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 Б20-514

Моисеенко Олеси Игоревны

Вариант № 15

Оценка: _____

Подпись: _____

2022 г.

ОТЧЕТ № 1

по теме «Проверка статистических гипотез»

Вариант № 15

ФИО студента Моисеенко Олеся Игоревна группа Б20-514

Оценка: _____ Подпись:_____

Результаты статистических тестов:

№ задания	Проверяемая гипотеза H_0	Критерий	Статистическое решение ($\alpha = 0.1$)	Вывод
4.1	$H_0 : F(x) \square N$	Хи-квадрат	H_0 отклоняется	распределение C7 не является нормальным
4.2	$H_0 : F(x) \square N$	Харке-Бера	H_0 отклоняется	распределение C7 не является нормальным
5.1	H_0: $F_X(\xi) = F_Y(\xi)$	знаков	H_0 отклоняется	C13 и C14 имеют различные распределения
5.2	H_0: $F_X(\xi) = F_Y(\xi)$	Хи-квадрат	H_0 отклоняется	C13 и C14 имеют различные распределения

Выводы:

В результате проведённого в п.4 статистического анализа обнаружено, что количество граммов жира, потребляемых в день пациентами, не является нормально распределённой случайной величиной.
В результате проведённого в п.5 статистического анализа обнаружено, что концентрации(ng/ml) в плазме бета-каротина и ретинола не являются случайными величинами, имеющими одинаковый закон распределения.

ОТЧЕТ № 2

по теме «Анализ статистических взаимосвязей»

Вариант №15

ФИО студента Моисеенко Олеся Игоревна группа Б20-514

Оценка: _____ Подпись:_____

Результаты статистических тестов:

№ задания	Проверяемая гипотеза H_0	Критерий	Статистическое решение ($\alpha = 0.1$)	Вывод
6	H_0: $F_Y(y\|X = x_1) = \ldots = F_Y(y\|X = x_n) = F_Y(y)$ H': $\neg H_0$	Хи-квадрат	H_0 принимается	Статистическая связь отсутствует
7	H_0: $F_{X_1}(x) = \ldots = F_{X_K}(x) = F_X(x)$ (H_0: $m_1 = \ldots = m_K$) H': $\neg H_0$	ANOVA	H_0 отклоняется	Статистическая связь присутствует

Выводы:

В результате проведённого в п.6 статистического анализа обнаружено, что пол пациентов не влияет на их отношение к курению.
В результате проведённого в п.7 статистического анализа обнаружено, что частота потребления витаминов пациентами оказывает слабое влияние на концентрацию бета-каротина(ng/ml) в плазме их крови.

ОТЧЕТ № 3

по теме «Основы регрессионного анализа»

Вариант №15

ФИО студента Моисеенко Олеся Игоревна группа Б20-514

Оценка: _____ Подпись:_____

Сводная таблица свойств различных регрессионных моделей:

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	2.2 %	6.4 %	23.8 %
Значимость	нет	нет	нет
Адекватность	-	-	-
Степень тесноты связи	Отсутствует	Отсутствует	Слабая

Выводы:

В результате проведённого в п.8 статистического анализа обнаружено, что между концентрациями(ng/ml) бета-каротина и ретинола в плазме крови нет зависимости. А также, что статистическая связь между потребляемыми пациентами количеством жира(grams per day), алкогольных напитков(number per week) и количеством диетического ретинола(mcg per day) отсутствует. Однако, из пункта б) видно, что между количеством жира(grams per day) и количеством диетического ретинола(mcg per day), потребляемых пациентами, зависимость есть, в то время как между потребляемыми количествами жира(grams per day) и алкогольных напитков(number per week), алкогольных напитков(number per week) и диетического ретинола(mcg per day) связь отсутствует.
В результате проведённого в п.9 статистического анализа обнаружено, что количество калорий, потребляемых пациентами в день, не влияет на концентрацию бета-каротина(ng/ml) в плазме их крови. Однако количество потребляемых бета-каротина с пищей(mcg per day) и калорий в день оказывает слабое воздействие на концентрацию бета-каротина(ng/ml).

1. Описательные статистики

1.1. Выборочные характеристики

Анализируемый признак 1 – C7

Анализируемый признак 2 – C9

Анализируемый признак 3 – C12

а) Привести формулы расчёта выборочных характеристик

Выборочная хар-ка	Формула расчета
Объём выборки	n
Среднее	$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$
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	77.03333333333333	3.2793650793650793	832.7142857142857
Выборочная дисперсия	1144.431210191083	151.8533626529168	347261.5614194723
Выборочное среднеквадратическое отклонение	33.82944294828224	12.32287964125743	589.2890304591392
Выборочный коэффициент асимметрии	1.0989962157492976	13.757134658078	4.452504551684713
Выборочный эксцесс	1.9647991849352238	217.81506388614636	37.44713035757864

1.2. Группировка и гистограммы частот

Анализируемый признак – C7

Объём выборки – 315

а) Выбрать число групп

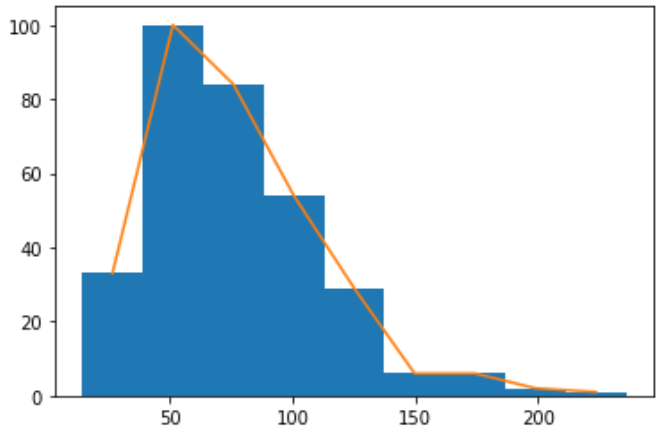
Число групп	Обоснование выбора числа групп	Ширина интервалов
9	По формуле Стерджесса: k=[1+log2n]	24

б) Построить таблицу частот

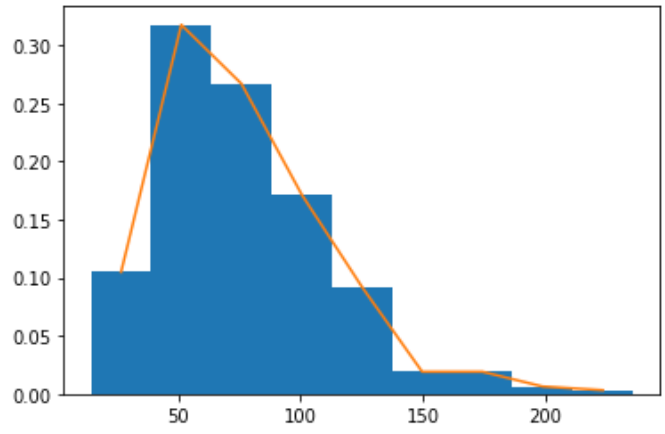
Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Накопл. частота	Относит. накопл. частота
1	14.4	39.01111111	33	0.1047619	33	0.1047619
2	39.01111111	63.62222222	100	0.31746032	133	0.42222222
3	63.62222222	88.23333333	84	0.26666667	217	0.68888889
4	88.23333333	112.84444444	54	0.17142857	271	0.86031746
5	112.84444444	137.45555556	29	0.09206349	300	0.95238095
6	137.45555556	162.06666667	6	0.01904762	306	0.97142857
7	162.06666667	186.67777778	6	0.01904762	312	0.99047619
8	186.67777778	211.28888889	2	0.00634921	314	0.9968254
9	211.28888889	235.9	1	0.0031746	315	1

