



Институт интеллектуальных кибернетических систем

КАФЕДРА КИБЕРНЕТИКИ

БДЗ

по курсу "Математическая статистика"

студента группы Б21-514

____Шамаева Сергея Денисовича____

Вариант № _19_

Оценка: _____

Подпись: _____

2023 г.

ОТЧЕТ № 1

по теме «Проверка статистических гипотез»

Вариант № 19

ФИО студента Шамаева Сергея Денисовича группа Б21-514

Оценка: _____ Подпись: _____

Результаты статистических тестов:

| № задания | Проверяемая гипотеза H_0 | Критерий | Статистическое решение ($\alpha = 0.1$) | Вывод |
|-----------|----------------------------|------------|---|-------|
| 4.1 | | Хи-квадрат | | |
| 4.2 | | Харке-Бера | | |
| 5.1 | | знаков | | |
| 5.2 | | Хи-квадрат | | |

Выводы:

В результате проведённого в п.4 статистического анализа обнаружено, что

В результате проведённого в п.5 статистического анализа обнаружено, что

| |
|--|
| |
|--|

ОТЧЕТ № 2

по теме «Анализ статистических взаимосвязей»

Вариант № _____

ФИО студента _____ **группа** _____

Оценка: _____ **Подпись:** _____

Результаты статистических тестов:

| № задания | Проверяемая гипотеза H_0 | Критерий | Статистическое решение ($\alpha = 0.1$) | Вывод |
|-----------|----------------------------|------------|---|-------|
| 6 | | Хи-квадрат | | |
| 7 | | ANOVA | | |

Выводы:

В результате проведённого в п.6 статистического анализа обнаружено, что

В результате проведённого в п.7 статистического анализа обнаружено, что

| |
|--|
| |
|--|

ОТЧЕТ № 3

по теме «Основы регрессионного анализа»

Вариант № _____

ФИО студента _____ группа _____

Оценка: _____ Подпись: _____

Сводная таблица свойств различных регрессионных моделей:

| Свойство | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
|-----------------------|-------------------------------|---|----------------------------------|
| Точность | | | |
| Значимость | | | |
| Адекватность | | | |
| Степень тесноты связи | | | |

Выводы:

В результате проведённого в п.8 статистического анализа обнаружено, что

В результате проведённого в п.9 статистического анализа обнаружено, что

1. Описательные статистики

1.1. Выборочные характеристики

Анализируемый признак 1 – С8

Анализируемый признак 2 – С10

Анализируемый признак 3 – С13

а) Привести формулы расчёта выборочных характеристик

| Выборочная хар-ка | Формула расчета |
|--|---|
| Объём выборки | n |
| Среднее | $\frac{1}{n} \sum_{i=1}^n x_i$ |
| Выборочная дисперсия | $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ |
| Выборочное среднеквадратическое отклонение | $\sigma_X^* = \sqrt{d_X^*}.$ |
| Выборочный коэффициент асимметрии | $\gamma_X^* = \frac{\mu_3^*}{(\sigma_X^*)^3},$ |
| Выборочный эксцесс | $\varepsilon_X^* = \frac{\mu_4^*}{(\sigma_X^*)^4} - 3,$ |

б) Рассчитать выборочные характеристики

| Выборочная хар-ка | Признак 1 | Признак 2 | Признак 3 |
|--|-----------|-----------|-----------|
| Среднее | 12.789 | 242.461 | 189.892 |
| Выборочная дисперсия | 28.411 | 17421.787 | 33489.294 |
| Выборочное среднеквадратическое отклонение | 5.330 | 131.992 | 183.001 |
| Выборочный коэффициент асимметрии | 1.153 | 1.480 | 3.564 |
| Выборочный эксцесс | 2.483 | 3.412 | 17.209 |

1.2. Группировка и гистограммы частот

Анализируемый признак – С8

Объём выборки – 315

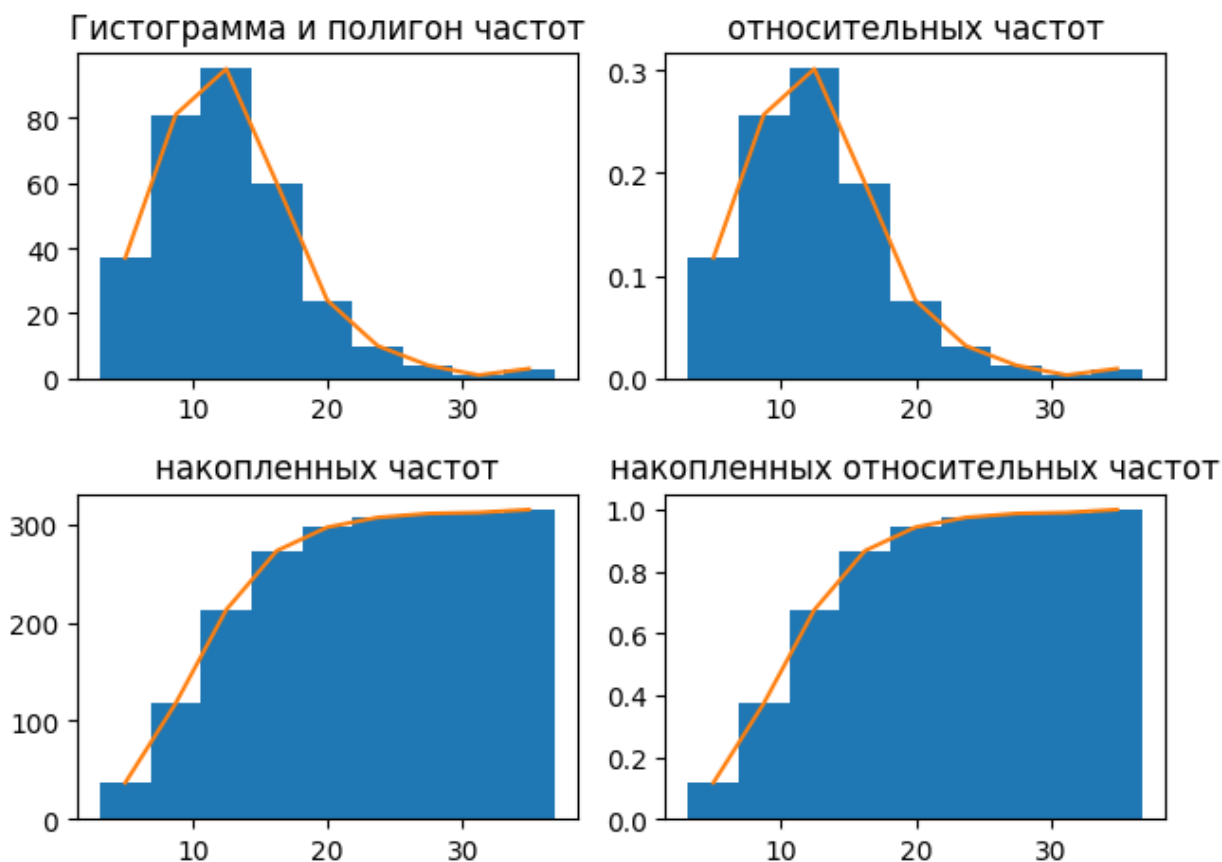
а) Выбрать число групп

| Число групп | Обоснование выбора числа групп | Ширина интервалов |
|-------------|--------------------------------|--------------------|
| 9 | Согласно формуле Стерджесса | 3.7444444444444444 |

