Chapter 6

Multiple Linear Regression (solutions to exercises)

Contents

Import Python packages

```
# Import all needed python packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import scipy.stats as stats
import statsmodels.formula.api as smf
import statsmodels.api as sm
```

6.1 Nitrate concentration

|||| Exercise 6.1 Nitrate concentration

In order to analyze the effect of reducing nitrate loading in a Danish fjord, it was decided to formulate a linear model that describes the nitrate concentration in the fjord as a function of nitrate loading, it was further decided to correct for fresh water runoff. The resulting model was

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \tag{6-1}$$

where Y_i is the natural logarithm of nitrate concentration, $x_{1,i}$ is the natural logarithm of nitrate loading, and $x_{2,i}$ is the natural logarithm of fresh water run off.

a) Which of the following statements are assumed fulfilled in the usual multiple linear regression model?

1) $\varepsilon_i = 0$ for all $i = 1, ..., n$, and β_j follows a normal distribution

2) $E[x_1] = E[x_2] = 0$ and $V[\varepsilon_i] = \beta_1^2$

3) $E[\varepsilon_i] = 0$ and $V[\varepsilon_i] = \beta_1^2$

4) ε_i is normally distributed with constant variance, and ε_i and ε_j are independent for $i \neq j$

5) $\varepsilon_i = 0$ for all $i = 1, ..., n$, and x_j follows a normal distribution for $j = \{1, 2\}$

The parameters in the model were estimated in Python and the following results are available (slightly modified output from summary):

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                     0.3438
Model:                            OLS   Adj. R-squared:                0.3382
No. Observations:                 240   F-statistic:                    62.07
Covariance Type:            nonrobust   Prob (F-statistic):           2.2e-16
==============================================================================
                coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -2.36500      0.222    -10.661     <2e-16           *           *
```

```
x1               0.4762      0.062      7.720   3.25e-13          *           *
x2               0.0827      0.070      1.185      0.273          *           *
==================================================================================
```

b) What are the parameter estimates for the model parameters ($\hat{\beta}_i$ and $\hat{\sigma}^2_{\beta_i}$) and how many degrees of freedom are there in the estimation?

c) Calculate the usual 95% confidence intervals for the parameters (β_0, β_1, and β_2).

d) On level $\alpha = 0.05$ which of the parameters are significantly different from 0, also find the p-values for the tests used for each of the parameters?

6.2 Multiple linear regression model

|||| Exercise 6.2 Multiple linear regression model

The following measurements have been obtained in a study:

No.	1	2	3	4	5	6	7	8	9	10	11	12	13
y	1.45	1.93	0.81	0.61	1.55	0.95	0.45	1.14	0.74	0.98	1.41	0.81	0.89
x_1	0.58	0.86	0.29	0.20	0.56	0.28	0.08	0.41	0.22	0.35	0.59	0.22	0.26
x_2	0.71	0.13	0.79	0.20	0.56	0.92	0.01	0.60	0.70	0.73	0.13	0.96	0.27

No.	14	15	16	17	18	19	20	21	22	23	24	25
y	0.68	1.39	1.53	0.91	1.49	1.38	1.73	1.11	1.68	0.66	0.69	1.98
x_1	0.12	0.65	0.70	0.30	0.70	0.39	0.72	0.45	0.81	0.04	0.20	0.95
x_2	0.21	0.88	0.30	0.15	0.09	0.17	0.25	0.30	0.32	0.82	0.98	0.00

It is expected that the response variable y can be described by the independent variables x_1 and x_2. This imply that the parameters of the following model should be estimated and tested

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2).$$

a) Calculate the parameter estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\sigma}^2$), in addition find the usual 95% confidence intervals for β_0, β_1, and β_2.
You can copy the following lines to Python to load the data:

```
df = pd.DataFrame({
    'x1': [0.58, 0.86, 0.29, 0.20, 0.56, 0.28, 0.08, 0.41, 0.22,
        0.35, 0.59, 0.22, 0.26, 0.12, 0.65, 0.70, 0.30, 0.70,
        0.39, 0.72, 0.45, 0.81, 0.04, 0.20, 0.95],
    'x2': [0.71, 0.13, 0.79, 0.20, 0.56, 0.92, 0.01, 0.60, 0.70,
        0.73, 0.13, 0.96, 0.27, 0.21, 0.88, 0.30, 0.15, 0.09,
        0.17, 0.25, 0.30, 0.32, 0.82, 0.98, 0.00],
    'y': [1.45, 1.93, 0.81, 0.61, 1.55, 0.95, 0.45, 1.14, 0.74,
        0.98, 1.41, 0.81, 0.89, 0.68, 1.39, 1.53, 0.91, 1.49,
        1.38, 1.73, 1.11, 1.68, 0.66, 0.69, 1.98]
})
```

b) Still using confidence level $\alpha = 0.05$ reduce the model if appropriate.

c) Carry out a residual analysis to check that the model assumptions are fulfilled.

d) Make a plot of the fitted line and 95% confidence and prediction intervals of the line for $x_1 \in [0, 1]$ (it is assumed that the model was reduced above).

6.3 MLR simulation exercise

|||| Exercise 6.3 MLR simulation exercise

The following measurements have been obtained in a study:

Nr.	1	2	3	4	5	6	7	8
y	9.29	12.67	12.42	0.38	20.77	9.52	2.38	7.46
x_1	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00
x_2	4.00	12.00	16.00	8.00	32.00	24.00	20.00	28.00

a) Plot the observed values of y as a function of x_1 and x_2. Does it seem reasonable that either x_1 or x_2 can describe the variation in y?
You may copy the following lines into Python to load the data

```
df = pd.DataFrame({
    'y': [9.29,12.67,12.42,0.38,20.77,9.52,2.38,7.46],
    'x1': [1.00,2.00,3.00,4.00,5.00,6.00,7.00,8.00],
    'x2': [4.00,12.00,16.00,8.00,32.00,24.00,20.00,28.00]
})
```

b) Estimate the parameters for the two models

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

and

$$Y_i = \beta_0 + \beta_1 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

and report the 95% confidence intervals for the parameters. Are any of the parameters significantly different from zero on a 5% confidence level?

c) Estimate the parameters for the model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i, \quad \varepsilon_i \sim (N(0, \sigma^2), \tag{6-2}$$

and go through the steps of Method 6.16 (use confidence level 0.05 in all tests).

d) Find the standard error for the line, and the confidence and prediction intervals for the line for the points $(\min(x_1), \min(x_2))$, $(\bar{x}_1, \bar{x}_2)$, $(\max(x_1), \max(x_2))$.

e) Plot the observed values together with the fitted values (e.g. as a function of x_1).