Обзор задач и используемых пакетов

В представленном коде решаются две главные задачи на основе статистического эксперимента:

1. **Задача 1** — исследование зависимости непрерывной отклика Y от ковариаты X путём построения
   - линейной модели $Y \sim X$;
   - квадратичной модели $Y \sim X + X^2$;
   - анализа остатков (гистограммы, тесты нормальности);
   - доверительных интервалов (ДИ) для коэффициентов;
   - проверки гипотез о линейности и о значимости коэффициентов;
   - сравнения моделей по информационным критериям AIC и BIC;
   - интерпретации полученных результатов (R^2 и выводы об адекватности).
2. **Задача 2** — двухфакторный дисперсионный анализ (ANOVA) отклика Y при факторах A, B:
   - полная модель с взаимодействием $Y \sim A * B$;
   - аддитивная модель $Y \sim A + B$;
   - модель только по фактору A;
   - оценка значимости эффектов ($A, B, A \times B$) через F-тест;
   - сравнение моделей по AIC и BIC;
   - профильные графики взаимодействия;
   - анализ остатков (гистограммы, тест Жарка–Бера);
   - итоговая интерпретация ($\hat{\sigma}^2$, лучшая модель).

Для расчётов применяются:

- `numpy`, `pandas` — работа с данными;
- `statsmodels` — МНК-оценка регрессий, ANOVA;
- `scipy.stats` — распределения F, t, χ^2, тест JB;
- `matplotlib` — графики (скаттеры, гистограммы, эллипсоиды).

Ниже подробно разберём каждый блок кода, формулы из `.tex`-файла и результаты.

1 Задача 1: зависимость Y от X

1.1 Исходные данные

Листинг 1: Фрагмент кода (данные задачи 1)

```
Y1 = [9.61, 9.22, 4.76, ..., 12.99]
X1 = [2, 4, 2, ..., 2]
```

- $n = 50$ — объём выборки;
- y_i, x_i — наблюдения отклика и коварианты;
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$.

Численные итоги:

$$\sum x_i = 99, \quad \sum y_i = 456.95, \quad \bar{x} = 1.98, \quad \bar{y} = 9.139, \quad S_{xx} = 40.98, \quad S_{xy} = -31.23.$$

1.2 Линейная регрессия

Модель

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

МНК-оценки

$$\hat{\beta}_2 = \frac{S_{xy}}{S_{xx}} = -0.762, \qquad \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x} = 10.648.$$

Качество $\mathrm{RSS}_{\mathrm{lin}} = 312.50,\ R^2_{\mathrm{lin}} = 0.071.$

Информационные критерии $AIC_{\mathrm{lin}} = 237.5,\ BIC_{\mathrm{lin}}$ (рассчитан кодом).

1.3 Квадратичная регрессия

Модель $y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \varepsilon_i.$

МНК-оценки $\hat{\beta}_1 = 11.004,\ \hat{\beta}_2 = -1.238,\ \hat{\beta}_3 = 0.124.$

Качество $\mathrm{RSS}_{\mathrm{quad}} = 309.92,\ R^2_{\mathrm{quad}} = 0.074.$

Информационные критерии $AIC_{\mathrm{quad}} = 239.36,\ \Delta AIC = 1.84 < 2 \Rightarrow$ модели почти эквивалентны, линейная предпочтительнее.

1.4 Нормальность остатков

Гистограммы + плотность $N(\hat{\mu}, \hat{\sigma}^2)$.

- $\chi^2 = 9.50,\ p = 0.091$ (нормальность *не* отвергается при $\alpha = 0.01$);
- Jarque–Bera: $JB = 6.72,\ p = 0.035$ (при $\alpha = 0.01$ также не отвергается).

1.5 Доверительные интервалы (99 %)

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{RSS_{\text{quad}}}{n-k} = 6.603, \\
t_{0.995}(47) &= 2.69, \\
CI_{99\%}(\beta_2) &= (-4.68;\, 2.20), \\
CI_{99\%}(\beta_3) &= (-0.72;\, 0.97).
\end{aligned}$$

Совместный эллипсоид: $(\beta - \hat{\beta})^\top (X^\top X)(\beta - \hat{\beta}) \le 67.43$.

1.6 Проверка гипотез

- Линейность ($H_0 : \beta_3 = 0$): $F = 0.153,\ p = 0.697$ — не отвергаем.
- Независимость ($H_0 : \beta_2 = 0$): $t = -1.912,\ p = 0.062$ — не отвергаем ($\alpha = 0.01$).

