Лабораторная работа 9
Линейная регрессия

Гузовская Александра Чеславовна
Б9123-01.03.02сп

9 июня 2025 г.

Уравнение регрессии

Имеется m признаков $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$
и зависящий от них целевой признак Y.
Уравнением регрессии Y на $\mathbf{X}$ называется уравнение

$$Y(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}) + \epsilon,$$

где $\epsilon \sim N(0; \sigma^2)$ – случайный остаток.
Так как вид функции $E(Y|X = x)$ неизвестен, то предполагают, что

$$E(Y|X = x) = a_1x_1 + a_2x_2 + \ldots + a_mx_m = \mathbf{a}^T\mathbf{x}.$$

Тогда получается уравнение линейной регрессии

$$Y(\mathbf{x}) = \mathbf{a}^T\mathbf{x} + \epsilon.$$

Модель определяется параметрами $\mathbf{a} = (a_1, a_2, \ldots, a_m)^T$, которые оцениваются с помощью выборки при условии $D(\epsilon) \to \min$.
Имеется выборка $(X_1^i, X_2^i, \ldots, X_m^i, Y_i)$, $i = \overline{1, n}$ из $(X_1, X_2, \ldots, X_m, Y)$. Тогда, подставляя значения в уравнение линейной регрессии получаем

$$Y_i = a^T\mathbf{X}^i + \epsilon_i.$$

Оценкой $D(\epsilon)$ является

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \to \min.$$

Оптимальные значения a находятся как

$$\hat{\mathbf{a}} = \text{argmin}_{\mathbf{a}} \sum_{i=1}^{n} (a^T \mathbf{X}^i - Y_i)^2.$$

Если обозначить $X = \begin{pmatrix} X_1^1 & X_2^1 & \cdots & X_m^1 \\ X_1^2 & X_2^2 & \cdots & X_m^2 \\ \vdots & \vdots & \ddots & \vdots \\ X_1^n & X_2^n & \cdots & X_m^n \end{pmatrix}$, $Y = (Y_1, Y_2, \dots, Y_n)^T$, $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$, то уравнение линейной регрессии принимает вид

$$Y = Xa + \epsilon,$$

тогда

$$\hat{a} = \underset{a}{\text{argmin}}\, \epsilon^T \epsilon = \underset{a}{\text{argmin}} (Y - Xa)^T (Y - Xa) = (X^T X)^{-1} X^T Y.$$

Оценки **a** параметров **a** являются несмещёнными, состоятельными и с наименьшей дисперсией при соблюдении условий **теоремы Гаусса-Маркова:**

1. $\epsilon_i \sim N\left(0; \sigma^2\right)$;
2. $\forall j < i \operatorname{Cov}\left(\epsilon_j, \epsilon_i\right) = 0$;
3. rang $X = m$.

Коэффициент детерминации

$$R^2 = 1 - \frac{D(\epsilon)}{D(Y)}$$

показывает долю дисперсии целевого признака, которую объясняет модель. Его оценка вычисляется по формуле

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \epsilon_i^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}.$$

