Chapter 4: Correlation and regression

Learning objectives of this chapter:

- Calculating covariance and correlation by hand
- Testing hypotheses about a correlation by hand
- Calculating and testing a correlation in R
- Testing hypotheses using linear regression in R
- Making predictions using linear regression in R

Assignment 4.1: Calculating covariance and correlation by hand

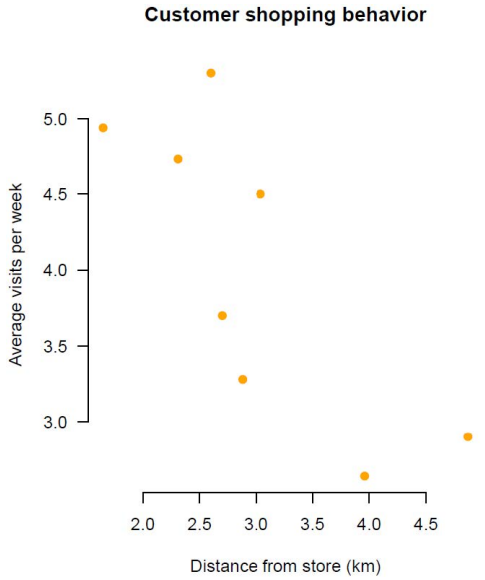
The manager of a local branch of a supermarket chain in the Netherlands has had a bad year last year and wants to improve their sales. Her co-worker has brought up the idea to spend more money on advertising in the neighborhoods that are further away from the store. The co-worker thinks that customers that live further away might visit the supermarket less frequently. He believes that making an effort to attract these people may prove to be profitable for the store. Realizing that her branch currently does not spend a notable amount of money on advertising at all, the manager is interested to know about the **relationship** between the distance a customer lives from the store and the average number of times they visit the store per week.

The manager requests her co-worker to ask the customers at his counter some questions when they check out at his counter. He is instructed to ask them a) how many times, on average, they visit the store per week and b) the number of kilometers they live from the store. The co-worker asks eight customers and reports the findings to his manager.

4.1 a) What kind of **relationship** do you expect to see? Formulate your answer in terms of the direction of the **relationship** .

Answer 4.1a:

The manager sits down with her co-worker, writes down the numbers, and creates a scatter plot of the data. She displays the distance a customer lives from the store (in kilometers) on the *x-axis* and their average number of visits per week on the *y-axis*.



i	x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$
1	4.87	2.90			
2	3.04	4.50			
3	1.65	4.94			
4	2.88	3.28			
5	2.31	4.73			
6	3.96	2.64			
7	2.70	3.70			
8	2.60	5.30			

$\bar{x} =$ ..... $\sum(x_i - \bar{x})(y_i - \bar{y}) =$ .....

$\bar{y} =$ ..... $n - 1 =$ .....

$s_{xy} =$ .....

4.1 b) Use the table above to calculate and fill in the **covariance** s_{xy} between the customer's distance from the store and their average number of visits per week.

Hint 4.1: you can find the complete formula for the ***covariance*** *in the formula sheet.*

4.1 c) Interpret the **covariance** from these data. What can you say about the **relationship** between the distance from the store and the average number of visits per week on the basis of this measure?

Answer 4.1c:

The **covariance** s_{xy} is a useful measure to get an idea about the extent to which two **variables** change together, but it also has some disadvantages when using it as a measure of the strength of a **relationship** .

4.4 d) Explain the disadvantage of using the **covariance** s_{xy} as a measure for the strength of this **relationship** . Consider in your answer what would have happened if the co-worker asked the customers how many meters they lived from the store.

Answer 4.1d:

To reliably measure the strength of the **relationship** , the manager wants to **standardize** the **covariance** s_{xy} to find out the **correlation** coefficient r_{xy}. To do this, she first needs to calculate the sample **standard deviation** of the distance from the store in kilometers (s_x) and the sample **standard deviation** of the average number of visits per week(s_y).

4.1 e) Calculate s_x and s_y for the data that the co-worker collected.

Hint 4.2: you can find the formula for the ***standard deviation*** *in the formula sheet.*

Answer 4.1e:

s_x : s_y :

4.1 f) Use the sample **standard deviations** and the sample **covariance** to calculate the sample **correlation** r_{xy}.