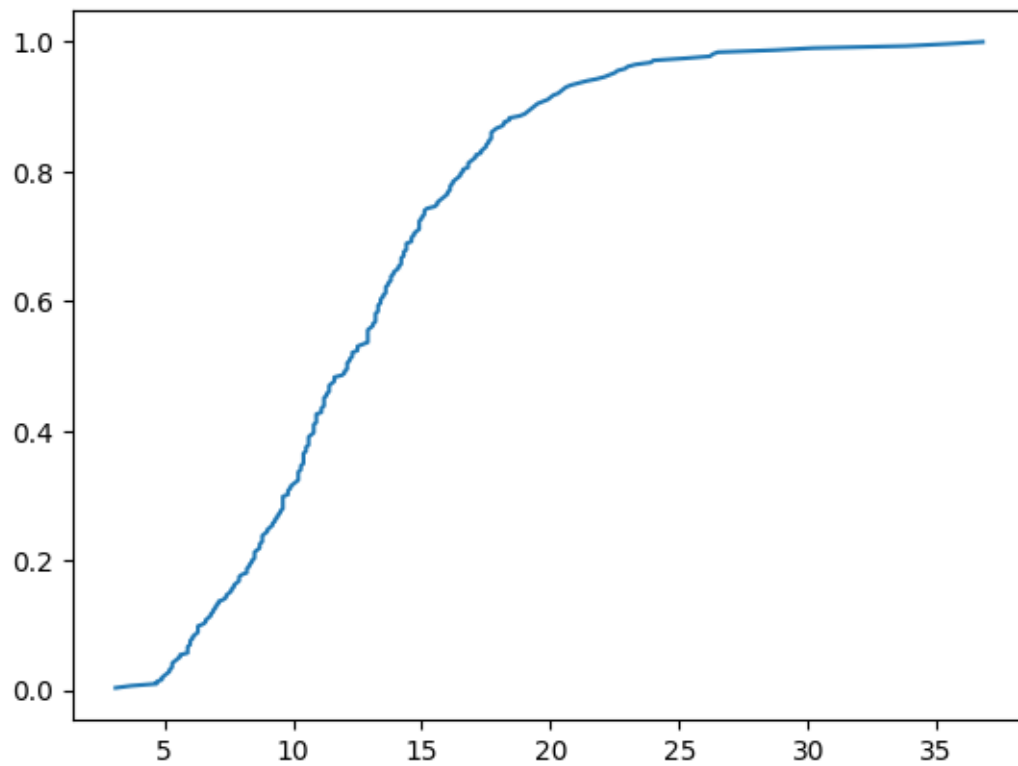
б) Построить таблицу частот

| Номер интервала | Нижняя граница | Верхняя граница | Частота | Относит. частота | Накопл. частота | Относит. накопл. частота |
|-----------------|----------------|-----------------|---------|------------------|-----------------|--------------------------|
| 1 | 3.10 | 6.84 | 37 | 0.117 | 37 | 0.117 |
| 2 | 6.84 | 10.59 | 81 | 0.257 | 118 | 0.375 |
| 3 | 10.59 | 14.33 | 95 | 0.302 | 213 | 0.676 |
| 4 | 14.33 | 18.08 | 60 | 0.190 | 273 | 0.867 |
| 5 | 18.08 | 21.82 | 24 | 0.076 | 297 | 0.943 |
| 6 | 21.82 | 25.56 | 10 | 0.032 | 307 | 0.975 |
| 7 | 25.56 | 29.31 | 4 | 0.013 | 311 | 0.987 |
| 8 | 29.31 | 33.05 | 1 | 0.003 | 312 | 0.990 |
| 9 | 33.05 | 36.80 | 3 | 0.009 | 315 | 1 |

в) Построить гистограммы частот и полигоны частот



г) Построить график эмпирической функции распределения



2. Интервальные оценки

2.1. Доверительные интервалы для мат. ожидания

Анализируемый признак – С8

Объём выборки – 315

Оцениваемый параметр – мат. ожидание

а) Привести формулы расчёта доверительных интервалов

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

б) Рассчитать доверительные интервалы

| Граница доверительного интервала | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|----------------------------------|-----------------|-----------------|----------------|
| Нижняя граница | 11.790 | 12.197 | 12.293 |
| Верхняя граница | 13.786 | 13.379 | 13.284 |

2.2. Доверительные интервалы для дисперсии

Анализируемый признак – С8

Объём выборки – 315

Оцениваемый параметр – дисперсия

а) Привести формулы расчёта доверительных интервалов

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

б) Рассчитать доверительные интервалы

| Граница доверительного интервала | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|----------------------------------|-----------------|-----------------|----------------|
| Нижняя граница | 22.133 | 24.442 | 25.036 |
| Верхняя граница | 37.478 | 33.436 | 32.564 |

2.3. Доверительные интервалы для разности мат. ожиданий

Анализируемый признак 1 – С11

Анализируемый признак 2 – С12

Объёмы выборок – 315

Оцениваемый параметр – разность мат. ожиданий

а) Привести формулы расчёта доверительных интервалов

| Граница доверительного интервала | Формула расчета |
|----------------------------------|--|
| Нижняя граница | $(\bar{X}_1 - \bar{X}_2) - \sqrt{S_1^2/n_1 + S_2^2/n_2} * t_{1-\alpha/2}(\lfloor 1/k \rfloor)$ |
| Верхняя граница | $(\bar{X}_1 - \bar{X}_2) + \sqrt{S_1^2/n_1 + S_2^2/n_2} * t_{\alpha/2}(\lfloor 1/k \rfloor)$ |

$$k = \frac{\left(\frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2} \right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2/n_1}{S_1^2/n_1 + S_2^2/n_2} \right)^2}{n_2 - 1}$$

б) Рассчитать доверительные интервалы

| Граница доверительного интервала | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|----------------------------------|-----------------|-----------------|----------------|
| Нижняя граница | 1120.416 | 1176.586 | 1205.131 |
| Верхняя граница | 1585.361 | 1529.190 | 1500.646 |

2.4. Доверительные интервалы для отношения дисперсий

Анализируемый признак 1 – С11

Анализируемый признак 2 – С12

Объёмы выборок – 315

Оцениваемый параметр – отношение дисперсий

а) Привести формулы расчёта доверительных интервалов

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

б) Рассчитать доверительные интервалы

| Граница доверительного интервала | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|----------------------------------|-----------------|-----------------|----------------|
| Нижняя граница | 4.309 | 5.012 | 5.194 |
| Верхняя граница | 9.081 | 7.807 | 7.53 |

3. Проверка статистических гипотез о математических ожиданиях и дисперсиях

3.1. Проверка статистических гипотез о математических ожиданиях

Анализируемый признак – С8

Объём выборки – 315

Статистическая гипотеза – $H_0: m = m_0$
 $H': m \neq m_0$

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

| | Выражение |
|--|---|
| Формула расчета статистики критерия | $Z = \frac{\bar{X} - m_0}{S/\sqrt{n}}$ |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $T(n-1)$ |
| Формулы расчета критических точек | $-t_{1-\alpha/2}(n-1), t_{1-\alpha/2}(n-1)$ |
| Формула расчета p -value | $2 * \min(T(Z, n-1), 1 - T(Z, n-1))$ |

б) Выбрать произвольные значения m_0 и проверить статистические гипотезы

| m_0 | Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|-------|--------------------|---|------------|------------------------|---|
| 5 | 0.1 | 25.934 | 4.7e-80 | H0 отвергается | $m \neq 5$ |
| 13 | 0.1 | -0.704 | 0.481 | H0 принимается | m можно принять 13 при данном уровне значимости |
| 20 | 0.1 | -24.012 | 4.6e-73 | H0 отвергается | $m \neq 20$ |

3.2. Проверка статистических гипотез о дисперсиях

Анализируемый признак – С8

Объём выборки – 315

Статистическая гипотеза – $H_0: \sigma = \sigma_0$
 $H': \sigma \neq \sigma_0$

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 | $Z _{H_0} \sim \chi^2(n-1)$ |
| Формулы расчета критических точек | $Z_1 = \chi_{\alpha/2}^2(n-1), Z_2 = \chi_{1-\alpha/2}^2(n-1), Z_1 < Z < Z_2$ |
| Формула расчета p -value | $p_value = 2 * \min\{F_z(Z), 1 - F_z(Z)\}$ |

б) Выбрать произвольные значения σ_0 и проверить статистические гипотезы

| σ_0 | Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|------------|--------------------|---|------------|------------------------|---|
| 2 | 0.1 | 2230.259 | 0.000 | Н0 отвергается | $\sigma \neq 2$ |
| 5 | 0.1 | 356.841 | 0.096 | Н0 принимается | σ можно принять 5 при данном уровне значимости |
| 20 | 0.1 | 22.302 | 7.0e-119 | Н0 отвергается | $\sigma \neq 20$ |

