Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Кафедра Систем управления и информатики

Дисциплина: Основы теории идентификации (Б.3.2.в.2-СУиИ)

Лабораторная работа 1:

Метод наименьших квадратов.

Вариант: 2.

Студенты: Дема Н.Ю.
Тюрин А.И.
Гирута М.Н.
Группа: Р3435
Преподаватель: Ведяков А.А.

«Санкт-Петербург 2016»

Задание 1:

В первом задании дается матрица наблюдений и вектор значений целевой функции (два набора таких значений). Так же предполагается заранее известной модель линейной регрессии:

$$y(t) = x_1(t)\theta_1 + x_2(t)\theta_2 + x_3(t)\theta_3 + \epsilon(t),$$

где $\theta_1, \theta_2, \theta_3$ — неизвестные параметры, а ϵ — шум измерений. Требуется по известным данным оценить значения неизвестных параметров и сделать выводы о достоверности полученных результатов.

Матричное представление модели имеет вид:

$$Y = X\Theta + \varepsilon$$

Вектор модельных оценок обозначим как:

$$\hat{Y} = X\hat{\Theta}$$

По методу наименьших квадратов найдем такой вектор параметров $\hat{\Theta}$, для которого критерий оценивания $J(\hat{\Theta})$, представляющий собой сумму квадратов ошибок оценивания, будет минимальным:

$$J(\hat{\Theta}) = \sum_{k=1}^{N} \big(y(k) - \hat{y}(k)\big)^2 = E^T E$$
$$\hat{\Theta}_{LSQ} = \underset{\hat{\Theta}}{\operatorname{argmin}}\, J(\hat{\Theta})$$

Для нахождения искомого вектора сначала преобразуем квадрат ошибки к следующему виду:

$$E^T E = (Y - X\hat{\Theta})^T (Y - X\hat{\Theta}) = Y^T Y - (X\hat{\Theta})^T Y - YX\hat{\Theta} + (X\hat{\Theta})^T X\hat{\Theta}$$
$$= Y^T Y - 2\hat{\Theta}^T X^T Y + \hat{\Theta}^T X^T X\hat{\Theta}$$

Чтобы решить задачу требуется найти минимум данного выражения. Дифференцируя полученную функцию по $\hat{\Theta}$:

$$\frac{\partial(Y^T Y - 2\hat{\Theta}^T X^T Y + \hat{\Theta}^T X^T X\hat{\Theta})}{\partial\hat{\Theta}} = 0$$
$$\frac{\partial(Y^T Y)}{\partial\hat{\Theta}} - \frac{\partial(2\hat{\Theta}^T X^T Y)}{\partial\hat{\Theta}} + \frac{\partial(\hat{\Theta}^T X^T X\hat{\Theta})}{\partial\hat{\Theta}} = 0$$
$$0 - 2X^T Y + 2X^T X\hat{\Theta} = 0$$
$$X^T X\hat{\Theta} = X^T Y$$
$$\hat{\Theta} = (X^T X)^{-1} X^T Y$$

В результате получаем для 1-го и 2-го наборов данных:

$$\hat{\Theta}_1 = \begin{bmatrix} -2.9973 \\ -7.9929 \\ 2.0066 \end{bmatrix}, \quad \hat{\Theta}_2 = \begin{bmatrix} -2.9919 \\ -7.9922 \\ 2.0129 \end{bmatrix}$$

Код программы, реализующий нахождение параметров представлен на листинге 1. Графики значений целевой функции и функции, полученной при использовании найденных параметров, для первого и второго наборов данных представлены соответственно на рисунках 1 и 2. График ошибки оценивания для каждого набора данных представлен на рисунках 3 и 4.

