Multiple linear regression

STATISTICAL DATA ANALYSIS

MANUELLA LECH CANTUARIA
VICTORIA BLANES-VIDAL
The Maersk Mc-Kinney Møller Institute
Applied AI and Data Science

1

1

LECTURE PLANNING

Lesson	Week	Date	TOPICS	Teacher
1	35	1/Sep	Introduction to the course Descriptive statistics – Part I	MLC
2	36	8/sep	Descriptive statistics – Part II	MLC
3	37	15/Sep	Probability distributions	MLC
4	38	22/Sep	Hypothesis testing (one sample)	VBV
5	39	29/Sep	Hypothesis testing (two samples)	VBV
6	40	6/Oct	ANOVA one-way	VBV
7	41	13/Oct	R class (Introduction to R and descriptive statistics) Point-giving activity (in class) - AT 13h10 in U45	MLC
-	42	20/Oct	NO CLASS (Autum holidays)	
8	43	27/Oct	R class (hypothesis testing + ANOVA)	MLC
9	44	3/Nov	ANOVA two-way	VBV
-	45	10/Nov	NO CLASS	
10	46	17/Nov	Regression analysis	VBV
11	47	24/Nov	Notions of experimental design and questions Point-giving activity (in class)	VBV+MLC
12	48	1/Dec	Multiple regression	MLC

Not using any software (lessons 1–6)

R is used for the analyses (lessons 7–12)

2

2

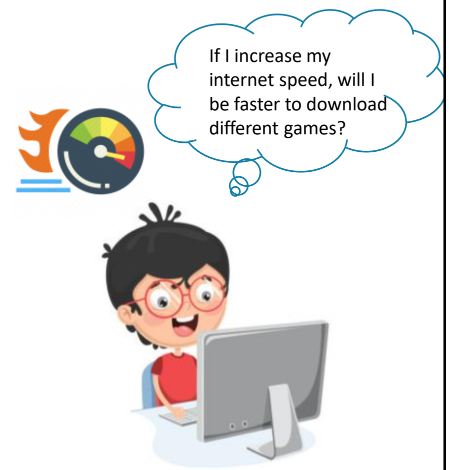
Content:

- What is a multiple regression?
- Simple linear regression vs. multiple linear regression
- Types of variables
- Multiple regression in R
- Prediction
- Questions related to the course/exam

Content:

- **What is a multiple regression?**
- Simple linear regression vs. multiple linear regression
- Types of variables
- Multiple regression in R
- Prediction

Regression analysis



Regression analysis: Statistical method to analyze the relationship between two or more variables.

- **Correlation** is a statistical method used to determine whether a relationship between two variables exists.
- **Regression** is a statistical method used to describe the nature of the relationship between variables (e.g. positive or negative, linear or nonlinear).

In regression, one variable is considered independent (=input) variable (X) and the other the dependent (=output) variable (Y).

Internet speed × Time to download games

Independent variable (X) Dependent variable (Y)

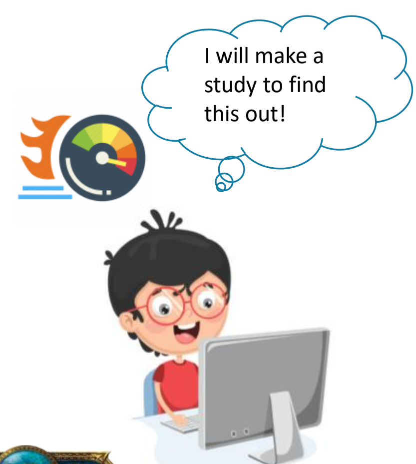
5

5

Simple linear regression

He therefore collects data from different school classmates on their internet speed and the time it took them to download the game World of Warcraft

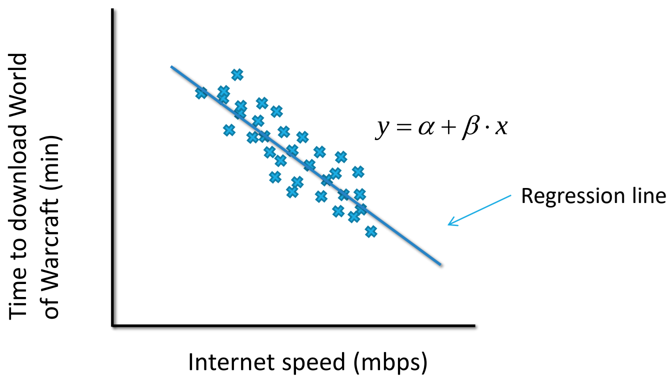
Person	Internet speed (mbps)	Time to download World of Warcraft (min)
1	1000	90.4
2	100	825.1
3	300	219.8
4	1000	85.4
5	500	150.2
6	2000	75.4
7	1200	100.1
8	600	145.1

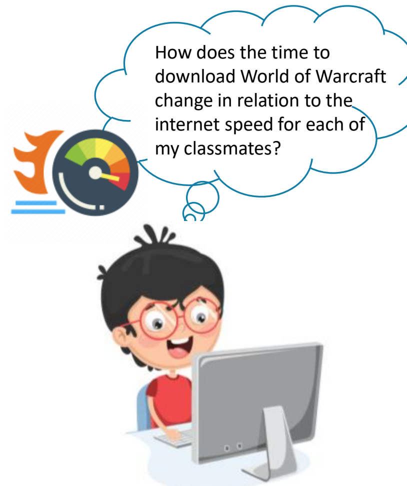


6

6

Simple linear regression

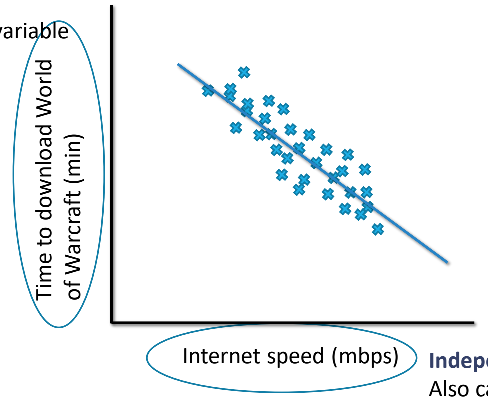




7

Simple linear regression

Dependent variables
Also called response variable

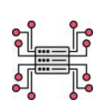


Independent variable
Also called explanatory variable or predictor variable

In a simple regression analysis, there is one independent variable and one dependent variable.

Other factors are not taken into account.

What should we do if we want to consider other factors in our model, for example:



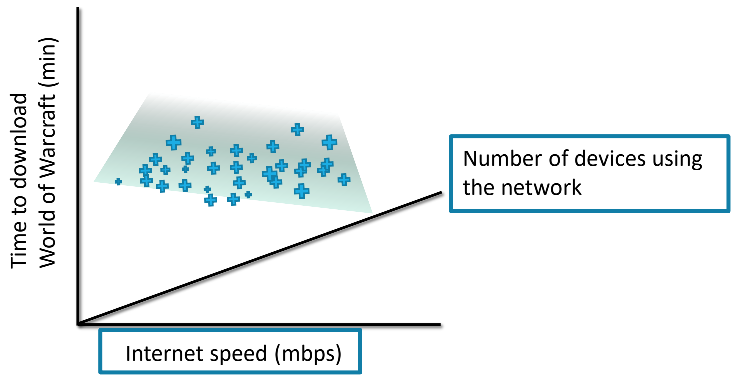
Network congestion (e.g. how many devices are using the same network?)

Computer multitasking (e.g. are you using other softwares on your computer?)

MULTIPLE LINEAR REGRESSION

8

Multiple linear regression



In **multilple linear regression**, there is not only one independent variable. There will be two or more independent variables.

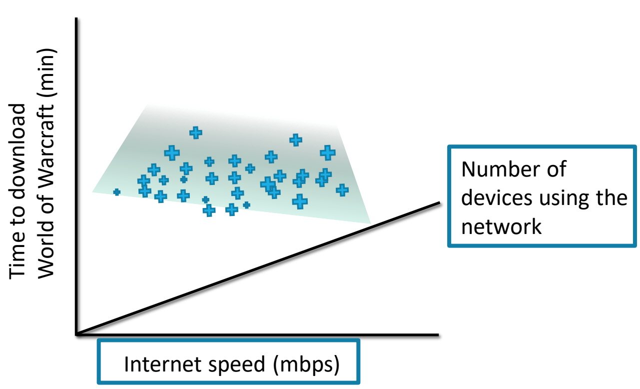
Instead of just modelling **download time** by **internet speed**, now we are modeling **download time** using **internet speed** and **network congestion (i.e. number of devices using the network at the same time)**.

It is also possible to add other factors, such as computer multitasking, server distance, etc.

9

9

Multiple linear regression



Multiple linear regression is used to estimate the relationship between **two or more independent variables** and **one dependent variable**.

You can use multiple linear regression when you want to know:

1.How strong the relationship is between two or more independent variables and one dependent variable

2.The value of the dependent variable at a certain value of the independent variables (e.g. what time will it take for me to download World of Warcraft, if I have an internet speed on e.g. 300 mbps and no one else is using my network?). **PREDICTION**

10

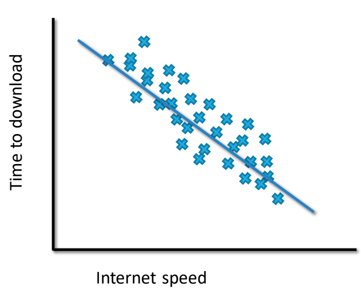
10

Content:

- What is a multiple regression?
- **Simple linear regression vs. multiple linear regression**
- Types of variables
- Multiple regression in R
- Prediction

Simple vs. multiple linear regression

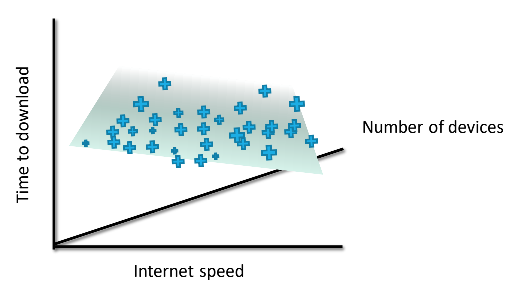
Simple linear regression



$$y = \alpha + \beta \cdot x$$

α is the intercept and β is the coefficient associated with internet speed

Multiple linear regression

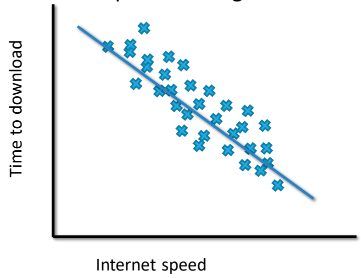


$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_n \cdot x_n$$

α is the intercept, β1 is the coefficient associated with the internet speed, and β2 is the coefficient associated with the number of devices

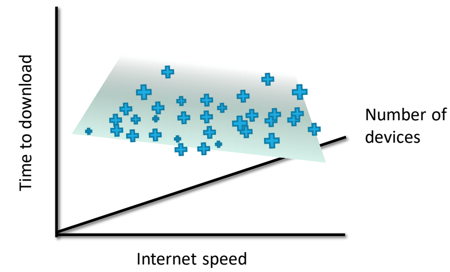
Simple vs. multiple linear regression

Simple linear regression



Coefficient of determination (R^2) is calculated the same way both for simple and multiple linear regression

Multiple linear regression



$$R^2 = \frac{Explained\ variation}{Total\ variation}$$

For example, a R^2=0.81 means that 81% of the variation in the dependent variable (y) is accounted for by the variations in the independent variable (x).

Every dependent variable included in the model will increase R^2.
A model with more terms may seem to have a better fit just for the fact that it has more terms.

$$R^2 = \frac{Explained\ variation}{Total\ variation}$$

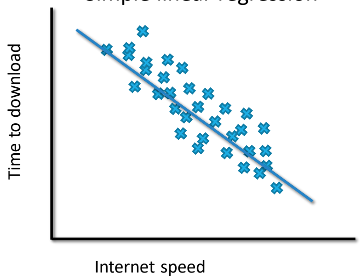
For example, a R^2=0.81 means that 81% of the variation in the dependent variable (y) is accounted for by the variations in the independent variables (x1, x2, x3, ….)

13

13

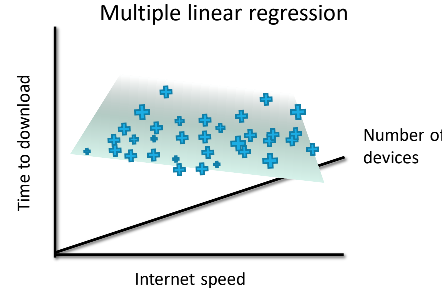
Simple vs. multiple linear regression

Simple linear regression



For multiple regression, R^2 is **adjusted** to compensate for the additional parameters in the equation

Multiple linear regression



$$R^2 = \frac{Explained\ variation}{Total\ variation}$$

Correction proposed by Ezekiel: $\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

p = number of independent variables, n = sample size

The **adjusted R^2** compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability.
With the **adjusted R^2** we can evaluate whether it is worth or not to collect more data.

14

14

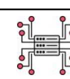
Content:

- What is a multiple regression?
- Simple linear regression vs. multiple linear regression
- **Types of variables**
- Multiple regression in R
- Prediction

Types of variables

In the previous example, we wanted to see how **body weight (kg)** changes in accordance to:

Internet speed (mbps) 

Number of devices using the same network (e.g., 1, 2, 3, 4) 

What do these variables have in common?

They are all continuous quantitative variables.

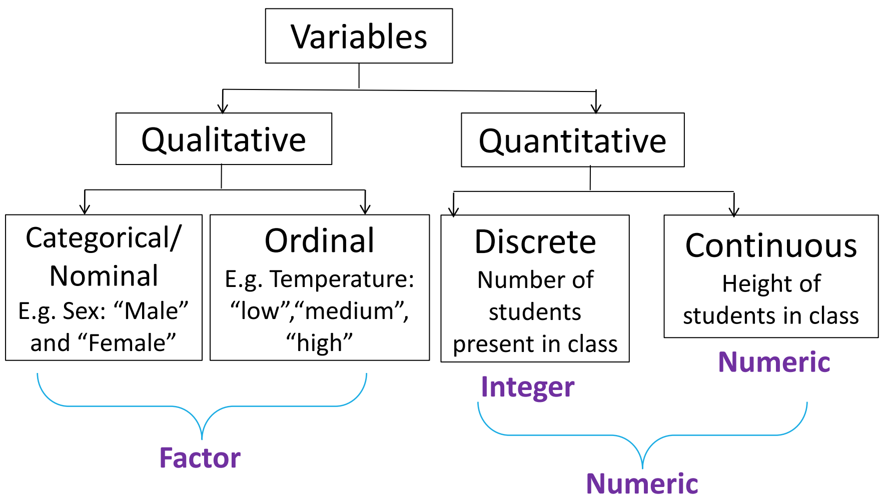
What if we have other types of variables, for example:

Internet type (Cable; Fiber) 

Multiple linear regression covers all type of independent variables, so they can all be included in the model.

We just need to make sure this is correctly done!

Types of variables – in R



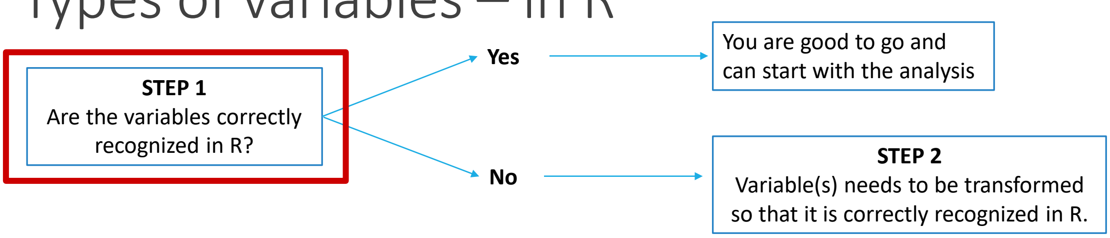
In R
There are also variables which are recognized as **CHARACTER**

A character vector is a vector consisting of characters.

17

17

Types of variables – in R



Person	Internet_speed	Time_downl	N_devices	Internet_type
1	1000	90.4	3	Cable
2	100	825.1	4	Cable
3	300	219.8	1	Fiber
4	1000	85.4	1	Cable
5	500	150.2	2	Fiber
6	2000	75.4	3	Cable
7	1200	100.1	5	Cable
8	600	145.1	2	Fiber

18

18

Types of variables – in R

STEP 1
Are the variables correctly recognized in R?

Yes → You are good to go and can start with the analysis

No → **STEP 2**
Variable(s) needs to be transformed so that it is correctly recognized in R.

The function ***str()*** can be used to see how the variables are recognized

```
> str(df)
tibble [8 × 5] (S3: tbl_df/tbl/data.frame)
 $ Person        : num [1:8] 1 2 3 4 5 6 7 8
 $ Internet_speed: num [1:8] 1000 100 300 1000 500 2000 1200 600
 $ Time_downl    : num [1:8] 90.4 825.1 219.8 85.4 150.2 ...
 $ N_devices     : num [1:8] 3 4 1 1 2 3 5 2
 $ Internet_type : chr [1:8] "Cable" "Cable" "Fiber" "Cable" ...
```

19

19

Types of variables – in R

STEP 1
Are the variables correctly recognized in R?

Yes → You are good to go and can start with the analysis

No → **STEP 2**
Variable(s) needs to be transformed so that it is correctly recognized in R.

As you learned in class 8, you can "transform" the variable, so it is correctly recognized by the software:

df$Internet_type <- as.factor(df$Internet_type)

```
> str(df)
tibble [8 × 5] (S3: tbl_df/tbl/data.frame)
 $ Person        : num [1:8] 1 2 3 4 5 6 7 8
 $ Internet_speed: num [1:8] 1000 100 300 1000 500 2000 1200 600
 $ Time_downl    : num [1:8] 90.4 825.1 219.8 85.4 150.2 ...
 $ N_devices     : num [1:8] 3 4 1 1 2 3 5 2
 $ Internet_type : Factor w/ 2 levels "Cable","Fiber": 1 1 2 1 2 1 1 2
```

In R:
as.factor()
as.integer()
as.numeric()
as.character()

20

20

Types of variables – in R

Example:

Person	Internet_speed	Time_downl	N_devices	Internet_type
1	1000	90.4	3	1
2	100	825.1	4	1
3	300	219.8	1	2
4	1000	85.4	1	1
5	500	150.2	2	2
6	2000	75.4	3	1
7	1200	100.1	5	1
8	600	145.1	2	2

Attention: Categorical variables can also be coded as numbers (e.g. cable = 1 and fiber = 2).
In this case, the same transformation procedure needs to be done.

```
> str(df2)
tibble [8 × 5] (S3: tbl_df/tbl/data.frame)
 $ Person        : num [1:8] 1 2 3 4 5 6 7 8
 $ Internet_speed: num [1:8] 1000 100 300 1000 500 2000 1200 600
 $ Time_downl    : num [1:8] 90.4 825.1 219.8 85.4 150.2 ...
 $ N_devices     : num [1:8] 3 4 1 1 2 3 5 2
 $ Internet_type : num [1:8] 1 1 2 1 2 1 1 2
```

df2$Internet_type <- as.factor(df2$Internet_type)

```
> str(df2)
tibble [8 × 5] (S3: tbl_df/tbl/data.frame)
 $ Person        : num [1:8] 1 2 3 4 5 6 7 8
 $ Internet_speed: num [1:8] 1000 100 300 1000 500 2000 1200 600
 $ Time_downl    : num [1:8] 90.4 825.1 219.8 85.4 150.2 ...
 $ N_devices     : num [1:8] 3 4 1 1 2 3 5 2
 $ Internet_type : Factor w/ 2 levels "1","2": 1 1 2 1 2 1 1 2
```


In-class exercise

A biologist wants to determine whether the oxygen level (measure in mg/L) and type of soil (“loam soil”, “sandy soil”, or “clay soil”) stimulate the plant growth (measured in cm). When building a regression model, what will be the role and type of each variable?

a) Oxygen level = independent continuous variable; Type of soil = independent nominal variable; Plant growth = dependent continuous variable.
b) Oxygen level = dependent continuous variable; Type of soil = dependent nominal variable; Plant growth = independent continuous variable.
c) Oxygen level = dependent discrete variable; Type of soil = independent ordinal variable; Plant growth = independent continuous variable.
d) Oxygen level = independent discrete variable; Type of soil = dependent nominal variable; Plant growth = dependent discrete variable.

