Отчёт по ИДЗ-3 «Регрессионный анализ» по дисциплине: Статистический анализ данных

Свиридов Артем, группа 5140201/50302
Вариант 18 (30225)

Задание 1

Результаты статистического эксперимента приведены в таблице 1. Требуется оценить характер зависимости наблюдаемой переменной Y от ковариаты X.

Таблица 1. $\alpha = 0.10; h = 1.60$.

No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Y	17.22	15.89	11.65	13.68	14.23	14.66	8.60	14.79	16.19	15.02	16.55	10.94	14.26	8.15	14.63	12.15	12.43
X	5	5	3	5	4	5	4	4	4	3	5	3	7	5	5	5	4

No	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
Y	10.29	16.15	13.42	8.54	9.73	17.70	11.31	10.73	17.95	15.09	17.06	15.45	12.28	15.45	12.08	13.47	8.62
X	5	6	6	7	3	9	5	7	6	5	5	3	6	4	5	4	3

No	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	
Y	10.47	13.64	13.95	15.31	14.29	13.90	9.22	7.99	14.59	14.63	11.87	12.64	14.99	12.53	15.75	13.56	
X	2	4	4	3	5	4	5	6	5	6	8	4	4	5	3	5	

Пункт (а). Линейная модель

Для аппроксимации зависимости переменной Y от ковариаты X используется простая линейная регрессионная модель:

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

где ошибки ε_i предполагаются независимыми и одинаково распределенными ($\mathrm{E}\varepsilon_i = 0, \mathrm{Var}(\varepsilon_i) = \sigma^2$). В матричной форме модель имеет вид $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, где $\mathbf{X}$ — матрица плана размерности $n \times 2$ с единичным первым столбцом.

Оценки параметров $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$ находятся методом наименьших квадратов (МНК) из условия минимизации суммы квадратов остатков $SS(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$. Решение системы нормальных уравнений дает:

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

Для случая парной регрессии эти оценки выражаются через выборочные характеристики:

$$\widehat{\beta}_2 = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}, \quad \widehat{\beta}_1 = \overline{Y} - \widehat{\beta}_2 \overline{X}.$$

Вычисленные значения оценок:

$$\widehat{\beta}_1 \approx 12.5457, \quad \widehat{\beta}_2 \approx 0.1614.$$

Уравнение полученной линии регрессии: $\widehat{Y} = 12.55 + 0.16X$.

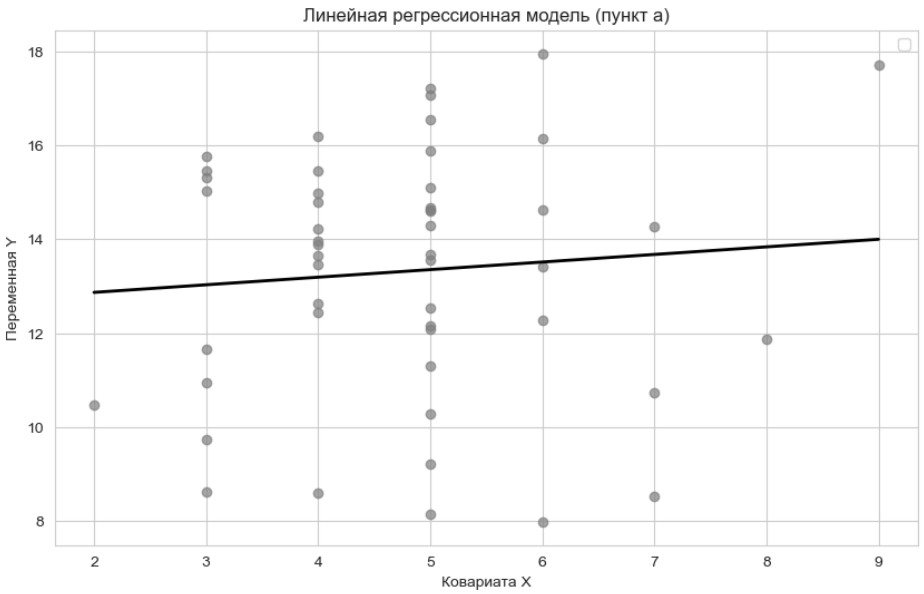

Рис. 1: Экспериментальные данные и линия регрессии.

Визуальный анализ графика (рис. 1) показывает несоответствие линейной модели наблюдаемым данным.