Листинг 1: Код программы нахождения параметров линейной модели.

```
1:  load('ident_lab1_v02.mat')
2:  X = [zad11.x1 zad11.x2 zad11.x3]
3:  if det(X'*X) = 0
4:      fpritnf('The estimation for X couldn't be calculated')
5:  else
6:      Y = zad11.y
7:      Theta = inv(X'*X)*X'*Y
8:      Yhat = X*Theta
9:      E=Y-Yhat
10:     E=abs(E)
11:     plot(E)
12: end
```

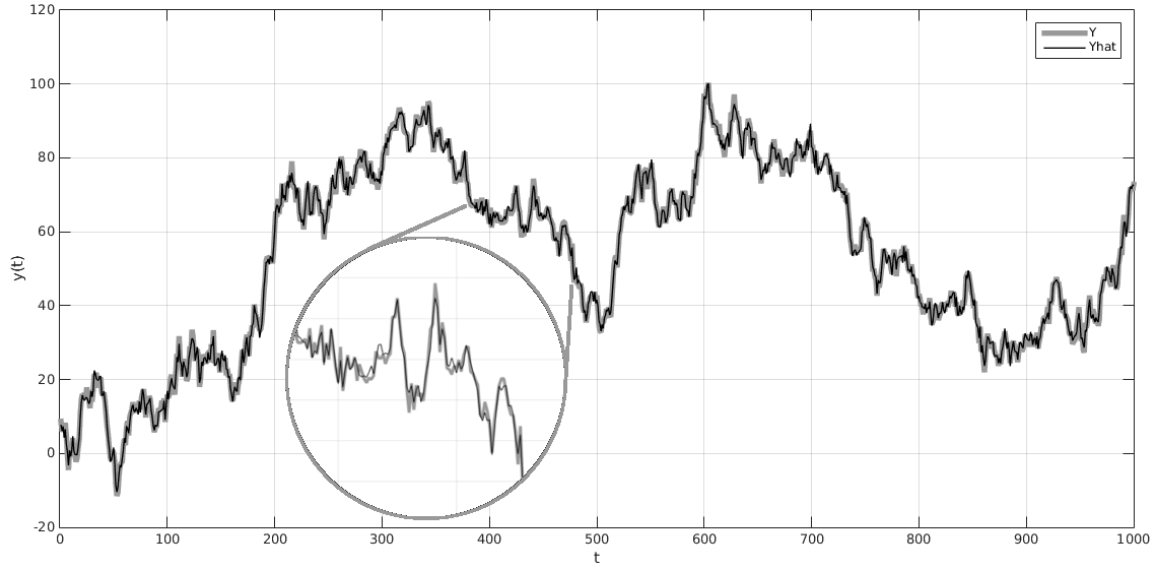

Рисунок 1: График целевой функции и модельных оценок первого набора данных.

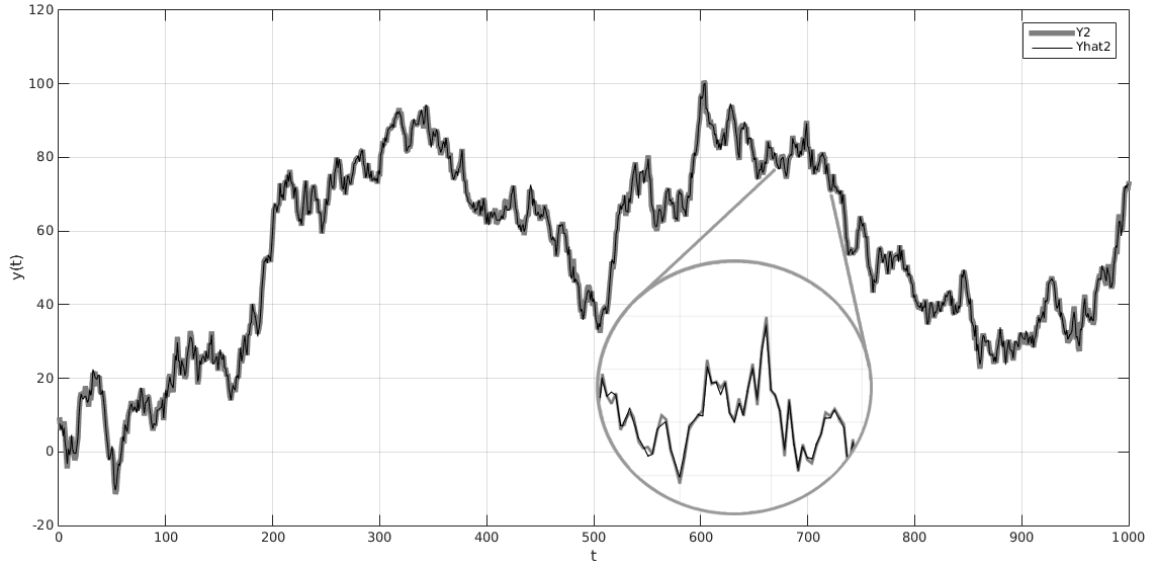

Рисунок 2: График целевой функции и модельных оценок второго набора данных.

