Chapter 5

Simple Linear regression (solutions to exercises)

Contents

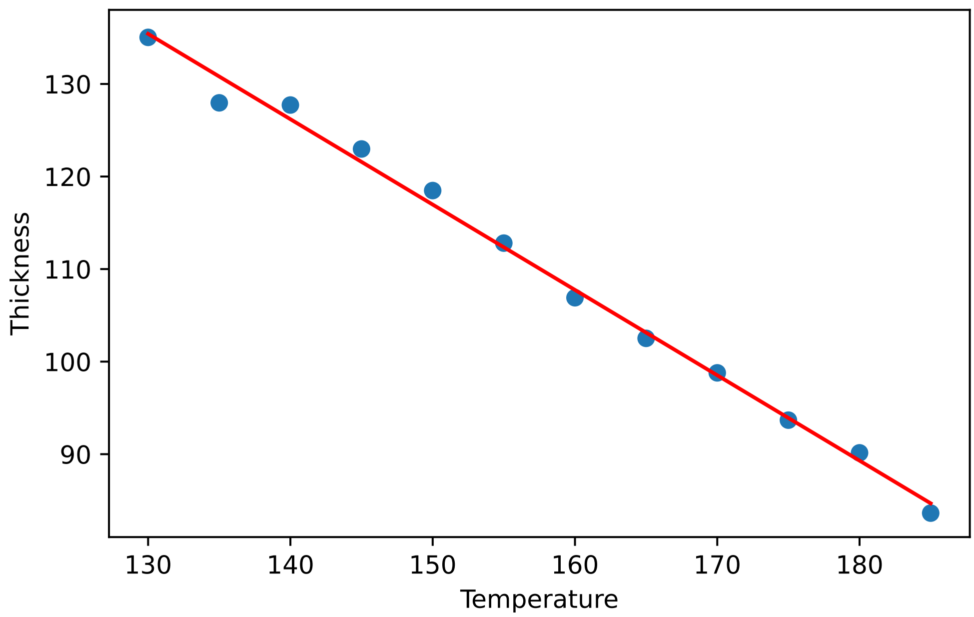
Import Python packages

```python
# 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
```

5.1 Plastic film folding machine

Exercise 5.1 Plastic film folding machine

On a machine that folds plastic film the temperature may be varied in the range of 130-185 °C. For obtaining, if possible, a model for the influence of temperature on the folding thickness, $n = 12$ related set of values of temperature and the fold thickness were measured that is illustrated in the following figure:



a) Determine by looking at the figure, which of the following sets of estimates for the parameters in the usual regression model is correct:

1) $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = -0.9$, $\hat{\sigma} = 36$
2) $\hat{\beta}_0 = 0$, $\hat{\beta}_1 = 0.9$, $\hat{\sigma} = 3.6$
3) $\hat{\beta}_0 = 252$, $\hat{\beta}_1 = -0.9$, $\hat{\sigma} = 3.6$
4) $\hat{\beta}_0 = -252$, $\hat{\beta}_1 = -0.9$, $\hat{\sigma} = 36$
5) $\hat{\beta}_0 = 252$, $\hat{\beta}_1 = -0.9$, $\hat{\sigma} = 36$

b) What is the only possible correct answer:

1) The proportion of explained variation is 50% and the correlation is 0.98
2) The proportion of explained variation is 0% and the correlation is -0.98
3) The proportion of explained variation is 96% and the correlation is -1
4) The proportion of explained variation is 96% and the correlation is 0.98
5) The proportion of explained variation is 96% and the correlation is -0.98

5.2 Linear regression life time model

Exercise 5.2 Linear regression life time model

A company manufactures an electronic device to be used in a very wide temperature range. The company knows that increased temperature shortens the life time of the device, and a study is therefore performed in which the life time is determined as a function of temperature. The following data is found:

Temperature in Celcius (t)	10	20	30	40	50	60	70	80	90
Life time in hours (y)	420	365	285	220	176	117	69	34	5

a) Calculate the 95% confidence interval for the slope in the usual linear regression model, which expresses the life time as a linear function of the temperature.

b) Can a relation between temperature and life time be documented on level 5%?

5.3 Yield of chemical process

Exercise 5.3 Yield of chemical process

The yield y of a chemical process is a random variable whose value is considered to be a linear function of the temperature x. The following data of corresponding values of x and y is found:

Temperature in °C (x)	0	25	50	75	100
Yield in grams (y)	14	38	54	76	95

The average and standard deviation of temperature and yield are

$$\bar{x} = 50,\ s_x = 39.52847,\ \bar{y} = 55.4,\ s_y = 31.66702,$$

In the exercise the usual linear regression model is used

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \quad i = 1, \dots, 5$$

a) Can a significant relationship between yield and temperature be documented on the usual significance level $\alpha = 0.05$?

b) Give the 95% confidence interval of the expected yield at a temperature of $x_{\text{new}} = 80$ °C.

c) What is the upper quartile of the residuals?

