак *x* – A9

Результативный признак *y* – A14

Объёмы выборок – 1073

Статистическая гипотеза – H0: $F_Y(y \text{ при } X = No) = F_Y(y \text{ при } X = Yes)$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z=\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}$	m_{ij} теоретические частоты, n_{ij} - наблюдаемые частоты
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z\sim\chi^2\big((k-1)(l-1)\big)$	k-число вариантов факторного признака, l - число вариантов результативного признака
Формула расчета критической точки	$\chi^2_{1-\alpha}\big((k-1)(l-1)\big)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

б) Построить эмпирическую таблицу сопряжённости

x \ *y*	N	Y	Σ
N	570	31	601
Y	22	450	472
Σ	592	481	1073

в) Построить теоретическую таблицу сопряжённости

x \ *y*	N	Y	Σ
N	331.586	269.414	601
Y	260.414	211.586	472
Σ	592	481	1073

г) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод

0.01	869.3	0.0	Отклоняем H0	Есть статистическая связь между случайными величинами
0.05			Отклоняем H0	Есть статистическая связь между случайными величинами
0.1			Отклоняем H0	Есть статистическая связь между случайными величинами

Вывод (в терминах предметной области)

В результате проведённого в п.6 статистического анализа обнаружено, что между «Средняя зарплата для всех категорий > средняя зарплата по колледжу» и «Средняя компенсация для всех категорий > средняя компенсация по колледжу» есть статистическая связь.

7. Дисперсионный анализ

Факторный признак *x* – A4

Результативный признак *y* – A8

Число вариантов факторного признака – 3

Объёмы выборок – 1073

Статистическая гипотеза – $m_1 = \cdots = m_k$

а) Рассчитать групповые выборочные характеристики

№ п/п	Вариант факторного признака	Объём выборки	Групповые средние	Групповые дисперсии
1	I	180	533.67	7652.25
2	IIA	359	440.82	4684.10
3	IIB	534	383.83	4975.63

б) Привести формулы расчёта показателей вариации, используемых в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$D^*_{\text{межгр}} = \frac{1}{n}\sum_{k=1}^{K} n_k(\overline{x_k} - \overline{x})^2$	K-1	$\frac{n}{K-1} D^*_{\text{межгр}}$
Остаточные признаки	$D^*_{\text{внутригр}} = \frac{1}{n}\sum_{k=1}^{K} n_k\, \sigma_k^{*2}$	n-K	$\frac{n}{n-K} D^*_{\text{внутригр}}$
Все признаки	$D^*_{\text{общ}} = \frac{1}{n}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left(x_i^{(k)} - \overline{x}\right)^2$	n-1	$\frac{n}{n-1} D^*_{\text{общ}}$

в) Рассчитать показатели вариации, используемые в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	2898.993	2	1555309.5673
Остаточные признаки	5310.970	1070	5325.8605
Все признаки	8209.963	1072	8217.6211

г) Проверить правило сложения дисперсий

Показатель	$D_{межгр}$	$D_{внутригр}$	$D_{общ}$	$D_{межгр} + D_{внутригр}$
Значение	2898.993	5310.970	8209.963	8209.963

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Эмпирический коэффициент детерминации	$\eta^2 = \frac{D^*_{\text{межгр}}}{D^*_{\text{общ}}}$	0.353
Эмпирическое корреляционное отношение	$\eta = \sqrt{\frac{D^*_{\text{межгр}}}{D^*_{\text{общ}}}}$	0.594

е) Охарактеризовать тип связи между факторным и результативным признаками

По шкале Чеддока: степень тесноты статистической связи между факторным и результативным признаками - заметная

ж) Указать формулы расчёта показателей, используемых при проверке статистической гипотезы дисперсионного анализа

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{D^*_{\text{межгр}}}{(k-1)} : \frac{D^*_{\text{внутригр}}}{(n-k)}$	n- объем выборки, k - число групп, $D^*_{\text{межгр}}$ межгрупповая дисперсия, $D^*_{\text{внутригр}}$ - внутригрупповая дисперсия
Закон распределения статистики критерия при условии истинности основной гипотезы	F(k-1, n-k)	
Формула расчета критической точки	$f_{1-\alpha}(k-1, n-k)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