3.3. Проверка статистических гипотез о равенстве математических ожиданий

Анализируемый признак 1 – С11

Анализируемый признак 2 – С12

Объёмы выборок – 315

Статистическая гипотеза – $H_0: m_1 = m_2$
 $H': m_1 \neq m_2$

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

| | Выражение |
|--|--|
| Формула расчета статистики критерия | $Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$ |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $k = \frac{\left(\frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2/n_1}{S_1^2/n_1 + S_2^2/n_2}\right)^2}{n_2 - 1}$ $Z _{H_0} \sim T(\lfloor 1/k \rfloor)$ |
| Формулы расчета критических точек | $Z_1 = t_{\alpha/2}(\lfloor 1/k \rfloor), Z_2 = t_{1-\alpha/2}(\lfloor 1/k \rfloor), Z_1 < Z < Z_2$ |
| Формула расчета p -value | $p_value = 2 * \min\{F_z(Z), 1 - F_z(Z)\}$ |

б) Проверить статистические гипотезы

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|-----------------------------|
| 0.01 | 15.126 | 1.9e-41 | H0 отвергается | Нельзя считать $m1 \neq m2$ |
| 0.05 | | | H0 отвергается | Нельзя считать $m1 \neq m2$ |
| 0.1 | | | H0 отвергается | Нельзя считать $m1 \neq m2$ |

3.4. Проверка статистических гипотез о равенстве дисперсий

Анализируемый признак 1 – C11

Анализируемый признак 2 – C12

Объёмы выборок – 315

Статистическая гипотеза – $H_0: \sigma_1 = \sigma_2$
 $H': \sigma_1 \neq \sigma_2$

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 | $Z _{H_0} \sim F(n_1 - 1, n_2 - 1)$ |
| Формулы расчета критических точек | $Z_1 = F_{\alpha/2}(n_1 - 1, n_2 - 1), Z_2 = F_{1-\alpha/2}(n_1 - 1, n_2 - 1), Z_1 < Z < Z_2$ |
| Формула расчета p -value | $p_value = 2 * \min\{F_z(Z), 1 - F_z(Z)\}$ |

б) Проверить статистические гипотезы

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|---|
| 0.01 | 6.255 | 2.2e-16 | H0 отвергается | Нельзя считать $\sigma_1 \neq \sigma_2$ |
| 0.05 | | | H0 отвергается | Нельзя считать $\sigma_1 \neq \sigma_2$ |
| 0.1 | | | H0 отвергается | Нельзя считать $\sigma_1 \neq \sigma_2$ |

4. Критерии согласия

Анализируемый признак – С8

Объём выборки – 315

4.1. Критерий хи-квадрат

Теоретическое распределение – нормальное.

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

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

| | Выражение | Пояснение использованных обозначений |
|--|--|---|
| Формула расчета статистики критерия | $Z = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$ | np_i - теоретическая частота попадания в i-й интервал n_i - наблюдаемая частота в i-м интервале |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $Z _{H_0} \sim \chi^2(k - r - 1)$ | k - количество отрезков разбиения r - количество неизвестных параметров проверяемого распределения |
| Формула расчета критической точки | $\chi^2_{1-\alpha}(k - r - 1)$ | |
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |

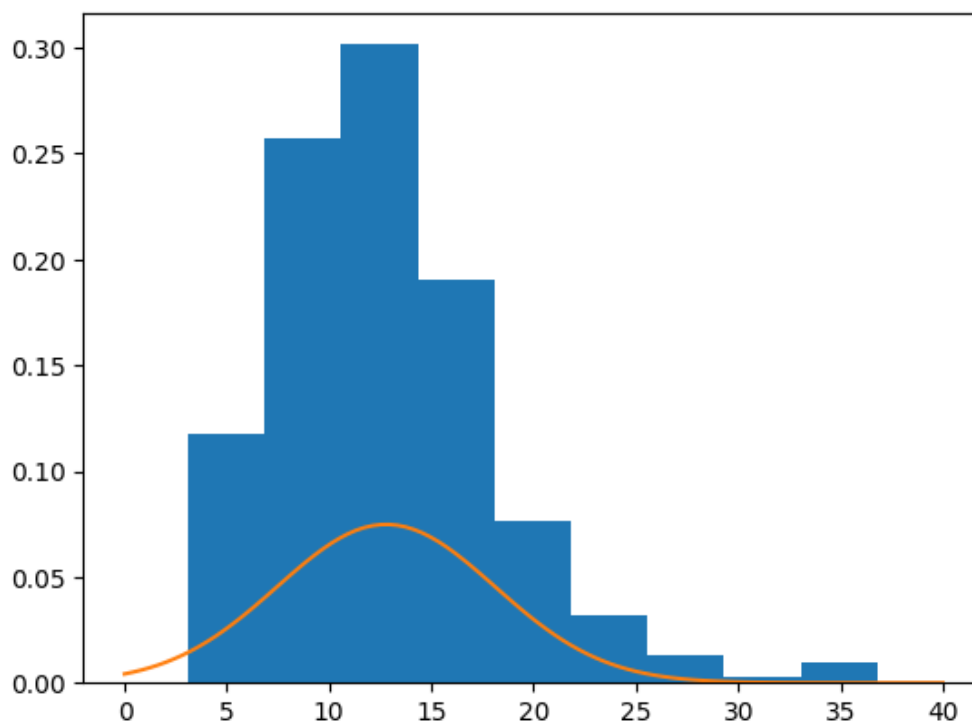
б) Выбрать число групп

| Число групп | Обоснование выбора числа групп | Ширина интервалов |
|-------------|--------------------------------|--------------------|
| 9 | Согласно формуле Стерджесса | 3.7444444444444444 |

в) Построить таблицу частот

| Номер интервала | Нижняя граница | Верхняя граница | Частота | Относит. частота | Вероятность попадания в интервал при условии истинности основной гипотезы |
|-----------------|----------------|-----------------|---------|------------------|---|
| 1 | 3.10 | 6.84 | 37 | 0.117 | 0.132 |
| 2 | 6.84 | 10.59 | 81 | 0.257 | 0.207 |
| 3 | 10.59 | 14.33 | 95 | 0.302 | 0.274 |
| 4 | 14.33 | 18.08 | 60 | 0.190 | 0.225 |
| 5 | 18.08 | 21.82 | 24 | 0.076 | 0.115 |
| 6 | 21.82 | 25.56 | 10 | 0.032 | 0.036 |
| 7 | 25.56 | 29.31 | 4 | 0.013 | 0.007 |
| 8 | 29.31 | 33.05 | 1 | 0.003 | 0.0008 |
| 9 | 33.05 | 36.80 | 3 | 0.009 | 7.1e-05 |

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 | 406.978 | 8.8e-85 | H0 отвергается | Нельзя считать распределение нормальным |
| 0.05 | 406.978 | 8.8e-85 | H0 отвергается | Нельзя считать распределение нормальным |
| 0.1 | 406.978 | 8.8e-85 | H0 отвергается | Нельзя считать распределение нормальным |

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

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

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

| | Выражение | Пояснение использованных обозначений |
|--|--|--|
| Формула расчета статистики критерия | $JB = n \left(\frac{S^2}{6} + \frac{(K)^2}{24} \right)$ | S – коэфф асимметрии, K–коэфф эксцесса |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $JB \sim \chi^2(2)$ | |
| Формула расчета критической точки | $Z_1 = \chi^2_{1-\alpha}(2), Z < Z_1$ | |
| Формула расчета <i>p-value</i> | $p_value = 1 - F_z(Z)$ | |

б) Проверить статистические гипотезы

| Уровень значимости | Выборочное значение статистики критерия | <i>p-value</i> | Статистическое решение | Вывод |
|--------------------|---|----------------|------------------------|---|
| 0.01 | 146.323 | 1.6e-32 | H0 отвергается | Нельзя считать распределение нормальным |
| 0.05 | | | H0 отвергается | Нельзя считать распределение нормальным |
| 0.1 | | | H0 отвергается | Нельзя считать распределение нормальным |

Вывод (в терминах предметной области)

В результате проведённого в п.4 статистического анализа обнаружено, что С8 (Граммы клетчатки, потребляемой в день) нельзя считать распределение нормальным.