Hint 4.3: you can find the formula for the ***correlation coefficient*** r_{xy} *in the formula sheet.*

Answer 4.1f:

r_{xy} :

4.1 g) Interpret the **correlation coefficient** . Is this a strong **relationship** ?

Answer 4.1g:

Assignment 4.2: Testing hypotheses about a correlation by hand

The store manager looks at the results of her calculations and sees some potential for increasing the sales of her branch. However, before she starts spending more money on advertising in these to hard-to-reach neighborhoods, she wants to gain reasonable assurance that her results can be generalized to all of her potential customers. Therefore, she wants to test the hypothesis that, with 95% confidence, this **correlation** is truly negative in the entire **population** of customers.

4.2 a) Formulate the **null hypothesis** H_0 and **alternative hypothesis** H_1 for the manager's test of the **population correlation coefficient** ρ_{xy}.

Hint 4.4: Think about whether this is a one-sided test or a two-sided test?

Answer 4.2a:

H_0: _____ H_1: _____

4.2 b) Write down the relevant information that appeared in assignment 4.1. Use the symbols N, n, r_{xy}, ρ_{xy}

Hint 4.5: Not all symbols are known.

Answer 4.2b:

N: _____ r_{xy}: _____

n: _____ ρ_{xy}: _____

4.2 c) Find out the **critical z-value** that is required to reject H_0 with 95% confidence.

Answer 4.2c:

$z_{critical}$: _____

4.2 d) Calculate the sample **z-value** for the **correlation coefficient** r_{xy}.

*Hint 4.6: You can find the formula for the **z-value** of a **correlation** in the formula sheet.*

Answer 4.2d:

z_{xy} : _____

4.2 e) Draw the conclusion for the manager. Include the following four elements:

- ☐ Show how the calculated **z-value** relates to the **critical z-value** .
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about ρ_{xy}.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 4.2e:

Assignment 4.3: Calculating and testing a correlation in R

The supermarket manager believes she has found a weakness in the distribution policy of the supermarket and believes that more stores should be built nationwide to be more active in other areas. To strengthen her case for the board of directors, she has asked permission to execute her survey in 100 stores in the Netherlands to gather data of 1000 customers. She wants to confirm her **hypothesis** that in the entire Netherlands, there is a negative **relationship** between the distance a customer lives from their supermarket, and the number of times they visit their supermarket.

For this assignment, you need the data file **localSupermarket.csv** , which contains a population of 1,000 observations.

4.3 a) Use the **read.csv()** function (and **setwd()** function if you prefer) to import the data set into a data frame called dataset6.

R code 4.3a:

```
```

4.3 b) Use the **cov()** and **cor()** functions to calculate the **covariance** s_{xy} and the **correlation** r_{xy} of the columns **Distance** and **AvgVisits** .

R code 4.3b:

```
```

Answer 4.3 b:

Covariance: _____

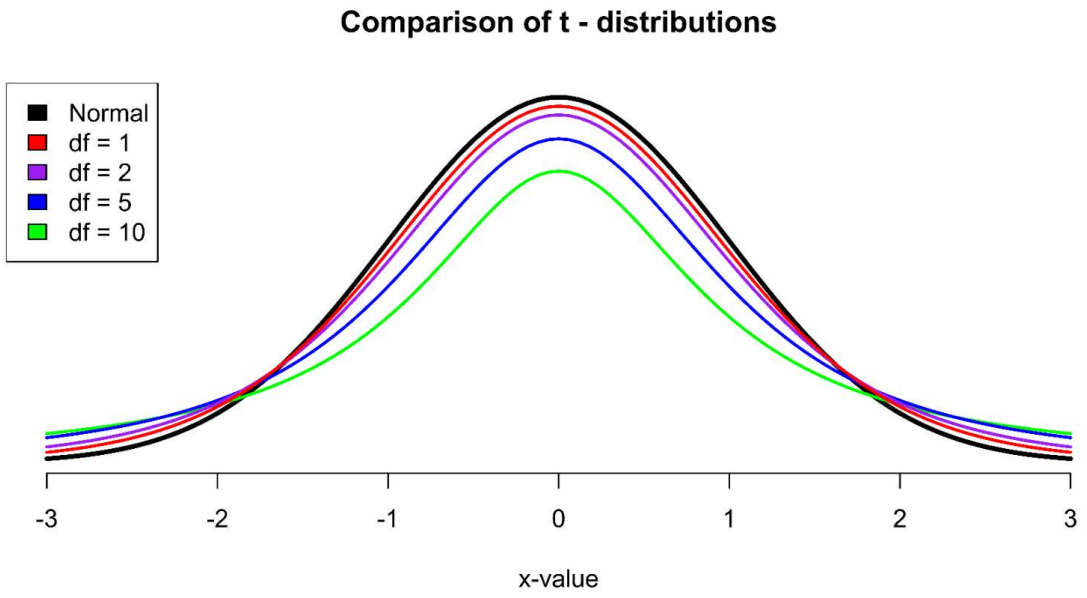
Correlation: _____

4.3 c) Use the function **cor.test()** to test for a negative **relationship** in this **population** . Use **?cor.test** to get help about the function arguments. Can you confirm the value of the **correlation** that you found in assignment 4.3b?

R code 4.3c:

```
```

Instead of using the **standard normal distribution** $N(\mu = 0, \sigma = 1)$ and a **z-value** for testing ρ_{xy} against any value, you can simplify the procedure when you are testing a **correlation coefficient** testing against the value zero (which is almost always the case). In such a case, you can use the **t-distribution** for calculating a **t-value** for the significance of ρ_{xy} against a value of zero. The **t-distribution** is almost identical to the **normal distribution** . However, where the **normal distribution** is defined by its **mean** (μ) and its **standard deviation** (σ), the **t-distribution** is defined by its degrees of freedom (df_{n-1}).



Use the following R code to create a figure of the **t-distribution** with $df = 3$.

```
curve(dnorm(x, mean = 0, sd = 1), from = -3, to = 3, ylab = 'Density')
curve(dt(x, df = 3), from = -3, to = 3, add = TRUE, col = 'red')
```

4.3 d) Which line represents the **normal distribution** in the figure that was drawn? And which line represents the **t-distribution** ? Can you explain what the difference between the two distributions is in terms of their shape? What happens when you increase the **degrees of freedom** (**df**) in the **dt()** function?

Answer 4.3d:

4.3 e) Calculate the **degrees of freedom** and the **t-value** for the **correlation coefficient** between **Distance** and AvgVisits.

*Hint 4.7: You can find the formula for the **degrees of freedom** and the **t-value** of a **correlation** in the formula sheet.*

R code 4.3e:

Answer 4.3e:

df : _____ t_{xy} : _____

4.3 f) Where do you find the **t-value** that you calculated in assignment 4.3e in the output of the **cor.test()** function from assignment 4.3c?

Answer 4.3f:

_____

4.3 g) What is the **p-value** for this hypothesis test? Interpret this **p-value** with respect to the confidence used and give a conclusion on the hypotheses. Include the following elements:

- ☐ Discuss what the **p-value** is for this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about ρ_{xy}.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 4.3g:

_____

_____

_____

_____

_____

_____

Assignment 4.4: Testing hypotheses using linear regression in R

A manager from a local supermarket wants to increase the sustainability of her store. Since the supermarket is located in a farm village that has easy access to milk, many milk cartons are not sold, expire, and have to be thrown away. From common sense, the manager suspects that decreasing the price of milk will result in more sales, and thus fewer milk cartons that get thrown away. In order to increase her sustainability through this method, the manager sets out to find out if a decrease in the price per carton of milk will result in fewer milk cartons thrown away on an average day. She also wants to predict her decrease in waste when she lowers the price of a milk carton by 30 cents.

The manager goes to the company headquarters and asks the prices of milk (they vary) in every branch of the supermarket chain in the Netherlands. She also asks the average number of milk cartons thrown away per day of each branch.

For this assignment, you need the data file **`nationalSupermarket.csv`**, which contains the full **`population`** of 200 stores of the supermarket chain in the Netherlands.

4.4 a) Use the **`read.csv()`** function (and **`setwd()`** function if you prefer) to import the data set into a data frame called **`dataset7`**.

R code 4.4a:

```
```

4.4 b) Create a scatter plot of the data. Put the price per milk carton (**`Price`**) on the *x-axis* and the average number of milk cartons that was thrown away (**`AvgWasted`**) on the *y-axis*.