Код программы

```
import numpy as np
import pandas as pd
```

```
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.stats.diagnostic import het_breuschpagan,
    acorr_breusch_godfrey
import scipy.stats as sps

def load_data():
    data = np.loadtxt('C:/Users/AliceWolf13/Documents/mathstats/
        pain9/regressia.txt', delimiter=';')
    df = pd.DataFrame({'y': data})
    df['x'] = df.index
    return df

def plot_data(x, y, title, color='blue', label='Данные', pred=None,
  pred_label=None, pred_color=None):
    plt.figure(figsize=(10, 6))
    plt.scatter(x, y, color=color, label=label)
    if pred is not None:
        plt.plot(x, pred, color=pred_color, label=pred_label)
    plt.title(title)
    plt.grid(True)
    plt.legend()
    plt.show()

def check_gauss_markov(model, X, model_name):
    print(f"\nПроверка условий теоремы Гаусса-Маркова для {model_name}:")
    residuals = model.resid

    shapiro_test = sps.shapiro(residuals)
    print_result("1.1 Тест на нормальность остатков (Шапиро-Уилк):",
                 shapiro_test[1] > 0.05,
                 f"Статистика: {shapiro_test[0]:.4f}, p-value =
                    {shapiro_test[1]:.4f}")

    bp_test = het_breuschpagan(residuals, X)
    print_result("\n1.2 Тест на неоднородность дисперсии (Бройша-Пагана):",
                 bp_test[1] > 0.05,
                 f"Статистика: {bp_test[0]:.4f}, p-value = {bp_test[1]:.4f}")

    t_stat, p_value = sps.ttest_1samp(residuals, 0)
    print_result("\n1.3 Проверка нулевого матожидания остатков:",
```

```
            p_value > 0.05,
            f"t-статистика: {t_stat:.4f}, p-value = {p_value:.4f}")

    bg_test = acorr_breusch_godfrey(model, nlags=1)
    print_result("\n2. Тест на автокорреляцию (Бройша-Годфри):",
            bg_test[1] > 0.05,
            f"p-value = {bg_test[1]:.4f}")

    rank = np.linalg.matrix_rank(X)
    print(f"\n3. Проверка полного ранга матрицы X:")
    print(f"Ранг матрицы X: {rank}, число параметров m: {model.df_model +
        1}")
    print("Условие rang(X) = m выполняется" if rank == model.df_model + 1
    else
        "    Столбцы X линейно зависимы, условие теоремы не выполняется")

def print_result(test_name, condition, stats):
    print(f"{test_name}\n    {stats}")
    print("Условие выполняется" if condition else "Условие не выполняется")

def main():
    df = load_data()
    plot_data(df['x'], df['y'], 'Диаграмма рассеяния')

    X_linear = sm.add_constant(df['x'])
    model_linear = sm.OLS(df['y'], X_linear).fit()
    print(f"Линейная модель:\ny = {model_linear.params['x']:.6f}x +
        {model_linear.params['const']:.6f}")
    plot_data(df['x'], df['y'], 'Линейная регрессия',
        pred=model_linear.predict(X_linear),
            pred_label='Линейная регрессия', pred_color='red')
    check_gauss_markov(model_linear, X_linear, "линейной модели")
    print(model_linear.summary())

    for d in range(2, 5):
        df[f'x^{d}'] = df['x']**d
    X_poly = sm.add_constant(df[['x', 'x^2', 'x^3', 'x^4']])
    model_poly = sm.OLS(df['y'], X_poly).fit()
    print("\nПолиномиальная модель 4-й степени:")
    equation = f"y = {model_poly.params['const']:.6f}"
    for d in range(1, 5):
```

```
        equation += f" + {model_poly.params[f'x^{d}'
            if d>1 else 'x']:.6f}x^{d}" if d>1 else f" +
                {model_poly.params['x']:.6f}x"
    print(equation)
    plot_data(df['x'], df['y'], 'Полиномиальная регрессия 4-й степени',
            pred=model_poly.predict(X_poly), pred_label='Полином 4-й
                степени', pred_color='green')
    check_gauss_markov(model_poly, X_poly, "полиномиальной модели
        4-й степени")
    print(model_poly.summary())

    plt.figure(figsize=(12, 8))
    plt.scatter(df['x'], df['y'], color='blue', s=50, label='Данные')
    plt.plot(df['x'], model_linear.predict(X_linear), color='red',
        label='Линейная')
    plt.plot(df['x'], model_poly.predict(X_poly), color='green',
        label='Полином 4-й степени')
    plt.title('Сравнение регрессионных моделей')
    plt.grid(True)
    plt.legend()
    plt.show()

if __name__ == "__main__":
    main()
```