Пункт (b). Полиномиальная модель

Рассмотрим полиномиальную регрессионную модель второго порядка для описания зависимости переменной Y от ковариаты X:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \varepsilon_i, \quad i = 1, \ldots, n,$$

где ε_i — независимые ошибки с $\mathrm{E}\varepsilon_i = 0$ и $\mathrm{Var}(\varepsilon_i) = \sigma^2$. Такая модель является частным случаем линейной регрессии по расширенному вектору регрессоров $x(X_i) = (1, X_i, X_i^2)^\top$.

Обозначим через

$$\mathbf{Y} = (Y_1, \ldots, Y_n)^\top, \quad \boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top,$$

а матрицу плана $\mathbf{X}$ запишем в виде

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 & X_1^2 \\ 1 & X_2 & X_2^2 \\ \vdots & \vdots & \vdots \\ 1 & X_n & X_n^2 \end{pmatrix}.$$

Тогда модель представляется в виде $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, и, согласно методу наименьших квадратов, оценка параметра регрессии $\widehat{\boldsymbol{\beta}}$ минимизирует сумму квадратов остатков $SS(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$, являясь решением нормальных уравнений:

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

В результате вычислений получены следующие оценки параметров (округленно):

$$\widehat{\beta}_1 \approx 11.9441, \quad \widehat{\beta}_2 \approx 0.4108, \quad \widehat{\beta}_3 \approx -0.0239.$$

Таким образом, уравнение эмпирической регрессии принимает вид:

$$\widehat{Y}(x) = 11.9441 + 0.4108\, x - 0.0239\, x^2.$$

На рис. 2 представлены экспериментальные данные, а также графики линейной (из пункта а) и полученной полиномиальной моделей.

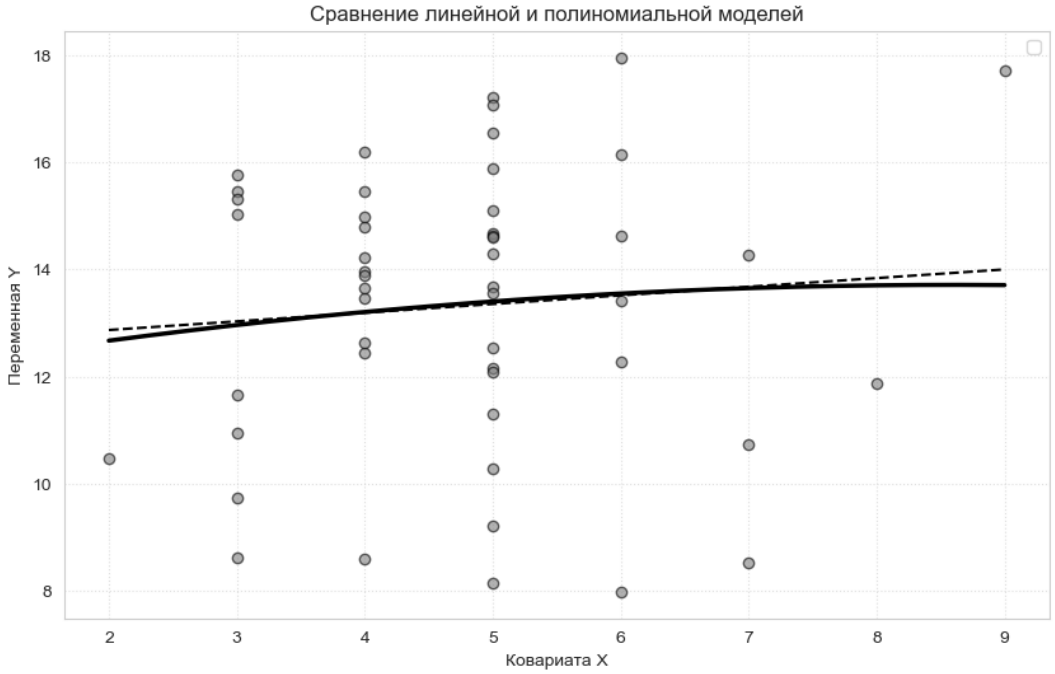

Рис. 2: Сравнение линейной и полиномиальной регрессионных моделей.

Графический анализ показывает, что полиномиальная модель визуально слабо отличается от линейной. Это объясняется малым абсолютным значением коэффициента при квадратичном члене ($\widehat{\beta}_3 \approx -0.02$), что делает параболу пологой. Для окончательного суждения о предпочтительности одной из моделей требуется дальнейший статистический анализ значимости коэффициентов и остатков.