R code 4.4b:

```
```

The **`lm()`** function is used to fit a linear model (linear regression) in R. It requires that you specify a **`formula`** that tells the function what the **`regression equation`** is. If you want to fit a **`model`** where you predict an outcome variable **`Y`** on the basis of one predictor variable **`X`**, the formula is as follows:

Formula on paper: $Y = \beta_0 + \beta_1 \times X$ **Formula in R:** **`Y ~ 1 + X`**

4.4 c) Write down the **`regression equation`** for a linear model where you predict the average number of thrown away milk cartons per day (**`AvgWasted`**) on the basis of the price per milk carton (**`Price`**).

Answer 4.4c:

AvgWasted = _____

A linear **model** with these variables can be fitted in R by calling the **lm()** function with the **formula** and the **dataset** (see the R code below). The variables **X** and **Y** should correspond to the names of the corresponding variables in your **dataset** .

```
lm(formula = Y ~ 1 + X, data = dataset)
```

4.4 d) Fit a linear **model** where you use the data in **dataset7** to predict the average number of thrown away milk cartons per day (**AvgWasted**) on the basis of the price per milk carton (**Price**). Store this **model** in an object called **lmfit** .

R code 4.4d:

Now run the following code in R:

```
summary(lmfit)
```

4.4 e) Use the summary to find out the β_0 and $_1$ parameters and write them down in the **regression equation** below.

Answer 4.4e:

AvgWasted = β_0 (_____) + β_1 (_____) × Price

4.4 f) Using the **abline()** function, paste the **regression equation** line into your scatter plot from assignment 4.4b.

R code 4.4f:

4.4 g) What is the R^2 of the **lmfit** regression **model** ? What is the interpretation of this value?

Answer 4.4g:

R^2 = _____

Interpretation:

_____

_____

From the regression line, the manager observes that there is indeed a positive **relationship** between the price of a carton of milk and the average number of milk cartons thrown away per day. To be sure she wants to test the hypothesis that, with 95% confidence, the price of milk is a good predictor of the average number of milk cartons that are thrown away per day, and that this **relationship** is truly positive.

4.4 h) Formulate the **null hypothesis** H_0 and the **alternative hypothesis** H_1 for the manager's test of the population coefficient β_1 for the price of a carton of milk.

Answer 4.4h:

H_0: _____ H_1: _____

4.4 i) Use the **summary** in R to draw the conclusion for the manager. Include the following four elements:

- ☐ Discuss what the **p-value** is for this test.
- ☐ Discuss whether H_0 is rejected or not.
- ☐ Describe what this tells us about β_1.
- ☐ Describe what type of error is relevant *(type-I or type-II)*.

Answer 4.4i:

Assignment 4.5: Making predictions using linear regression in R

The manager now wants to know what will happen to the number of milk cartons that she has to throw away when she lowers the price of a milk carton by 30 cents, from \$1.00 to \$0.70. Currently, the supermarket throws away 4 cartons of milk each day.

4.5 a) Create a new data frame that has only one column that includes the new value for the price of milk (the column has to be named exactly the same as in **dataset7**).

R code 4.5a:

```
```

The **predict()** function can be used to predict new data according to a **linear model** .

4.5 b) Use the **predict()** function to predict the number of milk cartons that the supermarket will have to throw away when the **Price** is \$0.70.

R code 4.5b:

```
```

Answer 4.5b:

```
Predicted value:  _____
```

4.5 c) Confirm this estimate by writing out the regression equation of the **linear model** from assignment 4.4e and filling in the new value of the price of milk.

Answer 4.5c:

```
Predicted value:  _____
```

The **predict()** function can also be used to construct a **confidence interval** for the predicted number of cartons thrown away by using **interval = 'prediction'** .

4.5 d) Create a 90% **confidence interval** for the predicted number of milk cartons thrown away.

R code 4.5d:

```
```

4.5 e) When the manager lowers her price from \$1.00 to \$0.70, will the supermarket throw away fewer cartons of milk? Incorporate the 90% **confidence interval** for the predicted value in your answer.

Answer 4.5e:

The supermarket **will** / **will not** throw away fewer cartons of milk.

Explanation: