Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт

**Высшая школа прикладной математики и вычислительной физики**

# Отчет по лабораторной работе №4 по дисциплине «Многомерный статистический анализ»

Вариант 11

Выполнила студентка группы 5030102/90401: О. А. Ковалёва

Преподаватель: Л. В. Павлова

Санкт-Петербург
2023

# Оглавление

Цель ..... 3
Описание эксперимента ..... 3
Ход работы ..... 3
0. Подготовим данные для работы с ними ..... 3
1. Построение регрессионной модели ..... 4
2. Нахождение МНК-оценки параметров регрессии ..... 4
3. Нахождение оценки дисперсии отклонений регрессионной модели ..... 5
4. Нахождение оценки матрицы ковариаций ..... 5
5. Нахождение стандартной ошибки оценки каждого коэффициента ..... 6
6. Построение матрицы корреляций ..... 6
7. Построение гистограммы остатков ..... 6
8. Построение графиков регрессии и оцененной регрессии ..... 7
9. Нахождение коэффициента детерминации ..... 7
10. Построение индивидуальных доверительных интервалов для коэффициентов регрессии 8
11. Построение обобщенной доверительной области для параметров модели ..... 9
12. Проверка основных линейных гипотез ..... 10
Гипотеза о равенстве нулю коэффициентов ..... 10
Гипотеза об адекватности модели среднего ..... 11
Гипотеза об идентичности двух регрессий ..... 11
13. Уменьшение числа параметров модели ..... 12
Оценки новых моделей ..... 12
Построение графиков оцененной регрессии ..... 14
Построение обобщенной доверительной области ..... 14
14. Прогноз наблюдения ..... 16
Выводы ..... 17

## Цель

Рассматривая химический эксперимент, построить и исследовать регрессионную модель.

## Описание эксперимента

Реакция вещества $B_1$ с веществом $B_2$ приводит к образованию другого вещества – $B_3$. Реакция происходит в присутствии катализатора $K$. Результат реакции – выход вещества $B_3$ – зависит от пропорции веществ $B_1$ и $B_2$; количество вещества $B_2$ во всех экспериментах одно и то же; количество вещества $B_1$ меняется. Проводится 15 наблюдений при определенной температуре и некотором количестве катализатора. Остальные условия проведения реакций во всех экспериментах остаются неизменными.

Данные о наблюдениях:

- Y_30.txt – значения вектора наблюдений $y$:

  $y_t$ – выход реакции (количество вещества $B_3$ в кг) в $t-$м эксперименте.
- X.txt – матрица $X$ условий экспериментов, столбцы которой:

  $x_{t1}$ – количество вещества $B_1$(в кг) в $t-$м эксперименте;

  $x_{t2}$ – температура ($°C$) в $t-$м эксперименте;

  $x_{t3}$ – количество катализатора $K$ (г) в $t-$м эксперименте.

$$t = \overline{1,15}$$

## Ход работы

### 0. Подготовим данные для работы с ними

Считаем данные из файла и перейдем от научной записи чисел к естественному формату, используя функцию:

```
def to_natural_format(num_str: str):
    mantissa, exponent = num_str.split("e")
    return float(mantissa) * 10 ** int(exponent)
```

Посмотрим на считанные данные:

```
n = 15, m = 4

Матрица X:
 [[252.36  96.67   8.37   1.  ]
 [262.54 100.07   9.07   1.  ]
 [285.7   96.78   9.35   1.  ]
 [277.52 101.3    9.67   1.  ]
 [307.95 100.35   9.45   1.  ]
 [322.44 104.8   10.12   1.  ]
 [334.88 106.17  10.35   1.  ]
 [350.11 109.2   11.03   1.  ]
 [346.1  104.48  10.38   1.  ]
 [374.91 106.88  12.15   1.  ]
 [378.49 113.14  12.98   1.  ]
 [397.48 112.38  11.34   1.  ]
 [378.39 109.07  10.95   1.  ]
 [393.44 114.45  12.89   1.  ]
 [403.84 115.23  13.71   1.  ]]

Вектор наблюдений y:
 [141.87305 149.40628 161.30455 160.22555 171.35829 179.62725 182.97827
 195.23086 191.41135 208.37234 215.25904 216.13533 204.68334 220.12231
 225.88958]
```

## 1. Построение регрессионной модели

Будем строить классическую линейную регрессию:

$$y_t = \alpha_1 x_{t1} + \alpha_2 x_{t2} + \alpha_3 x_{t3} + \alpha_4 x_{t4} + \varepsilon_t, \qquad t = \overline{1,n}, \qquad n = 15$$

где $x_1, \dots, x_m$ – ряд признаков, $m = 4$, $y$ – случайный вектор наблюдений, $\alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)^T$ – параметры регрессии, $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \dots, \varepsilon_n)$ – случайные отклонения.

В матричном виде:

$$\boldsymbol{y} = X\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} \end{pmatrix}, \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \dots \\ \alpha_4 \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

Последний столбец матрицы $X$ состоит из единиц и соответствует свободному члену уравнения регрессии – $\alpha_4$, введенному для уточнения модели.