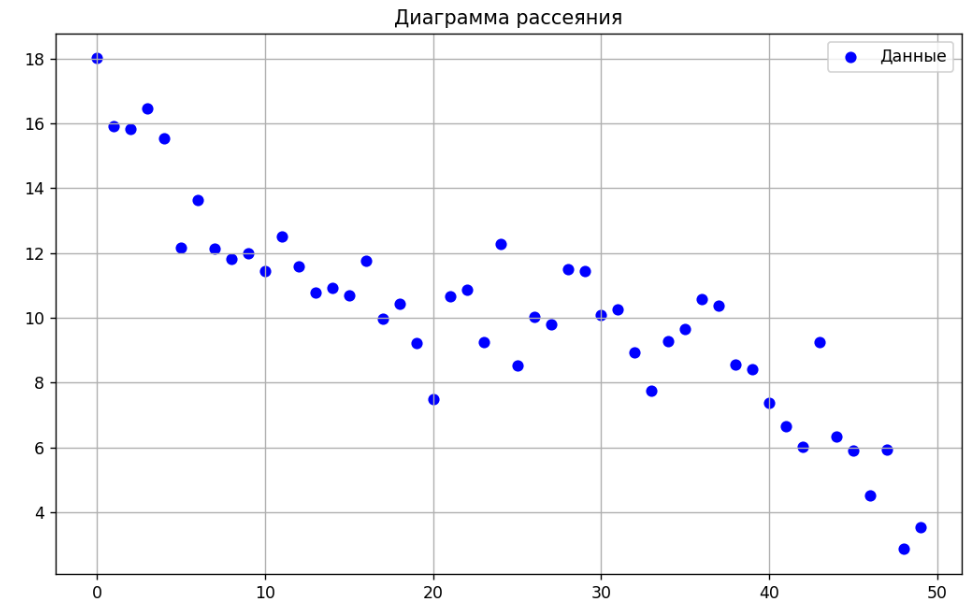
Вывод программы и графики

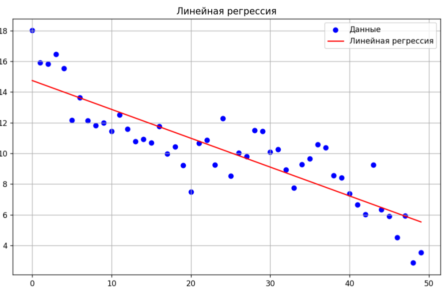
```
Линейная модель:
y = -0.188032x + 14.747349

Проверка условий теоремы Гаусса-Маркова для линейной модели:
1.1 Тест на нормальность остатков (Шапиро-Уилк):
    Статистика: 0.9738, p-value = 0.3287
    Условие выполняется

1.2 Тест на неоднородность дисперсии (Бройша-Пагана):
    Статистика: 0.0873, p-value = 0.7676
    Условие выполняется

1.3 Проверка нулевого матожидания остатков:
    t-статистика: 0.0000, p-value = 1.0000
    Условие выполняется

2. Тест на автокорреляцию (Бройша-Годфри):
    p-value = 0.0015
    Условие не выполняется

3. Проверка полного ранга матрицы X:
    Ранг матрицы X: 2, число параметров m: 2.0
    Условие rang(X) = m выполняется
```