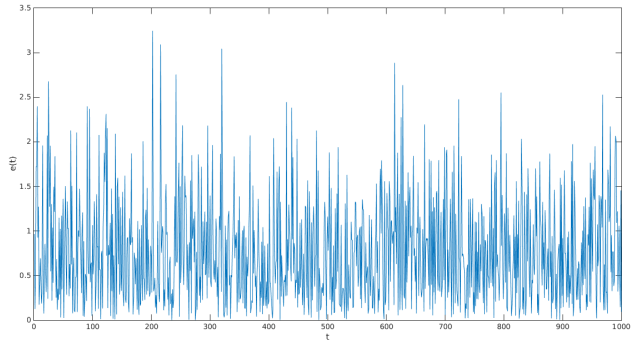

Рис. 3: График ошибки оценивания $\hat{y}(t)$ по первому набору эксперементальных данных

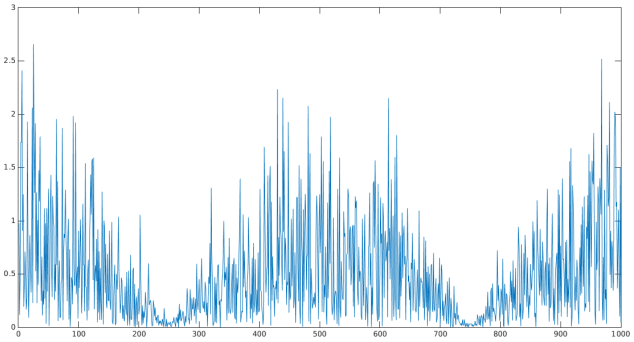

Рис. 4: График ошибки оценивания $\hat{y}(t)$ по второму набору эксперементальных данных

Задание 2:

В данном задании дается некоторая статическая зависимость функции V от T. Выдвигаются две гипотезы относительно способа аппроксимации:

- Линейная зависимость: $V = bT + c$.
- Квадратичная зависимость: $V = aT^2 + bT + c$.

Требуется для каждого набора данных записать гипотезы в форме линейной регрессии, найти параметры статической зависимости для каждой из них и сделать выводы о достоверности каждой из них.

Первую гипотезу можно представить как одномерную модель линейной регрессии:

$$y(t) = \theta_0 + x(t)\theta_1 + \epsilon(t),$$

где θ_0, θ_1 — неизвестные параметры, а ϵ — шум измерений. Вторую гипотезу, как регрессионную полиномиальную модель, где степень полинома равна двум:

$$y(t) = x^2(t)\theta_2 + x(t)\theta_1 + \theta_0 + \epsilon(t)$$

Принимая, как и в прошлом задании, что математическое ожидание случайных ошибок равно нулю, а сами они независимые случайные величины с постоянной дисперсией для всех наблюдений, то искомый вектор параметров $\hat{\Theta}$, как для первой, так и для второй модели можно найти как:

$$\hat{\Theta} = (X^T X)^{-1} X^T Y$$

Код программы представлен в листинге 2. На рисунках 5 и 6 представлены экспериментальные данные и оценки, полученные по каждой модели для первого и второго набора данных соответственно. Ошибки оценок для первого и второго набора данных для каждой модели представлены на рисунках 7 и 8.