в) Построить гистограммы частот и полигоны частот

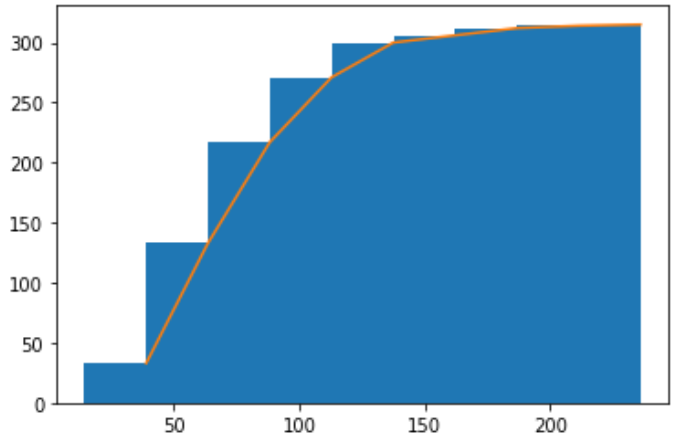
Гистограмма и полигон частот



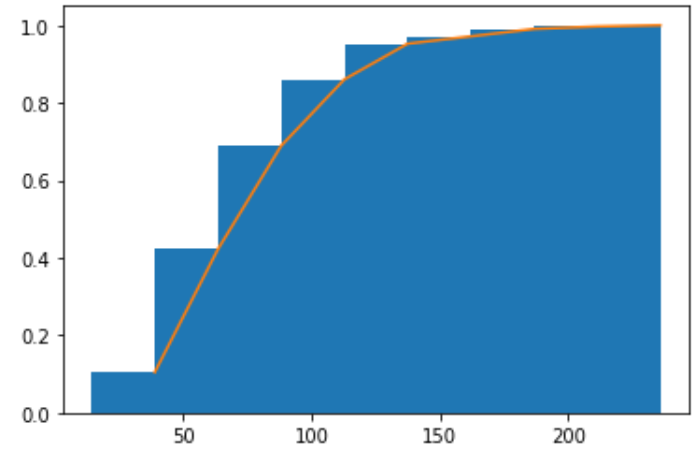
Гистограмма и полигон относительных частот



Гистограмма и полигон накопленных частот

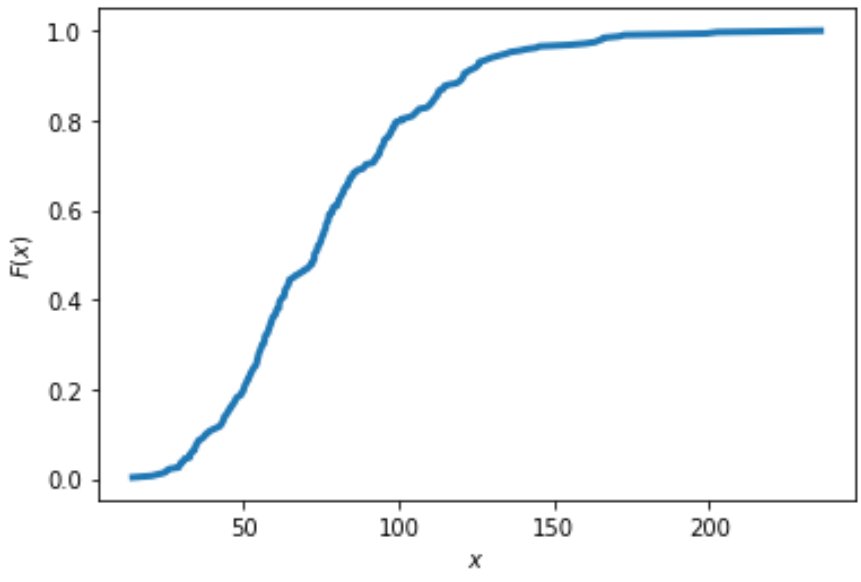


Гистограмма и полигон накопленных относительных частот



г) Построить график эмпирической функции распределения

Эмпирическая функция распределения



2. Интервальные оценки

2.1. Доверительные интервалы для мат. ожидания

Анализируемый признак – C7

Объём выборки – 315

Оцениваемый параметр – математическое 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	72.09359774349191	73.28304344314049	73.8888447091652
Верхняя граница	81.97306892317475	80.78362322352618	80.17782195750146

2.2. Доверительные интервалы для дисперсии

Анализируемый признак – C7

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	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	939.972084131803	984.571373041788	1008.4926769611158
Верхняя граница	1419.1945094555533	1346.872632211083	1311.7502714531515

2.3. Доверительные интервалы для разности мат. ожиданий

Анализируемый признак 1 – C13

Анализируемый признак 2 – C14

Объёмы выборок – 315

Оцениваемый параметр – разность математических ожиданий ($m_1 - m_2$)

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$(\bar{X}_1 - \bar{X}_2) - t_{1-\alpha/2}(n_1+n_2-2)S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
Верхняя граница	$(\bar{X}_1 - \bar{X}_2) + t_{1-\alpha/2}(n_1+n_2-2)S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	-453.3267319085694	-443.6263036888957	-438.6743954295366
Верхняя граница	-372.470093488256	-382.1705217079297	-387.1224299672888

2.4. Доверительные интервалы для отношения дисперсий

Анализируемый признак 1 – C13

Анализируемый признак 2 – C14

Объёмы выборок – 315

Оцениваемый параметр – отношение дисперсий σ_1^2/σ_2^2

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{S_1^2}{S_2^2} f_{\alpha/2}(n_2-1, n_1-1)$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
Верхняя граница	$\frac{S_1^2}{S_2^2} f_{1-\alpha/2}(n_2-1, n_1-1)$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	0.5734087875377081	0.6148952188376793	0.637234697462504
Верхняя граница	1.0271444406441992	0.9578439224965994	0.9242648755336237

3. Проверка статистических гипотез о математических ожиданиях и дисперсиях

3.1. Проверка статистических гипотез о математических ожиданиях

Анализируемый признак – C7

Объём выборки – 315

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	$Z = \frac{\overline{X} - m_0}{S/\sqrt{n}}$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z\|_{H_0} \sim T(n-1)$
Формулы расчета критических точек	$\pm t_{1-\alpha/2}(n-1)$
Формула расчета *p-value*	$2\min\left(F_Z(z \mid H_0), 1 - F_Z(z \mid H_0)\right)$

