Санкт-Петербургский Политехнический Университет имени Петра Великого
Физико-Механический Институт

Отчет по лабораторной работе №4
по дисциплине
"Многомерный Статистический Анализ"
Вариант 12

Выполнил студент:
Тырыкин Я. А.
группа 5030102/80401
Преподаватель:
Павлова Л. В.

Санкт-Петербург
2022

Содержание

1 Формулировка задачи

В данной лабораторной работе требуется построить регрессионную модель для реальных данных, полученных в результате проведения химических экспериментов. Также требуется построить ее МНК оценки, проверить основные линейные гипотезы, построить доверительные интервалы для ее коэффициентов и визуализировать все полученные результаты.

2 Описание экспериментальных данных

Проводится химический эксперимент, заключающийся в изучении субстрата **B3**, полученного при реакции веществ **B1** и **B2**. Реакция **B1** с **B2** происходит при участии катализатора **K**. Результат реакции - выход вещества **B3** - зависит от пропорции веществ **B1** и **B2**.

Количество вещества **B2** остается неизменным, меняется лишь концентрация вещества **B1**. Проводится 15 экспериментов, каждый эксперимент проводится при некоторой температуре и некотором количестве катализатора. Остальные условия проведения реакции во всех экспериментах (давление, среда проведения и т.д.) остаются неизменными.

В файле "$Y.txt$"располагаются данные о выходах всех 15 экспериментов (количество вещества B_3 в кг). Замер выхода реакции производится по истечении некоторого временного интервала τ.

В файле "$X.txt$"располагаются данные в виде таблицы размером 15x4, которая состоит из следующих столбцов:

Последний параметр является фиктивным, введенным искусственно, и в каждому наблюдаемом эксперименте принимает значения, равные 1.

- x_{t1} - количество вещества B_1 (в кг) в $t-$ом эксперименте
- x_{t2} - температура (C°) в $t-$ом эксперименте
- x_{t3} - количество катализатора K (в г) в $t-$ом эксперименте

3 Построение регрессионной модели

Пусть $x_1, ..., x_m$ - ряд признаков, определяющих случайную величину **y**. В случае данного исследования $m = 4$, а случайной величиной **y** будем считать количество вещества B_3 (в кг) на выходе каждого эксперимента.

Будем строить стохастическую зависимость случайной величины **y** от 4 предикторов x_1, x_2, x_3, x_4 в виде линейной регрессии, представимой следующим образом:

$$y_t = \alpha_1 x_{t1} + \alpha_2 x_{t2} + \alpha_3 x_{t3} + \alpha_4 x_{t4} + \epsilon_t$$

где $t = 1, ..., n(n = 15)$ - число наблюдений, $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$ - параметры регрессии, ϵ - случайные отклонения.

Можно представить линейную регрессию в следующем матричном виде:

$$y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{pmatrix}, y = X\alpha + \epsilon$$

4 Оценки параметров модели

Для оценивания описанных в предыдущем пункте параметров модели воспользуемся методом наименьших квадратов (МНК).

Сначала оценим параметры регрессии по следующей формуле:

$$a = \hat{\alpha} = (X^T X)^{-1} X y$$

где $X-$ матрица предикторов, в нашем случае размерности 15x4, X^T - ее транспонированый вариант, $y-$ вектор значений случайной величины. Результат вычисления данного параметра на наших экспериментальных данных:

$$a = \begin{pmatrix} 0.39747108 \\ 0.67599168 \\ 2.28282375 \\ -43.73977273 \end{pmatrix}$$

Далее найдем оценку дисперсии отклонений регрессионной модели $\sigma^2 = D[\epsilon_t], t = 1, ..., n, \overline{e} = 0$ по следующей формуле:

$$s^2 = \frac{\sum_{t=1}^{n} e_t^2}{n-m} = \frac{e^T e}{n-m} = \frac{(y-\hat{y})^T(y-\hat{y})}{n-m} = \frac{\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}{n-m}$$