Пункт (c). Анализ нормальности остатков

Для проверки предположения о законе распределения ошибок сформулируем статистические гипотезы:

$$H_0 : \varepsilon_i \sim N(\mu, \sigma^2) \quad \text{(остатки подчиняются нормальному закону)},$$

$$H_1 : \varepsilon_i \nsim N(\mu, \sigma^2).$$

В соответствии с заданием, на базе ошибок полиномиальной модели ($e_i = Y_i - \widehat{Y}_i$) была построена гистограмма частот с шагом группировки $h = 1.60$ (рис. 3).

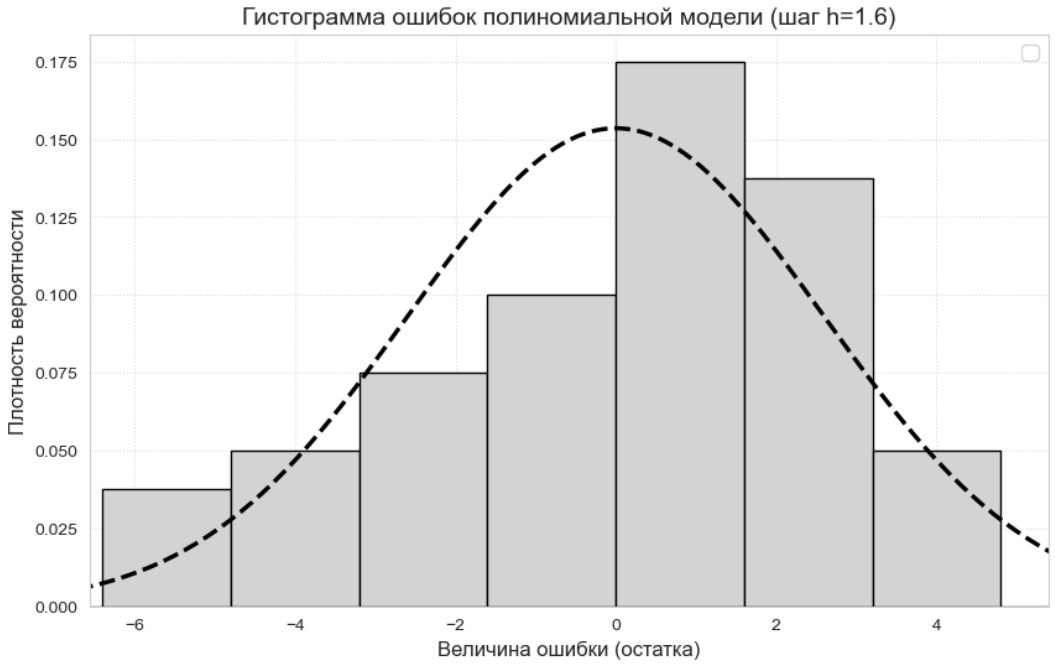

Рис. 3: Гистограмма ошибок полиномиальной модели (шаг $h = 1.60$).

Для проверки гипотезы о нормальном распределении остатков использован критерий согласия χ^2 Пирсона. Чтобы обеспечить корректность применения критерия (ожидаемые частоты в интервалах не менее 5), произведено объединение соседних интервалов с малыми частотами. Результат группировки представлен на рис. 4.