## 2. Нахождение МНК-оценки параметров регрессии

МНК-оценка вектора параметров $\boldsymbol{\alpha}$:

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

```
rank = np.linalg.matrix_rank(X)  # должен быть равен 4
print("Ранг матрицы X: ", rank)
X_T = X.T
X_T_X_inv = np.linalg.inv(np.matmul(X_T, X))
a_est = np.matmul(X_T_X_inv, X_T).dot(y)
print("Оценка параметров регрессии:", a_est)
```

```
Ранг матрицы X:  4
Оценка параметров регрессии: [0.41562877 0.01176349 3.94271963 4.10352525]
```

Получаем регрессионную модель:

$$\hat{y_t} = 0.4156x_{t1} + 0.0118x_{t2} + 3.9427x_{t3} + 4.1035x_{t4}$$

Далее будем оценивать характеристики полученной модели.

3. Нахождение оценки дисперсии отклонений регрессионной модели

$$s^2 = \frac{\sum_t (y_t - \hat{y_t})^2}{n - m}$$

```
y_est = X.dot(a_est)
s_2_est = sum((y[i] - y_est[i])**2 for i in range(y.shape[0])) / (n - m)
print("Оценка дисперсии:", s_2_est)
```

```
Оценка дисперсии: 1.8020646675870131
```

4. Нахождение оценки матрицы ковариаций

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

```
cov_matrix_est = s_2_est * X_T_X_inv
print("Оценка матрицы ковариаций:\n", cov_matrix_est)
```

```
Оценка матрицы ковариаций:
 [[ 4.40168712e-04 -2.53169041e-03 -3.68477638e-03  1.59607763e-01]
 [-2.53169041e-03  3.94459165e-02 -6.95113053e-02 -2.57891480e+00]
 [-3.68477638e-03 -6.95113053e-02  4.13375217e-01  4.15798583e+00]
 [ 1.59607763e-01 -2.57891480e+00  4.15798583e+00  1.74891839e+02]]
```

### 5. Нахождение стандартной ошибки оценки каждого коэффициента

$$s_i(a_i) = \sqrt{\widehat{cov_{ii}}(\boldsymbol{a})}$$

```
s_i = np.diag(cov_matrix_est)**0.5
print("Стандартная ошибка оценки коэффициентов:", s_i)
```

```
Стандартная ошибка оценки коэффициентов: [ 0.0209802   0.19860996  0.64294262 13.22466784]
```

### 6. Построение матрицы корреляций

Элементы матрицы имеют вид:

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

```
corr_matrix_est = np.ones((m, m))
for i in range(m):
    for j in range(m):
        corr_matrix_est[i, j] = X_T_X_inv[i, j] / (X_T_X_inv[i, i] * X_T_X_inv[j, j])**0.5

print("Матрица корреляций:\n", corr_matrix_est)
```

```
Матрица корреляций:
 [[ 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.        ]]
```

Видим, что имеет место большая корреляция между 2-м и 4-м параметром модели.

### 7. Построение гистограммы остатков

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

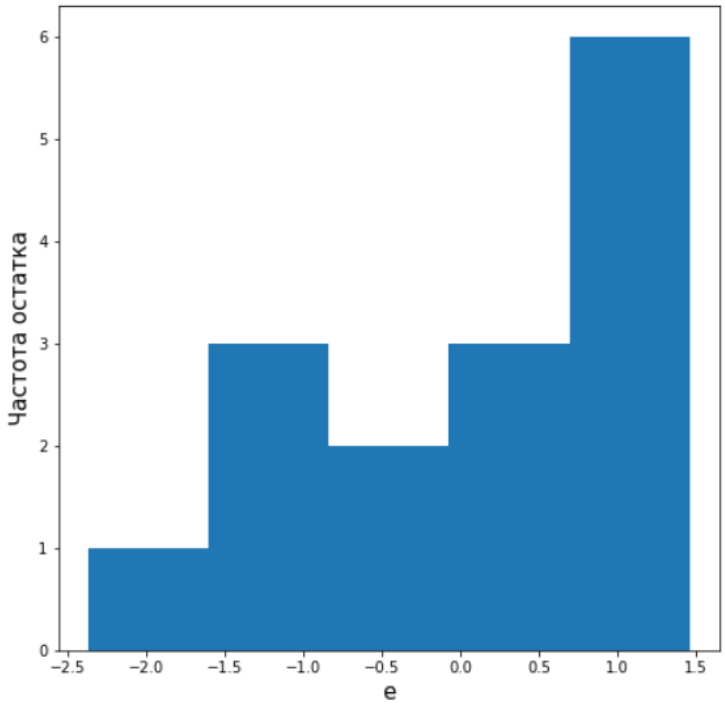



*Рисунок 1. Гистограмма остатков*

```
Вектор остатков: [-1.2562913  -0.75406177  0.45298633  1.4589884   0.8227186   0.37524805
 -2.36709533  0.8387758   1.30422859 -0.71589236  1.33675991  0.79526003
 -1.14577896  0.32581441 -1.47166041]
```

Заметим, что остатки не распределены нормально. Однако полученные значения остатков малы, в сравнении со значениями вектора наблюдений, поэтому можем сделать вывод о том, что модель достаточно хорошо описывает данные.