Получим следующее значение:

$$s^2 = 4.927065$$

После нахождения данной оценки можем оценить матрицу ковариации оценки параметров регрессии $\hat{cov}(a)$ по следующей формуле:

$$\hat{cov}(a) = s^2(X^T X)^{-1}$$

Численно получаем следующее:

$$\hat{cov}(a) = \begin{pmatrix} 1.20347482e-03 & -6.92194963e-03 & -1.00746270e-02 & 4.36387044e-01 \\ -6.92194963e-03 & 1.07849935e-01 & -1.90052366e-01 & -7.05106685e+00 \\ -1.00746270e-02 & -1.90052366e-01 & 1.13021814e+00 & 1.13684392e+01 \\ 4.36387044e-01 & -7.05106685e+00 & 1.13684392e+01 & 4.78175569e+02 \end{pmatrix}$$

Кроме того, построим стандартные ошибки оценок каждого из параметров регрессии $s(a_i)$:

$$s(a_i) = s\big(X^T X\big)^{-\frac{1}{2}}_{ii}, i \in \overline{1,4}$$

Численно получим следующие значения:

- $s(a_1) = 0.00728941$
- $s(a_2) = 0.06024949$
- $s(a_3) = 0.93837527$
- $s(a_4) = 21.85918285$

Также посчитаем матрицу корреляций $corr(a)$ с ячейками, представимыми как:

$$corr_{ij}(a) = \frac{(X^T X)^{-1}_{ij}}{\sqrt{(X^T X)^{-1}_{ii}(X^T X)^{-1}_{jj}}}$$

где $(X^T X)^{-1}_{ij}$ - элемент в $i-$ой строке и $j-$ом столбце обратной матрицы к произведению $(X^T X)$.

Численно получим:

$$corr(a) = \begin{pmatrix} 1. & -0.60757513 & -0.27316769 & 0.57525401 \\ -0.60757513 & 1. & -0.54435498 & -0.98186368 \\ -0.27316769 & -0.54435498 & 1. & 0.48901932 \\ 0.57525401 & -0.98186368 & 0.48901932 & 1. \end{pmatrix}$$

После построения оценок визуально оценим качество полученной модели, построив графики предсказанных выходов реакций и их точных значений:

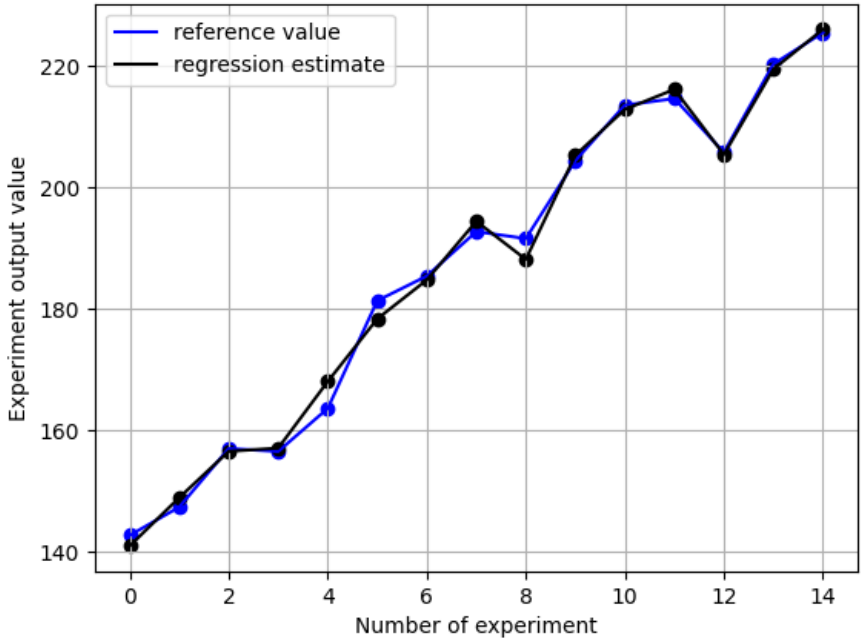

Рис. 1: Сравнение построенной регрессионной модели с точными данными выходов в каждом эксперименте

Гистограмма отстатков (ошибок предсказания модели относительно точных данных о выходах экспериментальных реакций):