5. Проверка однородности выборок

Анализируемый признак 1 – С11

Анализируемый признак 2 – С12

Объёмы выборок – 315

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 расчёта статистики критерия | $Z = 2\sqrt{n}(H - 1/2)$ | $H = K/n$ - частота "успеха" (знак +) K - количество положительных разностей среди $x_1 - y_1, \dots, x_n - y_n$ |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $Z _{H_0} \sim N(0, 1)$ | |
| Формула расчёта критической точки | $Z_1 = N_{\alpha/2}(0, 1)$ $Z_2 = N_{1-\alpha/2}(0, 1)$ | |
| Формула расчёта p -value | $p_value = 2 * \min\{F_z(Z), 1 - F_z(Z)\}$ | |

б) Проверить статистические гипотезы

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|-----------------------|
| 0.01 | 124.5 | 1.9e-50 | Н0 отвергается | Признаки не однородны |
| 0.05 | | | Н0 отвергается | Признаки не однородны |
| 0.1 | | | Н0 отвергается | Признаки не однородны |

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 | $\sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$ | np_i - теоретическая частота попадания в i-й интервал n_i - наблюдаемая частота в i-м интервале |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $Z _{H_0} \sim \chi^2(k - r - 1)$ | k - количество отрезков разбиения r - количество неизвестных параметров проверяемого распределения |
| Формула расчета критической точки | $\chi^2_{1-\alpha}(k - r - 1)$ | |
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |

б) Выбрать число групп

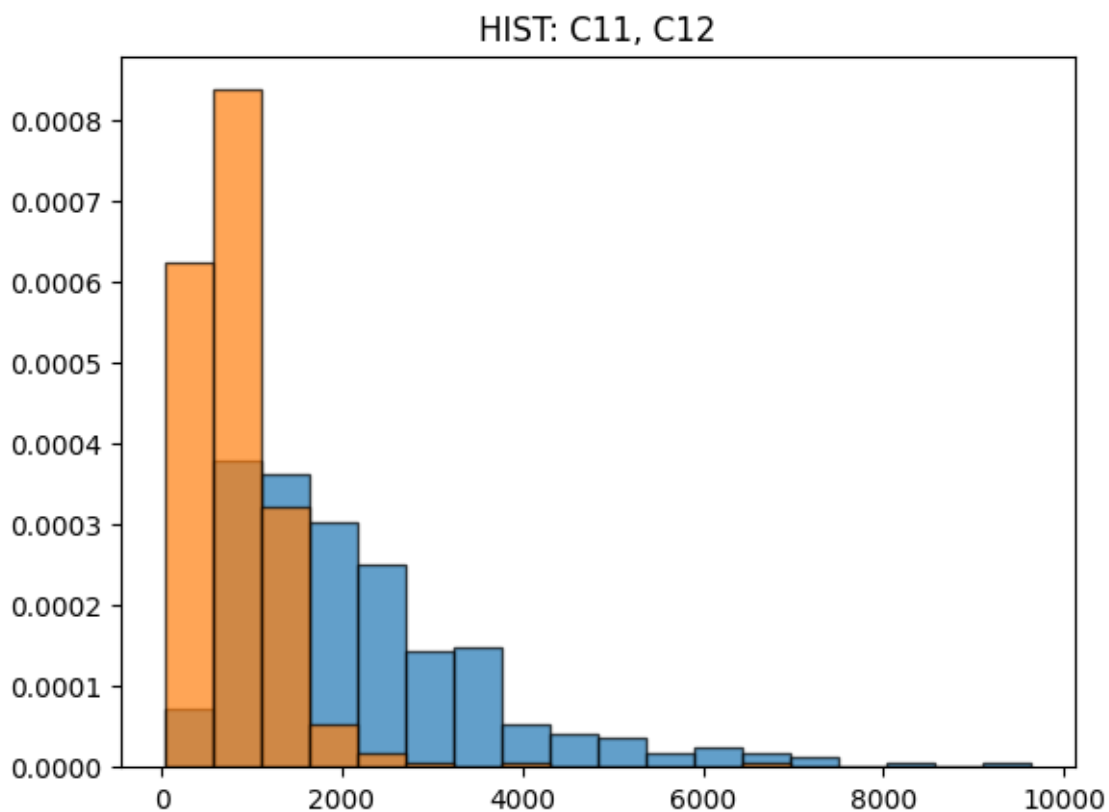
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
|-------------|--------------------------------|-------------------|
| 12 | По заполненности | 534 |

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 | 30 | 564 | 12 | 105 | 0.038 | 0.333 |
| 2 | 564 | 1098 | 64 | 141 | 0.203 | 0.448 |
| 3 | 1098 | 1632 | 61 | 54 | 0.194 | 0.171 |
| 4 | 1632 | 2166 | 51 | 9 | 0.162 | 0.029 |
| 5 | 2166 | 2700 | 42 | 3 | 0.133 | 0.01 |
| 6 | 2700 | 3234 | 24 | 1 | 0.076 | 0.003 |
| 7 | 3234 | 3768 | 25 | 0 | 0.079 | 0.0 |
| 8 | 3768 | 4302 | 9 | 1 | 0.029 | 0.003 |
| 9 | 4302 | 4836 | 7 | 0 | 0.022 | 0.0 |
| 10 | 4836 | 5370 | 6 | 0 | 0.019 | 0.0 |

| | | | | | | |
|----|------|------|----|---|-------|-------|
| 11 | 5370 | 5904 | 3 | 0 | 0.01 | 0.0 |
| 12 | 5904 | 9642 | 11 | 1 | 0.035 | 0.003 |

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 | 243.364 | 6.8e-46 | H0 отвергается | Признаки не однородны |
| 0.05 | | | H0 отвергается | Признаки не однородны |
| 0.1 | | | H0 отвергается | Признаки не однородны |

Вывод (в терминах предметной области)

В результате проведённого в п.5 статистического анализа обнаружено, что признаки С11(Потребляемый с пищей бета-каротин (мкг в день)) и С12 (Потребляемый с пищей ретинол (мкг в день)) не однородны.

6. Таблицы сопряжённости

Факторный признак x – С2

Результативный признак y – С5

Объёмы выборок – 315

Статистическая гипотеза – $H_0: F_Y(y|X = x_1) = \dots = F_Y(y|X = X_k) = F_Y(y)$

$$H': \neg H_0$$

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

| | Выражение | Пояснение использованных обозначений |
|--|--|--|
| Формула расчета статистики критерия | $Z = \sum_{i=1}^k \sum_{j=1}^l \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$ | n – наблюдаемые частоты, m – теоретические |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $\chi^2((k - 1) * (l - 1))$ | |
| Формула расчета критической точки | $\chi^2_{1-\alpha}(k - 1)(l - 1)$ | |
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |

б) Построить эмпирическую таблицу сопряжённости

| $x \backslash y$ | No | Not often | Often | Σ |
|------------------|-----|-----------|-------|----------|
| Female | 87 | 77 | 109 | 273 |
| Male | 24 | 5 | 13 | 42 |
| Σ | 111 | 82 | 122 | 315 |

в) Построить теоретическую таблицу сопряжённости

| $x \backslash y$ | No | Not often | Often | Σ |
|------------------|------|-----------|---------|----------|
| Female | 96.2 | 71.067 | 105.733 | 273 |

| | | | | |
|----------|------|--------|--------|-----|
| Male | 14.8 | 10.933 | 16.267 | 42 |
| Σ | 111 | 82 | 122 | 315 |

г) Проверить статистические гипотезы

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|---|
| 0.01 | 11.071 | 0.0039 | H0 принимается | Можно считать, что признаки независимы |
| 0.05 | | | H0 отклоняется | Нельзя считать, что признаки независимы |
| 0.1 | | | H0 отклоняется | Нельзя считать, что признаки независимы |

Вывод (в терминах предметной области)

В результате проведённого в п.6 статистического анализа обнаружено, что распределение признака С2 (пол) слабо зависит от признака С5 (употребление витаминов), при уровне значимости 0.01 можно считать что зависимости нет, а при уровнях 0.05 и 0.01 можно предполагать что статистическая связь имеется.