з) Проверить статистическую гипотезу дисперсионного анализа

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	292.0297	*0.0*	Отклоняем H0	Присутствует статистическая связь
0.05			Отклоняем H0	Присутствует статистическая связь
0.1			Отклоняем H0	Присутствует статистическая связь

Вывод (в терминах предметной области)

В результате проведённого в п.7 статистического анализа обнаружено, что средняя заработная плата всех должностей зависит от типа учебного заведения.

8. Корреляционный анализ

8.1. Расчёт парных коэффициентов корреляции

Анализируемый признак 1 – A5

Анализируемый признак 2 –A8

Объёмы выборок – 1073

а) Рассчитать точечные оценки коэффициентов корреляции

	Формула расчета	Значение
Линейный коэффициент корреляции	$\overline{\rho_{XY}} = \frac{\overline{k_{XY}}}{\overline{\sigma_X} \cdot \overline{\sigma_Y}}$	0.968
Ранговый коэффициент корреляции по Спирмену	$\overline{\rho_{XY}} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n}(r_i - \overline{r})(s_i - \overline{s})\right)}{\overline{\sigma_R} \cdot \overline{\sigma_S}}$	0.964
Ранговый коэффициент корреляции по Кендаллу	$\overline{\tau_{XY}} = \frac{4R}{n(n-1)} - 1,\ R = \sum_{i=1}^{n-1} R_i\ ,\ R_i = \sum_{j=i+1}^{n}[s_j > s_i]$	0.841

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

Граница доверительного интервала	Формула расчета
Нижняя граница	$\overline{\rho_{XY}} + \frac{\overline{\rho}_{XY}\left(1 - \overline{\rho}_{XY}^2\right)}{2n} - U_{1-\frac{\alpha}{2}} \cdot \frac{\left(1 - \overline{\rho_{XY}}^2\right)}{\sqrt{n}}$
Верхняя граница	$\overline{\rho}_{XY} + \frac{\overline{\rho_{XY}}\left(1 - \overline{\rho}_{XY}^2\right)}{2n} + U_{1-\frac{\alpha}{2}} \cdot \frac{\left(1 - \overline{\rho_{XY}}^2\right)}{\sqrt{n}}$

в) Рассчитать доверительные интервалы для линейного коэффициента корреляции

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	0.9627	0.9639	0.9645
Верхняя граница	0.9727	0.9714	0.9709

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициентов корреляции

Статистическая гипотеза	Формула расчета статистики критерия	Закон распределения статистики критерия при условии истинности основной гипотезы
$H_0 : \rho = 0$ $H' : \rho \neq 0$	$Z = \frac{\overline{\rho_{XY}}}{\sqrt{1 - (\overline{\rho_{XY}})^2}} \cdot \sqrt{n-2}$	T(n-2)

$H_0 : r^{(сп)} = 0$ $H' : r^{(сп)} \neq 0$	$Z = \frac{\overline{\rho_{XY}}^{(sp)}}{\sqrt{1 - \left(\overline{\rho_{XY}}^{(sp)}\right)^2}} \cdot \sqrt{n-2}$	T(n-2)
$H_0 : r^{(кен)} = 0$ $H' : r^{(кен)} \neq 0$	$Z = \overline{\tau_{XY}} \cdot \sqrt{\frac{9n(n+1)}{2(2n+5)}}$	N(0,1)

д) Проверить значимость коэффициентов корреляции

Статистическая гипотеза	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
$H_0 : \rho = 0$ $H' : \rho \neq 0$	0.1	125.516	*0.0*	Отклоняем H0	$\rho \neq 0$ Статистическая связь присутствует
$H_0 : r^{(сп)} = 0$ $H' : r^{(сп)} \neq 0$	0.1	119.027	*0.0*	Отклоняем H0	$\rho^{(sp)} \neq 0$ Статистическая связь присутствует
$H_0 : r^{(кен)} = 0$ $H' : r^{(кен)} \neq 0$	0.1	41.271	*0.0*	Отклоняем H0	$\tau \neq 0$ Статистическая связь присутствует