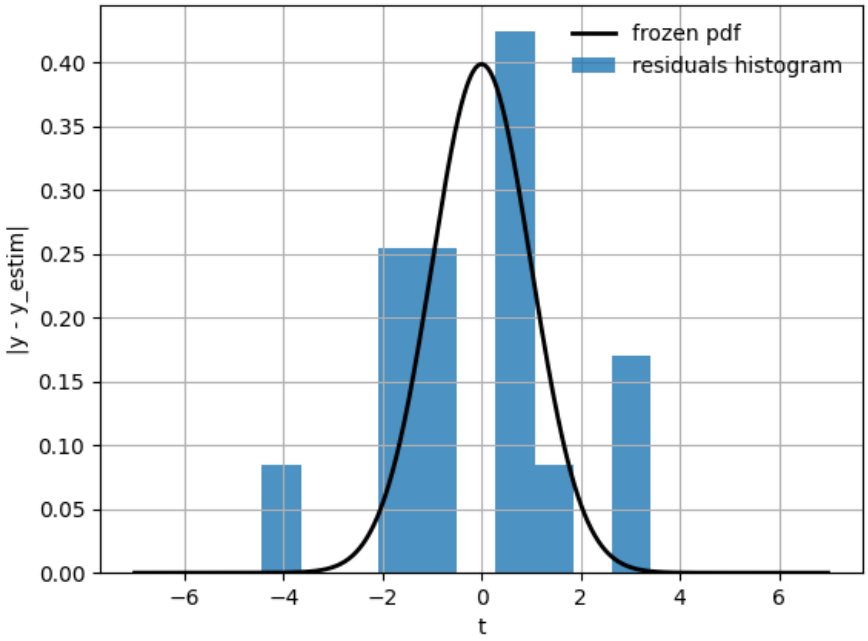

Рис. 2: Сравнение гистограммы остатков построенной линейной регрессии с нормальным распределением $N(0,1)$

5 Построение детерминированных оценок

Построим коэффициент детерминации модели по двум формулам:

$$R^2 = 1 - \frac{\left[\sum_{t=1}^{n} e_t^2\right]}{\left[\sum_{t=1}^{n} (y_t - \overline{y}_t)^2\right]}, n = 15, m = 4$$

$$R_H^2 = 1 - \frac{\left[\frac{\sum_{t=1}^{n} e_t^2}{n-m}\right]}{\left[\frac{\sum_{t=1}^{n} (y_t - \overline{y}_t)^2}{n-1}\right]}, n = 15, m = 4$$

Коэффициент детерминации показывает, насколько построенная регрессионная модель лучше модели среднего. Второй коэффициент при этом еще и является несмещенной оценкой. Приведем полученные численные значения данных коэффициентов:

$$R^2 = 0.99492, R_H^2 = 0.99354$$

Видно, что оба значения близки к 1, значит, модель существенно лучше модели среднего.

6 Проверка гипотез

6.1 Построение индивидуальных доверительных интервалов для коэффициентов регрессии

Доверительные интервалы строим по формуле:

$$D_i = \left\{[a_i - s_i t_{(1+\gamma)/2,S(n-m)}, a_i + s_i t_{(1+\gamma)/2,S(n-m)}]\right\}$$

где $t_{(1+\gamma)/2,S(n-m)}$ - квантиль уровня $\frac{1+\gamma}{2}$ распределния Стьюдента с $n-m$ степенями свободы, a_i - оценка $i-$ го параметра регрессии, s_i - средняя квадратичная ошибка $i-$ го коэффициента регрессии. Численно высчитывая данное значение для каждого из наших 4 параметров регрессии, получим:

- $D_1 = \left\{[0.39176145, 0.40318071]\right\}$
- $D_2 = \left\{[0.62879963, 0.72318373]\right\}$
- $D_3 = \left\{[1.54781587, 3.01783164]\right\}$
- $D_4 = \left\{[-60.8615708, -26.61797467]\right\}$

При вычислении данных величин был использован параметр $\alpha = 0.45$, $\frac{1+\gamma}{2} = 1-\frac{\alpha}{2}$.

6.2 Построение обобщенной доверительной области

Обобщенная доверительная область строится в форме обобщенного прямоугольного параллелепипеда согласно принципу Тьюки и неравенству Чебышева по следующей формуле:

$$D|D_i = \left\{a_i - \tau s_i, a_i + \tau s_i\right\}, i = \overline{1,m}, m = 4$$

Интервалы D_i ограничивают размеры $m-$мерного параллелепипеда, являющегося обобщенной доверительной областью. Их числовые значения равны:

- $D_1 = \left\{[0.38660467, 0.40833749]\right\}$
- $D_2 = \left\{[0.58617704, 0.76580632]\right\}$