7. Дисперсионный анализ

Факторный признак x – СЗ

Результативный признак y – С14

Число вариантов факторного признака – 3

Объёмы выборок – 315

Статистическая гипотеза –

$$H_0 : F_{X_1}(x) = \dots = F_{X_K}(x) = F_X(x)$$

а) Рассчитать групповые выборочные характеристики

| № п/п | Вариант факторного признака | Объём выборки | Групповые средние | Групповые дисперсии |
|-------|-----------------------------|---------------|-------------------|---------------------|
| 1 | Current Smoker | 43 | 563.069 | 42674.399 |
| 2 | Former | 115 | 644.243 | 53438.466 |
| 3 | Never | 157 | 583.305 | 35209.931 |

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

| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
|---------------------|---|------------------------|-----------------------|
| Факторный признак | $D_b^* = \frac{1}{n} \sum_{k=1}^K n_k (\bar{x}_k - \bar{x})^2$ | $K - 1$ | $\frac{n}{K-1} D_b^*$ |
| Остаточные признаки | $D_w^* = \frac{1}{n} \sum_{k=1}^K n_k \tilde{\sigma}_k^2$ | $N - K$ | $\frac{n}{n-K} D_w^*$ |
| Все признаки | $D_X^* = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} (x_i^{(k)} - \bar{x})^2$ | $N - 1$ | $\frac{n}{n-1} D_X^*$ |

в) Рассчитать показатели вариации, используемые в дисперсионном анализе

| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
|-------------------|---------------------|------------------------|--------------------|
| Факторный признак | 1031.932 | 2 | 162529.436 |

| | | | |
|---------------------|-----------|-----|-----------|
| Остаточные признаки | 42883.752 | 312 | 43296.096 |
| Все признаки | 43637.319 | 314 | 44056.908 |

г) Проверить правило сложения дисперсий

| | | | | |
|------------|--------------------|-----------------------|------------------|--|
| Показатель | $D_{\text{межгр}}$ | $D_{\text{внутригр}}$ | $D_{\text{общ}}$ | $D_{\text{межгр}} + D_{\text{внутригр}}$ |
| Значение | 1031.932 | 42883.752 | 43637.319 | 43915.684 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

| | | |
|---------------------------------------|-------------------------------------|----------|
| Показатель | Формула расчета | Значение |
| Эмпирический коэффициент детерминации | $\eta^2 = \frac{D_b^*}{D_x^*}$ | 0.024 |
| Эмпирическое корреляционное отношение | $\eta = \sqrt{\frac{D_b^*}{D_x^*}}$ | 0.155 |

е) Охарактеризовать тип связи между факторным и результативным признаками

практически отсутствует

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

| | | |
|--|---------------------------------------|---|
| | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия | $F = \frac{D_b^*/(K-1)}{D_w^*/(n-K)}$ | K – количество вариантов факторного признака, n – размер. D_b^* – межгрупповая дисперсия, D_w^* – внутригрупповая дисперсия |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $F(K-1, n-K)$ | |
| Формула расчета критической точки | $F_{1-\alpha}(k-1, n-k)$ | |

| | | |
|----------------------------|-------------------------|--|
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |
|----------------------------|-------------------------|--|

з) Проверить статистическую гипотезу дисперсионного анализа

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|---|
| 0.01 | 3.790 | 0.023 | H0 принимается | Можно считать, что признаки независимы |
| 0.05 | | | H0 отклоняется | Нельзя считать, что признаки независимы |
| 0.1 | | | H0 отклоняется | Нельзя считать, что признаки независимы |

Вывод (в терминах предметной области)

В результате проведённого в п.7 статистического анализа обнаружено, что распределение признака С3 (статус курения) слабо зависит от признака С14 (Ретинол плазмы (нг/мл)), при уровне значимости 0.01 можно считать что зависимости нет, а при уровнях 0.05 и 0.01 можно предполагать что статистическая связь имеется.

8. Корреляционный анализ

8.1. Расчёт парных коэффициентов корреляции

Анализируемый признак 1 – С11

Анализируемый признак 2 –С12

Объёмы выборок – 315

а) Рассчитать точечные оценки коэффициентов корреляции

| | Формула расчета | Значение |
|---|---|----------|
| Линейный коэффициент корреляции | $\rho_{XY}^* = \frac{k_{XY}^*}{\sigma_X^* \sigma_Y^*}$ $k_{XY}^* = \frac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}).$ | 0.052 |
| Ранговый коэффициент корреляции по Спирмену | $\rho_{XY}^{(sp)} = \frac{\sum_{i=1}^n (R_i - \bar{r})(S_i - \bar{s})}{\sqrt{\sum_{i=1}^n (R_i - \bar{r})^2 \sum_{i=1}^n (S_i - \bar{s})^2}}$ | 0.196 |
| Ранговый коэффициент корреляции по Кендаллу | $\tau_{XY} = \frac{N^+ - N^-}{\frac{1}{2}n(n-1)}$ | 0.132 |

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

| Граница доверительного интервала | Формула расчета |
|----------------------------------|---|
| Нижняя граница | $\left(\rho_{XY}^* + \frac{\rho_{XY}^*(1-(\rho_{XY}^*)^2)}{2n} - u_{1-\alpha/2} \frac{1-(\rho_{XY}^*)^2}{\sqrt{n}}; \right.$ |
| Верхняя граница | $\left. \rho_{XY}^* + \frac{\rho_{XY}^*(1-(\rho_{XY}^*)^2)}{2n} + u_{1-\alpha/2} \frac{1-(\rho_{XY}^*)^2}{\sqrt{n}} \right)$ |

в) Рассчитать доверительные интервалы для линейного коэффициента корреляции

| Граница доверительного интервала | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|----------------------------------|-----------------|-----------------|----------------|
| Нижняя граница | -0.091 | -0.057 | -0.039 |
| Верхняя граница | 0.197 | 0.163 | 0.145 |

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 |
|--|--|--|
| $H_0: \rho = 0$ $H': \rho \neq 0$ | $Z = \frac{\rho_{XY}^*}{\sqrt{1 - (\rho_{XY}^*)^2}} \sqrt{n - 2}$ | $f_Z(z H_0) \sim T(n - 2)$ |
| $H_0: r^{(cn)} = 0$ $H': r^{(cn)} \neq 0$ | $Z = \frac{\tilde{\rho}_{XY}^{(sp)}}{\sqrt{1 - \tilde{\rho}_{XY}^{(sp)2}}} \sqrt{n - 2}$ | $f_Z(z H_0) \sim T(n - 2)$ |
| $H_0: r^{(кен)} = 0$ $H': r^{(кен)} \neq 0$ | $Z = \tilde{\tau}_{XY} \sqrt{\frac{9n(n+1)}{2(2n+5)}}$ | $f_Z(z H_0) \sim N(0, 1)$ |

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 | 0.936 | 0.349 | Н0 принимается | Можно считать, что признаки независимы |
| $H_0: r^{(cn)} = 0$ $H': r^{(cn)} \neq 0$ | 0.1 | 3.541 | 0.000 | Н0 отклоняется | Нельзя считать, что признаки независимы |
| $H_0: r^{(кен)} = 0$ $H': r^{(кен)} \neq 0$ | 0.1 | 3.517 | 0.000 | Н0 отклоняется | Нельзя считать, что признаки независимы |

8.2. Расчёт множественных коэффициентов корреляции

Анализируемый признак 1 – С8

Анализируемый признак 2 – С10

Анализируемый признак 3 – С13

Объёмы выборок – 315

а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу

| Признак \ Признак | C8 | C10 | C13 |
|-------------------|-------|--------|--------|
| C8 | 1 | 0.148 | 0.128 |
| C10 | 0.148 | 1 | -0.092 |
| C13 | 0.128 | -0.092 | 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 | C8 | C10 | C13 |
|-------------------|---------|---------|--------|
| C8 | – | 8.7e-05 | 0.0007 |
| C10 | 8.7e-05 | – | 0.014 |
| C13 | 0.0007 | 0.014 | – |