8.2. Расчёт множественных коэффициентов корреляции

Анализируемый признак 1 – A5

Анализируемый признак 2 – A6

Анализируемый признак 3 – A8

Объёмы выборок – 1073

а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу

Признак / Признак	A5	A6	A8
A5	1	0.821	0.841
A6	0.821	1	0.813
A8	0.841	0.813	1

б) Рассчитать матрицу значений p-value для ранговых коэффициентов корреляции по Кендаллу (статистическая гипотеза $H_0 : r^{(кен)} = 0,\ H' : r^{(кен)} \neq 0$)

Признак / Признак	A5	A6	A8
A5	–	0.0	0.0
A6	0.0	–	0.0
A8	0.0	0.0	–

в) Рассчитать точечную оценку коэффициента конкордации

	Формула расчета	Значение
Коэффициент конкордации	$W = \frac{12}{k^2(n^3-n)} \sum_{i=1}^{n} \left(\sum_{j=1}^{k} R_{ij} - \frac{k(n+1)}{2} \right)^2$ Где $R_{ij} \in \{1, \dots, n\}$ - ранг i-ого элемента в X_j выборке	0.971

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициента конкордации

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = n(k-1)W$	n-размер выборки, W - коэффициент конкордации, k-число выборок
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(n-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(n-1)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

д) Проверить значимость коэффициента конкордации

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	3121.937	*0.0*	Отклоняем H0	Статистическая связь присутствует
0.05			Отклоняем H0	Статистическая связь присутствует
0.1			Отклоняем H0	Статистическая связь присутствует

Вывод (в терминах предметной области)

В результате проведённого в п.8 статистического анализа обнаружено, что между средней заработной платой всех должностей и средней заработной платой всех профессоров есть зависимость, также присутствует статистическая связь между средними заработными платами всех профессоров, всех должностей и всех доцентов.

9. Регрессионный анализ

9.1 Простейшая линейная регрессионная модель

Факторный признак *x* – A15

Результативный признак *y* – A13

Уравнение регрессии – $f(x) = \beta_0 + \beta_1 x$

9.1.1. Точечные оценки линейной регрессионной модели

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
β_0	$\beta_0^* = m_Y^* - \rho_{XY}^* \cdot \frac{\sigma_Y^*}{\sigma_X^*} m_X^*$	488.993
β_1	$\beta_1^* = \rho_{XY}^* \cdot \frac{\sigma_Y^*}{\sigma_X^*}$	0.465

б) Записать точечную оценку уравнения регрессии

f(*x*) = 488.993+ 0.465*x

в) Привести формулы расчёта показателей вариации, используемых в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$D_{\text{регр}}^* = \frac{1}{n}\sum_{i=1}^{n}(f(x_i,\beta_0,\beta_1) - \overline{y})^2$	k-1	$\frac{n}{k-1} D_{\text{регр}}^*$
Остаточные признаки	$D_{\text{ост}}^* = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i,\beta_0,\beta_1)\right)^2$	n-k	$\frac{n}{n-k} D_{\text{ост}}^*$
Все признаки	$D_{\text{общ}}^* = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$	n-1	$\frac{n}{n-1} D_{\text{общ}}^*$

г) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	4589.956	1	4925022.788
Остаточные признаки	9440.535	1071	9458.164
Все признаки	14030.492	1072	14043.580

д) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	4589.956	9440.535	14030.492	14030.492

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$\frac{D^*_{\text{регр}}}{D^*_{\text{общ}}}$	0.327
Корреляционное отношение	$\sqrt{\frac{D^*_{\text{регр}}}{D^*_{\text{общ}}}}$	0.572

ж) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

Между факторным и результативным признаками присутствует заметная связь.