б) Выбрать произвольные значения m_0 и проверить статистические гипотезы

m_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
77	0.1	0.01748796098163266	0.9860584437550612	H_0 принимается	m = 77
250	0.1	-90.74502953369704	2.350509132273903e-227	H_0 отклоняется	m ≠250
860	0.1	-410.774715497593	0.0	H_0 отклоняется	m ≠860

3.2. Проверка статистических гипотез о дисперсиях

Анализируемый признак – C7

Объём выборки – 315

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	$Z = \frac{(n-1)S^2}{\sigma_0^2}$ $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z\big\vert_{H_0} \sim \chi^2(n-1)$
Формулы расчета критических точек	$\chi^2_{\alpha/2}(n\text{-}1), \chi^2_{1-\alpha/2}(n\text{-}1)$
Формула расчета *p-value*	$2\min\left(F_Z(z \mid H_0), 1 - F_Z(z \mid H_0)\right)$

б) Выбрать произвольные значения σ_0 и проверить статистические гипотезы

σ_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
34	0.1	310.857612456746 74	0.9209902160950361	H_0 принимается	$\sigma = 34$
340	0.1	3.108576124567475	4.2802552281498377 e-249	H_0 отклоняется	$\sigma \neq 340$
760	0.1	0.6221457756232689	0.0	H_0 отклоняется	$\sigma \neq 760$

3.3. Проверка статистических гипотез о равенстве математических ожиданий

Анализируемый признак 1 – C13

Анализируемый признак 2 – C14

Объёмы выборок – 315

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	Для более точной проверки поставленной в этой задаче статистической гипотезы необходимо сначала проверить гипотезу пункта 3.4: 1)Если $\sigma_1=\sigma_2$: $$Z = \frac{\overline{X}_1 - \overline{X}_2}{S/\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$ $$S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$$ 2) Если $\sigma_1 \neq \sigma_2$: $$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$
Закон распределения статистики критерия при условии истинности основной гипотезы	1) $$Z|_{H_0} \sim T(n_1 + n_2 - 2)$$ 2) $Z|_{H0} \sim T([1 / k])$ $$k = \frac{\left(\frac{S_1^2/n_1}{S_1^2/n_1+S_2^2/n_2}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2/n_2}{S_1^2/n_1+S_2^2/n_2}\right)^2}{n_2-1}$$
Формулы расчета критических точек	1)$\pm t_{1-\alpha}(n_1+n_2-2)$ 2) $\pm t_{1-\alpha}([1 / k])$
Формула расчета *p-value*	$$2\min\left(F_Z(z \mid H_0), 1-F_Z(z \mid H_0)\right)$$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	-26.387385396640326	2.7044815438723324 e-103	H_0 отклоняется	$m_1 \neq m_2$
0.05			H_0 отклоняется	$m_1 \neq m_2$
0.1			H_0 отклоняется	$m_1 \neq m_2$

3.4. Проверка статистических гипотез о равенстве дисперсий

Анализируемый признак 1 – C13

Анализируемый признак 2 – C14

Объёмы выборок – 315

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	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ $Z = \frac{S_1^2}{S_2^2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z\|_{H_0} \sim F(n_1 - 1, n_2 - 1)$
Формулы расчета критических точек	$f_{\alpha/2}(n_1\text{-}1, n_2\text{-}1);\ f_{1-\alpha/2}(n_1\text{-}1, n_2\text{-}1)$
Формула расчета *p-value*	$2\min\left(F_Z(z \mid H_0), 1 - F_Z(z \mid H_0)\right)$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.7674461859543552	0.01929088453087414	H_0 принимается	$\sigma_1=\sigma_2$ (но при малых значениях α более вероятна ошибка 2-го рода)
0.05			H_0 отклоняется	$\sigma_1\neq\sigma_2$
0.1			H_0 отклоняется	$\sigma_1\neq\sigma_2$

4. Критерии согласия

Анализируемый признак – C7

Объём выборки – 315

4.1. Критерий хи-квадрат

Теоретическое распределение – нормальное

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	$Z = n\sum_{i=1}^{k}\frac{(\tilde{p}_i - p_i)^2}{p_i} = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i}$	k – число интервалов; n_i – число элементов в i-м интервале; p_i – вероятности попадания в i-й интервал при условии истинности H_0; n – объём выборки; r – число неизвестных параметров распределения.
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(k-r-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(k - r - 1)$	
Формула расчета *p-value*	$p\text{-}value = 1 - F_Z(z \mid H_0).$	

б) Выбрать число групп

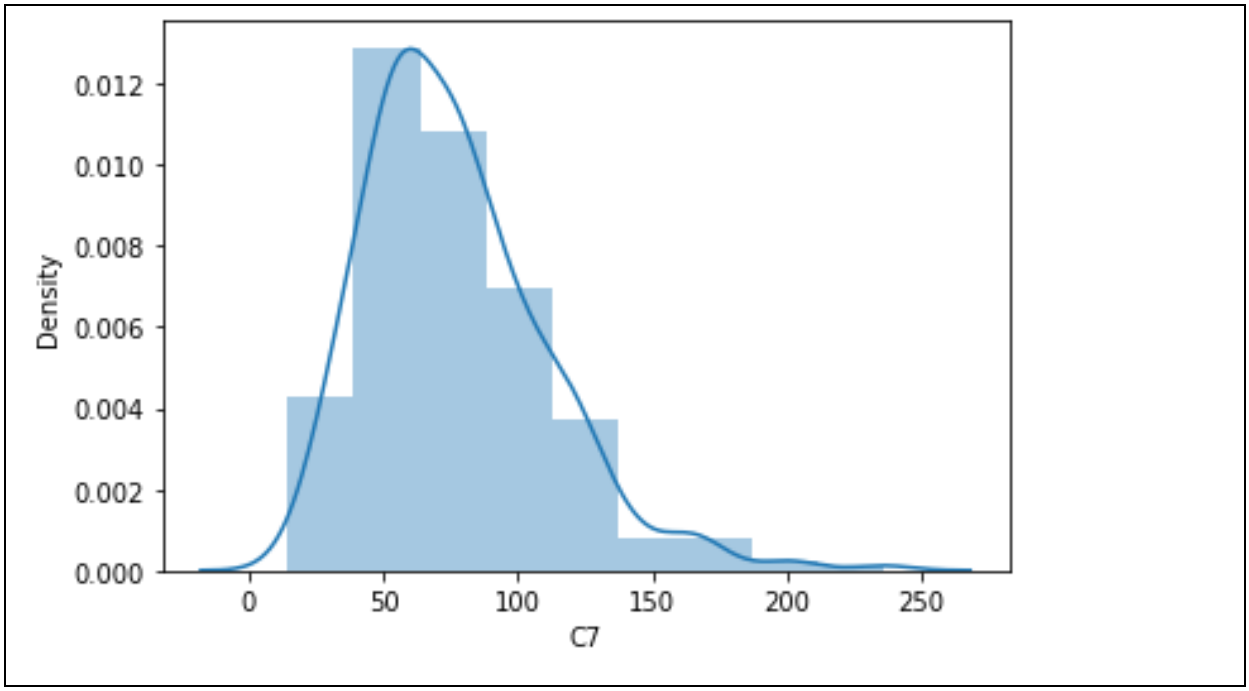
Число групп	Обоснование выбора числа групп	Ширина интервалов
9	По формуле Стерджесса k=[1+log2n]	24