Значения параметров для одномерной и полиномиальной модели для первого и второго набора данных соответственно:

$$1 : \hat{\Theta}_1 = \begin{bmatrix} -2.9973 \\ 2.0066 \end{bmatrix}, \quad \hat{\Theta}_2 = \begin{bmatrix} -269.3676 \\ 23.6268 \\ -0.1336 \end{bmatrix}$$

$$2 : \hat{\Theta}_1 = \begin{bmatrix} 38.5527 \\ 10.3959 \end{bmatrix}, \quad \hat{\Theta}_2 = \begin{bmatrix} 26.6665 \\ 9.8848 \\ 0.0115 \end{bmatrix}$$

Листинг 2: Программа нахождения параметров для одномерной и полиномиальной моделей.

```
1:  O = ones(14, 1)
2:  X1 = [O zad21.T ]
3:  if det(X1'*X1) = 0
4:      fpritnf('The estimation for X1 couldn't be calculated')
5:  else
6:      Y1 = zad21.V
7:      Theta1 = inv(X1'*X1)*X1'*Y1
8:  end
9:  Yhat1 = X1*Theta1
10: X2 = [O zad21.T zad21.T.^2]
11: if det(X2'*X2) = 0
12:     fpritnf('The estimation for X2 couldn't be calculated')
13: else
14:     Theta2 = inv(X2'*X2)*X2'*Y1
15:     Yhat2 = X2*Theta2
16: end
17: plot(zad21.T, zad21.V, zad21.T,  Yhat1, zad21.T, Yhat2)
```

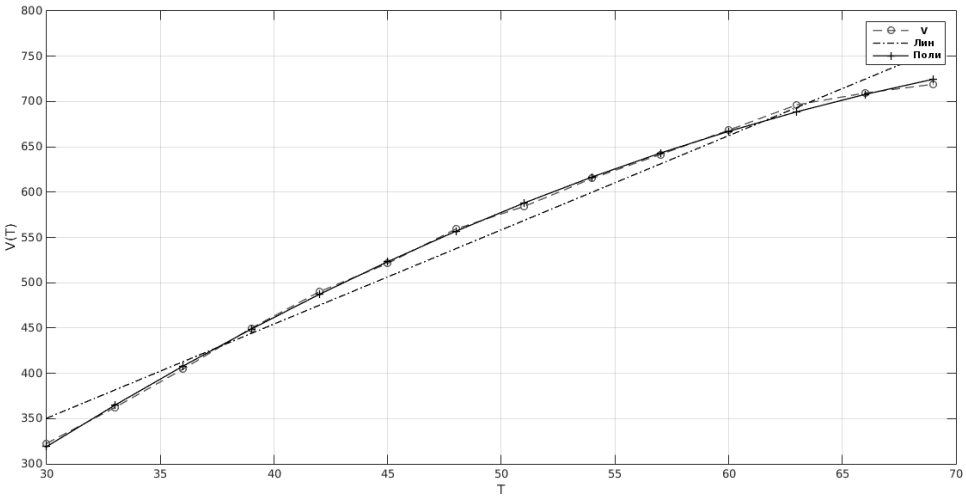

Рисунок 5: График модельных оценок первого набора данных.

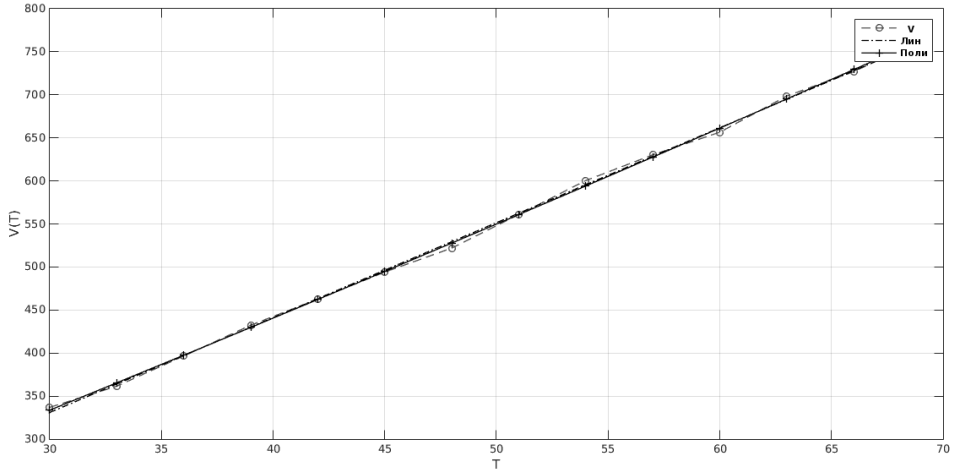

Рисунок 6: График модельных оценок второго набора данных.