9.1.2. Интервальные оценки линейной регрессионной модели

а) Привести формулы расчёта доверительных интервалов для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	Формула расчета
β_0	Нижняя граница	$\beta_0^* - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D^*_{\text{ост}}} \cdot \sqrt{\frac{(\sum_{i=1}^n x_i^2)}{n^2 D^*_{\text{общ}}}}$
	Верхняя граница	$\beta_0^* + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D^*_{\text{ост}}} \cdot \sqrt{\frac{(\sum_{i=1}^n x_i^2)}{n^2 D^*_{\text{общ}}}}$
β_1	Нижняя граница	$\beta_1^* - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D^*_{\text{ост}}} \cdot \sqrt{\frac{1}{n^2 D^*_{\text{общ}}}}$
	Верхняя граница	$\beta_1^* + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D^*_{\text{ост}}} \cdot \sqrt{\frac{1}{n^2 D^*_{\text{общ}}}}$

б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

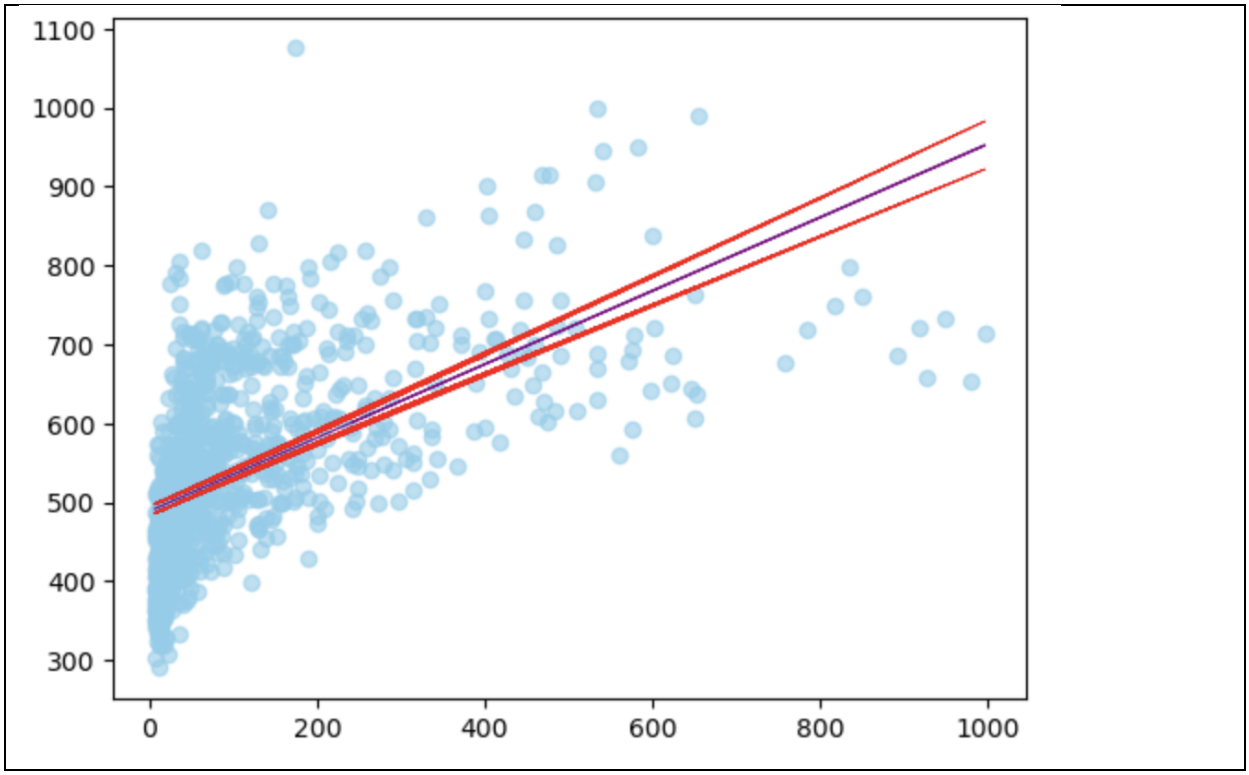
Параметр	Границы доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
β_0	Нижняя граница	479.629	481.873	483.019