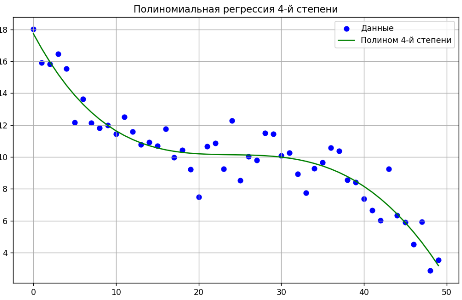
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.757
Model:                            OLS   Adj. R-squared:                  0.752
Method:                 Least Squares   F-statistic:                     149.6
Date:                Sat, 07 Jun 2025   Prob (F-statistic):           2.34e-16
Time:                        08:28:41   Log-Likelihood:                -92.445
No. Observations:                  50   AIC:                             188.9
Df Residuals:                      48   BIC:                             192.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         14.7473      0.437     33.733      0.000      13.868      15.626
x             -0.1880      0.015    -12.230      0.000      -0.219      -0.157
==============================================================================
Omnibus:                        1.246   Durbin-Watson:                   0.993
Prob(Omnibus):                  0.536   Jarque-Bera (JB):                1.067
Skew:                           0.146   Prob(JB):                        0.587
Kurtosis:                       2.347   Cond. No.                         56.1
==============================================================================
```

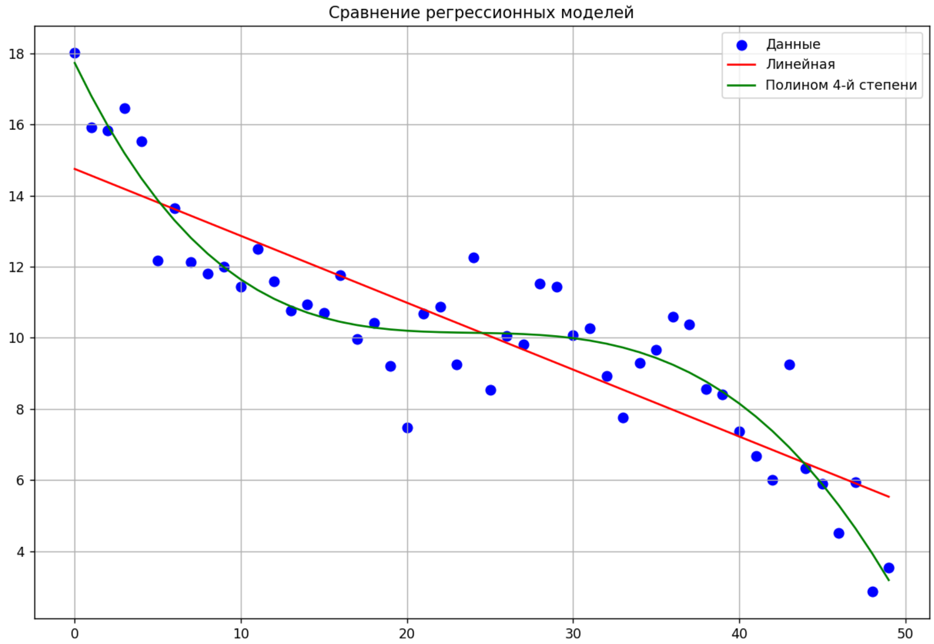
```
Полиномиальная модель 4-й степени:
y = 17.726984 + -0.968816x + 0.042852x^2 + -0.000708x^3 + 0.000002x^4

Проверка условий теоремы Гаусса-Маркова для полиномиальной модели 4-й степени:
1.1 Тест на нормальность остатков (Шапиро-Уилк):
    Статистика: 0.9849, p-value = 0.7687
    Условие выполняется

1.2 Тест на неоднородность дисперсии (Бройша-Пагана):
    Статистика: 1.6187, p-value = 0.8054
    Условие выполняется

1.3 Проверка нулевого матожидания остатков:
    t-статистика: 0.0000, p-value = 1.0000
    Условие выполняется

2. Тест на автокорреляцию (Бройша-Годфри):
    p-value = 0.6419
    Условие выполняется

3. Проверка полного ранга матрицы X:
    Ранг матрицы X: 5, число параметров m: 5.0
    Условие rang(X) = m выполняется
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.890
Model:                            OLS   Adj. R-squared:                  0.881
Method:                 Least Squares   F-statistic:                     91.35
Date:                Sat, 07 Jun 2025   Prob (F-statistic):           5.29e-21
Time:                        08:29:26   Log-Likelihood:                -72.556
No. Observations:                  50   AIC:                             155.1
Df Residuals:                      45   BIC:                             164.7
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         17.7270      0.686     25.845      0.000      16.345      19.108
x             -0.9688      0.198     -4.890      0.000      -1.368      -0.570
x^2            0.0429      0.017      2.568      0.014       0.009       0.076
x^3           -0.0007      0.001     -1.376      0.176      -0.002       0.000
x^4         2.307e-06     5.2e-06      0.443      0.660   -8.17e-06    1.28e-05
==============================================================================
Omnibus:                        0.224   Durbin-Watson:                   2.127
Prob(Omnibus):                  0.894   Jarque-Bera (JB):                0.037
Skew:                          -0.067   Prob(JB):                        0.981
Kurtosis:                       3.000   Cond. No.                     9.13e+06
==============================================================================
```




Линейная регрессия
Данные
Линейная регрессия


Полиномиальная регрессия 4-й степени
Данные
Полином 4-й степени


Сравнение регрессионных моделей
Данные
Линейная
Полином 4-й степени