в) Рассчитать точечную оценку коэффициента конкордации

| | Формула расчета | Значение |
|-------------------------|--|----------|
| Коэффициент конкордации | $W = \frac{12}{k^2(n^3 - n)} \sum_{i=1}^n \left(\sum_{j=1}^k R_{ij} - \frac{k(n+1)}{2} \right)^2$ | 0.393 |

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

| | Выражение | Пояснение использованных обозначений |
|--|--------------------------|--|
| Формула расчета статистики критерия | $n(k-1)W$ | N – размер выборок, k – количество выборок, W – выборочный коэффициент конкордации |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $\chi^2(n-1)$ | |
| Формула расчета критической точки | $\chi^2_{1-\alpha}(n-1)$ | |

| | | |
|----------------------------|-------------------------|--|
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |
|----------------------------|-------------------------|--|

д) Проверить значимость коэффициента конкордации

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|--|
| 0.01 | 248.04 | 0 | H0 отклоняется | Нельзя считать, что признаки независимы в совокупности |
| 0.05 | | | H0 отклоняется | Нельзя считать, что признаки независимы в совокупности |
| 0.1 | | | H0 отклоняется | Нельзя считать, что признаки независимы в совокупности |

Вывод (в терминах предметной области)

В результате проведённого в п.8 статистического анализа обнаружено, что признаки С8 (Граммы клетчатки, потребляемой в день), С10 (Потребляемый холестерин (мг в день)), С13(Бета-каротин плазмы (нг/мл)) имеют какую-то взаимосвязь в совокупности.

9. Регрессионный анализ

9.1 Простейшая линейная регрессионная модель

Факторный признак x – С6

Результативный признак y – С14

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 | $\tilde{\beta}_0 = \bar{y} - \rho_{XY}^* \frac{\sigma_Y^*}{\sigma_X^*} \bar{x}$ | 643.242 |
| β_1 | $\tilde{\beta}_1 = \rho_{XY}^* \frac{\sigma_Y^*}{\sigma_X^*}$ | -0.022 |

б) Записать точечную оценку уравнения регрессии

$$f(x) = 643.2421313446065 * x - 0.02251498706688208$$

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

| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
|---------------------|---|------------------------|----------------------------|
| Факторный признак | $D_{Y X} = M \left[(f(X) - m_Y)^2 \right]$ | $k-1$ | $\frac{n}{k-1} D_{Y X}^*$ |
| Остаточные признаки | $D_{resY}^* = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i, \beta_0, \dots, \beta_{k-1}))^2$ | $n-k$ | $\frac{n}{n-k} D_{resY}^*$ |
| Все признаки | $D_Y^* = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ | $n-1$ | $\frac{n}{n-1} D_Y^*$ |

г) Рассчитать показатели вариации, используемые в регрессионном анализе

| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
|-------------------|---------------------|------------------------|--------------------|
| Факторный признак | 233.897 | 1 | 73677.444 |

| | | | |
|---------------------|-----------|-----|-----------|
| Остаточные признаки | 43264.891 | 313 | 43541.344 |
| Все признаки | 43637.319 | 314 | 43776.291 |

д) Проверить правило сложения дисперсий

| | | | | |
|------------|------------|-----------|-----------|----------------------|
| Показатель | $D_{регp}$ | $D_{ост}$ | $D_{общ}$ | $D_{регp} + D_{ост}$ |
| Значение | 233.897 | 43264.891 | 43637.319 | 43498.787 |

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

| Показатель | Формула расчета | Значение |
|--------------------------|--|----------|
| Коэффициент детерминации | $R_{Y X}^2 = \frac{D_{Y X}}{D_Y}$ | 0.005 |
| Корреляционное отношение | $R_{Y X} = \sqrt{\frac{D_{Y X}}{D_Y}}$ | 0.073 |

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

практически отсутствует

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 | $\tilde{\beta}_0 - t_{1-\alpha/2}(n-2) \sqrt{\tilde{D}_{resY} \sqrt{\frac{\sum_{i=1}^n x_i^2}{n^2 D_X^*}}}$ |

| | | |
|-----------|-----------------|---|
| | Верхняя граница | $\tilde{\beta}_0 + t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{resY}}\sqrt{\frac{\sum_{i=1}^n x_i^2}{n^2 D_X^*}}$ |
| β_1 | Нижняя граница | $\tilde{\beta}_1 - t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{resY}}\sqrt{\frac{1}{n D_X^*}} ;$ |
| | Верхняя граница | $\tilde{\beta}_1 + t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{resY}}\sqrt{\frac{1}{n D_X^*}} ,$ |

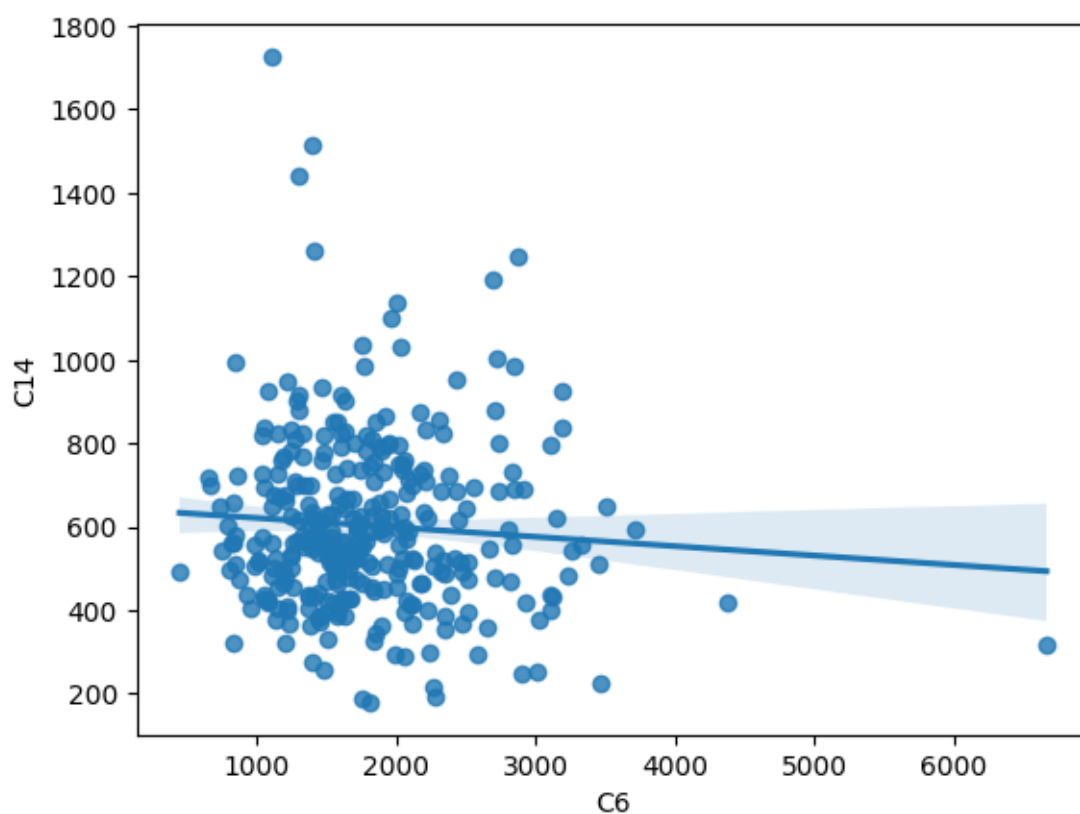
б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

| Параметр | Границы доверительного интервала | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.1$ |
|-----------|----------------------------------|-----------------|-----------------|----------------|
| β_0 | Нижняя граница | 557.219 | 577.933 | 588.483 |
| | Верхняя граница | 729.264 | 708.550 | 698.001 |
| β_1 | Нижняя граница | 643.197 | 643.208 | 643.213 |
| | Верхняя граница | 643.286 | 643.276 | 643.270 |