	Верхняя граница	498.357	496.114	494.967
β_1	Нижняя граница	0.463	0.463	0.464
	Верхняя граница	0.466	0.466	0.466

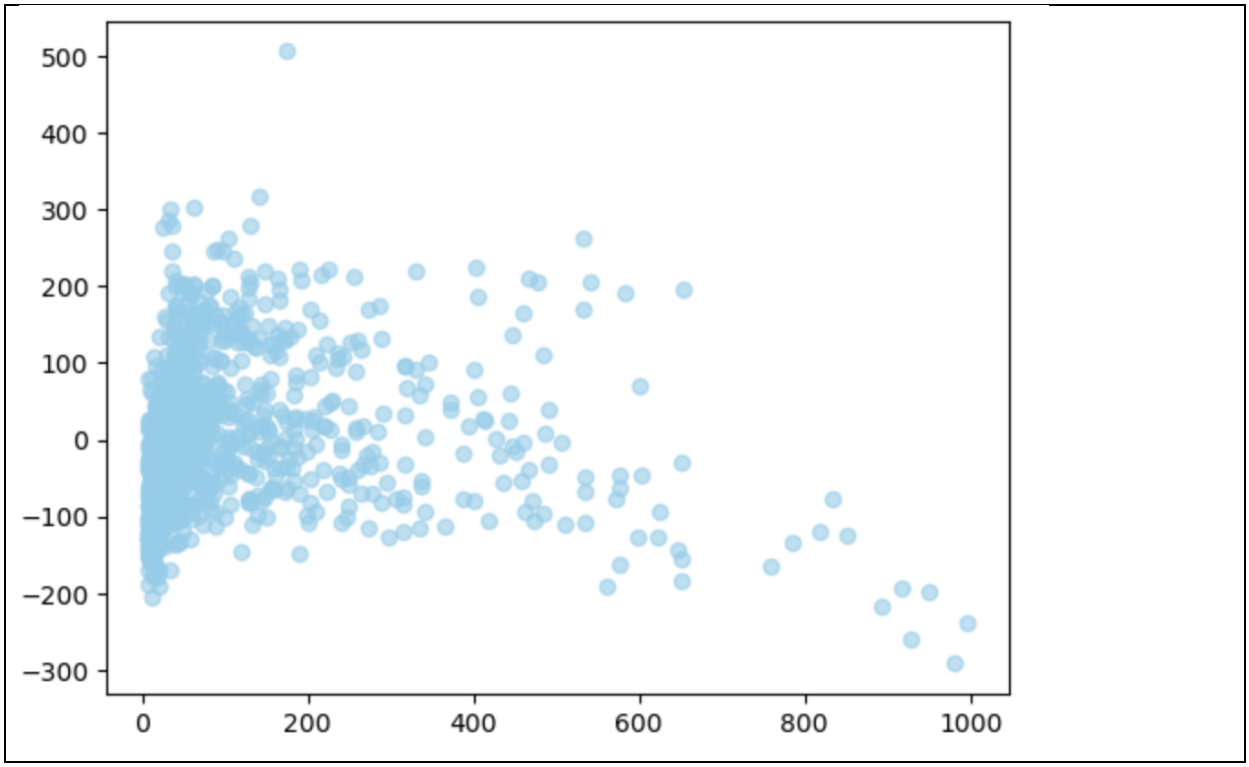
в) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$f^*(x) - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D^*_{\text{ост}}} \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{n^2 D^*_{\text{общ}}}}$
Верхняя граница $f_{high}(x)$	$f^*(x) + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{D^*_{\text{ост}}} \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{n^2 D^*_{\text{общ}}}}$

г) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии $f_{low}(x)$ и $f_{high}(x)$ на уровне значимости $\alpha = 0.1$