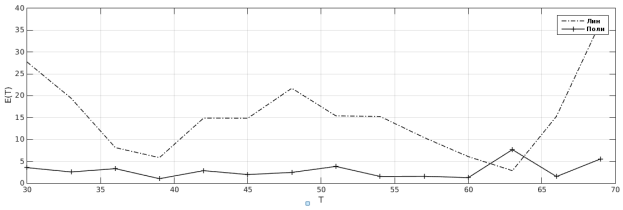

Рис. 7: График ошибки оценивания $\hat{y}(t)$ по первому набору эксперементальных данных

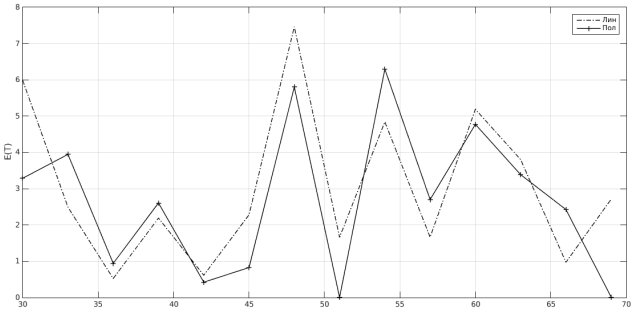

Рис. 8: График ошибки оценивания $\hat{y}(t)$ по второму набору эксперементальных данных

Из полученных графиков видно, что полиномиальная модель гораздо лучше подходит под заданную экспериментальную зависимость для первого набора данных. Для второго набора данных о том, какая и моделей лучше подходит ничего сказать нельзя.

Задание 3:

В данном задании дается символьная запись двух функций и соответствующие наборы данных. Требуется записать функции в форме линейной регрессии и оценить значения неизвестных параметров для каждой из них.

Первая функция имеет вид:

$$y(x) = p_1 \sin(11x + p_2)$$

Преобразуем ее к форме линейной регрессии:

$$y(x) = p_1 \sin(11x + p_2)$$
$$y(x) = p_1 \cos(p_2) \sin(11x) + p_1 \sin(p_2) \cos(11x)$$
$$y(x) = \theta_1 x_1 + \theta_2 x_2,$$
$$\text{где } \theta_1 = p_1 \cos(p_2);\ \theta_2 = p_1 \sin(p_2);\ x_1 = \sin(11x);\ x_2 = \cos(11x).$$

Найденные значения параметров:

$$\hat{\Theta} = \begin{bmatrix} 14.4663 \\ 5.2806 \end{bmatrix}$$

Найдем параметры p_1 и p_2:

$$p_2 = atan(\frac{x_2}{x_1}) = 70^o$$
$$p_1 = \frac{x_2}{\sin(p_2)} = 5.6186$$

На рисунке 9 представлены графики зависимости отчетов набора данных и $y = (x,\ p_1,\ p_2)$ для полученных оценок параметров. В листинге 3 код программы.

Листинг 3: Код программы нахождения параметров.

```
1:  x = zad31.x'
2:  X1 = sin( 11*x )
3:  X2 = cos( 11*x )
4:  X = [ X1 X2 ]
5:  if det(X'*X) = 0
6:      fpritnf('The estimation for X couldn't be calculated')
7:  else
8:      Y1 = zad31.y'
9:      Theta = inv(X'*X)*X'*Y
10:     Yhat = X*Theta
11:     plot(zad31.x, zad31.y, zad31.x,  Yhat)
12: end
```