- $D_3 = \{[0.8839765, 3.68167101]\}$
- $D_4 = \{[-76.3255186, -11.15402687]\}$

При вычислении данных величин был использован параметр $\alpha = 0.45$. Второй же параметр, τ, брался согласно принципу Тьюки:

$$1 - \frac{\alpha}{m} = 1 - \frac{1}{m\tau^2} \to \tau = \sqrt{\frac{1}{\alpha}} \approx 1.4907$$

Стоит отметить, что при уменьшении τ доверительные интервалы, составляющие обобщенную доверительную область, сужаются, и границы ближе сходятся к значениям оценок параметров регрессии. Например, при $\tau = 0.231$ получаем следующий обобщенный прямоугольный параллелепипеда:

- $D_1 = \{[0.39579452, 0.39914765]\}$
- $D_2 = \{[0.6621343, 0.68984907]\}$
- $D_3 = \{[2.06699744, 2.49865006]\}$
- $D_4 = \{[-48.76738479, -38.71216067]\}$

6.3 Проверка гипотезы о равенстве отдельных коэффициентов регрессии нулю

Проверим гипотезу H_0 для каждого из параметров регрессии $\mathbf{a}$, предполагающую равенство α_i нулю:

$$H_0 : \alpha_i = 0$$

Далее задаем статистику t уровнем значимости $\alpha = 0.45$ (но, в целом, можно брать любое значение $\alpha \in (0, 1)$) и на ее основе строим следующее критическое множество:

$$\Im_{1\alpha} = \left(t | t = \frac{(|a_i|)}{s(a_i)} > t_\alpha \right), t_\alpha = t_{1-\frac{\alpha}{2}, S(n-m)}, i = \overline{1, m}, m = 4$$

где t_α - квантиль распределения Стьюдента уровня $1 - \frac{\alpha}{2}$ с $n - m$ степенями свободы, $s(a_i) = s(X^T X)^{-\frac{1}{2}}_{ii}$ - оценка стандартного отклонения оценки параметра регрессии

a_i. В случае, если неравенство, стоящее в определении множества выполняется, гипотеза H_0 отвергается, следовательно, полагаем параметр регрессии a_i отличным от нуля.

Высчитывая численно критическое множество, получаем следующие значения статистик для параметров и значение квантиля распределения Стьюдента:

- $t_1 = 54.52718663$
- $t_2 = 11.21987322$
- $t_3 = 2.43274075$
- $t_4 = 2.00097931$

$$t_\alpha = t_{1-\frac{\alpha}{2},S(n-m)} = 0.783277$$

Для каждого параметра в силу того, что неравенство в определении критического множества выполняется, гипотезу H_0 отвергаем. Утверждаем с этого момента, что ни один параметр регрессии нельзя считать нулевым в случае данной работы.

6.4 Проверка гипотезы об идентичности двух моделей

Для проверки данной гипотезы разобьем исходные данные на две выборки размеров $n_1, n_2 = 8, 7$ и построим регрессии на каждом из полученных наборов данных:

$$y_1 = X_1\alpha_1 + \epsilon_1, n_1$$

$$y_2 = X_2\alpha_2 + \epsilon_2, n_2$$

$$\alpha_1, \alpha_2 \in R^m$$

Проверим гипотезу H_0 об идентичности построенных регрессионных моделей: $\alpha_1 = \alpha_2$. Для этого вычислим обычную сумму квадратов отклонений регрессии и сумму квадратов отклонений регрессии от оцененной регрессии ($\hat{Q}$ и $\hat{Q}_R$ соответственно). Приведем промежуточные математические выкладки вычислительного процесса:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix} \alpha + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \to y = X\alpha + \epsilon$$

$$a_R = (X^T X)^{-1} X^T y$$

$$\hat{Q}_R = (y - Xa_R)^T (y - Xa_R)$$

$$a_i = (X_i^T X_i)^{-1} X_i^T y_i, \hat{Q}_i = (y_i - X_i a_i)^T (y_i - X_i a_i), i \in \overline{1,2}$$

$$\hat{Q} = \hat{Q}_1 + \hat{Q}_2, \hat{Q}_R - \hat{Q} = \hat{Q}_R - \hat{Q}_1 - \hat{Q}_2$$