1.7 Итог задачи 1

Линейная модель достаточна, но объясняет лишь $\approx 7\%$ вариации Y; β_2, β_3 статистически незначимы.

2 Задача 2: двухфакторный ANOVA (A, B)

2.1 Модели

- Полная: $Y \sim A * B$ ($k = 16$);
- Аддитивная: $Y \sim A + B$ ($k = 7$);
- Только A ($k = 4$).

2.2 ANOVA-таблица (ручной расчёт)

Источник	SS	df	MS	F
A	979.34	3	326.45	118.30
B	513.23	3	171.08	61.99
$A \times B$	1047.67	9	116.41	42.18
	88.30	32	2.76	

Все $p \ll 0.01$ — эффекты значимы.

2.3 Информационные критерии

Модель	k	$\hat{\sigma}^2$	AIC	BIC
$A * B$	16	2.76	197.5	227.4
$A + B$	7	10.45	302.1	315.2
A	4	21.66	314.0	321.5

Лучшая — полная модель $A * B$.

2.4 Нормальность остатков

Jarque–Bera: $JB = 1.10,\ p = 0.576$ ($\alpha = 0.10$ — нормальность принимается).

2.5 Итог задачи 2

- Значимы главные эффекты A, B и взаимодействие $A \times B$.
- Лучшая модель: $Y \sim A * B$.
- Остатки нормальны, $\hat{\sigma}^2 = 2.76$.

Сводка ключевых понятий

Понятие	Обозначение/переменная	Смысл	Формула/примечание
x_i, y_i	`df1.X`, `df1.Y`	наблюдения	$n = 50$ пар (x_i, y_i)
$\bar{x}, \bar{y}$	`mean()`	выборочные средние	$\bar{x} = \frac{1}{n} \sum x_i$ и т.д.
S_{xx}, S_{xy}	`Sxx,Sxy`	суммарные квадраты/-произведения	$S_{xx} = 40.98$, $S_{xy} = -31.23$
$\hat{\beta}_1, \hat{\beta}_2$	`lin.params`	МНК-оценки лин. модели	$\hat{\beta}_2 = S_{xy}/S_{xx}$ и т.д.
RSS	`.ssr`	сумма квадратов остатков	$RSS_{\text{lin}} = 312.50$
R^2	`.rsquared`	коэффициент детерминации	0.071 лин., 0.074 квадр.
AIC,BIC	`.aic,.bic`	инфо-критерии	$n \ln(\hat{\sigma}^2) + 2k$ и $+k \ln n$
F-тест	`F_lin,p_lin`	значимость доп. параметров	см. формулу в тексте
t-тест	`lin.tvalues,pvalues`	значимость β_2	$t = -1.912,\ p = 0.062$
99% ДИ	`ci_b2,ci_b3`	интервалы для $\beta_{2,3}$	$(-4.68; 2.20)$, $(-0.72; 0.97)$
Эллипсоид	график	совместный ДИ	$(\beta - \hat{\beta})^\top (X^\top X)(\beta - \hat{\beta}) \leq 67.43$
JB-тест	см. код	нормальность остатков	$JB = 6.72,\ p = 0.035$ (lin./quad.)

Фактор A	`C(A)`	4 уровня	$\sum \alpha_i = 0$
Взаим. $A \times B$	`C(A)*C(B)`	пересечение эффектов	см. ANOVA

Заключение

1. В задаче 1 зависимость Y от X слаба ($R^2 \approx 7\%$); линейная модель достаточна, но статистически незначима при $\alpha = 0.01$.
2. В задаче 2 факторы A, B и их взаимодействие оказывают сильное влияние; лучшая модель — полная $A * B$; остатки нормальны, $\hat{\sigma}^2 = 2.76$.

