Regression

Master Data Science
Summer 2023

Prof. Dr. Marlene Müller
marlene.mueller@bht-berlin.de

Exercises 1

Exercise 1

(a) Two (continuous) variables have a correlation of -0.4 and a covariance of -1.84. One of the variables has a variance of 4. Calculate the variance of the other variable.

(b) The transformation between Celsius ($^\circ C$) and Fahrenheit ($^\circ F$) degrees for temperatures is given by:

$$X_{^\circ F} = X_{^\circ C} \cdot 1.8 + 32$$

Assume that the average temperature in summer in Berlin is $25^\circ C$ where we have a standard deviation of $3^\circ C$.

How could you transform these values into $^\circ F$? Do also calculate the variances in both degree measures.

(a) we know,

$$r_{x.y} = \frac{s_{x.y}}{s_x . s_y}$$

$$\Rightarrow s_y = \frac{s_{xy}}{r_{xy} \cdot s_x}$$

$$\Rightarrow s_y^2 = \left(\frac{s_{xy}}{r_{xy}}\right)^2 \times \frac{1}{s_x^2}$$

$$\boxed{\therefore s_y^2 = 5.29}$$

here,

$r_{xy} = -0.4$

$s_{xy} = -1.84$

$s_x^2 = 4$

$\boxed{s_y^2 = ?}$

(b) here, $X_{^\circ F} = 1.8 \cdot X_{^\circ C} + 32$

$T_{^\circ C} = 25^\circ C$

$S_{t^\circ C} = 3^\circ C$

As the equation is a linear function.

So, $\bar{X}_{°F} = 1.8 \cdot \bar{X}_{°C} + 32$

$= 1.8 \times 25 + 32$

$= 77°F$

$X_{°F} = 1.8 \times S_{x°C}$

$\boxed{X_{°F} = 5.4°F}$

For the variance:

$S_x^2 = (3°C)^2$ or $(5.4°F)^2$

$= 9°C$ or $29.16\,F$

Exercise 2

Consider the observations of two variables (quite artificial data! ☺):

x_i	1	1	2	2	3	3
y_i	1	3	2	4	3	5

The following tasks should be solved without R:

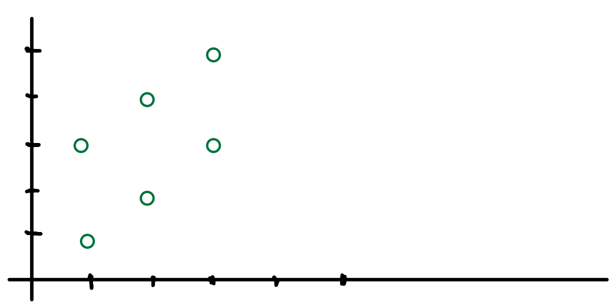
(a) Draw a scatterplot.

(b) Calculate the regression line.

(c) Check that the line goes through $(\bar{x}, \bar{y})$. (Is that always the case?)

(d) Calculate the correlation. How do you obtain the coefficient of determination?

a



ⓑ

x_i	y_i	$\bar{x}$	$\bar{y}$	$x_i-\bar{x}$	$y_i-\bar{y}$
1	1	2	3	-1	-2
1	3	2	3	-1	0
2	2	2	3	0	-1
2	4	2	3	0	1
3	3	2	3	1	0
3	5	2	3	1	2

$$\hat{b}_1 = \frac{cov(x,y)}{var(x)} =$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$= \left((-1)^2+(-1)^2+(0)^2+(0)^2+(1)^2+(1)^2\right)$$

$$= 1+1+0+0+1+1$$

$$= 4/5$$

$$S_y^2 = \frac{1}{n-1}\left((-2)^2+(0)^2+(-1)^2+(1)^2+(0)^2+(2)^2\right)$$

$$= 4+0+1+1+0+4$$

$$= 10/5 = 2$$

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\times(y_i-\bar{y})$$

$$= \left((-1)\cdot(-2) + (-1)\times 0 + 0\times(-1) + 0\cdot 1 + 1\cdot 0 + 1\cdot 2\right)$$

$$= 2+0+0+0+0+2 = 4/5$$

$$\hat{b}_1 = \frac{S_{xy}}{S_x^2}$$

$$= \frac{4}{4} = 1$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \cdot \bar{X}$$

$$= 3 - 1 \cdot 2$$

$$= 1$$

$\therefore \hat{b}_0, \hat{b}_1 = (1, 1)$

For regression line:

slope = 1 and interceptor = 1.

$$Y = X + 1$$

ⓒ $\bar{x} = 2$, $\bar{Y} = 3$

$\therefore$ if $x = 2$, $Y = 2 + 1 = 3$

$\therefore$ The regression line passes through $(2, 3)$.

ⓓ correlation, $r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$

$$= \frac{4/5}{\sqrt{4/5}\sqrt{10/5}}$$

$$\boxed{\therefore r_{xy} = 0.63}$$

coefficient of determination:

$$R^2 = 1 - \frac{RSS}{TSS}$$

where, TSS = total variance
ESS = Explained Variance

$$= 1 - \frac{TSS - ESS}{TSS}$$

$$= 1 - \frac{TSS}{TSS} + \frac{ESS}{TSS} = \frac{ESS}{TSS}$$

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

$\hat{y}_i$ is the predicted value calculated from the regression line

$$= \frac{\hat{S}_y^2}{S_y^2}$$

Exercise 3

Assume we have observations $x_1, \ldots, x_n$ of a variable X. Determine the value a, which minimizes the following criterion:

$$Q(a) = \sum_{i=1}^{n} (x_i - a)^2$$

$$\frac{d(Q(a))}{d(a)} = \frac{d}{da} \left(\sum_{i=1}^{n} (x_i^2 - 2ax_i + a^2) \right)$$

$$= \frac{d}{da} \left(\sum_{i=1}^{n} (x_i^2) - \sum_{i=1}^{n} 2ax_i + \sum_{i=1}^{n} a^2 \right)$$

$$= 0 - 2\sum_{i=1}^{n} x_i + n.2a$$

$$= 2an - 2\sum_{i=1}^{n} x_i$$

$$= 2an - 2n\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= 2an - 2n \cdot E(X)$$

To minimize the equation.

$$\frac{d(Q(a))}{da} = 0$$

$$\therefore 2an - 2n E(X) = 0$$

$$\Rightarrow a - E(X) = 0 \qquad \boxed{\therefore a = E(X).}$$

Exercise 4

Consider $X \sim N(2, 9)$. What does that mean?

Calculate the following probabilities (with and without R):

(a) $P(X \leq 0)$ (b) $P(X \leq -1)$

(c) $P(X \geq 5)$ (d) $P(-2 \leq X \leq 2)$

$X \sim N(2,9)$ means, X is a Random varieble which follow the normal distribution where the sum is 2 and variance of 9.

(a) $P(X \leq 0)$ $\qquad z = \dfrac{0-2}{3} = -0.6667.$

$\therefore P(X \leq 0) \approx P(z \leq -0.6667)$

Pdf of normal distribution $= \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$= \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot(z)^2}$

$\because z = \dfrac{x-\mu}{\sigma}$

here, μ = mean, σ = std. dev.

$= \dfrac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1(z)^2}{2}}$

$X \sim N(2, 9)$

$P(X \leq 0) = P\left(\dfrac{X-2}{3} \leq \dfrac{0-2}{3}\right) = \phi\left(-\frac{2}{3}\right)$

$= Q(-0.\bar{6})$

$= 0.251$

Exercise 5
Consider the following data for the speed versus braking distance (of a car):

Speed X (in km/h)	20	25	30	35	40	45	50	55	60	65	70
Braking distance Y (in m)	18	26	33	40	46	59	72	85	97	120	141

Here are some (possibly) useful values when calculating without R:

$$\overline{x} = 45, \quad \overline{y} = 67, \quad s_X^2 = 275, \quad s_Y^2 = 1600.6, \quad s_{XY} = 649.5$$

(a) Explain the meaning of: $\overline{x}$, $\overline{y}$, s_X^2, s_Y^2 and s_{XY}.

(b) What are the R functions to calculate them? (Check with R!)

(c) Display the data in R using a scatterplot.

(d) Calculate the correlation and check with R. What could we conclude from this value?

(e) Calculate the linear regression coefficients (first without R). Do the same using R and display the line in the scatterplot.

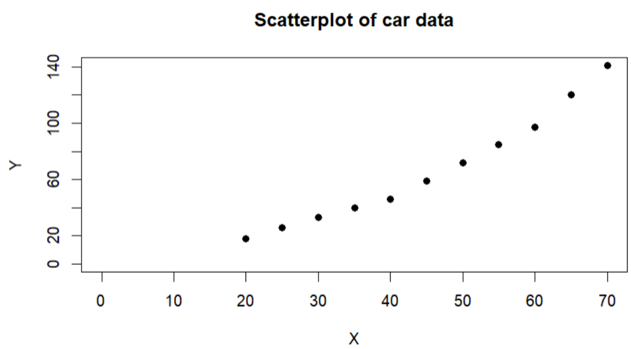
(a) $\overline{x}$ is the mean of speed X.
$\overline{y}$ is the mean of braking distance Y.
s_x^2 is the variance of speed X.
s_y^2 is the variance of
s_{xy} is the covariance of speed X and braking distance Y.

(b)

```
# Solution (b)

X <- c(20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70)
Y <- c(18, 26, 33, 40, 46, 59, 72, 85, 97, 120, 141)

X_mean <- mean(X); X_mean    # 45

Y_mean <- mean(Y); Y_mean    # 67

Sx2 <- var(X);    Sx2        # 275

Sy2 <- var(Y);    Sy2        # 1600.6

Sxy <- cov(X,Y); Sxy         # 649.5
```

(c)

```r
# Solution (c)
plot(X, Y, xlim = c(0, max(X)), ylim = c(0, max(Y)),
     pch=19, main = 'Scatterplot of car data')
```



(d) $r_{xy} = \dfrac{S_{xy}}{S_x \cdot S_y} = \dfrac{649.5}{\sqrt{275} \cdot \sqrt{1600.6}} = 0.9789$

```r
# Solution (d)
cor(X, Y)    # 0.9789746
```

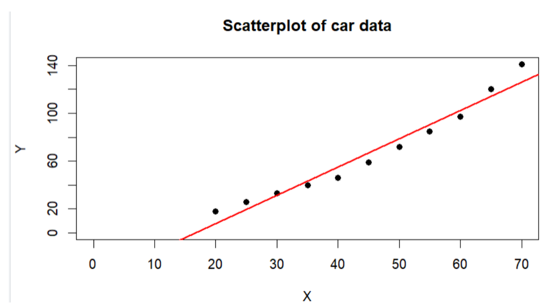
Conclusion: Variable X and Y are highly correlated.

(e) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_x^2} = \dfrac{649.5}{275} = 2.3618$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 67 - (45 \times 2.3618) = -39.28$

```r
# Solution (e)
lreg_model <- lm(Y ~ X)

summary(lreg_model)

abline(lreg_model, col = 'red', lwd = 2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -39.282      7.824  -5.021 0.000719 ***
X              2.362      0.164  14.398 1.61e-07 ***
```



Exercise 6

On Moodle you find the file `MunichRent2003.csv`. This is a sample of 2053 appartments in Munich from 2003 (Munich was already an expensive city at this time ...). The variables are coded as follows:

Variable	Meaning	Values
`netrent`	net rent (per month)	in Euro
`netrent.sqm`	net rent per square metre	in Euro
`living.space`	living space	in square metres
`rooms`	no. of rooms	1,...,6
`year`	year of construction	year
`district`	Munich district	1,...,25
`location`	quality of location	best, good, simple
`warm.water`	warm water provided	yes, no
`central.heating`	central heating	yes, no
`bath.tiles`	bath room with tiles	yes, no
`bath.extras`	bath room with extras	yes, no
`upscale.kitchen`	kitchen with upscale equipment	yes, no

Original source: https://doi.org/10.5282/ubm/data.2

(a) For which pairs of variables would it be useful to estimate a simple linear regression model? (Choose at least two different examples. Explain which of the variables do you consider the dependent and the independent one.)

(b) Load the data into R. If you don't know what to do, check `?read.csv`.
(Extra task: Try also to load the original data into R!)

(c) Estimate the model that you have chosen in (a), i.e. calculate the coefficients, draw scatterplots and regression lines, determine R^2.

(a) netrent and living.space where living.space is independent variable.

netrent and year also may be usefull for regression estimation, where year would be independent variable.

(b)

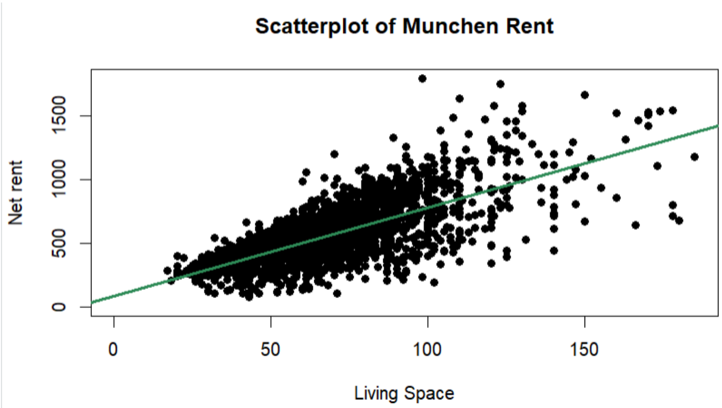
```
# Solution (b)

data <- read.csv('Dataset/MunichRent2003.csv')

# Snapshot of the data
str(data)
```

```
'data.frame':   2053 obs. of  12 variables:
 $ netrent        : num  741 716 528 554 698 ...
 $ netrent.sqm    : num  10.9 11.01 8.38 8.52 6.98 ...
 $ living.space   : int  68 65 63 65 100 81 55 79 52 77 ...
 $ rooms          : int  2 2 3 3 4 4 2 3 1 3 ...
 $ year           : num  1918 1995 1918 1983 1995 ...
 $ district       : int  2 2 2 16 16 16 6 6 6 6 ...
 $ location       : chr  "good" "good" "good" "simple" ...
 $ warm.water     : chr  "yes" "yes" "yes" "yes" ...
 $ central.heating: chr  "yes" "yes" "yes" "yes" ...
 $ bath.tiles     : chr  "yes" "yes" "yes" "yes" ...
 $ bath.extras    : chr  "no" "no" "no" "yes" ...
 $ upscale.kitchen: chr  "no" "no" "no" "no" ...
```

c

Regression Analysis of living space and net rent.

```
# Solution (c)

X <- living.space
Y <- netrent

# variance
Sx2 <- var(X)
Sy2 <- var(Y)

# Covariance
Sxy <- cov(X, Y)

# Regression Co-efficient
b1 <- Sxy / Sx2
b0 <- mean(Y) - (b1 * mean(X))

cat('b0= ', b0, ' | b1= ', b1)
# b0=  89.84691  | b1=  6.90056

# Plotting

plot(X, Y, xlim = c(0, max(X)), ylim = c(0, max(Y)),
     pch=19, main = 'Scatterplot of Munchen Rent',
     xlab = 'Living Space', ylab = 'Net rent')

abline(b0, b1,  col = 'seagreen', lwd = 3)
```

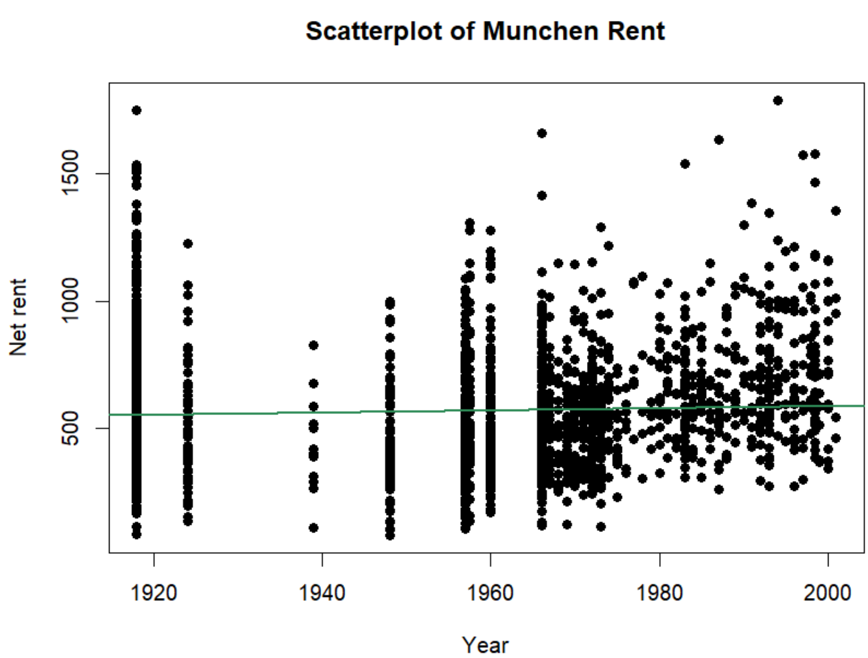



```
# R2 Score calculation
Y_hat <- b0 + (b1 * X)

Sy_hat2 <- var(Y_hat)

r2 <- Sy_hat2 / Sy2
r2 # 0.5005034
```

Use regression with netrent and year.

```
# Regression analysis of net rent and year

ll_model <- lm(netrent ~ year)
summary(ll_model)

plot(year, netrent, pch=19, main = 'Scatterplot of Munchen Rent',
     xlab = 'Year', ylab = 'Net rent')

abline(ll_model, col = 'seagreen', lwd = 2)
```

```
Call:
lm(formula = netrent ~ year)

Residuals:
    Min      1Q  Median      3Q     Max 
-488.15 -179.30  -36.78  127.38 1202.74 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -338.8859   426.0683  -0.795    0.426  
year           0.4642     0.2176   2.134    0.033 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 245.2 on 2051 degrees of freedom
Multiple R-squared:  0.002215,	Adjusted R-squared:  0.001728 
F-statistic: 4.552 on 1 and 2051 DF,  p-value: 0.033
```



Comment: Too low R2-score, which observed in the plot.