### 8. Построение графиков регрессии и оцененной регрессии

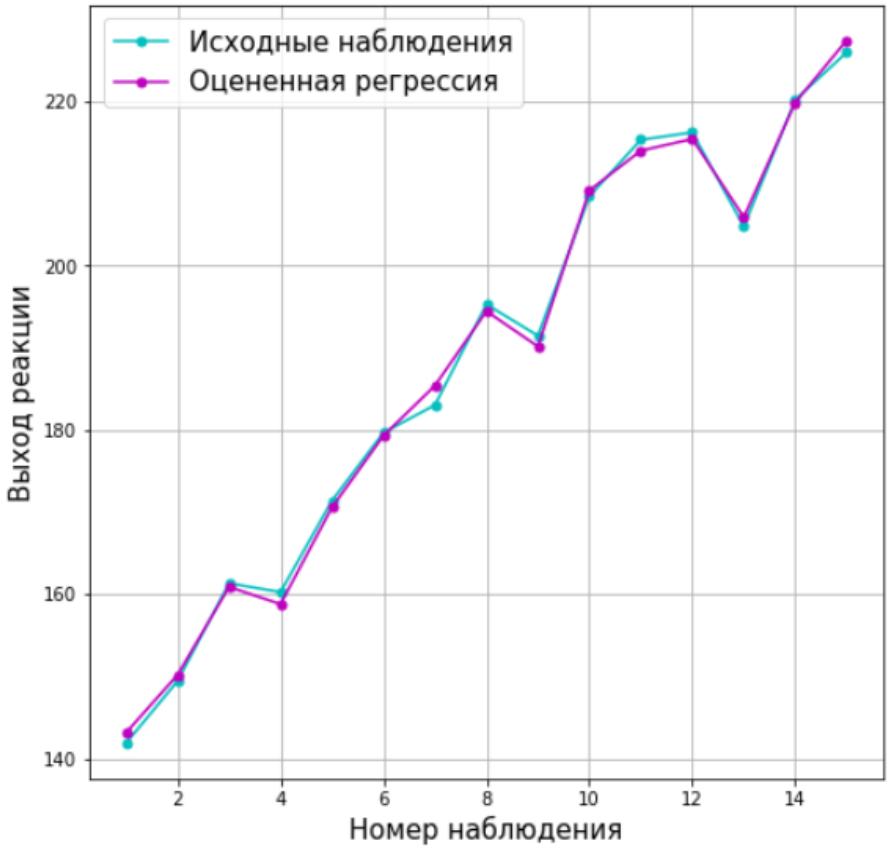



*Рисунок 2. Регрессия и оцененная регрессия*

По графику видно, что кривые очень близки.

### 9. Нахождение коэффициента детерминации

Коэффициент детерминации:

$$R^2 = 1 - \frac{\left[\sum_t \frac{e_t^2}{n}\right]}{\left[\sum_t \frac{(y_t - \bar{y})^2}{n}\right]} = 1 - \frac{[\sum_t e_t^2]}{[\sum_t (y_t - \bar{y})^2]},$$

где $e_t$ – компоненты вектора остатков, $\bar{y}$ – среднее значение $\boldsymbol{y}$.

Формула для несмещенного коэффициента детерминации:

$$R_H^2 = 1 - \frac{[\sum_t \frac{e_t^2}{n-m}]}{[\sum_t \frac{(y_t - \bar{y})^2}{n-1}]}$$

```
y_mean = np.mean(y)
R = 1 - np.sum(e**2)/np.sum((y - y_mean)**2)
R_n = 1 - np.sum(e**2)/(n-m)/(np.sum((y - y_mean)**2)/(n-1))
print("Коэффициент детерминации")
print("\tСмещенный:", R)
print("\tНесмещенный:", R_n)
```

```
Коэффициент детерминации
        Смещенный: 0.9980626571363289
        Несмещенный: 0.9975342909007823
```

Коэффициент детерминации показывает, насколько регрессионная модель лучше модели среднего. Видим, что коэффициент почти равен единице, значит наша модель гораздо лучше модели среднего.

## 10. Построение индивидуальных доверительных интервалов для коэффициентов регрессии

Построим индивидуальные доверительные интервалы с доверительной вероятностью 0.9 (здесь и далее уровень значимости $\alpha = 0.1$) по формуле:

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

где $t_{1-\frac{\alpha}{2},\ S(n-m)}$ – квантиль уровня $1 - \frac{\alpha}{2}$ распределения Стьюдента с $n - m$ степенями свободы.

```
alpha = 0.1

t_quantile = t.ppf(1 - alpha/2, df=n - m)
print(f"Значение квантиля уровня {1-alpha/2} распределения Стьюдента: {t_quantile}")
a_intervals_l = np.array([a_est - s_i*t_quantile])[0]
a_intervals_r = np.array([a_est + s_i*t_quantile])[0]
print("\nИндивидуальные доверительные интервалы")
for i in range(m):
    print(f"{i+1}-й коэффициент: [{a_intervals_l[i], a_intervals_r[i]}]")
```

Получим следующие интервалы:

```
Значение квантиля уровня 0.95 распределения Стьюдента: 1.7958848187036691

Индивидуальные доверительные интервалы
1-й коэффициент: [0.37795074968623615, 0.4533067881813631]
2-й коэффициент: [-0.3449171224005289, 0.36844410386679755]
3-й коэффициент: [2.7880687295137587, 5.097370523008928]
4-й коэффициент: [-19.646454954802216, 27.853505446655795]
```

Ранее были получены параметры модели:

```
[0.41562877 0.01176349 3.94271963 4.10352525]
```

Видим, что все коэффициенты попали в свои интервалы.

## 11. Построение обобщенной доверительной области для параметров модели

Для построения обобщенной доверительной области используем принцип Тьюки. По теореме Тьюки: доверительные интервалы $D_i$ для параметров $\alpha_i$, $i = 1, \dots, m$, с доверительной вероятностью $\left(1 - \frac{\alpha}{m}\right)$ будут совместными с вероятностью не менее $(1 - \alpha)$.