Ключевые статистические термины и понятия, встречающиеся в отчёте

Обозначение	Расшифровка	Что показывает / как вычисляется
p**-value**	«уровень значимости, достигнутый данными»	Вероятность при нулевой гипотезе H_0 получить наблюдаемую (или ещё более экстремальную) статистику. Считается по табличному распределению (например, F, t, χ^2). Если $p < \alpha$ — отвергаем H_0.
SS (Sum of Squares)	Сумма квадратов отклонений	В ANOVA — 4 источника: SS_A, SS_B, SS_{AB}, SS_E. В регрессии — SS_{Reg} и RSS.
df (degrees of freedom)	Степени свободы	Число «независимых кусков информации» после учёта оценённых параметров. Примеры: $df_A = a - 1$, $df_E = n - k$.
MS (Mean Square)	Средний квадрат	$MS = SS/df$. Нужно, чтобы привести все суммы квадратов к одной шкале.
F	Статистика Фишера	Отношение двух средних квадратов: $F = MS_H/MS_E$ (ANOVA) или формула $\frac{(RSS_R - RSS_F)/q}{RSS_F/df_F}$ (вложенные регрессии).
k	Число оценённых параметров в модели	В линейной регрессии «константа + угловые коэффициенты», в ANOVA включает эффекты и их условия идентификации.
$\widehat{\sigma}^2$	Оценка дисперсии ошибок	В регрессии — $\widehat{\sigma}^2 = RSS/(n-k)$; в полной ANOVA — MS_E.
AIC / BIC	Информационные критерии Акаике / Байеса	$AIC = n \ln \widehat{\sigma}^2 + 2k$, $BIC = n \ln \widehat{\sigma}^2 + k \ln n$. Чем ниже, тем модель лучше (учёт «качество – сложность»).
α	Заданный уровень риска ошибки I рода	В задаче 2 преподаватель потребовал $\alpha = 0{,}10$ для проверки нормальности JB, поэтому используем именно 0.10.
$X^\top$	Транспонированная матрица признаков	В МНК решение $\hat{\beta} = (X^\top X)^{-1} X^\top y$.
β	Истинный (неизвестный) параметр модели	Например, наклон прямой.
$\hat{\beta}$	Оценка β по данным	Выдаёт алгоритм МНК.

Модели факторного анализа

Обозначение	Формула	Смысл
A	$Y \sim A$	Учитываем только главный эффект фактора A.
$A+B$	$Y \sim A+B$	Два главных эффекта, **без** взаимодействия. Аддитивная модель — влияние каждого фактора независимо суммируется.
$A*B$	$Y \sim A+B+A{:}B$	Главные эффекты и взаимодействие $A \times B$. Полная модель.

Аддитивная модель «предполагает параллельность» профилей (линии не пересекутся). *Модель с взаимодействием* допускает, что эффект A зависит от того, какой уровень B выбран (линии пересекаются).

Как решаем, что нужна $A * B$

1. F-тест сравнивает $A + B$ (огр.) и $A * B$ (полная).
2. Если $p \ll \alpha$ (в примере $p_{AB} = 3 \cdot 10^{-15}$), взаимодействие значимо → берём $A * B$.
3. AIC/BIC у $A * B$ минимальны → дополнительно подтверждает выбор.

Доверительный эллипсоид

- Для двух коэффициентов (β_2, β_3) строит **совместный** 99 %-интервал: все точки внутри эллипса — правдоподобные истинные значения обеих β одновременно.
- Полезен, когда одиночные интервалы широкие: визуально показывает, какие комбинации β_2, β_3 ещё допускаются данными.

Гипотезы, проверяемые в отчёте

№	Гипотеза H_0	Статистика	Где встречается
1	$\beta_3 = 0$ (квадратичный член не нужен)	F — сравнение лин. и квадр. моделей	Задача 1 «линейность»
2	$\beta_2 = 0$ (Y не зависит от X)	t-тест	Задача 1
3	Остатки $\sim N(0, \sigma^2)$	JB, χ^2	Задачи 1 и 2
4	Нет взаимодействия $A \times B$	F — сравнение $A + B$ и $A * B$	Задача 2
5	Нет эффекта A	$F_A = MS_A / MS_E$	Задача 2
6	Нет эффекта B	$F_B = MS_B / MS_E$	Задача 2

Почему в JB-тесте $\alpha = 0.10$

Преподаватель в условии задачи 2 поставил именно такое требование (доверие 90 %). Мы подчиняемся: тест нормальности проверяется при $\alpha = 0{,}10$. Для остальных проверок (эффекты факторов) использовалось $\alpha = 0{,}01$.

Итоговое понимание

- **p-value** мы \emph{сами вычисляем} по наблюдаемой статистике и табличному распределению.
- SS, df, MS, F — кирпичики ANOVA и F-тестов.
- $k, \widehat{\sigma}^2, AIC, BIC$ — нужны для «штрафа за сложность» и выбора лучшей модели.
- **Модели** A, $A + B$, $A * B$ различаются тем, позволяют ли факторам «взаимодействовать».
- **Доверительный эллипсоид** даёт совместную область допустимых β.
- **Проверяемые гипотезы** сводятся к трем классам: значимость параметров, необходимость взаимодействий, нормальность ошибок.