д) Построить график остатков $\varepsilon(x) = y - f(x)$



9.1.3. Проверка значимости линейной регрессионной модели

Статистическая гипотеза – $\begin{matrix} H_0 : \beta_1 = 0 \\ H' : \beta_1 \neq 0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{R^{2*}_{Y\|X}}{\left(1-R^{2*}_{Y\|X}\right)/(n-2)}$	n – объем выборки
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(1,\, n-2)$	
Формула расчета критической точки	$f_{1-\alpha}(1,\, n-2)$	α - уровень значимости
Формула расчета *p-value*	$p = 1 - F_Z(z)$ (в условиях истинности H0)	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	520.716	*0.0*	Отклоняем H0	Модель значимая

0.05			Отклоняем H0	Модель значимая
0.1			Отклоняем H0	Модель значимая

9.2 Линейная регрессионная модель общего вида

Факторный признак *x* – A15

Результативный признак *y* – A13

Уравнение регрессии – квадратичное по *x*: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

9.2.1. Точечные оценки линейной регрессионной модели

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
β_0	$\beta^* = (F^T F)^{-1} F^T y$	460.118
β_1		1.022
β_2		-0.001

б) Записать точечную оценку уравнения регрессии

$f(x) = 460.118+1.022*x-0.001*x^2$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	5844.452	2	3135548.498
Остаточные признаки	8186.039	1070	8208.991
Все признаки	14030.491	1072	14043.579

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	5844.452	8186.039	14030.491	14030.491

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение

Коэффициент детерминации	$\frac{D^*_{\text{регр}}}{D^*_{\text{общ}}}$	0.417
Корреляционное отношение	$\sqrt{\frac{D^*_{\text{регр}}}{D^*_{\text{общ}}}}$	0.645

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

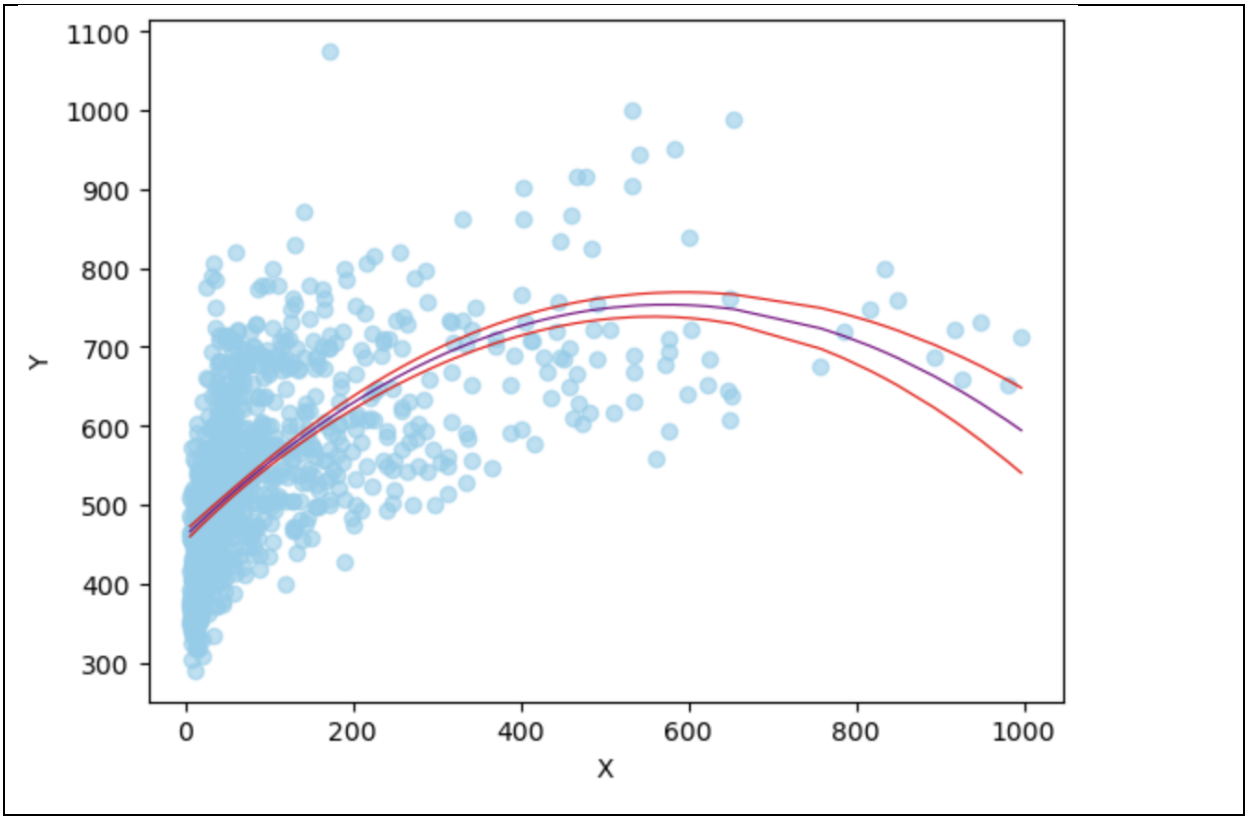
Между факторным и результативным признаками присутствует заметная связь.