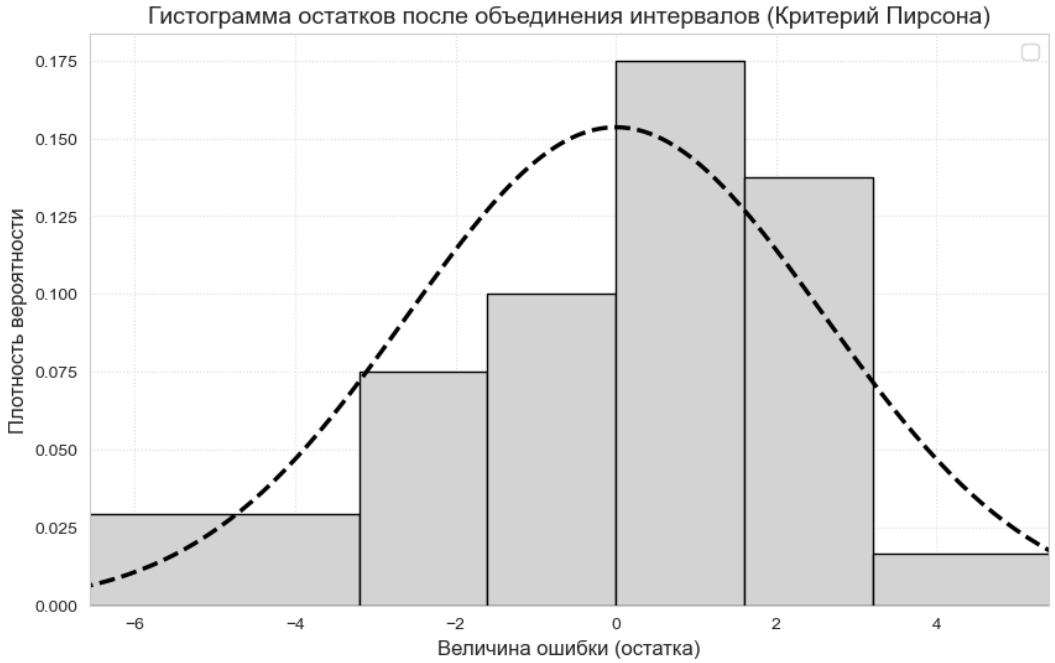

Рис. 4: Гистограмма остатков после объединения интервалов для проверки по критерию χ^2.

Значение статистики критерия χ^2 вычисляется по формуле Пирсона:

$$\chi^2_{\text{набл}} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j},$$

где:

- O_j — наблюдаемая частота (количество остатков) в j-м интервале $[a_j, b_j)$ после объединения;
- $E_j = n \cdot P_j$ — теоретическая (ожидаемая) частота попадания случайной величины в j-й интервал при нормальном распределении $N(\widehat{\mu}, \widehat{\sigma}^2)$.

Вероятность P_j вычисляется как:

$$P_j = \Phi\left(\frac{b_j - \widehat{\mu}}{\widehat{\sigma}}\right) - \Phi\left(\frac{a_j - \widehat{\mu}}{\widehat{\sigma}}\right),$$

где $\Phi(z)$ — функция стандартного нормального распределения.

Детализация расчета теоретических вероятностей:

Для расчетов использованы следующие оценки параметров распределения остатков, полученные по выборке:

$$\widehat{\mu} \approx 0, \quad \widehat{\sigma} \approx 2.5978.$$

Значения нормированных границ интервалов ($z = \frac{x-\widehat{\mu}}{\widehat{\sigma}}$), соответствующих вероятностей P_j и ожидаемых частот E_j приведены в таблице 1.

Интервал $[a_j; b_j)$	z_{a_j}	z_{b_j}	Вероятность P_j	Ожидаемая частота E_j
$[-8.00; -3.20)$	-3.080	-1.232	0.1082	5.41
$[-3.20; -1.60)$	-1.232	-0.616	0.1603	8.02
$[-1.60; 0.00)$	-0.616	0.000	0.2315	11.58
$[0.00; 1.60)$	0.000	0.616	0.2315	11.58
$[1.60; 3.20)$	0.616	1.232	0.1603	8.02
$[3.20; 8.00)$	1.232	3.080	0.1082	5.41
Сумма	—	—	1.0000	50.00

Таблица 1: Расчет вероятностей попадания в интервалы для нормального распределения $N(0, 2.5978^2)$.

Результаты проверки значимости:

- Наблюдаемое значение статистики критерия: $\chi^2_{\text{набл}} \approx 4.0653$.

- Число степеней свободы: Так как среднее значение остатков регрессии тождественно равно нулю ($\bar{e} \equiv 0$), оцениваемым параметром является только дисперсия (σ^2), то есть $p = 1$.

$$df = k - 1 - p = 6 - 1 - 1 = 4.$$

- Критическое значение при уровне значимости $\alpha = 0.10$ и $df = 4$: $\chi^2_{\text{крит}}(0.10; 4) \approx 7.7794$.
- Достигаемый уровень значимости (P-value): $p \approx 0.3972$.

Так как $\chi^2_{\text{набл}} < \chi^2_{\text{крит}}$ ($4.0653 < 7.7794$), нет оснований отвергнуть нулевую гипотезу. Гипотеза о нормальности остатков принимается.

Пункт (d). Частные и совместные доверительные интервалы для β_2 и β_3

В предположении нормальности ошибок полиномиальной модели

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

МНК-оценки параметров $\widehat{\beta}_2$ и $\widehat{\beta}_3$ имеют (при фиксированных X_i) распределение Стьюдента, и для них могут быть построены доверительные интервалы стандартного вида.