Доверительная область имеет вид:

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

```
t_quantile_turky = t.ppf(1 - alpha/(2*m), df=n - m)
print(f"Значение квантиля уровня {1-alpha/(m)} распределения Стьюдента: {t_quantile_turky}")
a_intervals_turky_l = np.array([a_est - s_i*t_quantile_turky])[0]
a_intervals_turky_r = np.array([a_est + s_i*t_quantile_turky])[0]
print("\nОбобщенная доверительная область")
for i in range(m):
    print(f"{i+1}-й коэффициент: [{a_intervals_turky_l[i]}, {a_intervals_turky_r[i]}]")
```

```
Значение квантиля уровня 0.975 распределения Стьюдента: 2.5930926824101492

Обобщенная доверительная область
1-й коэффициент: [0.36122517078484334, 0.4700323670827559]
2-й коэффициент: [-0.5032505448229411, 0.5267775262892097]
3-й коэффициент: [2.2755098141890224, 5.609929438333665]
4-й коэффициент: [-30.189264149303533, 38.39631464115711]
```

Заметим, что границы интервалов расширились. Следовательно, значения параметров модели также лежат внутри полученных интервалов.

Изобразим полученную обобщенную доверительную область.

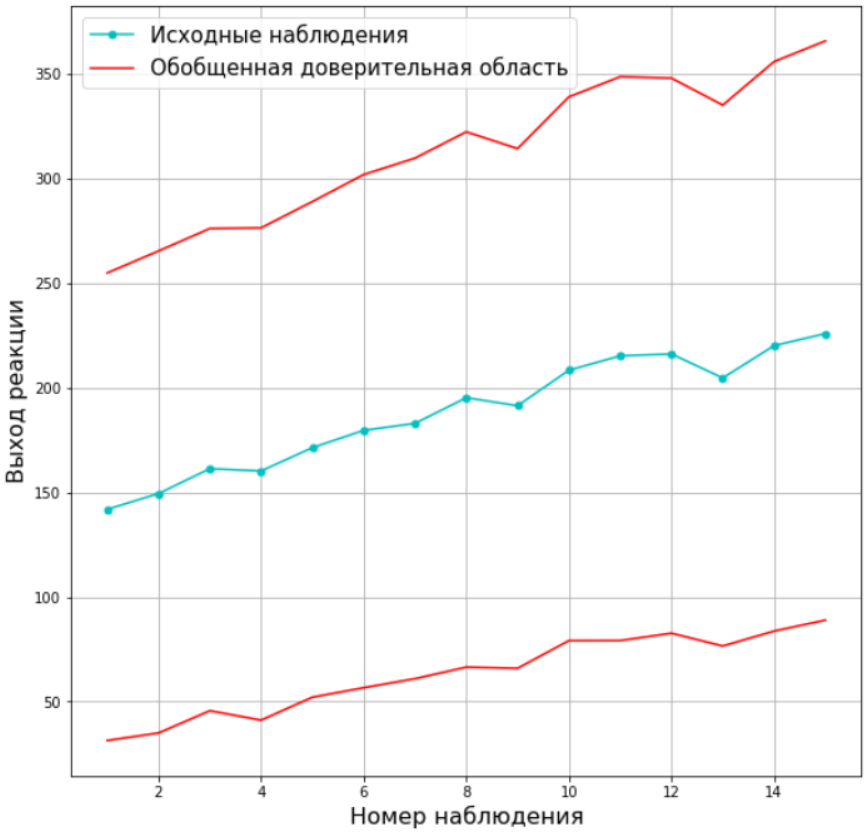



*Рисунок 3. Обобщенная доверительная область параметров модели*

Видно, что полученная область слишком широка для какого-либо адекватного оценивания параметров регрессионной модели.

## 12. Проверка основных линейных гипотез

### Гипотеза о равенстве нулю коэффициентов

Проверим гипотезу $H_0$: $\alpha_i = \beta$, для $\beta = 0$.

Область принятия гипотезы имеет вид:

$$\tau_{0\alpha} = \left\{ t: t = \frac{|\alpha_i|}{s_i} \leq t_{1-\frac{\alpha}{2},S(n-m)} \right\},$$

$t_{1-\frac{\alpha}{2},S(n-m)}$ – по-прежнему квантиль уровня $1-\frac{\alpha}{2}$ распределения Стьюдента с $n-m$ степенями свободы.

```
print(f"Значение квантиля уровня {1-alpha/2} распределения Стьюдента: {t_quantile}")

t_zero = np.abs(a_est)/s_i <= t_quantile
print("\nПроверка принятия гипотезы")
for i in range(m):
    print(f"\t{np.abs(a_est[i])/s_i[i]} <= {t_quantile}? - {t_zero[i]}")
```

```
Значение квантиля уровня 0.95 распределения Стьюдента: 1.7958848187036691

Проверка принятия гипотезы
        19.8105264355948 <= 1.7958848187036691? - False
        0.05922910762374534 <= 1.7958848187036691? - True
        6.132304007343302 <= 1.7958848187036691? - False
        0.3102932562487339 <= 1.7958848187036691? - True
```

Видим, что для второго и четвертого коэффициента гипотеза $H_0$ принимается. Следовательно, можно попробовать исключить эти коэффициенты из модели. Вернемся к этому исследованию позже.