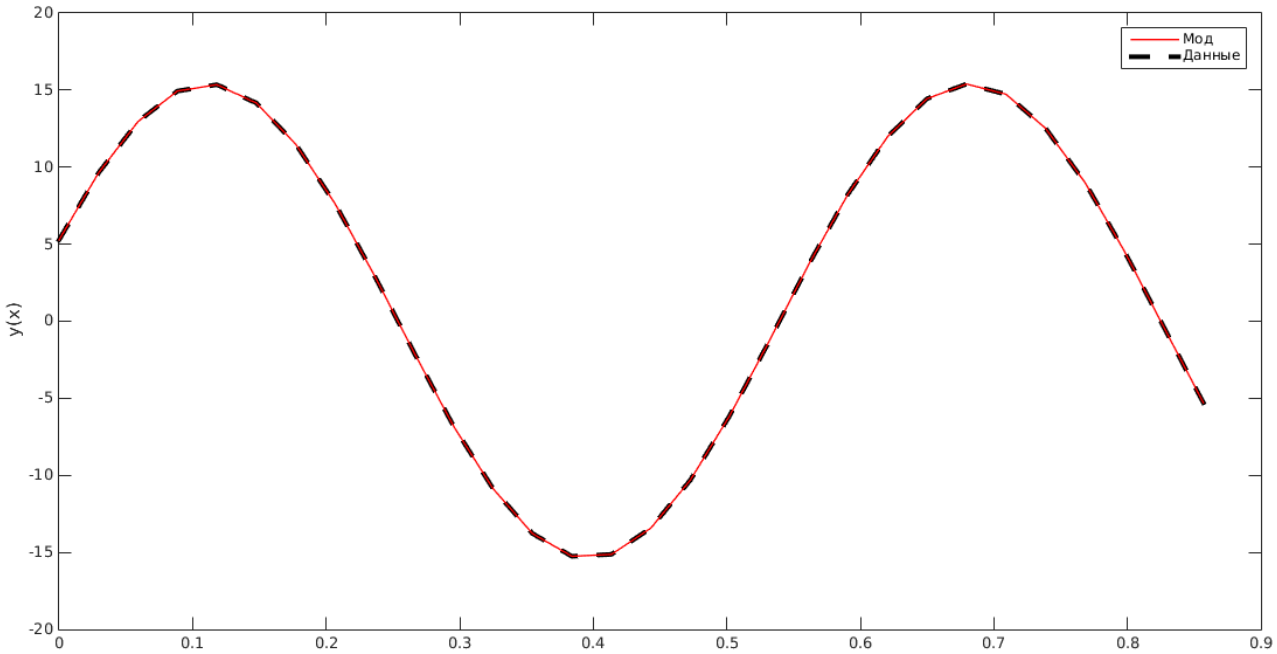

Рисунок 9: График экспериментальных данных и график $y(p_1, p_2, x)$ для полученных оценок.

Вторая функция имеет вид:

$$y(x) = \frac{x + p_1}{x + p_2}$$

Преобразуем ее к форме линейной регрессии:

$$\frac{1}{y-1} = \frac{x + p_2}{p_2 + p_1}$$

$$y_l = ax + b,$$

$$\text{где } a = \frac{1}{p_2 + p_1};\ b = \frac{p_2}{p_2 + p_1};\ y_l = \frac{1}{y-1}.$$

Найденные значения параметров модели:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} -1.5526 \\ -0.2632 \end{bmatrix}$$

Найдем параметры p_1 и p_2:

$$p_2 = \frac{b}{a} = 0.1695$$
$$p_1 = \frac{1}{a} - p_2 = -0.8136$$

На рисунке 10 представлены графики зависимости отчетов набора данных и $y = (x,\ p_1,\ p_2)$ для полученных оценок параметров. В листинге 4 код программы.
Листинг 4: Код программы нахождения параметров.

```
1:  O = ones(30, 1)
2:  X = zad32.x'
3:  X = [O X]
4:  if det(X'*X) = 0
5:      fpritnf('The estimation for X couldn't be calculated')
6:  else
7:      Y = zad32.y'
8:      Y_ = 1./(Y-1)
9:      Theta = inv(X'*X)*X'*Y_
10:     Y_hat = X*Theta
11:     X = X(:, 2)
12:     YHAT = 1./Y_hat + 1
13:     plot(zad32.x, Y, zad32.x,  YHAT)
14: end
```