1. p-value

Для любой тестовой статистики $T_{\text{набл}}$ с известным при H_0 распределением $F_T(t) = P(T \le t)$

$$\boxed{p = \ P\big(T \ge T_{\text{набл}} \mid H_0\big) = \ 1 - F_T\big(T_{\text{набл}}\big)}.$$

- Если тест двусторонний для статистики симметричного закона (например, t),

$$p_{\text{двустор}} = 2\ \big[1 - F_T\big(|T_{\text{набл}}|\big)\big].$$

---

2. Матрица признаков и транспонирование

Для выборки $\{x_i\}_{i=1}^n$ с полиномиальной регрессией второго порядка:

$$X \ = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}, \qquad \boxed{X^\top \text{ — просто та же матрица «повёрнутая» на } 90^\circ}$$

т. е. $X^\top$ имеет размер $k \times n$ (здесь $k = 3$).

---

3. t-тест для одного коэффициента

Гипотезы

$$H_0: \ \beta_j = 0 \quad (\text{или любое фиксированное } \beta_0), \qquad H_1: \ \beta_j \ne 0.$$

Статистика

$$\boxed{t = \frac{\hat{\beta}_j - \beta_0}{\mathrm{SE}(\hat{\beta}_j)}}, \qquad \mathrm{SE}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2\ \big[(X^\top X)^{-1}\big]_{jj}},$$

$$\hat{\sigma}^2 = \frac{RSS}{n-k}, \qquad RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \qquad t \ \sim \ t_{n-k} \ \text{ при } H_0.$$

p-value (двусторонний)

$$p = \ 2\Big[1 - F_{t_{n-k}}\big(|t_{\text{набл}}|\big)\Big].$$

4. Остатки и RSS

$$\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, \qquad \boxed{r_i = y_i - \hat{y}_i} \quad (i = 1, \dots, n),$$

$$RSS = \sum_{i=1}^{n} r_i^2 = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2,$$

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{y}.$$

От остатков r_i напрямую зависят:

- $\hat{\sigma}^2 = RSS/(n-k)$;
- все стандартные ошибки $\mathrm{SE}(\hat{\beta}_j)$;
- статистики t, F, тесты нормальности и пр.

Эти формулы полностью описывают:

- вычисление p-значения,
- построение $X^\top$,
- статистику t и её стандартную ошибку,
- получение остатков и RSS.

Думал на протяжении 5 секунд >

Перечень всех статистических гипотез, проверенных в работе

(формула ↔ словесное описание ↔ какой критерий используется)

№	Формальная запись нулевой гипотезы H_0	Словами	Проверочный тест / статистика
Задача 1 (регрессия Y от X)			
1	$H_0:\ \beta_3 = 0$	квадратичный член **не** нужен, зависимость линейна	F-тест для вложенных моделей $F = \dfrac{(RSS_{\text{lin}} - RSS_{\text{quad}})/1}{RSS_{\text{quad}}/(n-3)}$
2	$H_0:\ \beta_2 = 0$	наклон равен нулю ⇒ Y **не зависит** от X	t-тест: $t = \dfrac{\hat{\beta}_2}{\mathrm{SE}(\hat{\beta}_2)}$
3	$H_0:\ \varepsilon_i \sim N(0, \sigma^2)$	ошибки регрессии нормальны	(a) χ² по бинам, (b) Jarque–Bera $JB = \frac{n}{6}\left(s^2 + \frac{1}{4}k^2\right)$
Задача 2 (ANOVA по факторам A, B)			
4	$H_0:\ (\alpha\beta)_{ij} = 0\ \forall i, j$	**нет взаимодействия** факторов $A \times B$	F-тест: $F_{AB} = \dfrac{MS_{AB}}{MS_E}$ с $df = (a-1)(b-1)$, df_E
5	$H_0:\ \alpha_i = 0\ \forall i$	фактор A не влияет	F-тест: $F_A = MS_A / MS_E$
6	$H_0:\ \beta_j = 0\ \forall j$	фактор B не влияет	F-тест: $F_B = MS_B / MS_E$
7	$H_0:\ \varepsilon_{ijk} \sim N(0, \sigma^2)$	ошибки полной модели нормальны	Jarque–Bera $JB = \frac{n}{6}(s^2 + \frac{1}{4}k^2)$ (здесь $\alpha = 0.10$ — так задано в условии)

Пояснения к каждой группе гипотез

- **№ 1 – 2 (коэффициенты β)**
  Проверяем, нужны ли конкретные регрессоры.
  - Если $H_0 : \beta_3 = 0$ не отвергнут, оставляем только линейный член.
  - Если $H_0 : \beta_2 = 0$ не отвергнут, считаем, что X вообще не объясняет Y.
- **№ 3 и 7 (нормальность ошибок)**
  Нужна для корректности доверительных интервалов, F- и t-тестов.
  В задаче 2 преподаватель потребовал уровень $\alpha = 0{,}10$ — поэтому нормальность там проверяется именно при 10 %-м риске ошибки I рода.
- **№ 4 (взаимодействие)**
  Сравниваем аддитивную модель $A + B$ и полную $A * B$.
  Если H_0 отвергнут, используем $A * B$ (линии на interaction-plot пересекаются).
- **№ 5–6 (главные эффекты A, B)**
  Стандартные F-тесты «средний квадрат эффекта / средний квадрат ошибки».
  Критические значения берутся из распределения $F(df_H, df_E)$.

Остатки, RSS и стандартная ошибка

$$\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, \qquad r_i = y_i - \hat{y}_i, \qquad RSS = \sum_{i=1}^{n} r_i^2, \qquad \hat{\sigma}^2 = RSS/(n-k).$$

Именно $\hat{\sigma}^2$ входит в t-, F-, AIC, BIC и формулы доверительных интервалов.

Доверительный эллипсоид

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top (X^\top X)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \ \le\ 2\,\hat{\sigma}^2\, F_{2,n-k}(1-\alpha).$$

Даёт \textbf{совместную} область доверия для пары (β_2, β_3);
если точка $(0, 0)$ лежала бы вне эллипса, обе коэффициента сразу были бы значимы.

№	Название проверки	Формулировка гипотез	Тест-статистика (теория)	Наблюд-ное значение	p-value	α	Решение
Задача 1: регрессия Y от X							
1	**Линейность**	$H_0 : \beta_3 = 0$ (квадратичный член не нужен) $H_1 : \beta_3 \neq 0$	$F = \dfrac{(RSS_R - RSS_F)/q}{RSS_F/df_F}$, $q = 1$	$F = 0{,}153;\ df_1 = 1,\ df_2 = 47$	0,697	0,01	$p > \alpha$ → не отвергаем H_0
2	**Значимость наклона**	$H_0 : \beta_2 = 0$ (-нет связи $Y \leftrightarrow X$) $H_1 : \beta_2 \neq 0$	$t = \dfrac{\hat{\beta}_2}{\mathrm{SE}(\hat{\beta}_2)} \sim t_{n-k}$	$t = -1{,}912;\ df = 48$	0,062	0,01	Не отвергаем
3-a	**Нормальность остатков (χ^2)**	$H_0 : \varepsilon_i \sim N(0, \sigma^2)$	$\chi^2 = \sum \dfrac{(O-E)^2}{E}$	$\chi^2 = 9{,}50$	0,091	0,01	Не отвергаем
3-b	**Нормальность (Jarque–Bera)**	то же H_0	$JB = \dfrac{n}{6}(s^2 + k^2/4) \sim \chi^2_2$	$JB = 6{,}72$	0,035	0,01	Не отвергаем
Задача 2: двухфакторный ANOVA (A, B)							
4	**Взаимодействие $A \times B$**	$H_0 : (\alpha\beta)_{ij} = 0\ \forall i, j$ $H_1 : \exists\, (\alpha\beta)_{ij} \neq 0$	$F_{AB} = \dfrac{MS_{AB}}{MS_E}$, $df_1 = (a-1)(b-1),\ df_2 = df_E$	$F_{AB} = 42{,}18;\ df_1 = 9,\ df_2 = 32$	$3 \cdot 10^{-15}$	0,01	$p < \alpha$ → отвергаем H_0 ⇒ взаимодействие ЕСТЬ
5	**Эффект фактора A**	$H_0 : \alpha_i = 0\ \forall i$	$F_A = MS_A/MS_E$	$F_A = 118{,}30;\ df_1 = 3,\ df_2 = 32$	$< 10^{-15}$	0,01	Отвергаем ⇒ фактор A значим
6	**Эффект фактора B**	$H_0 : \beta_j = 0\ \forall j$	$F_B = MS_B/MS_E$	$F_B = 61{,}99;\ df_1 = 3,\ df_2 = 32$	$< 10^{-12}$	0,01	Отвергаем ⇒ фактор B значим
7	**Нормальность остатков полной модели**	$H_0 : \varepsilon_{ijk} \sim N(0, \sigma^2)$	Jarque–Bera $JB = \dfrac{n}{6}(s^2 + k^2/4)$	$JB = 1{,}10$	0,576	0,10*	$p > \alpha$ → не отвергаем H_0