### Гипотеза об адекватности модели среднего

Рассматриваем гипотезу

$$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_{m-1} = 0, \qquad y_t = (\beta_m + \xi_t), t = 1, \dots, n$$

$$=> \ H_0: \widehat{Q_R} = \sum_t (y_t - \bar{y})^2$$

Область принятия гипотезы будет иметь вид:

$$\tau_{0\alpha} = \left\{ t: t = \frac{R^2}{1 - R^2} \cdot \frac{n - m}{m - 1} \le t_{1-\alpha, F(m-1, n-m)} \right\},$$

где $t_{1-\alpha,F(m-1,n-m)}$ – квантиль уровня $1 - \alpha$ распределения Фишера с параметрами $m - 1, \ n - m$.

```python
f_quantile = f.ppf(1 - alpha, dfn=m-1, dfd=n-m)
print(f"Значение квантиля уровня {1-alpha} распределения Фишера: {f_quantile}")
t_mean = R**2/(1 - R**2)*(n - m)/(m - 1)
print("\nПроверка принятия гипотезы")
print(f"\t{t_mean} <= {f_quantile}? - {t_mean <= f_quantile}")
```

```
Значение квантиля уровня 0.9 распределения Фишера: 2.6602286837653133

Проверка принятия гипотезы
        943.5641964394616 <= 2.6602286837653133? - False
```

Конечно, гипотеза отвергается.

### Гипотеза об идентичности двух регрессий

Разделим данные на две подвыборки и построим две регрессионные модели:

$$\boldsymbol{y}_1 = X_1 \boldsymbol{\alpha}_1 + \boldsymbol{\varepsilon}_1, n_1;$$

$$\boldsymbol{y}_2 = X_2\boldsymbol{\alpha}_2 + \boldsymbol{\varepsilon}_2, n_2;$$

$\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in R^m, \varepsilon_1, \varepsilon_2$ – независимы.

Проверим гипотезу $H_0: \boldsymbol{\alpha_1} = \boldsymbol{\alpha_2}$

Область принятия гипотезы:

$$\tau_{0\alpha} = \left\{ t: t = \frac{(\hat{Q}_R - \hat{Q}_1 - \hat{Q}_2)/m}{(\hat{Q}_1 + \hat{Q}_2)/(n_1 + n_2 - 2m)} \leq t_{1-\alpha,\ F(m,\, n_1+n_2-2m)} \right\},$$

где $t_{1-\alpha,F(m,n_1+n_2-2m)}$ – квантиль уровня $1-\alpha$ распределения Фишера с параметрами $m,\ n_1 + n_2 - 2m$.

```
x1 = np.copy(X[:7, :])
y1 = np.copy(y[:7])
n1 = x1.shape[0]

x2 = np.copy(X[7:, :])
y2 = np.copy(y[7:])
n2 = x2.shape[0]

a1 =  np.matmul(np.linalg.inv(np.matmul(x1.T,x1)),x1.T).dot(y1)
y1_est = x1.dot(a1)
a2 = np.matmul(np.linalg.inv(np.matmul(x2.T,x2)),x2.T).dot(y2)
y2_est = x2.dot(a2)

Q_r = np.matmul((y - y_est).T, (y - y_est))
Q_1 = np.matmul((y1 - y1_est).T, (y1 - y1_est))
Q_2 = np.matmul((y2 - y2_est).T, (y2 - y2_est))

t_ident = (Q_r - Q_1 - Q_2)/m/((Q_1 + Q_2)/(n1 + n2 - 2*m))
f_quantile_ident = f.ppf(1 - alpha, dfn=m, dfd=n1 + n2 - 2*m)
print(f"Значение квантиля уровня {1-alpha} распределения Фишера: {f_quantile_ident}")
print(f"{t_ident} <= {f_quantile_ident}? {t_ident <= f_quantile_ident}")
```

```
Значение квантиля уровня 0.9 распределения Фишера: 2.9605340887350957
1.1896443223100994 <= 2.9605340887350957? True
```

Гипотеза об идентичности двух регрессий принимается.

## 13. Уменьшение числа параметров модели

### Оценки новых моделей

В соответствии с гипотезой о равенстве нулю коэффициентов, будем приравнивать к нулю 2-й и 4-й коэффициенты. Посмотрим, получится ли таким образом уточнить модель.

Рассмотрим три модели: уберем оба коэффициента, уберем только 2-й, уберем только 4-й. Для каждой модели находим все оценки и характеристики, вычисляемые ранее.

```
Оценка параметров регрессии(без 2го и 4го коэффициентов): [0.420965   4.26307861]
Оценка параметров регрессии(без 2го коэффициента): [0.41638377 3.96344916 4.87260461]
Оценка параметров регрессии(без 4го коэффициента): [0.41188386 0.07227313 3.8451599 ]

Оценка дисперсии (полная модель): 1.8020646675870131
Оценка дисперсии(без 2го и 4го коэффициентов): 2.04885876014246
Оценка дисперсии(без 2го коэффициента): 1.6524194286192488
Оценка дисперсии(без 4го коэффициента): 1.666351463529992
```