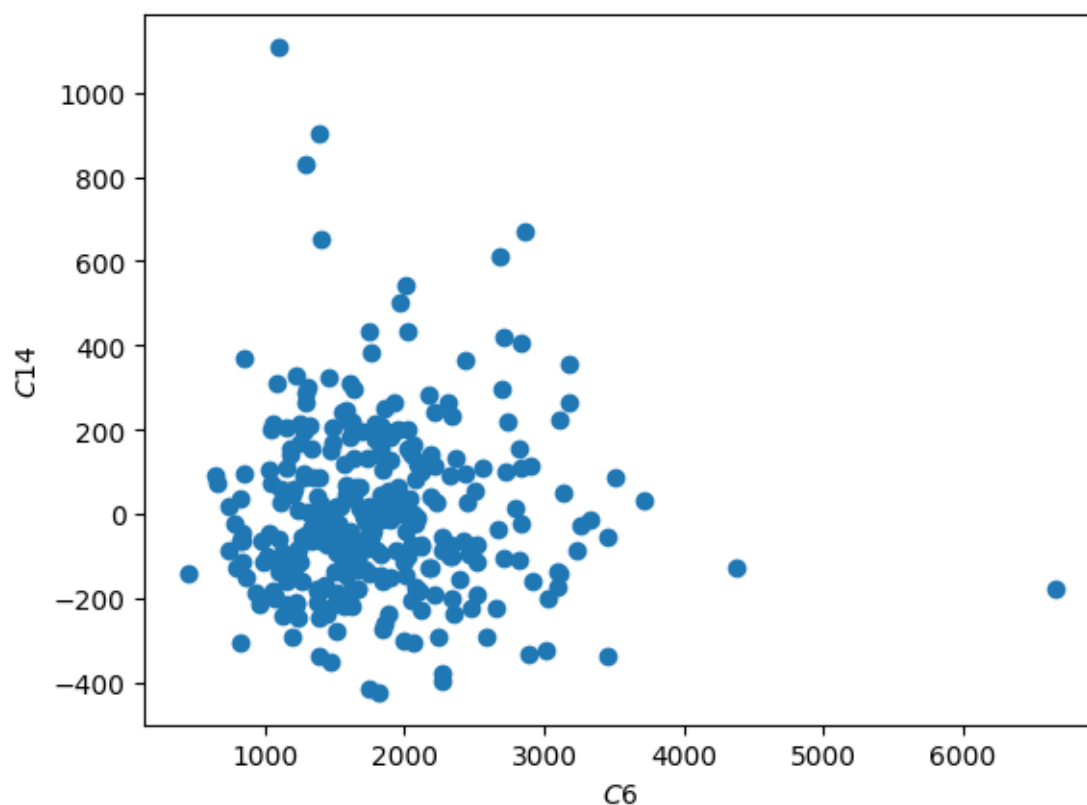
в) Привести формулы расчёта доверительного интервала для значений регрессии $f(x)$

| Границы доверительного интервала | Формула расчета |
|----------------------------------|--|
| Нижняя граница $f_{low}(x)$ | $\tilde{f}(x) - t_{1-\alpha/2}(n-2) \sqrt{\tilde{D}_{resY}} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{nD_X^*}}$ |
| Верхняя граница $f_{high}(x)$ | $\tilde{f}(x) + t_{1-\alpha/2}(n-2) \sqrt{\tilde{D}_{resY}} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{nD_X^*}}$ |

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 – $H_0: \beta_1 = 0$
 $H': \beta_1 \neq 0$

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

| | Выражение | Пояснение использованных обозначений |
|--|--|--------------------------------------|
| Формула расчета статистики критерия | $\frac{\overline{D}_{\text{регр}}}{\frac{\overline{D}_{\text{ост}}}{(n-2)}}$ | $R_{Y X}$ – коэффициент детерминации |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $f_Z(z H_0) \sim F(1, n - 2)$ | |
| Формула расчета критической точки | $Z_1 = F_{1-\alpha}(n_1 - 1, n_2 - 1), Z < Z_1$ | |

| | | |
|----------------------------|-------------------------|--|
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |
|----------------------------|-------------------------|--|

б) Проверить значимость линейной регрессионной модели

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|--------------------------------|
| 0.01 | 1.692 | 0.194 | Н0 принимается | Можно считать, что $b1 \neq 0$ |
| 0.05 | | | Н0 принимается | Можно считать, что $b1 \neq 0$ |
| 0.1 | | | Н0 принимается | Можно считать, что $b1 \neq 0$ |

9.2 Линейная регрессионная модель общего вида

Факторный признак x – С6

Результативный признак y – С14

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 | $F = \begin{pmatrix} \varphi_0(x_1) & \dots & \varphi_{k-1}(x_1) \\ \varphi_0(x_2) & \dots & \varphi_{k-1}(x_2) \\ \dots & \dots & \dots \\ \varphi_0(x_n) & \dots & \varphi_{k-1}(x_n) \end{pmatrix}$ | 5.911 e+02 |
| β_1 | $\beta = (F^T F)^{-1} F^T y$ | 2.838 e-02 |
| β_2 | $F = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}$ | -1.065 e-05 |

б) Записать точечную оценку уравнения регрессии

$$f(x) = 5.91101645e+02 + 2.83858107e-02 * x - 1.06551135e-05 * x ** 2$$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
|---------------------|---------------------|------------------------|--------------------|
| Факторный признак | 401.697 | 2 | 253069.181 |
| Остаточные признаки | 43097.090 | 312 | 43511.485 |
| Все признаки | 43637.319 | 314 | 43776.291 |

г) Проверить правило сложения дисперсий

| Показатель | $D_{рег}$ | $D_{ост}$ | $D_{общ}$ | $D_{рег} + D_{ост}$ |
|------------|-----------|------------|-----------|---------------------|
| Значение | 401.697 | 43097.0901 | 43637.319 | 43498.787 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

| Показатель | Формула расчета | Значение |
|--------------------------|--|----------|
| Коэффициент детерминации | $R^2_{Y X} = \frac{D_{Y X}}{D_Y}$ | 0.009 |
| Корреляционное отношение | $R_{Y X} = \sqrt{\frac{D_{Y X}}{D_Y}}$ | 0.095 |

