

Chapter 5

Simple linear regression

Exercises

Contents

5 Simple linear regression

Exercises	1
5.1 Plastic film folding machine	3
5.2 Linear regression life time model	5
5.3 Yield of chemical process	7
5.4 Plastic material	8
5.5 Water pollution	9
5.6 Membrane pressure drop	10
5.7 Membrane pressure drop (matrix form)	15
5.8 Independence and correlation	17

5.1 Plastic film folding machine

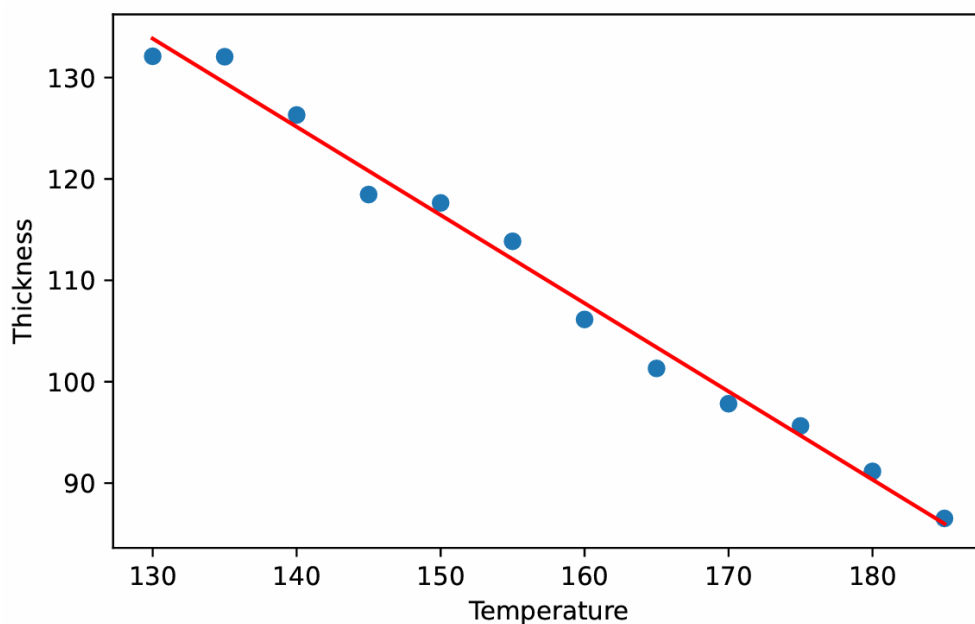
|||| Exercise 5.1 Plastic film folding machine

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 sets of values of temperature and the fold thickness were measured.

The following regression model was fitted to the data:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

The data and the estimated model are illustrated in the following figure:



Exam-style questions

- a) Determine by looking at the figure which of the following sets of estimates for the parameters in the 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) The proportion of the variation in fold thickness that is explained by the regression model is given by the coefficient of determination, R^2 .

Which of the following statements is correct?

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

In this exercise you should use Python for computations, data visualization etc.

The data can be stored in a Python DataFrame named `df` by copy-pasting the following code:

```
df = pd.DataFrame({
    't': [10, 20, 30, 40, 50, 60, 70, 80, 90],
    'y': [420, 365, 285, 220, 176, 117, 69, 34, 5]
})
```

- a) Analyse the relationship between Temperature and Life time:
 - Visualize the data
 - Specify a relevant statistical model
 - Estimate the model parameters
 - Interpret the results
- b) Compute a 95% confidence interval for the slope in the linear regression model. What do you conclude from this confidence interval?
- c) Compute a 99% confidence interval for the slope in the linear regression model. What do you conclude from this confidence interval?

Use Python to compute quantiles and probabilities from relevant probability distribution(s).

d) Compute the p -value for the slope.

Use Python to compute quantiles and probabilities from relevant probability distribution(s).

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

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

- Can a significant relationship between yield and temperature be documented on the usual significance level $\alpha = 0.05$?
- Give the 95% confidence interval of the expected yield at a temperature of $x_{\text{new}} = 80$ °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:

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

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

|||| Exercise 5.5 Water pollution

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

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

The data can be stored in a Python DataFrame named `df` by copy-pasting the following code:

```
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) Visualize the data and estimate the sample correlation coefficient between pressure and flux. Use Python for both tasks.

Does the resulting correlation match your expectations from inspecting the data visually?

The following regression model is fitted to the data:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where x_i is the pressure and Y_i is the flux. The resulting output table is shown here:

```
fit = smf.ols('flux ~ pressure', data=df).fit()
print(fit.summary(slim=True))
```

```

=====
                        OLS Regression Results
=====
Dep. Variable:          flux      R-squared:                0.929
Model:                  OLS      Adj. R-squared:            0.920
No. Observations:       10      F-statistic:           104.6
Covariance Type:        nonrobust Prob (F-statistic):      7.18e-06
=====
                        coef      std err          t      P>|t|      [0.025      0.975]
-----
Intercept      -0.1886      0.442      -0.427      0.681      -1.207      0.830
pressure        0.7225      0.071     10.227      0.000      0.560      0.885
=====

```

- b) What is a 90% confidence interval for the slope β_1 ? Compare this 90% confidence interval to the 95% confidence interval shown in the regression table.

Use Python to compute quantiles and probabilities from relevant probability distribution(s).

- c) Let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

be the predicted (or *fitted*) values of flux corresponding to pressure x_i (values on the regression line).

Now compute the following measures of variability of the flux (use Python for computations).

1. The sample variance of the flux.
2. The variability **explained** by the regression line, i.e.:

$$\frac{1}{n-1} \sum_{i=1}^{10} (\hat{y}_i - \bar{y})^2$$

Explain in words what the *variability explained by the regression line* expresses.

3. The variability **NOT explained** by the regression line, i.e.:

$$\frac{1}{n-1} \sum_{i=1}^{10} (y_i - \hat{y}_i)^2$$

Explain in words what the *variability NOT explained by the regression line* expresses.

What is the proportion of the flux-variation that is explained by the model, and what is the proportion that is NOT explained? How does this result compare to the R-squared value printed in the regression table and to the sample correlation coefficient calculated above?

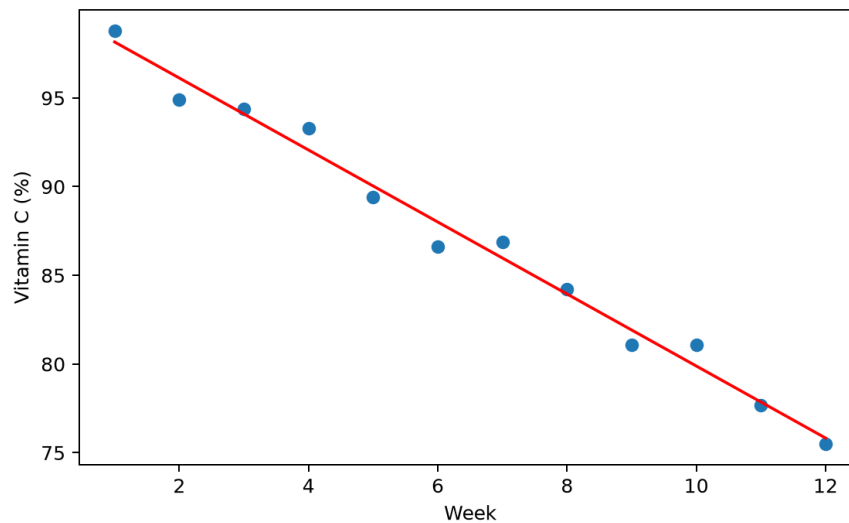
- d) Can you (at significance level $\alpha = 0.05$) reject the hypothesis that the line passes through $(0,0)$?
- e) What is the expected (mean) flux at pressure 5.0 and what is the 95% confidence interval for this expected flux?
- f) At pressure 5.0, what is the 95% prediction interval? What is the interpretation of this prediction interval?

Exam-style questions

A food company examines how the vitamin C content in bottled orange juice decreases during storage. In each of the weeks $1, 2, \dots, 12$ after bottling, one bottle is opened and the remaining vitamin C content is measured (in percent of the content at bottling). The following regression model is fitted to the data:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where x_i is the storage time (in weeks) and Y_i is the vitamin C content (in %). The data and the estimated regression line are shown in the following figure:



The regression table from Python is:

```
fit = smf.ols('vitaminC ~ week', data=df).fit()
print(fit.summary(slim=True))
```

OLS Regression Results						
Dep. Variable:	vitaminC		R-squared:	0.985		
Model:	OLS		Adj. R-squared:	0.983		
No. Observations:	12		F-statistic:	647.2		
Covariance Type:	nonrobust		Prob (F-statistic):	2.02e-10		
	coef	std err	t	P> t	[0.025	0.975]
Intercept	100.1848	0.587	170.614	0.000	98.876	101.493
week	-2.0297	0.080	-25.440	0.000	-2.207	-1.852

Furthermore, the following values have been printed:

```
print(df['week'].mean())
6.5
print(df['week'].var(ddof=1))
13.0
print(np.sqrt(fit.scale))
0.9541
print(stats.t.ppf(0.95, df=10))
1.812
print(stats.t.ppf(0.975, df=10))
2.228
```

g) What is the 95% confidence interval for the expected vitamin C content after 10 weeks of storage?

1) [79.27; 80.50]

- 2) [77.59; 82.19]
- 3) [79.01; 80.76]
- 4) [77.73; 82.04]
- 5) [79.18; 80.60]

h) Which of the following statements is correct?

- 1) Since $R^2 = 0.985$, 98.5% of the observations are expected to lie exactly on the regression line.
- 2) The p -value for the slope (0.000) is the probability that the vitamin C content does not change with storage time.
- 3) The 95% confidence interval for the expected (mean) vitamin C content is narrowest in week 1 and week 12, since these are the endpoints of the data.
- 4) If the experiment were repeated with many more bottles, the 95% prediction interval for a single new bottle after 10 weeks would still not shrink towards zero width.
- 5) The model can safely be used to predict the vitamin C content after 60 weeks of storage.

5.7 Membrane pressure drop (matrix form)

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

This exercise uses the same data and regression model as in Exercise 5.6 above. 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

The data can be stored in a Python DataFrame named `df` by copy-pasting the following code:

```
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]
})
```

The following regression model is fitted to the data:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2),$$

where x_i is the pressure and Y_i is the flux. The regression model can also be written in matrix form:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}),$$

where $\mathbf{X}$ is the design matrix, $\mathbf{Y}$ is the vector of outcome values (flux) and $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is the vector of parameters. The resulting output table is shown here:

```
fit = smf.ols('flux ~ pressure', data=df).fit()
print(fit.summary(slim=True))
```

OLS Regression Results						
=====						
Dep. Variable:	flux	R-squared:	0.929			
Model:	OLS	Adj. R-squared:	0.920			
No. Observations:	10	F-statistic:	104.6			
Covariance Type:	nonrobust	Prob (F-statistic):	7.18e-06			
=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	-0.1886	0.442	-0.427	0.681	-1.207	0.830
pressure	0.7225	0.071	10.227	0.000	0.560	0.885
=====						

- a) Write down (by hand) the design matrix X and the vector of outcome values Y corresponding to the linear regression model given above. How many rows and columns do they have, and what does each column of X contain?

See also Example 5.24.

- b) Reproduce the parameter estimates and standard errors from the output table above, but using matrix-vector calculations in Python.

Start by implementing the design matrix X and the vector Y from question a) in Python, and print them to check that they match your sketch.

You will need some matrix notation in Python:

- A column of n ones: `np.ones(n)`
- Combine two columns `x1` and `x2` into a matrix:
`np.column_stack((x1, x2))`
- A column from the data, e.g. the fluxes: `df['flux']`
- Matrix multiplication (XY): `np.dot(X, Y)` or `X@Y`
- Matrix transpose (X^T): `X.T`
- Matrix inverse (X^{-1}): `np.linalg.inv(X)`

See also Theorem 5.23 and 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

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

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 pros and cons for the two layouts.