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Вероятность попадания в интервал при условии истинности основной гипотезы
1	14.4	39.01111111	33	0.1047619	0.09846668101283465
2	39.01111111	63.62222222	100	0.31746032	0.2153732710742174
3	63.62222222	88.23333333	84	0.26666667	0.2838123839037856
4	88.23333333	112.84444444	54	0.17142857	0.22539888430188793

5	112.84444444	137.45555556	29	0.09206349	0.10785311768451489
6	137.45555556	162.06666667	6	0.01904762	0.031067367978185212
7	162.06666667	186.67777778	6	0.01904762	0.00538011500744251
8	186.67777778	211.28888889	2	0.00634921	0.0005592091710248104
9	211.28888889	235.9	1	0.0031746	3.482209569194428e-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	136.11572105304234	6.612017938543149e-27	H_0 отклоняется	Распределение не является нормальным
0.05			H_0 отклоняется	Распределение не является нормальным
0.1			H_0 отклоняется	Распределение не является нормальным

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

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

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \left(\gamma^*_{ст}\right)^2 + \left(\varepsilon^*_{ст}\right)^2 = \frac{n}{6}\left((\gamma^*)^2 + \frac{(\varepsilon^*)^2}{4}\right)$	n – объем выборки; $\gamma = \mu 3 / \sigma^3$ – коэффициент асимметрии; $\varepsilon = \mu 4 / \sigma^4 - 3$ – коэффициент эксцесса; μ_i = i-ый центральный момент; σ – среднеквадратичное отклонение (корень из второго центрального момента).
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(2)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(2)$	
Формула расчета *p-value*	$p\text{-}value = 1 - F_Z(z \mid H_0).$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	114.07733617936984	0.0	H_0 отклоняется	Распределение не является нормальным
0.05			H_0 отклоняется	Распределение не является нормальным
0.1			H_0 отклоняется	Распределение не является нормальным

Вывод (в терминах предметной области)

В результате проведённого в п.4 статистического анализа обнаружено, что количество граммов жира, потребляемых в день пациентами, не является нормально распределённой случайной величиной.

5. Проверка однородности выборок

Анализируемый признак 1 – C13

Анализируемый признак 2 – C14

Объёмы выборок – 315

5.1 Критерий знаков

Статистическая гипотеза – H_0: $F_X(\xi) = F_Y(\xi)$

H': $F_X(\xi) \neq F_Y(\xi)$

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{H - 1/2}{\sqrt{\frac{1}{4n}}} = 2\sqrt{n}(H - 1/2)$	H = K/n – частота успеха; K – число знаков «+» в последовательности знаков разностей X_i-Y_i; n – объём выборок.
Закон распределения статистики критерия при условии истинности основной гипотезы	$f_Z(z \mid H_0) \sim N(0,1)$	
Формула расчета критической точки	$\pm N_{1-\alpha/2}(0,1)$	
Формула расчета *p-value*	$2\min\left(F_Z(z \mid H_0), 1 - F_Z(z \mid H_0)\right)$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	-146.5	1.9772043529779528e-75	H_0 отклоняется	$F_X(\xi) \neq F_Y(\xi)$
0.05			H_0 отклоняется	$F_X(\xi) \neq F_Y(\xi)$
0.1			H_0 отклоняется	$F_X(\xi) \neq F_Y(\xi)$

5.2. Критерий хи-квадрат

Статистическая гипотеза – H_0: $F_X(\xi) = F_Y(\xi)$

H': $F_X(\xi) \neq F_Y(\xi)$

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z_{n_X,n_Y} = n_X n_Y \sum_{i=1}^{k} \frac{1}{m_i^{(X)}+m_i^{(Y)}} \left(\frac{m_i^{(X)}}{n_X} - \frac{m_i^{(Y)}}{n_Y} \right)^2$	k – число интервалов; n_X, n_Y – число элементов в выборках; $m_i^{(X)}$, $m_i^{(Y)}$ – частота 1 и 2 выборок в i-й группе.
Закон распределения статистики критерия при условии истинности основной гипотезы	$Z\|_{H_0} \sim \chi^2(k-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(k-1)$	
Формула расчета *p-value*	$p\text{-}value = 1 - F_Z(z \mid H_0).$	

б) Выбрать число групп

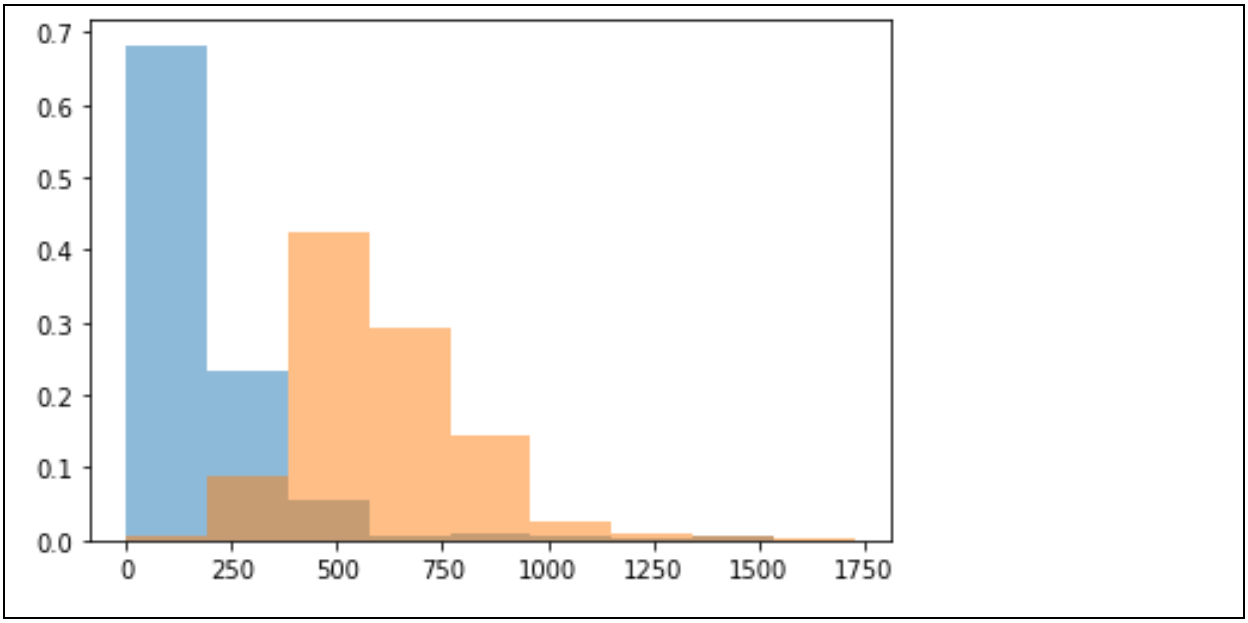
Число групп	Обоснование выбора числа групп	Ширина интервалов
9	По формуле Стерджесса $k=[1+\log_2 n]$	157

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	0.	191.88888889	215	2	0.68253968	0.00634921
2	191.88888889	383.77777778	73	28	0.23174603	0.08888889
3	383.77777778	575.66666667	17	134	0.05396825	0.42539683
4	575.66666667	767.55555556	2	92	0.00634921	0.29206349
5	767.55555556	959.44444444	3	45	0.00952381	0.14285714
6	959.44444444	1151.33333333	2	8	0.00634921	0.02539683
7	1151.33333333	1343.22222222	1	3	0.0031746	0.00952381
8	1343.22222222	1535.11111111	2	2	0.00634921	0.00634921

9	1535.11111111	1727.	0	1	0.	0.0031746

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	72.99382525899497	9.928006143298864e-14	H_0 отклоняется	$F_X(\xi) \neq F_Y(\xi)$
0.05			H_0 отклоняется	$F_X(\xi) \neq F_Y(\xi)$
0.1			H_0 отклоняется	$F_X(\xi) \neq F_Y(\xi)$

Вывод (в терминах предметной области)

В результате проведённого в п.5 статистического анализа обнаружено, что концентрации(ng/ml) в плазме бета-каротина и ретинола не являются случайными величинами, имеющими одинаковый закон распределения.

6. Таблицы сопряжённости

Факторный признак *x* – C2

Результативный признак *y* – C3

Объёмы выборок – 315

Статистическая гипотеза – H_0: $F_Y(y|X = x_1) = \ldots = F_Y(y|X = x_n) = F_Y(y)$

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_{ij} – наблюдаемые (эмпирические) частоты, m_{ij} – теоретические частоты, k – число вариантов факторного признака, l – число вариантов результативного признака.
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 - $F_z(z|H_0)$	

б) Построить эмпирическую таблицу сопряжённости

x \ *y*	Current Smoker	Former	Never	Σ
Female	36	93	144	273
Male	7	22	13	42
Σ	43	115	157	315

в) Построить теоретическую таблицу сопряжённости

x \ *y*	Current Smoker	Former	Never	Σ
Female	37.26666667	99.66666667	136.06666667	273
Male	5.73333333	15.33333333	20.93333333	42
Σ	43	115	157	315

г) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	7.136512415601504	0.30840701536046683	H_0 принимается	Статистическая связь отсутствует
0.05			H_0 принимается	Статистическая связь отсутствует
0.1			H_0 принимается	Статистическая связь отсутствует

Вывод (в терминах предметной области)

В результате проведённого в п.6 статистического анализа обнаружено, что пол пациентов не влияет на их отношение к курению.

7. Дисперсионный анализ

Факторный признак *x* – C5

Результативный признак *y* – C13

Число вариантов факторного признака – 3

Объёмы выборок – 315

Статистическая гипотеза – H_0: $F_{X_1}(x) = \ldots = F_{X_K}(x) = F_X(x)$

эквивалентная гипотеза – H_0: $m_1 = \ldots = m_K$

H': $\neg H_0$

а) Рассчитать групповые выборочные характеристики

№ п/п	Вариант факторного признака	Объём выборки	Групповые средние	Групповые дисперсии
1	No	111	136.8918918918919	8493.660933660934
2	Not often	82	185.65853658536585	20775.388136103582
3	Often	122	240.95901639344262	60058.70078580138

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	2004.5282328560947	2	315713.19667483494
Остаточные признаки	31662.030751316273	312	31966.473354694313
Все признаки	33382.978825900725	314	33489.2940450915

г) Проверить правило сложения дисперсий

Показатель	$D_{межгр}$	$D_{внутригр}$	$D_{общ}$	$D_{межгр} + D_{внутригр}$
Значение	2004.5282328560947	31662.030751316273	33382.978825900725	33666.55898417237

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Эмпирический коэффициент детерминации	$\eta^2 = \frac{D_b^*}{D_X^*}$	0.060046415968752584
Эмпирическое корреляционное отношение	$\eta = \sqrt{\frac{D_b^*}{D_X^*}}$	0.24504370216096677

е) Охарактеризовать тип связи между факторным и результативным признаками

По шкале Чеддока значение эмпирического коэффициента корреляции попадает в диапазон 0,1–0,3, характеризующий слабую степень тесноты статистической связи между факторным и результативным признаками.

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 номинального группировочного(факторного)признака, n – объем выборки, 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 - $F_z(z\|H_0)$	

з) Проверить статистическую гипотезу дисперсионного анализа

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	9.96564197452329	6.376754182611269 e-05	H_0 отклоняется	Статистическая связь присутствует

0.05			H_0 отклоняется	Статистическая связь присутствует
0.1			H_0 отклоняется	Статистическая связь присутствует

Вывод (в терминах предметной области)

В результате проведённого в п.7 статистического анализа обнаружено, что частота потребления витаминов пациентами оказывает слабое влияние на концентрацию бета-каротина(ng/ml) в плазме их крови.

8. Корреляционный анализ

8.1. Расчёт парных коэффициентов корреляции

Анализируемый признак 1 – C13

Анализируемый признак 2 – C14

Объёмы выборок – 315

а) Рассчитать точечные оценки коэффициентов корреляции

	Формула расчета	Значение
Линейный коэффициент корреляции	$\rho^*_{XY} = \frac{k^*_{XY}}{\sigma^*_X \sigma^*_Y}$ $k^*_{XY} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$	0.07157724015217476
Ранговый коэффициент корреляции по Спирмену	$\tilde{\rho}^{(sp)}_{XY} = \frac{\sum_{i=1}^{n}(r_i-\bar{r})(s_i-\bar{s})}{\sqrt{\sum_{i=1}^{n}(r_i-\bar{r})^2 \sum_{i=1}^{n}(s_i-\bar{s})^2}} = \frac{\mu^*_{RS}}{\sigma^*_S \sigma^*_R}$. $\bar{r} = \bar{s} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}$	0.13062133340920742
Ранговый коэффициент корреляции по Кендаллу	$\tilde{\tau}_{XY} = \frac{4Q}{n(n-1)} - 1$, $Q = \sum_{i=1}^{n-1} Q_i$; $Q_i = \sum_{j=i+1}^{n} [s_j > s_i]$	0.0857942303995992

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

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

в) Рассчитать доверительные интервалы для линейного коэффициента корреляции

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	-2.621884255204232	-1.9660189361953855	-1.6009085786068036
Верхняя граница	2.6418842552042316	2.0660189361953853	1.8009085786068038

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

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	1.2695870624981909	0.20517488254189867	H_0 принимается	Статистическая связь отсутствует
$H_0 : r^{(cn)} = 0$ $H' : r^{(cn)} \neq 0$	0.1	2.330897651015813	0.20517488254189867	H_0 принимается	Статистическая связь отсутствует
$H_0 : r^{(кен)} = 0$ $H' : r^{(кен)} \neq 0$	0.1	2.2786430348594875	0.023439852661112 17	H_0 отклоняется	Статистическая связь присутствует