$$s^2 = \frac{\hat{Q}_1 + \hat{Q}_2}{n_1 + n_2 - 2m}$$

Далее задаем статистику t уровнем значимости $\alpha = 0.45$ (но, в целом, можно брать любое значение $\alpha \in (0, 1)$) и на ее основе строим следующее критическое множество:

$$\Im_{1\alpha} = \left(t | t = \frac{\frac{\hat{Q}_R - \hat{Q}_1 - \hat{Q}_2}{m}}{\frac{\hat{Q}_1 + \hat{Q}_2}{n_1 + n_2 - 2m}} > t_\alpha \right), t_\alpha = t_{1-\alpha, F(m, n_1+n_2-2m)}, i = \overline{1,m}, m = 4$$

где t_α - квантиль распределения Фишера уровня $1-\alpha$ с m и n_1+n_2-2m степенями свободы. В случае, если неравенство, стоящее в определении множества выполняется, гипотеза H_0 отвергается, следовательно, полагаем дисперсии не идентичными друг другу.

Проверим гипотезу H_0 дважды при разном разделении исходных данных: в первом случае выборка X_1 содержит все наблюдения с четными индексами, а выборка X_2 - все наблюдения с нечетными индексами; во второй раз в выборку X_1 выбираются случаным образом $n_1 = 8$ наблюдений, остальные отправляются в выборку X_2.

Вычислим параметры, не зависящие от разделения исходной выборки:

$$a_R = [0.39747108, 0.67599168, 2.28282375, -43.73977]$$

$$\hat{Q}_R = 54.19770473$$

Проверим гипотезу H_0 в первом случае разделения данных:

$$a_1 = [0.38263, 0.92499, 1.7300, -58.75487]$$

$$a_2 = [0.38755, 0.70619, 2.61733, -47.583932]$$

$$\hat{Q}_1 = 34.39507, \hat{Q}_2 = 17.24537038$$

$$s^2 = 7.3772$$

$$t_\alpha = 1.04104, t = 0.0866$$

Проверим гипотезу H_0 во втором случае разделения данных:

$$a_1 = [0.40523, 0.4463, 1.6312, -14.146]$$

$$a_2 = [0.3704, 0.6498, 3.5443, -46.4458]$$

$$\hat{Q}_1 = 15.8757, \hat{Q}_2 = 17.41683$$

$$s^2 = 4.75608$$

$$t_\alpha = 1.04104, t = 1.09886$$

По построенным параметрам видно, что в случае случайного разбиения исходной выборки гипотеза об идентичности регрессий α_1 и α_2, построенных на наборах X_1 и X_2, отвергается, в то время как в случае разделения данных по четности номера наблюдения мы наоборот гипотезу H_0 принимаем и считаем полученные регрессии равными. Помимо словесного описания экспериментов, ниже приведено графическое сравнение полученных регрессий.

Сначала сравним качество построенных регрессий на выборках, разделенных по четности номера наблюдения:

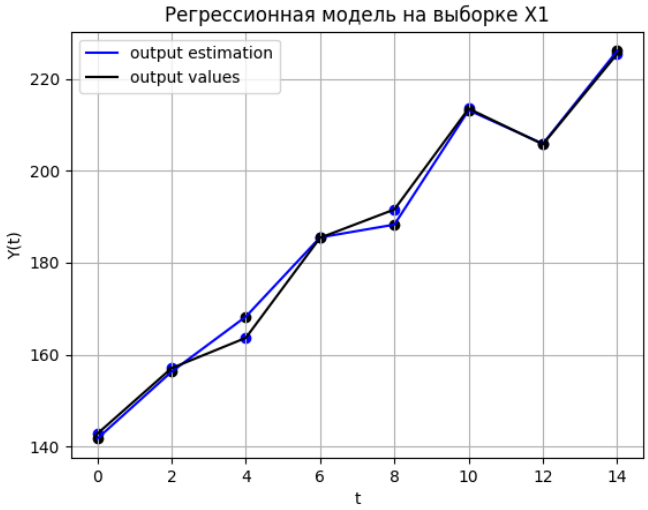

Рис. 3: Сравнение регрессионной модели, построенной на выборке X_1, с точными значениями откликов