Видим, что оценка дисперсии отклонений моделей, из которых было исключено по одному параметру, уменьшилась. Для модели, в которой осталось всего два параметра, значение дисперсии, наоборот, выросло.

```
Оценка матрицы ковариаций(без 2го и 4го коэффициентов):
 [[ 0.00030939 -0.00967493]
 [-0.00967493  0.3036936 ]]
Оценка матрицы ковариаций(без 2го коэффициента):
 [[ 2.54622568e-04 -7.46964224e-03 -5.41955395e-03]
 [-7.46964224e-03  2.66727698e-01 -3.54458723e-01]
 [-5.41955395e-03 -3.54458723e-01  5.76424391e+00]]
Оценка матрицы ковариаций(без 4го коэффициента):
 [[ 2.72329875e-04 -1.64734557e-04 -6.91611758e-03]
 [-1.64734557e-04  1.31105578e-03 -7.58116803e-03]
 [-6.91611758e-03 -7.58116803e-03  2.90834208e-01]]

Стандартная ошибка оценки коэффициентов(без 2го и 4го коэффициентов): [0.01758955 0.55108402]
Стандартная ошибка оценки коэффициентов(без 2го коэффициента): [0.0159569  0.51645687 2.40088398]
Стандартная ошибка оценки коэффициентов(без 4го коэффициента): [0.01650242 0.0362085  0.53929047]
```

Можно заметить, что ошибка оценки для четвертого коэффициента для модели, из которой исключен второй параметр, сильно уменьшилась.

```
Матрица корреляций(без 2го и 4го коэффициентов):
 [[ 1.         -0.99810336]
 [-0.99810336  1.        ]]

Матрица корреляций(без 2го коэффициента):
 [[ 1.         -0.90639457 -0.14146334]
 [-0.90639457  1.         -0.28586464]
 [-0.14146334 -0.28586464  1.        ]]

Матрица корреляций(без 4го коэффициента):
 [[ 1.         -0.27569347 -0.77712697]
 [-0.27569347  1.         -0.38824225]
 [-0.77712697 -0.38824225  1.        ]]
```

Построение графиков оцененной регрессии

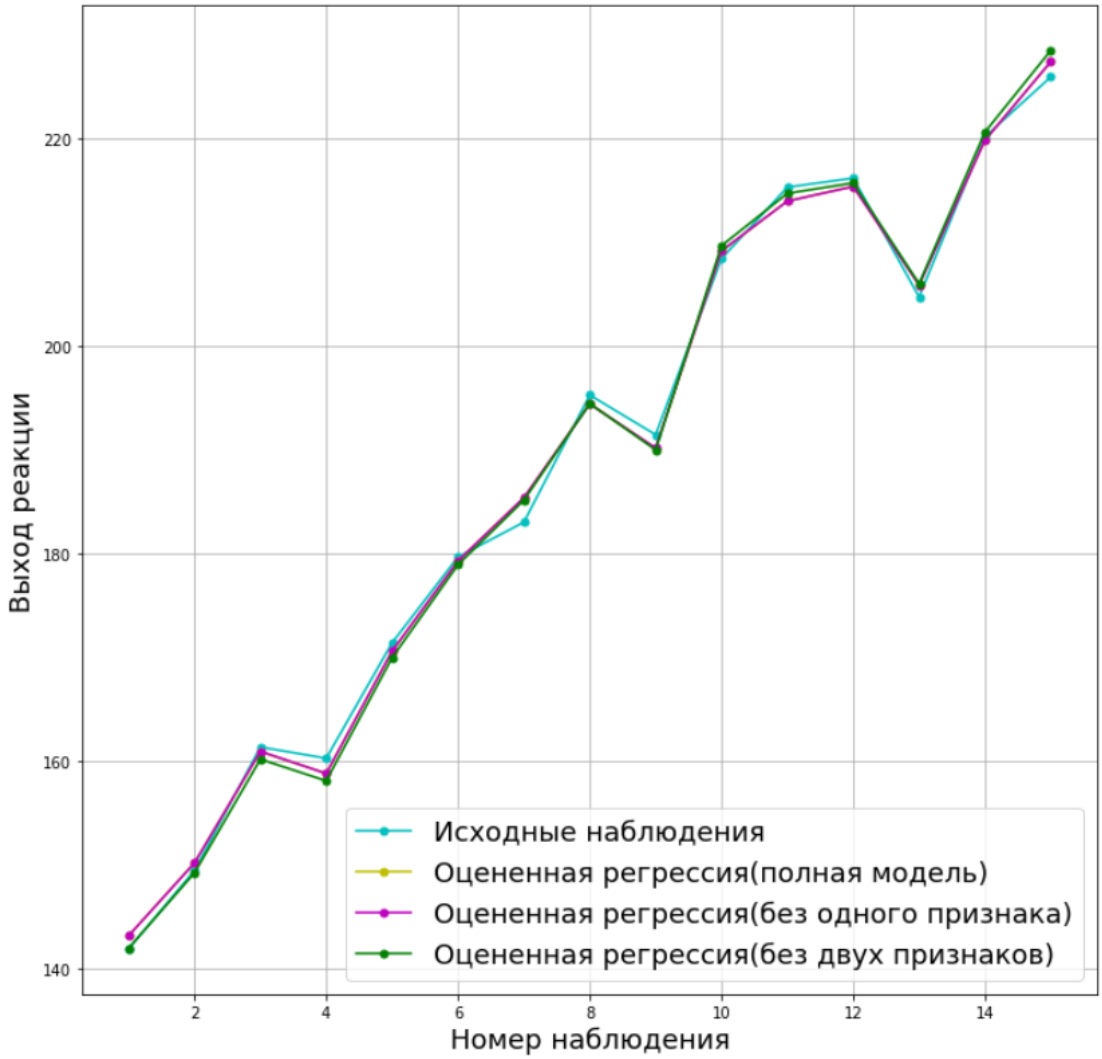



*Рисунок 4. Оцененная регрессия, модели с меньшим числом параметров*

Видим, что кривые, соответствующие полной модели и модели с одним исключенным параметром (вторым), полностью наложились. Можно также сказать, что кривая, соответствующая модели, из которой исключены оба признака, находится ближе к графику исходных наблюдений и на некоторых участках практически совпадает с ним.