8.2. Расчёт множественных коэффициентов корреляции

Анализируемый признак 1 – C7

Анализируемый признак 2 – C9

Анализируемый признак 3 – C12

Объёмы выборок – 315

а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу

Признак / Признак	C7	C9	C12
C7	1.000000	0.040055	0.351902
C9	0.040055	1.000000	-0.029812
C12	0.351902	-0.029812	1.000000

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	C7	C9	C12
C7	–	0.313979	1.295163e-20
C9	0.313979	–	0.453599
C12	1.295163e-20	0.453599	–

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 $R_{ij} \in \{1, \cdots, n\}$ - ранг i-го элемента в X_j выборке.	0.44417366595468666

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	χ^2(n-1)	
Формула расчета критической точки	χ^2_α(n – 1)	
Формула расчета *p-value*	1 - χ^2(Z, n – 1)	

д) Проверить значимость коэффициента конкордации

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	279.8294095514526	0.9176686807978349	H_0 принимается	Статистическая связь отсутствует
0.05			H_0 принимается	Статистическая связь отсутствует
0.1			H_0 принимается	Статистическая связь отсутствует

Вывод (в терминах предметной области)

В результате проведённого в п.8 статистического анализа обнаружено, что между концентрациями(ng/ml) бета-каротина и ретинола в плазме крови нет зависимости.
А также, что статистическая связь между потребляемыми пациентами количеством жира(grams per day), алкогольных напитков(number per week) и количеством диетического ретинола(mcg per day) отсутствует. Однако, из пункта б) видно, что между количеством жира(grams per day) и количеством диетического ретинола(mcg per day), потребляемых пациентами, зависимость есть, в то время как между потребляемыми количествами жира(grams per day) и алкогольных напитков(number per week), алкогольных напитков(number per week) и диетического ретинола(mcg per day) связь отсутствует.

9. Регрессионный анализ

9.1 Простейшая линейная регрессионная модель

Факторный признак *x* – C6

Результативный признак *y* – C13

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	$\tilde{\beta}_0 = \bar{y} - \rho^*_{XY}\frac{\sigma^*_Y}{\sigma^*_X}\bar{x}$	200.6239454749055
β_1	$\tilde{\beta}_1 = \rho^*_{XY}\frac{\sigma^*_Y}{\sigma^*_X}$	-0.005973258278958724

б) Записать точечную оценку уравнения регрессии

f(*x*) = 200.6239454749055 - 0.005973258278958724 * x

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

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$D^*_{\text{регр } Y\|X} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i, \beta_0, \dots, \beta_{k-1}) - \bar{y})^2$	k – 1	$\frac{n}{k-1}D^*_{\text{регр } Y\|X}$
Остаточные признаки	$D^*_{\text{ост } Y} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, \beta_0, \dots, \beta_{k-1}))^2$	n – k	$\frac{n}{n-k}D^*_{\text{ост } Y}$
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	16.462780398473043	1	5185.775825519008
Остаточные признаки	33366.516045502256	313	33579.720620872875
Все признаки	33382.97882590072	314	33489.29404509148

д) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	16.462780398473043	33366.516045502256	33382.97882590072	33382.978825900725

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$R^{2*}_{Y\|X} = \frac{D^*_{\text{регр } Y\|X}}{D^*_Y} = 1 - \frac{D^*_{\text{ост } Y}}{D^*_Y}$	0.000493148933303104
Корреляционное отношение	$R^*_{Y\|X} = \sqrt{\frac{D^*_{\text{регр } Y\|X}}{D^*_Y}} = \sqrt{1 - \frac{D^*_{\text{ост } Y}}{D^*_Y}}$	0.022206956867232033

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

Связь между факторным и результативным признаками 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	$\tilde{\beta}_1 - t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{resY}}\sqrt{\frac{1}{nD^*_X}}$;
	Верхняя граница	$\tilde{\beta}_1 + t_{1-\alpha/2}(n-2)\sqrt{\tilde{D}_{resY}}\sqrt{\frac{1}{nD^*_X}}$

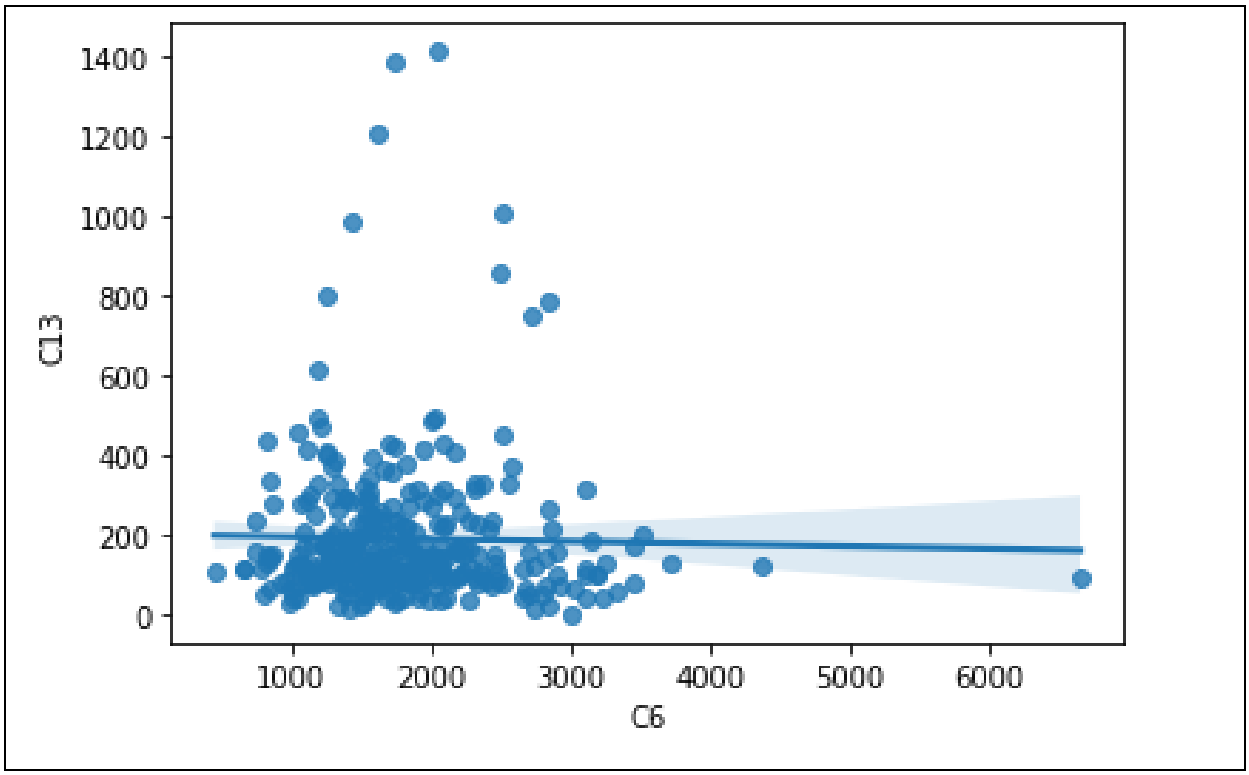
б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
β_0	Нижняя граница	125.31992251590073	143.4528909519943	152.68812898968702
	Верхняя граница	275.9279684339103	257.7949999978167	248.55976196012398
β_1	Нижняя граница	-0.045178308606084266	-0.03573785758922106	-0.03092977486238141
	Верхняя граница	0.033231792048166814	0.02379134103130361	0.018983258304463964