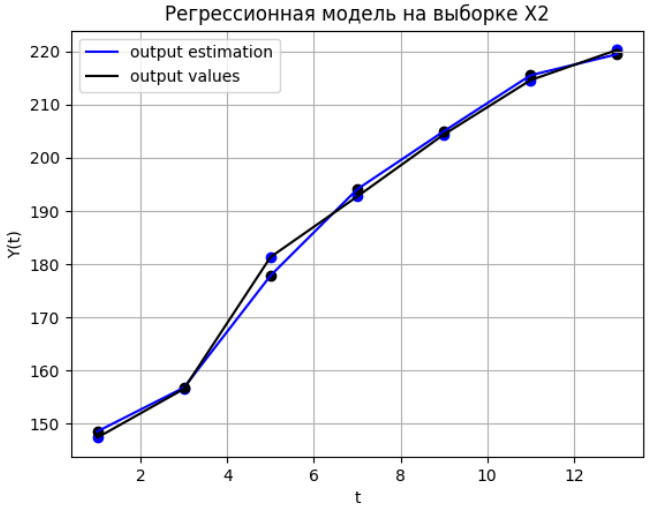

Рис. 4: Сравнение регрессионной модели, построенной на выборке X_2, с точными значениями откликов

А теперь оценим качество моделей на выборках, разделенных случайным образом:

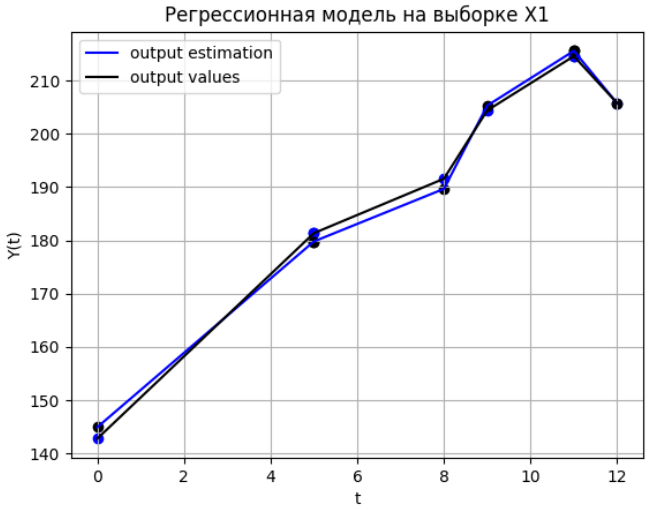

Рис. 5: Сравнение регрессионной модели, построенной на выборке X_1, с точными значениями откликов

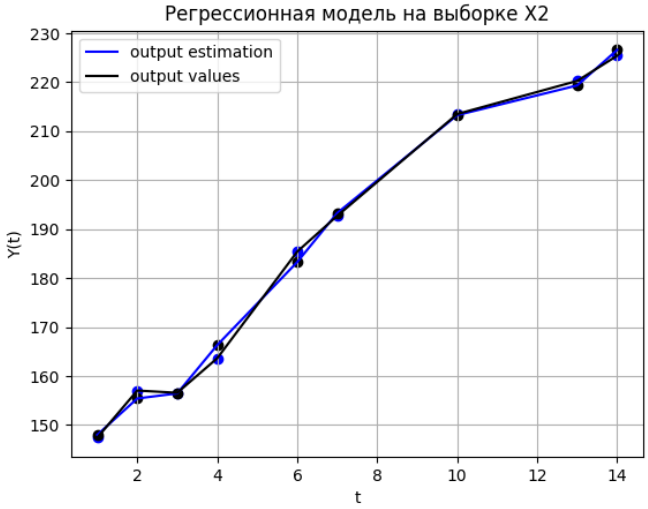

Рис. 6: Сравнение регрессионной модели, построенной на выборке X_2, с точными значениями откликов

В целом на графиках видно, что при случайном разделении наших 15 экспериментов на 2 выборки отклонения прогнозов регрессионной модели от точных значений значения и встречаются чаще.