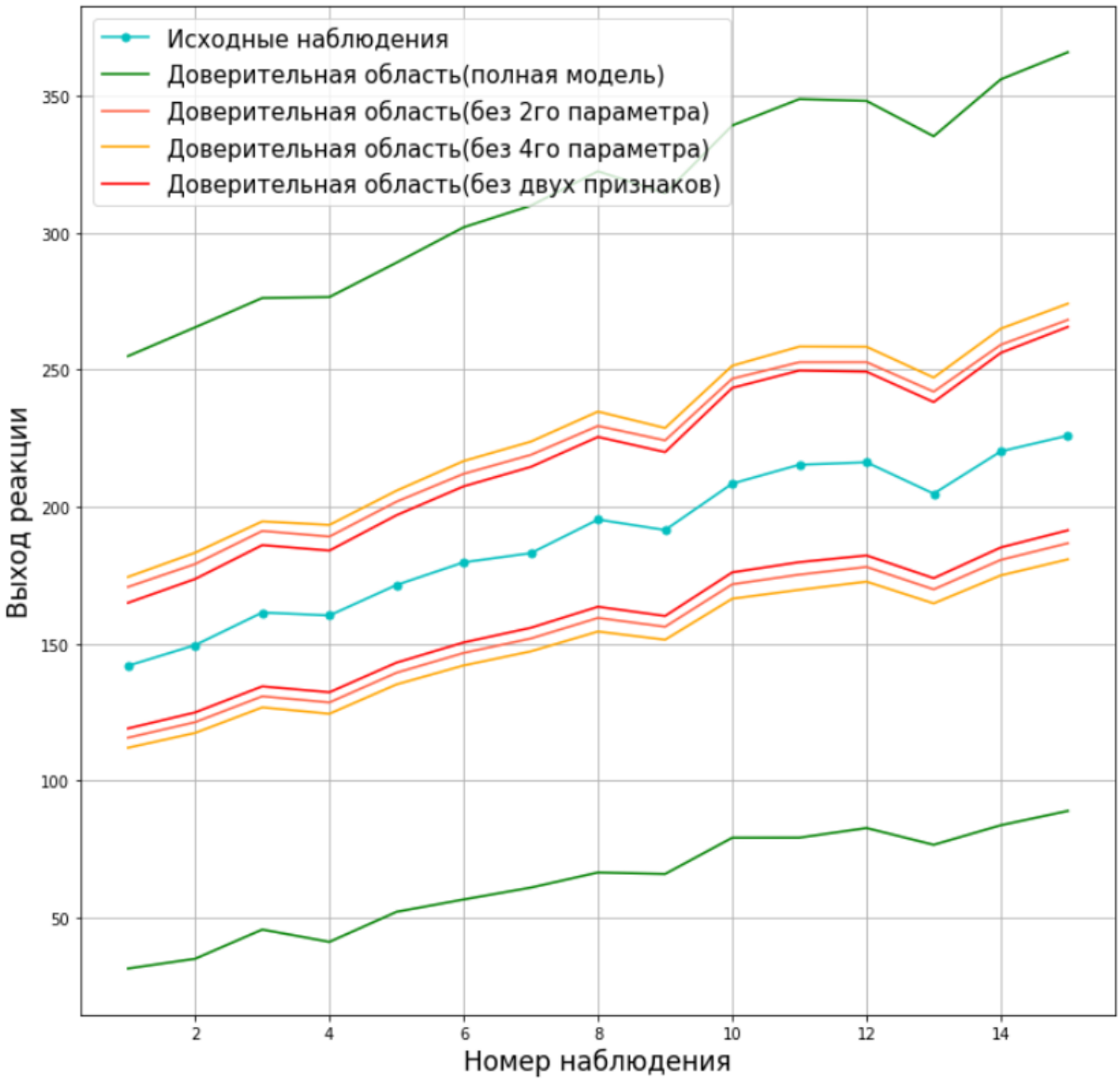
Построение обобщенной доверительной области

Найдем доверительные области параметров моделей. Число степеней свободы распределения Стьюдента изменяется, в соответствии с числом исключенных параметров.

```
Обобщенная доверительная область(без 2го и 4го коэффициента)
Значение квантиля уровня 0.9875 распределения Стьюдента: 2.5326378146335955
1-й коэффициент: [0.37641703976799373, 0.46551295619818506]
2-й коэффициент: [2.867382377563025, 5.658774840698264]
```

```
Обобщенная доверительная область(без 2го коэффициента)
Значение квантиля уровня 0.9875 распределения Стьюдента: 2.5600329593015543
1-й коэффициент: [0.37553358242554796, 0.45723394782652343]
2-й коэффициент: [2.641302554689167, 5.28559577239964]
4-й коэффициент: [-1.2737375204917463, 11.01894674096604]

Обобщенная доверительная область(без 4го коэффициента)
Значение квантиля уровня 0.9875 распределения Стьюдента: 2.5600329593015543
1-й коэффициент: [0.36963711745126393, 0.4541305971490496]
2-й коэффициент: [-0.020421834911961825, 0.16496809352532504]
3-й коэффициент: [2.4645585313715923, 5.22576127050276]
```




По графику видно, что области значительно сузились для моделей, из которых были исключены параметры. Причем ширина областей для моделей, из которых был исключен один параметр, практически не отличается от ширины области, полученной для модели без двух параметров.