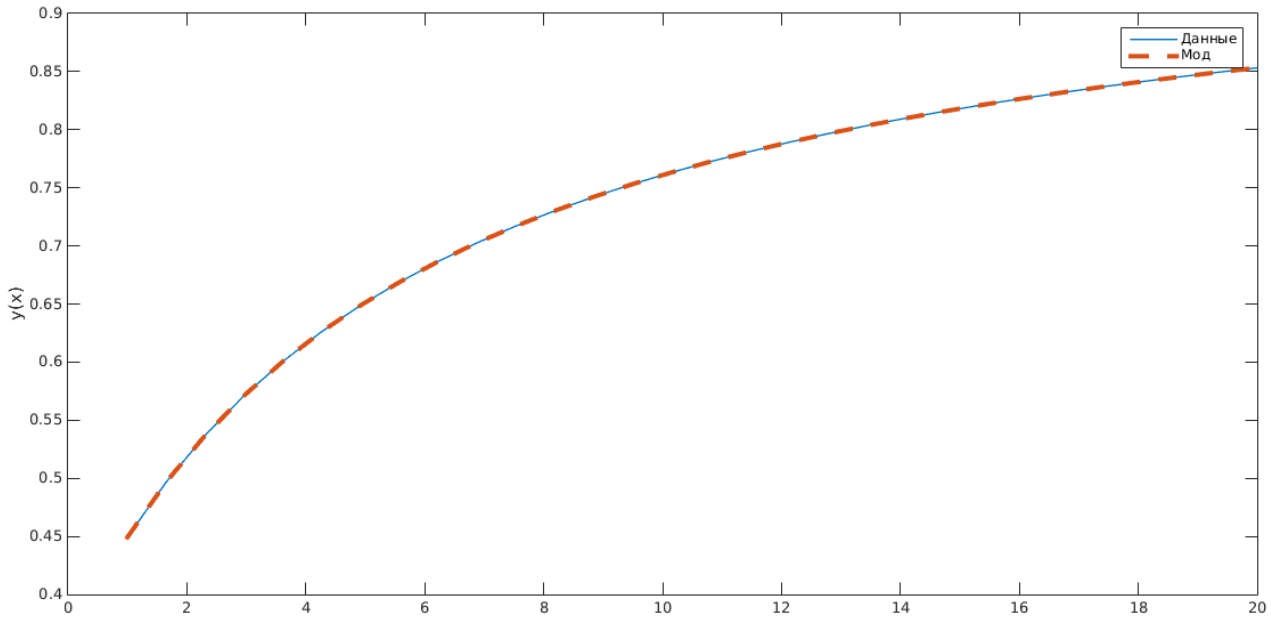

Рисунок 10: График экспериментальных данных и график $y(p_1, p_2, x)$ для полученных оценок.

Вывод:

В ходе выполнения лабораторной работы мы получили практические навыки использования метода наименьших квадратов для определения параметров различных моделей.

Проверка принятых допущений:

В регрессионном анализе предполагается, что зависимая переменная есть сумма значений некоторой регрессионной модели и случайной величины(ошибки). Относительно характера распределения этой величины делаются предположения, называемые гипотезой порождения данных. Для получения информации о соответствии данным условиям выполняются статистические тесты, называемые анализом остатков (вектор $\varepsilon = Y - \hat{Y},\ \varepsilon \in \mathbb{R}^n$), который заключается в проверке следующих гипотез:

- $E[\varepsilon_i] = 0,\ i = 1..n$
- $D[\varepsilon_i] = \sigma^2 = const,\ i = 1..n$ (гомоскедастичность)
- $\varepsilon_i \sim \mathbf{N}(0, \sigma),\ i = 1..n$
- $\forall i, j,\ i \neq j\ cov(\varepsilon_i, \varepsilon_j) = 0)$

Проверим данные гипотезы на данных из первого задания для первого набора данных. Для проверки первой гипотезы воспользуемся критерием знаков. Данный критерий можно использовать как непараметрический статистический критерий для для проверки гипотезы равенства медианы какому-либо заданному значению (в нашем случае — нулю).