5.4 Plastic material

|||| Exercise 5.4 Plastic material

In the manufacturing of a plastic material, it is believed that the cooling time has an influence on the impact strength. Therefore a study is carried out in which plastic material impact strength is determined for 4 different cooling times. The results of this experiment are shown in the following table:

Cooling times in seconds (x)	15	25	35	40
Impact strength in kJ/m^2 (y)	42.1	36.0	31.8	28.7

The following statistics may be used:

$$\bar{x} = 28.75, \; \bar{y} = 34.65, \; S_{xx} = 368.75.$$

a) What is the 95% confidence interval for the slope of the regression model, expressing the impact strength as a linear function of the cooling time?

b) Can you conclude that there is a relation between the impact strength and the cooling time at significance level $\alpha = 5\%$?

c) For a similar plastic material the tabulated value for the linear relation between temperature and impact strength (i.e the slope) is -0.30. If the following hypothesis is tested (at level $\alpha = 0.05$)

$$H_0 : \beta_1 = -0.30$$
$$H_1 : \beta_1 \neq -0.30$$

with the usual t-test statistic for such a test, what is the range (for t) within which the hypothesis is accepted?

5.5 Water polution

|||| Exercise 5.5 Water polution

In a study of pollution in a water stream, the concentration of pollution is measured at 5 different locations. The locations are at different distances to the pollution source. In the table below, these distances and the average pollution are given:

Distance to the pollution source (in km)	2	4	6	8	10
Average concentration	11.5	10.2	10.3	9.68	9.32

a) What are the parameter estimates for the three unknown parameters in the usual linear regression model: 1) The intercept (β_0), 2) the slope (β_1) and 3) error standard deviation (σ)?

b) How large a part of the variation in concentration can be explained by the distance?

c) What is a 95%-confidence interval for the expected pollution concentration 7 km from the pollution source?

5.6 Membrane pressure drop

Exercise 5.6 Membrane pressure drop

When purifying drinking water you can use a so-called membrane filtration. In an experiment one wishes to examine the relationship between the pressure drop across a membrane and the flux (flow per area) through the membrane. We observe the following 10 related values of pressure (x) and flux (y):

	1	2	3	4	5	6	7	8	9	10
Pressure (x)	1.02	2.08	2.89	4.01	5.32	5.83	7.26	7.96	9.11	9.99
Flux (y)	1.15	0.85	1.56	1.72	4.32	5.07	5.00	5.31	6.17	7.04

Copy this into Python to avoid typing in the data:

```
df = pd.DataFrame({
    'pressure': [1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99],
    'flux': [1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04]
})
```

a) What is the empirical correlation between pressure and flux estimated to? Give also an interpretation of the correlation.

b) What is a 90% confidence interval for the slope β_1 in the usual regression model?

c) How large a part of the flux-variation ($\sum_{i=1}^{10}(y_i - \bar{y})^2$) is not explained by pressure differences?

d) Can you at significance level $\alpha = 0.05$ reject the hypothesis that the line passes through $(0, 0)$?

e) A confidence interval for the line at three different pressure levels: $x_{\text{new}}^{\text{A}} = 3.5$, $x_{\text{new}}^{\text{B}} = 5.0$ and $x_{\text{new}}^{\text{C}} = 9.5$ will look as follows:

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot x_{\text{new}}^{U} \pm C_U$$

where U then is either A, B or C. Write the constants C_U in increasing order.

5.7 Membrane pressure drop (matrix form)

|||| Exercise 5.7 Membrane pressure drop (matrix form)

This exercise uses the data presented in Exercise 6 above.

a) Find parameters values, standard errors, t-test statistics, and p-values for the standard hypotheses tests.

   Copy this into Python to avoid typing in the data:

```
df = pd.DataFrame({
    'pressure': [1.02,2.08,2.89,4.01,5.32,5.83,7.26,7.96,9.11,9.99],
    'flux': [1.15,0.85,1.56,1.72,4.32,5.07,5.00,5.31,6.17,7.04]
})
```

b) Reproduce the above numbers by matrix vector calculations. You will need some matrix notation in Python:
   - Matrix multiplication (XY): `np.dot(X,Y) or X@Y`
   - Matrix transpose (X^T): `X.T`
   - Matrix inverse (X^{-1}): `np.linalg.inv(X)`
   - Make a matrix from vectors ($X = [x_1^T; x_2^T]$): `np.column_stack((x1,x2))`

   See also Example 5.24.

5.8 Independence and correlation

|||| Exercise 5.8 Independence and correlation

Consider the layout of independent variable in Example 5.11,

a) Show that $S_{xx} = \frac{n \cdot (n+1)}{12 \cdot (n-1)}$.

Hint: you can use the following relations

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2},$$
$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

b) Show that the asymptotic correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\lim_{n \to \infty} \rho_n(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sqrt{3}}{2}.$$

Consider a layout of the independent variable where $n = 2k$ and $x_i = 0$ for $i \leq k$ and $x_i = 1$ for $k < i \leq n$.

c) Find S_{xx} for the new layout of x.

d) Compare S_{xx} for the two layouts of x.

e) What is the consequence for the parameter variance in the two layouts?

f) Discuss pro's and cons for the two layouts.