9.2.2. Интервальные оценки линейной регрессионной модели

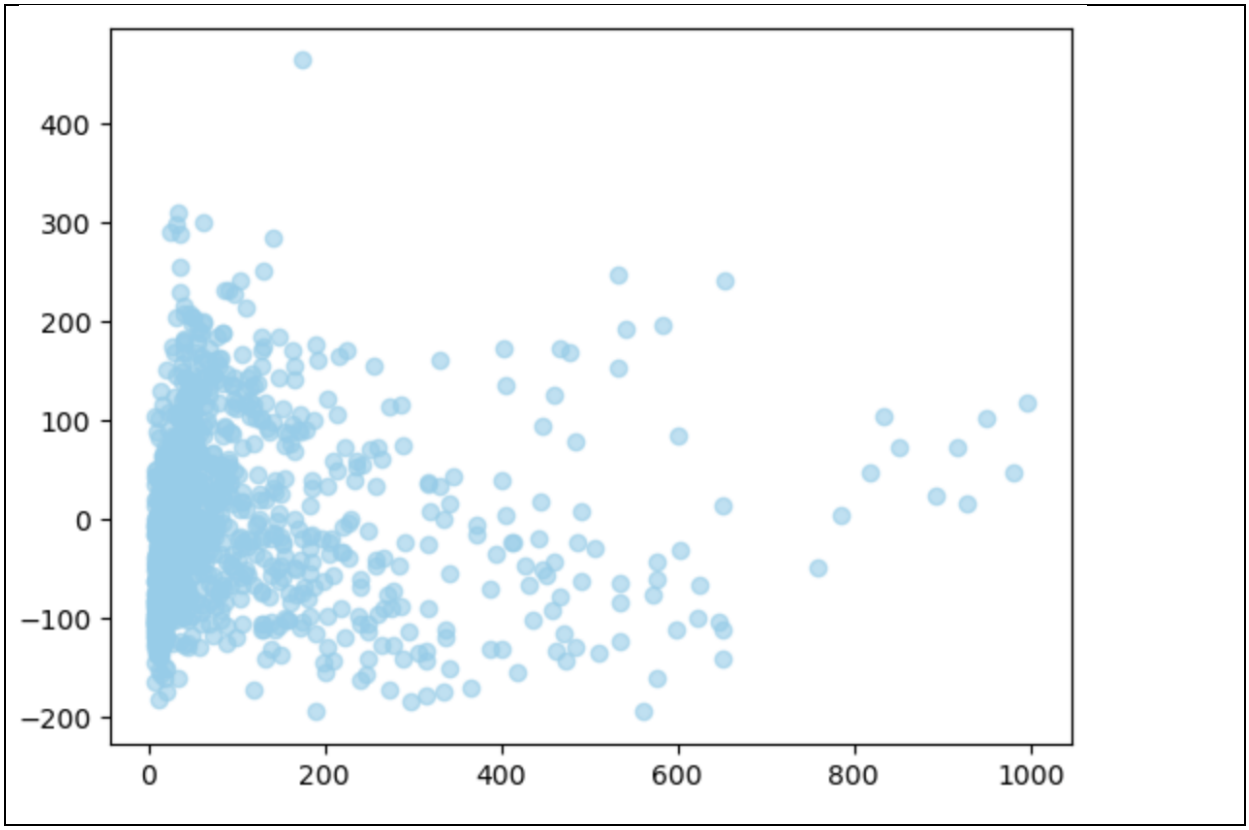
а) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$\left(\tilde{f}(x) - t_{1-\alpha/2}(n-k)\sqrt{\tilde{D}_{resY}}\sqrt{\varphi^T(x)(F^TF)^{-1}\varphi(x)}\right.;$
Верхняя граница $f_{high}(x)$	$\tilde{f}(x) + t_{1-\alpha/2}(n-k)\sqrt{\tilde{D}_{resY}}\sqrt{\varphi^T(x)(F^TF)^{-1}\varphi(x)}$

б) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии $f_{low}(x)$ и $f_{high}(x)$ на уровне значимости $\alpha = 0.1$



в) Построить график остатков $\varepsilon(x) = y - f(x)$



9.2.3. Проверка значимости линейной регрессионной модели

Статистическая гипотеза – $\begin{matrix} H_0 : \beta_1 = \beta_2 = 0 \\ H' : \textit{не } H_0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{R^{2*}_{Y\|X}/(k-1)}{\left(1-R^{2*}_{Y\|X}\right)/(n-k)}$	n-объем выборки
Закон распределения статистики критерия при условии истинности основной гипотезы	F(k-1, n-k)	
Формула расчета критической точки	$f_{1-\alpha}(k-1,\, n-k)$	α - уровень значимости
Формула расчета *p-value*	(в условиях истинности $p = 1 - F_Z(z)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	981.965	*0.0*	Отклоняем H0	Модель значимая
0.05			Отклоняем H0	Модель значимая
0.1			Отклоняем H0	Модель значимая

9.3 Множественная линейная регрессионная модель

Факторный признак 1 x_1 – A15

Факторный признак 2 x_2 – A5

Результативный признак y –A13

Уравнение регрессии – $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
β_0	$\beta^* = (F^T F)^{-1} F^T y$	40.180
β_1		0.037
β_2		0.936

б) Записать точечную оценку уравнения регрессии

$f(x) = 40.180 + 0.037x + 0.936x^2$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	12834.70	2	6885816.550
Остаточные признаки	1195.79	1070	1199.142
Все признаки	14030.492	1072	14043.578

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	12834.70	1195.79	14030.49	14030.49

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Множественный коэффициент детерминации	$\frac{D^*_{\text{регр}}}{D^*_{\text{общ}}}$	0.915
Множественное корреляционное отношение	$\sqrt{\frac{D^*_{\text{регр}}}{D^*_{\text{общ}}}}$	0.956

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

Между факторным и результативным признаками присутствует сильная связь.

9.4. Выводы

а) Сводная таблица показателей вариации для различных регрессионных моделей

Источник вариации	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Факторный признак	4589.956	5844.452	12834.70
Остаточные признаки	9440.535	8186.039	1195.79
Все признаки	14030.492	14030.491	14030.492

б) Сводная таблица свойств различных регрессионных моделей

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	32,7%	41,7%	91,5%
Значимость	да	да	да
Адекватность	неадекватная	неадекватная	адекватная
Степень тесноты связи	заметная	заметная	сильная

Вывод (в терминах предметной области)

В результате проведённого в п.9 статистического анализа обнаружено, что количество профессоров заметно влияет на среднюю компенсацию для всех должностей. Однако количество профессоров в учебном заведении и средняя заработная плата всех профессоров сильно влияет на среднюю компенсацию для всех должностей.