Задана выборка $\varepsilon^m = \{\varepsilon_1, .., \varepsilon_1\},\ \varepsilon_i \in \mathbb{R}^n$.
Нулевая гипотеза $H_0 : P\{\varepsilon < 0\} = 1/2$.
Перейдем к бинарной выборке: $b_i = [x_i < a],\ i = 1, .., n$.
Тогда нулевая гипотеза будет иметь вид: $H_0 : P\{b = 1\} = 1/2$.
Посчитаем статистику критерия: $k = \sum_{i=0}^{n} b_i = 493$
При уровне значимости $\alpha = 0.05$ нулевая гипотеза не отвергается если: $Bin_p(n,\ k) \in [\alpha/2,\ 1-\alpha/2]$,
где $Bin_p(n,\ k) = 0.5^n \sum_{i=0}^{k} C_n^i = \dfrac{3.4035301^{300}}{9.33264^{302}} = 0.318$

Полученное значение является P-value данного критерия (левый хвост биномиального распределения с p=1/2). Гипотеза не отвергается, так как это значение больше выбранного уровня значимости.

Проверку на гомоскедастичность проведем эвристическим методом суть которого состоит в ранжировании псевдодисперсий и анализе полученной гистограммы. Под пвсевдодисперсией понимаются величины:

$$dis_i = \varepsilon_i^2 - \frac{1}{n}\sum_{i=0}^{n} \varepsilon_i$$

Построенная по ним гистограмма, состоящая из 10 интервалов представлена на рисунке 11. Проводится сравнение количества элементов на первых двух интервалах: $710/155 = 4.6$. Обычно для гетероскедастичного случая это отношение больше семи, отсюда делаем вывод, что наш ряд гомоскедастичен.

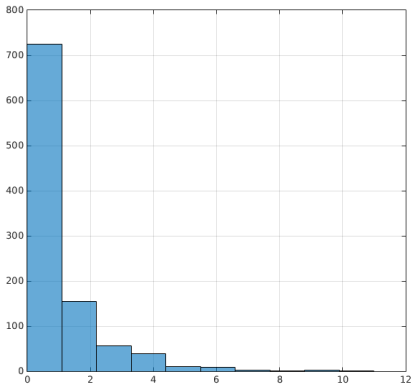

Рис. 11: Гисторгамма псеводисперсии.

Проверку нормальности распределения осуществим с помощью критерия хи-квадрат. Выдвигаемая нами гипотеза H_0: случайная величина ε подчиняется нормальному закону распределения. В данном распределении на интервал $A_1 = [-\sigma,\ \sigma]$ приходится приблизительно 68.2% всех измерений, на период $A_2 = [-2\sigma,\ -\sigma] \cup [\sigma,\ 2\sigma]$ — 27.2% и на $A_3 = [-3\sigma,\ -2\sigma] \cup [2\sigma,\ 3\sigma]$ — 4.2, следовательно, имея тысячу элементов выборке мы должны были получить следующее распределение: $|A_1| \approx 682;\ |A_2| \approx 272;\ |A_3| \approx 42$. Для исследуемой выборки $|E_1| = 678;\ |E_2| = 278;\ |E_3| = 41$. Найдем критерий согласия Пирсона:

$$\chi^2 = \sum_{j=1}^{k} \frac{(|A_j| - |E_j|)^2}{|E_j|} = 0.0236 + 0.1295 + 0.0244 = 0.1775$$

При уровне значимости $\alpha = 0.05$ и степени свободы k, равной 2 проверяемая гипотеза выполняется.

Проверку последнего условия реализуем с помощью статистики Дарбина-Уотсона.

$$d = \frac{\sum_{i=2}^{n}(\varepsilon_i - \varepsilon_{i-1})^2}{\sum_{i=1}^{n}\varepsilon_i^2} = \frac{1.7945^3}{926.3} = 1.94$$

При уровне значимости $\alpha = 0.05$ табличное значение $d_{U,0.05} = 1.673$. Так как $d > d_{U,0.05}$ и $(4 - d) > d_{U,0.05}$, то можно сделать вывод об отсутствии как положительной, так и отрицательной корреляций остатков.