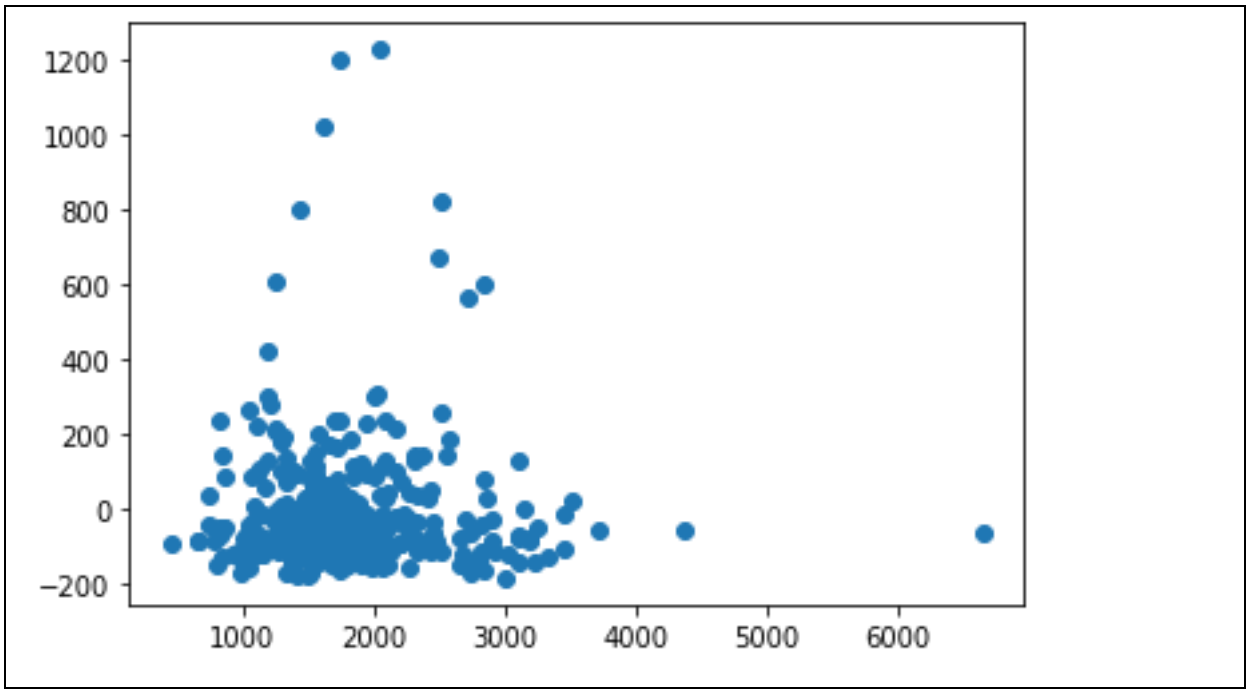
в) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

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

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{R^{2*}_{Y\|X}}{\left(1-R^{2*}_{Y\|X}\right)/(n-2)}$	$R^{2*}_{Y\|X} = \frac{D^*_{\text{регр } Y\|X}}{D^*_Y}$; n – объем выборки.
Закон распределения статистики критерия при условии истинности основной гипотезы	$f_Z(z\|H_0) \sim F(1, n-2)$	
Формула расчета критической точки	$f_{1-\alpha}(1, n-2)$	
Формула расчета *p-value*	$1 - F(Z, 1, n-2)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.15443177398848207	0.694603179135282	H_0 принимается	$\beta_1 = 0$
0.05			H_0 принимается	$\beta_1 = 0$
0.1			H_0 принимается	$\beta_1 = 0$

9.2 Линейная регрессионная модель общего вида

Факторный признак *x* – С6

Результативный признак *y* – С13

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	$\tilde{\beta} = (F^T F)^{-1} F^T y$; $y = (y_1, \ldots, y_n)^T$; $\beta = (\beta_0, \beta_1, \beta_2)^T$; $F = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 \end{pmatrix}$	156.531322
β_1		0.0370710218
β_2		-9.01050102e-06

б) Записать точечную оценку уравнения регрессии

$f(x)$ = 156.531322 + 0.0370710218 * x - 9.01050102e-06 * x^2

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	136.46102080306048	2	21492.610776482026
Остаточные признаки	33246.517805097676	312	33566.19586091593
Все признаки	33382.97882590072	314	33489.29404509148

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	136.46102080306048	33246.517805097676	33382.97882590072	33382.97882590073

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$R^{2^*}_{Y\|X} = \frac{D^*_{\text{регр } Y\|X}}{D^*_Y} = 1 - \frac{D^*_{\text{ост } Y}}{D^*_Y}$	0.004087742484417988
Корреляционное отношение	$R^*_{Y\|X} = \sqrt{\frac{D^*_{\text{регр } Y\|X}}{D^*_Y}} = \sqrt{1 - \frac{D^*_{\text{ост } Y}}{D^*_Y}}$	0.06393545561281305

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

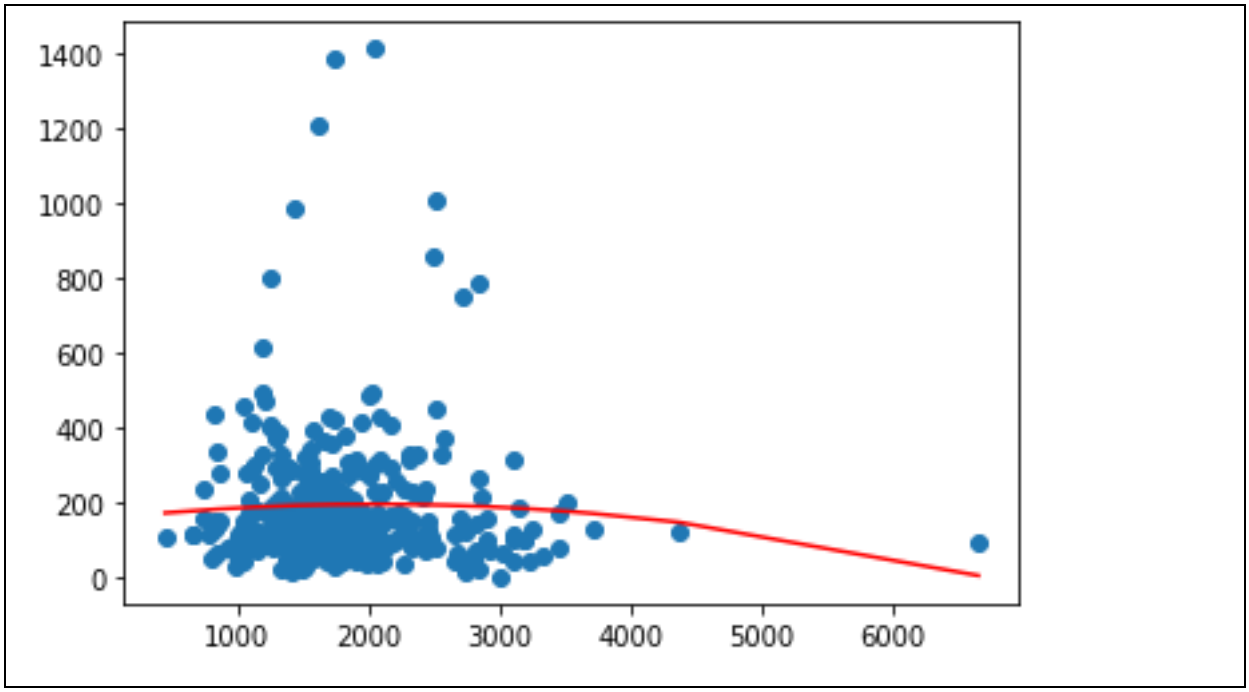
Связь между факторным и результативным признаками отсутствует.

9.2.2. Интервальные оценки линейной регрессионной модели

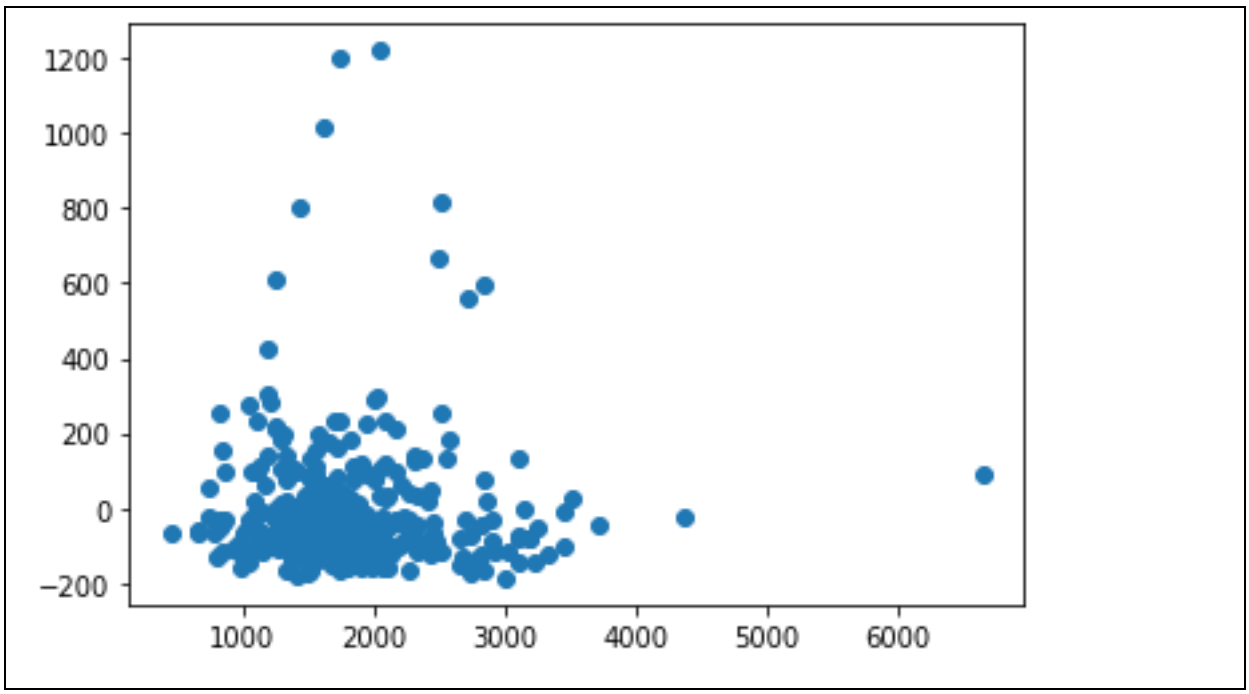
а) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

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

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

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{R^{2*}_{Y\|X}/(k-1)}{\left(1-R^{2*}_{Y\|X}\right)/(n-k)}$	$R^{2^*}_{Y\|X} = \frac{D^*_{регр\ Y\|X}}{D^*_Y}$; n – объём выборки; k = 3.
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	1 – F(Z, k-1, n–k)	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.6403052304627638	0.5278227554063295	H_0 принимается	$\beta_1 = \beta_2 = 0$
0.05			H_0 принимается	$\beta_1 = \beta_2 = 0$
0.1			H_0 принимается	$\beta_1 = \beta_2 = 0$

9.3 Множественная линейная регрессионная модель

Факторный признак 1 x_1 – C6

Факторный признак 2 x_2 – C11

Результативный признак y – C13

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	$\tilde{\beta} = (F^T F)^{-1} F^T y$; $y = (y_1, \ldots, y_n)^T$; $\beta = (\beta_0, \beta_1, \beta_2)^T$; $F = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \ldots & \ldots & \ldots \\ 1 & x_n & x_n^2 \end{pmatrix}$	163.004105
β_1		-0.0219907501
β_2		0.0303795958

б) Записать точечную оценку уравнения регрессии

$f(x)$ = 163.004105 - 0.0219907501 * x_1 + 0.0303795958 * x_2

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	1896.6175210691147	2	298717.25956838555
Остаточные признаки	31486.361304831622	312	31789.114778916544
Все признаки	33382.97882590073	314	33489.2940450915

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	1896.6175210691147	31486.361304831622	33382.97882590073	33382.97882590074

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Множественный коэффициент детерминации	$R^{2^*}_{Y\|X_1 X_2} = \frac{D^*_{\text{регр } Y\|X_1\ X_2}}{D^*_Y} = 1 - \frac{D^*_{\text{ост } Y}}{D^*_Y}$	0.05681390899716813
Множественное корреляционное отношение	$R^*_{Y\|X_1 X_2} = \sqrt{\frac{D^*_{\text{регр } Y\|X_1\ X_2}}{D^*_Y}} = \sqrt{1 - \frac{D^*_{\text{ост } Y}}{D^*_Y}}$	0.23835668439791682

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

Между факторным и результативным признаками присутствует слабая 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	16.462780398473043	136.46102080306048	1896.6175210691147
Остаточные признаки	33366.516045502256	33246.517805097676	31486.361304831622
Все признаки	33382.97882590072	33382.978825900072	33382.978825900073

б) Сводная таблица свойств различных регрессионных моделей

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	2.2 %	6.4 %	23.8 %
Значимость	нет	нет	нет
Адекватность	-	-	-
Степень тесноты связи	Отсутствует	Отсутствует	Слабая

Вывод (в терминах предметной области)

В результате проведённого в п.9 статистического анализа обнаружено, что количество калорий, потребляемых пациентами в день, не влияет на концентрацию бета-каротина(ng/ml) в плазме их крови. Однако количество потребляемых бета-каротина с пищей(mcg per day) и калорий в день оказывает слабое воздействие на концентрацию бета-каротина(ng/ml).