## 14. Прогноз наблюдения

Находим прогноз для $\boldsymbol{y}_\tau$:

$$\widehat{y_\tau} = \boldsymbol{x}_\tau^T \boldsymbol{a}$$

Оценка дисперсии прогноза:

$$s_\tau^2 = s^2 \ (\boldsymbol{x}_\tau^T \ (X^T X)^{-1} \boldsymbol{x}_\tau \ + \ 1)$$

Интервальный прогноз:

$$D_\tau = \{\widehat{y_\tau} - t_\tau \cdot s_\tau \leq y_\tau \leq \widehat{y_\tau} + t_\tau \cdot s_\tau\}, t_\gamma = t_{\frac{1+\gamma}{2}, S(n-m)},$$

где $\gamma = 1 - \alpha$ – доверительная вероятность.

Построим прогноз для первого наблюдения:

```
Полная модель(4 параметра):
Прогноз выхода реакции: 150.51872008655297; исходное наблюдение: 149.40628; отклонение: 1.1124400865529651
Оценка дисперсии: 1.8983862803155767

Интервальная оценка прогноза:
Прогноз выхода реакции: 150.51872008655297: [148.2223632008409, 152.81507697226505]
==========================================================================================================


Модель без 2го и 4го коэффициента:
Прогноз выхода реакции: 149.15814754780502; исходное наблюдение: 149.40628; отклонение: 0.24813245219499436
Оценка дисперсии: 2.215047761935859

Интервальная оценка прогноза:
Прогноз выхода реакции: 149.15814754780502: [147.01454176119185, 151.30175333441818]
==========================================================================================================


Модель без 2го коэффициента:
Прогноз выхода реакции: 150.37810399493017; исходное наблюдение: 149.40628; отклонение: 0.9718239949301619
Оценка дисперсии: 1.7379510415548831

Интервальная оценка прогноза:
Прогноз выхода реакции: 150.37810399493017: [148.30734286447372, 152.44886512538662]
==========================================================================================================


Модель без 4го коэффициента:
Прогноз выхода реакции: 150.5727326288976; исходное наблюдение: 149.40628; отклонение: 1.1664526288975878
Оценка дисперсии: 1.7290093851184762

Интервальная оценка прогноза:
Прогноз выхода реакции: 150.5727326288976: [148.45716926201197, 152.68829599578322]
```

Видим, что по модели, из которой были исключены второй и четвертый параметры, получается самое маленькое отклонение прогноза от истинного значения.

Построим прогноз для случайно выбранного второго наблюдения.

```
Полная модель(4 параметра):
Прогноз выхода реакции: 160.55459937963485; исходное наблюдение: 161.30454999999998; отклонение: 0.7499506203651265
Оценка дисперсии: 1.948299396350666

Интервальная оценка прогноза:
Прогноз выхода реакции: 160.55459937963485: [158.090184222701, 163.0190145365687]
==========================================================================================================


Модель без 2го и 4го коэффициента:
Прогноз выхода реакции: 160.05997593479086; исходное наблюдение: 161.30454999999998; отклонение: 1.2445740652091217
Оценка дисперсии: 2.0977256964415116

Интервальная оценка прогноза:
Прогноз выхода реакции: 160.05997593479086: [158.03843261152895, 162.08151925805277]
==========================================================================================================


Модель без 2го коэффициента:
Прогноз выхода реакции: 160.82381148002807; исходное наблюдение: 161.30454999999998; отклонение: 0.48073851997190786
Оценка дисперсии: 1.7845962105202693

Интервальная оценка прогноза:
Прогноз выхода реакции: 160.82381148002807: [158.8583744439604, 162.78924851609574]
==========================================================================================================


Модель без 4го коэффициента:
Прогноз выхода реакции: 160.55257761969008; исходное наблюдение: 161.30454999999998; отклонение: 0.7519723803098941
Оценка дисперсии: 1.7711819404340015

Интервальная оценка прогноза:
Прогноз выхода реакции: 160.55257761969008: [158.64792298309624, 162.45723225628393]
```

Минимальное отклонение при этом получено по модели, из которой исключен второй коэффициент. А наибольшему отклонению соответствует результат, полученный по модели без второго и четвертого коэффициента.

## Выводы

По результатам работы можно сказать, что классическая модель линейной регрессии может быть успешно использована для восстановления зависимости целевой переменной от некоторого набора нецелевых признаков. Регрессионная модель обладает значительным преимуществом, по сравнению

с моделью среднего, что было показано применением соответствующей гипотезы и было замечено по значению коэффициента детерминации.

Большим преимуществом является возможность перейти посредством дополнительных исследований к регрессионной модели, построенной на меньшем числе параметров, а, следовательно, уменьшить вычислительную сложность. Так, применение гипотезы о равенстве нулю коэффициентов модели позволило при незначительной потере точности прогноза получить гораздо более точную интервальную оценку параметров модели.