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

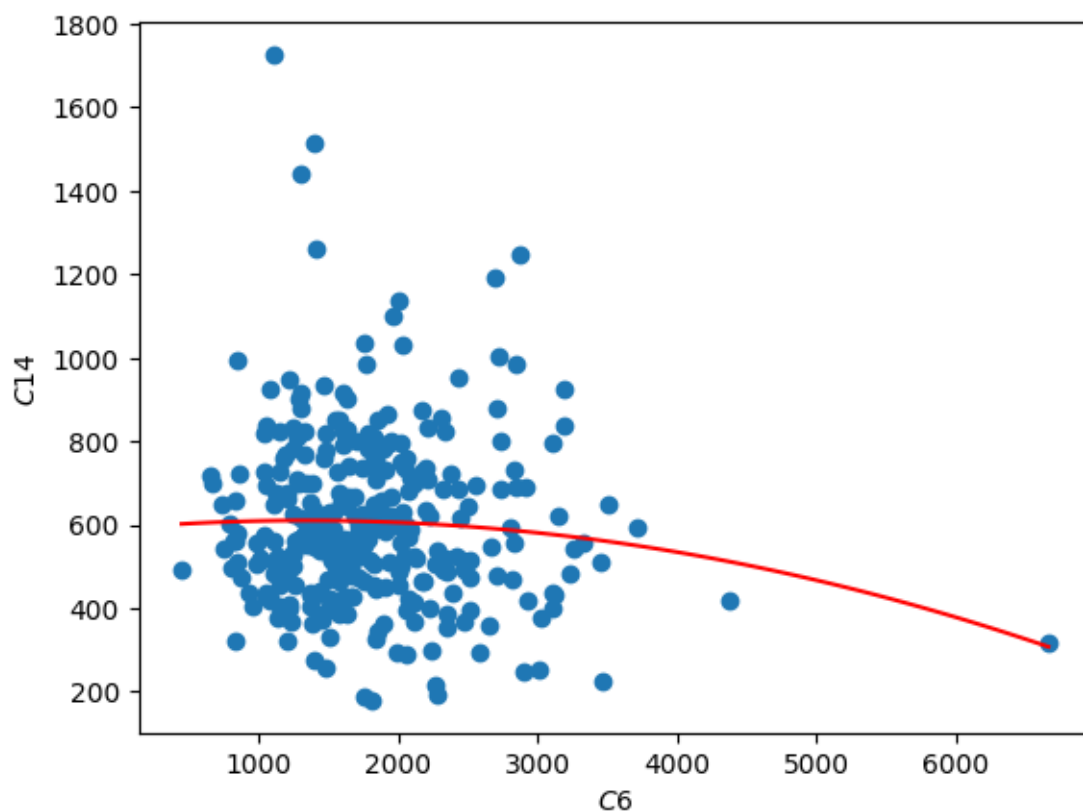
практически отсутствует

9.2.2. Интервальные оценки линейной регрессионной модели

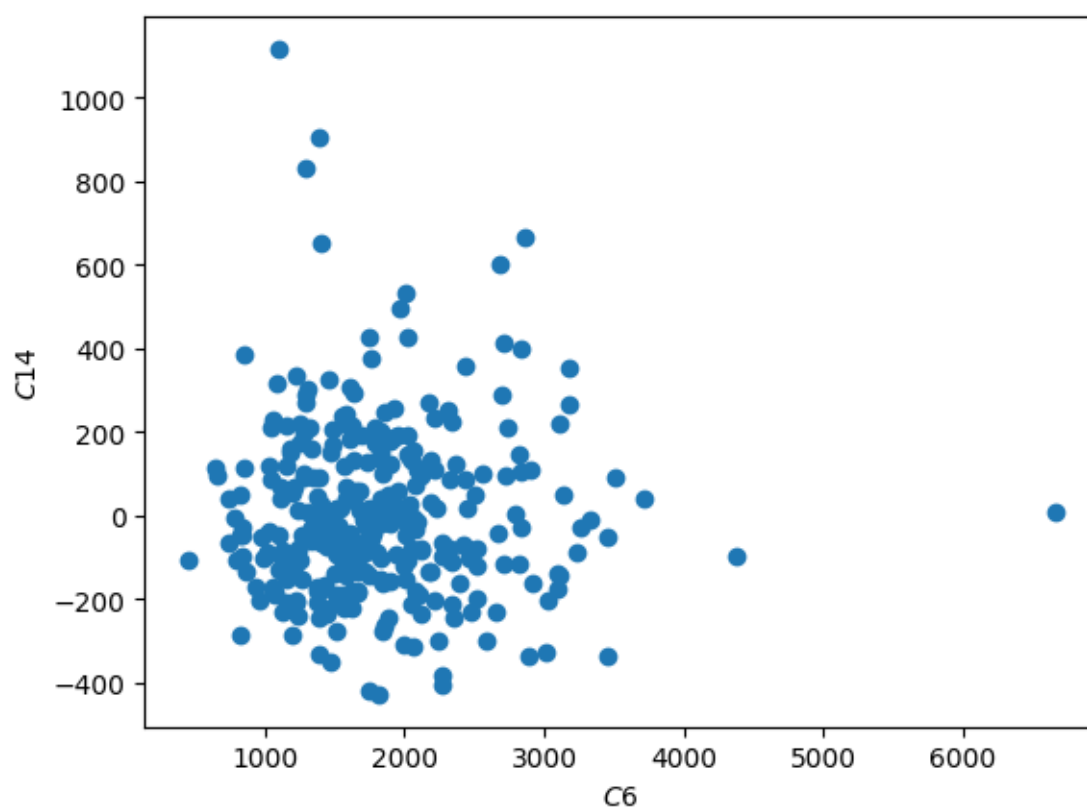
а) Привести формулы расчёта доверительного интервала для значений регрессии $f(x)$

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

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 – $H_0 : \beta_1 = \beta_2 = 0$
 $H' : \text{не } H_0$

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

| | Выражение | Пояснение использованных обозначений |
|--|---|--------------------------------------|
| Формула расчета статистики критерия | $Z = \frac{R_{Y X}^{2*}/(k-1)}{(1-R_{Y X}^{2*})/(n-k)}$ | |
| Закон распределения статистики критерия при условии истинности основной гипотезы | $f_Z(z H_0) \sim F(k-1, n-k)$ | |
| Формула расчета критической точки | $f_{1-\alpha}(1, n-2)$ | |
| Формула расчета p -value | $p_value = 1 - F_z(Z)$ | |

б) Проверить значимость линейной регрессионной модели

| Уровень значимости | Выборочное значение статистики критерия | p -value | Статистическое решение | Вывод |
|--------------------|---|------------|------------------------|-----------------------------------|
| 0.01 | 1.449 | 0.236 | Н0 принимается | Можно считать, что коэфф $\neq 0$ |
| 0.05 | | | Н0 принимается | Можно считать, что коэфф $\neq 0$ |
| 0.1 | | | Н0 принимается | Можно считать, что коэфф $\neq 0$ |

9.3 Множественная линейная регрессионная модель

Факторный признак 1 x_1 – С6

Факторный признак 2 x_2 – С12

Результативный признак y – С14

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 | $F = \begin{pmatrix} 1 & x_1^{(1)} & x_1^{(2)} \\ 1 & x_2^{(1)} & x_2^{(2)} \\ \dots & \dots & \dots \\ 1 & x_n^{(1)} & x_n^{(2)} \end{pmatrix}$ | 6.461e+02 |
| β_1 | $\tilde{\beta} = (F^T F)^{-1} y,$ | -1.760-02 |
| β_2 | | -1.408e-02 |

б) Записать точечную оценку уравнения регрессии

$$f(x_1, x_2) = 6.46147981e+02 + -1.76059169e-02 * x_1 + -1.40813648e-02 * x_2$$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
|---------------------|---------------------|------------------------|--------------------|
| Факторный признак | 291.415 | 2 | 183591.712 |
| Остаточные признаки | 43207.372 | 312 | 43622.827 |
| Все признаки | 43637.319 | 314 | 43776.291 |

г) Проверить правило сложения дисперсий

| Показатель | $D_{рег}$ | $D_{ост}$ | $D_{общ}$ | $D_{рег} + D_{ост}$ |
|------------|-----------|-----------|-----------|---------------------|
| Значение | 291.415 | 43207.372 | 43637.319 | 43498.787 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

| Показатель | Формула расчета | Значение |
|--|--|----------|
| Множественный коэффициент детерминации | $R_{Y X_1, \dots, X_m}^{2*} = \frac{D_{Y X_1, \dots, X_m}^*}{D_Y^*} = 1 - \frac{D_{resY}^*}{D_Y^*}$ | 0.006 |
| Множественное корреляционное отношение | $R_{Y X_1, X_2}^* = \sqrt{\frac{(\rho_{YX_1}^*)^2 + (\rho_{YX_2}^*)^2 - 2\rho_{YX_1}^* \rho_{YX_2}^* \rho_{X_1X_2}^*}{1 - (\rho_{X_1X_2}^*)^2}}$ | 0.081 |

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

| |
|-------------------------|
| практически отсутствует |
|-------------------------|

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 | 233.897 | 401.697 | 401.697 |
| Остаточные признаки | 43264.891 | 43097.090 | 43097.090 |
| Все признаки | 43637.319 | 43637.319 | 43637.319 |

б) Сводная таблица свойств различных регрессионных моделей

| Свойство | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
|-----------------------|----------------------------|---------------------------------------|-------------------------------|
| Точность | не точная | не точная | не точная |
| Значимость | значима | значима | значима |
| Адекватность | адекватна | адекватна | адекватна |
| Степень тесноты связи | практически отсутствует | практически отсутствует | практически отсутствует |

Вывод (в терминах предметной области)

В результате проведённого в п.9 статистического анализа обнаружено, что для С6 и С14 нет подходящей линейной модели (так как признаки практически не имеют зависимости).