7 Прогнозирование на построенной регрессионной модели

Выделим один элемент из исходного набора данных X для прогнозирования его значения с помощью регрессионной модели. Случайным образом выберем один из элементов исходной выборки. В нашем случае был получен индекс 0, поэтому прогноз строился для первого эксперимента. Полученная модель на основании оставшихся наблюдений и ее параметры:

$$a = \left[0.403966, 0.65791, 2.32479, -44.6281\right]$$

$$s^2 = 5.00174002$$

$$s(a) = \left[0.0081, 0.06881, 0.94784, 22.04572\right]$$

$$\hat{cov}(a) = \begin{pmatrix} 1.27218412e-03 & -7.16733980e-03 & -9.90115082e-03 & 4.36097234e-01 \\ -7.16733980e-03 & 1.09875558e-01 & -1.93840737e-01 & -7.13871823 \\ -9.90115082e-03 & -1.93840737e-01 & 1.14945600 & 1.14961247e+01 \\ 4.36097234e-01 & -7.13871823 & 1.14961247e+01 & 4.86367323e+02 \end{pmatrix}$$

Спронозируем значение отклика в выделенном 0-вом эксперименте и оценим его:

$$\hat{y}_\tau = 140.3756, e = y_\tau - \hat{y}_\tau = 2.3927(1.6\% \text{ от значения отклика})$$

Также найдем интервальную оценку прогноза по следующей формуле:

$$D_\tau = \left\{\hat{y}_\tau - t_\gamma s_\tau, \hat{y}_\tau + t_\gamma s_\tau\right\}$$

$$s_\tau^2 = s^2\left[x_\tau^T (X^T X)^{-1} x_\tau\right]$$

В данном случае X - матрица меньшей размерности, нежели в предыдущих экспериментах, в силу того, что одно из наблюдений мы теперь используем для прогнозирования. t_γ - квантиль уровня $1-\frac{\alpha}{2}$ распределения Стьюдента с $n-1-m$ степенями свободы, $n = 15, m = 4$. Численно получим следующий интервал оценки прогноза:

$$D_\tau = \left[138.31807225, 142.43332372\right]$$

График сравнения точных значений откликов, регрессионной модели, построенной на полном множестве экспериментальных данных, и регрессионной модели, от-

куда убрано 0-ое наблюдение:

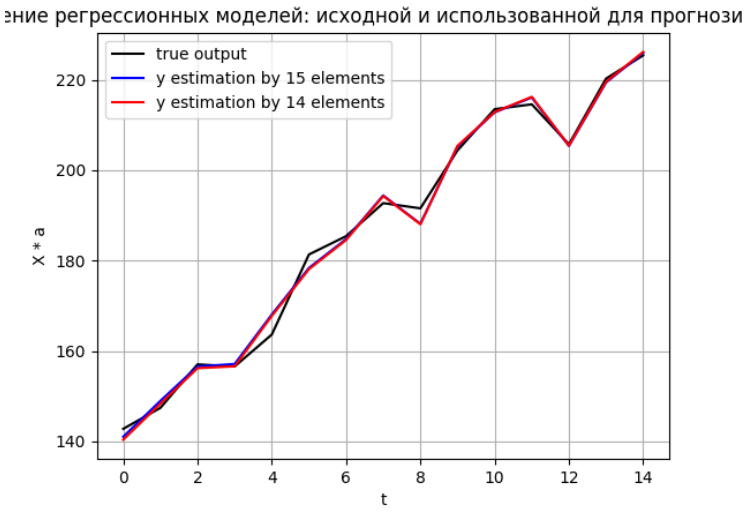

Рис. 7: Сравнение двух регрессионных моделей с точным значением откликов экспериментов

По результатам прогнозирования можно сделать вывод, что с удалением одного наблюдения из исходного набора построенная регрессионная модель не сильно теряет в качестве. В целом прогноз является достаточно точным, учитывая что ошибка составляет всего ли 1.6% от точного значения выхода химического эксперимента.

8 Модульная структура программы

Построение регрессионных моделей, вычисление их оценок, проверки линейных гипотез и визуализация результатов в среде разработки PyCharm на языке программирования Python версии 3.10.1 с применением таких фреймворков, как **numpy**, **scipy**, **matplotlib** .

Репозиторий с кодом данной лабораторной работы находится по данной **ссылке**.

9 Заключение

Линейная регрессия показывает хорошие результаты прогнозирования на экспериментальных данных, на которых проводились все исследования данной работы. В целом, данный метод является мощным прогностическим инструментом, но нельзя

утверждать, что на других данных, имеющих закономерность, сильнее отличающуюся от линейной (в отличие от набора 15 наблюдений, использованных в данной работе), мы получим такое же качество предсказаний. Для этого стоит провести отдельное исследование, выходящее за рамки данной работы.