По результатам оценки модели получены:

$$\widehat{\beta}_2 = 0.4108, \qquad \widehat{\beta}_3 = -0.0239,$$

стандартные ошибки

$$s_{\widehat{\beta}_2} = 1.4099, \qquad s_{\widehat{\beta}_3} = 0.1324,$$

число наблюдений $n = 50$, число параметров модели $p = 3$, поэтому число степеней свободы $df = n - p = 47$.

Частные доверительные интервалы уровня доверия $1 - \alpha$:

Частный доверительный интервал для параметра β_j имеет вид

$$\widehat{\beta}_j \pm t_{1-\alpha/2,\, df}\, s_{\widehat{\beta}_j},$$

где $t_{1-\alpha/2,\, df}$ — квантиль распределения Стьюдента. При $\alpha = 0.10$ и $df = 47$:

$$t_{1-\alpha/2,\, 47} = t_{0.95,\, 47} \approx 1.6779.$$

Тогда

$$\beta_2 \in \left[\, 0.4108 \pm 1.6779 \cdot 1.4099 \,\right] = [-1.9549;\, 2.7765],$$

$$\beta_3 \in \left[\, -0.0239 \pm 1.6779 \cdot 0.1324 \,\right] = [-0.2460;\ 0.1983].$$

Совместные доверительные интервалы для пары (β_2, β_3) (метод Бонферрони):

Для получения совместной доверительной области уровня $1-\alpha$ для двух параметров β_2 и β_3 используем приём Бонферрони. Строим интервалы для каждого параметра с уровнем доверия $1-\alpha/2$:

$$\widehat{\beta}_j \pm t_{1-\alpha/(2m),\,df}\, s_{\widehat{\beta}_j}, \qquad m = 2.$$

При $\alpha = 0.10$ и $m = 2$ получаем

$$t_{1-\alpha/(2m),\,47} = t_{0.975,\,47} \approx 2.0117.$$

Отсюда совместные (бонферрониевские) интервалы:

$$\beta_2 \in \left[\, 0.4108 \pm 2.0117 \cdot 1.4099 \,\right] = [-2.4255;\ 3.2472],$$

$$\beta_3 \in \left[\, -0.0239 \pm 2.0117 \cdot 0.1324 \,\right] = [-0.2902;\ 0.2424].$$

Как частные, так и совместные доверительные интервалы для β_2 и β_3 содержат нулевое значение. Это означает, что на уровне значимости $\alpha = 0.10$ нет статистически значимых доказательств отличия соответствующих коэффициентов регрессии от нуля.

Пункт (e). Проверка линейности и независимости

Рассматривается полиномиальная модель

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2),$$

в которой по данным выборки получены оценки

$$\widehat{\beta}_2 = 0.4108, \quad s_{\widehat{\beta}_2} = 1.4099, \qquad \widehat{\beta}_3 = -0.0239, \quad s_{\widehat{\beta}_3} = 0.1324,$$

число наблюдений $n = 50$, число параметров $p = 3$, число степеней свободы $df = n - p = 47$.

1. Проверка гипотезы линейности зависимости. Гипотеза линейности соответствует отсутствию квадратичного члена:

$$H_0^{\text{lin}} : \beta_3 = 0, \qquad H_1^{\text{lin}} : \beta_3 \neq 0.$$

В рамках классической линейной регрессии в предположении нормальности ошибок в качестве критерия используется t-статистика

$$T_{\beta_3} = \frac{\widehat{\beta_3}}{s_{\widehat{\beta_3}}} \sim t_{df} \quad \text{при } H_0^{\text{lin}}.$$

Подставляя численные значения, получаем

$$T_{\beta_3} = \frac{-0.0239}{0.1324} \approx -0.18.$$

Критическое значение при уровне значимости $\alpha = 0.10$ и $df = 47$:

$$t_{1-\alpha/2,47} = t_{0.95,47} \approx 1.6779.$$

Достигаемый уровень значимости (P-value):

$$p_{\text{value}} = 2 \cdot P(t_{47} > |-0.18|) \approx 0.8576.$$

Так как $|T_{\beta_3}| < t_{\text{крит}}$ (и $p_{\text{value}} > 0.10$), нет оснований отвергнуть гипотезу H_0^{lin}. Следовательно, введение квадратичного члена X^2 статистически не обосновано.

2. Проверка гипотезы независимости Y от X. Независимость наблюдаемой переменной Y от ковариаты X в полиномиальной модели означает незначимость регрессии в целом (равенство нулю всех коэффициентов при регрессорах):

$$H_0^{\text{ind}} : \beta_2 = 0,\ \beta_3 = 0, \qquad H_1^{\text{ind}} :\ \beta_2^2 + \beta_3^2 > 0.$$

Для проверки данной гипотезы используем F-критерий Фишера. Статистика критерия:

$$F = \frac{(SS_{total} - SS_{res})/(p-1)}{SS_{res}/(n-p)} \sim F_{p-1,n-p},$$

где $p - 1 = 2$ — число проверяемых ограничений. Наблюдаемое значение статистики:

$$F_{\text{набл}} \approx 0.185.$$

Критическое значение F-распределения при $\alpha = 0.10$:

$$F_{\text{крит}}(0.10; 2, 47) \approx 2.42.$$

Достигаемый уровень значимости:

$$p_{\text{value}} = P(F_{2,47} > 0.185) \approx 0.832.$$

Так как $F_{\text{набл}} < F_{\text{крит}}$ $(0.185 < 2.42)$, гипотеза H_0^{ind} принимается.

Вывод: Статистический анализ не выявил существенного отклонения от линейной формы зависимости (пункт 1) и не даёт оснований утверждать наличие какой-либо значимой зависимости Y от X (пункт 2).

Пункт (f). Выбор модели (AIC, BIC)

Рассматриваются три модели:

$$M_0 : Y = \beta_1 + \varepsilon, \qquad M_1 : Y = \beta_1 + \beta_2 X + \varepsilon, \qquad M_2 : Y = \beta_1 + \beta_2 X + \beta_3 X^2 + \varepsilon,$$

где $\varepsilon \sim N(0, \sigma^2)$, объём выборки $n = 50$.

Информационные критерии вычисляются по формулам

$$\text{AIC} = -2 \ln L(\widehat{\theta}) + 2k, \qquad \text{BIC} = -2 \ln L(\widehat{\theta}) + k \ln n,$$

где k — число параметров регрессии (коэффициентов β), n — объём выборки.

Число параметров для каждой модели:

$$k_0 = 1 \quad (\beta_1), \qquad k_1 = 2 \quad (\beta_1, \beta_2), \qquad k_2 = 3 \quad (\beta_1, \beta_2, \beta_3).$$

По результатам оценивания получены следующие значения критериев:

Модель	AIC	BIC
M_0 (Нулевая)	239.75	241.66
M_1 (Линейная)	241.40	245.22
M_2 (Полиномиальная)	243.36	249.10

Видно, что

$$\text{AIC}(M_0) < \text{AIC}(M_1) < \text{AIC}(M_2), \qquad \text{BIC}(M_0) < \text{BIC}(M_1) < \text{BIC}(M_2).$$

Оба критерия минимизируются на нулевой модели M_0, то есть добавление линейного и квадратичного членов X, X^2 не улучшает модель с учётом штрафа за сложность. Следовательно, среди рассмотренных моделей по AIC и BIC предпочтительной является нулевая модель M_0.

Общий вывод

Пункт (g). Интерпретация результатов

Проведённый регрессионный анализ показал, что как линейная, так и полиномиальная модели дают очень близкую подгонку к данным: добавление квадратичного члена X^2 практически не изменяет значения критерия качества, а оценка $\widehat{\beta}_3$ статистически незначима на уровне $\alpha = 0.10$. Проверка нормальности остатков полиномиальной модели с помощью критерия согласия χ^2 не выявила существенных отклонений от нормального закона, что подтверждает корректность применения МНК и асимптотических процедур проверки гипотез.

Доверительные интервалы для коэффициентов β_2 и β_3 (как частные, так и совместные по методу Бонферрони) содержат нулевое значение, а t- и F-критерии не дают оснований отвергать гипотезы линейности и независимости Y от X на уровне значимости $\alpha = 0.10$. Информационные критерии AIC и BIC минимизируются на нулевой модели $M_0 \colon Y = \beta_1 + \varepsilon$, то есть включение регрессоров X и X^2 не улучшает модель с учётом штрафа за сложность. Таким образом, на основе имеющейся выборки статистически значимой зависимости Y от X не обнаружено, и адекватным описанием данных служит модель с постоянным средним значением без включения ковариаты X.