Content:

- What is a multiple regression?
- Simple linear regression vs. multiple linear regression
- Types of variables
- **Multiple regression in R**
- Prediction

23

23

Example

We want to investigate different factors that may increase the systolic blood pressure in patients. We therefore measure the blood pressure (in mmHg) of 15 patients and collect data on their BMI (in kg/m^2), age, and whether they are male or female.

ID	Blood.pressure	BMI	Age	Sex
1	127.3	28.2	36	1
2	121.2	17	39	1
3	154.3	32.2	87	2
4	95.7	24.1	45	1
5	152.4	30.4	81	2
6	144.3	27.2	52	2
7	111.9	22.9	49	1
8	99.8	21.4	56	1
9	167.3	32.5	68	2
10	141.5	25	48	2
11	111.2	18.7	41	1
12	132.4	29.1	67	1
13	161.4	34.5	79	2
14	129.8	26.4	59	1
15	128.7	22.5	48	2

Sex:
1 = Female
2 = Male

24

24

Multiple regression in R

We want to investigate different factors that may increase the systolic blood pressure in patients. We therefore measure the blood pressure (in mmHg) of 15 patients and collect data on their BMI (in kg/m^2), age, and whether they are male or female.

ID	Blood.pressure	BMI	Age	Sex
1	127.3	28.2	36	1
2	121.2	17	39	1
3	154.3	32.2	87	2
4	95.7	24.1	45	1
5	152.4	30.4	81	2
6	144.3	27.2	52	2
7	111.9			
8	99.8			
9	167.3			
10	141.5			
11	111.2			
12	132.4			
13	161.4	34.5	79	2
14	129.8	26.4	59	1
15	128.7	22.5	48	2

Sex:
1 = Female
2 = Male

What is the dependent variable?

What are the independent variables?

Are the variables correctly recognized by in R?

```
> str(data_blood)
tibble [15 x 5] (S3: tbl_df/tbl/data.frame)
 $ ID            : num [1:15] 1 2 3 4 5 6 7 8 9 10 ...
 $ Blood.pressure: num [1:15] 127.3 121.2 154.3 95.7 152.4 ...
 $ BMI           : num [1:15] 28.2 17 32.2 24.1 30.4 27.2 22.9 21.4 32.5 25 ...
 $ Age           : num [1:15] 36 39 87 45 81 52 49 56 68 48 ...
 $ Sex           : num [1:15] 1 1 2 1 2 2 1 1 2 2 ...
```

25

25

Multiple regression in R

```
> str(data_blood)
tibble [15 x 5] (S3: tbl_df/tbl/data.frame)
 $ ID            : num [1:15] 1 2 3 4 5 6 7 8 9 10 ...
 $ Blood.pressure: num [1:15] 127.3 121.2 154.3 95.7 152.4 ...
 $ BMI           : num [1:15] 28.2 17 32.2 24.1 30.4 27.2 22.9 21.4 32.5 25 ...
 $ Age           : num [1:15] 36 39 87 45 81 52 49 56 68 48 ...
 $ Sex           : num [1:15] 1 1 2 1 2 2 1 1 2 2 ...
```

data_blood$Sex <- as.factor(data_blood$Sex)

```
> str(data_blood)
tibble [15 x 5] (S3: tbl_df/tbl/data.frame)
 $ ID            : num [1:15] 1 2 3 4 5 6 7 8 9 10 ...
 $ Blood.pressure: num [1:15] 127.3 121.2 154.3 95.7 152.4 ...
 $ BMI           : num [1:15] 28.2 17 32.2 24.1 30.4 27.2 22.9 21.4 32.5 25 ...
 $ Age           : num [1:15] 36 39 87 45 81 52 49 56 68 48 ...
 $ Sex           : Factor w/ 2 levels "1","2": 1 1 2 1 2 2 1 1 2 2 ...
```

26

26

Multiple regression in R

Using only BMI and age, build and interpret a multiple linear regression model

model1 <- lm(Blood.pressure ~ BMI + Age, data = data_blood)
summary(model1)

```
Call:
lm(formula = Blood.pressure ~ BMI + Age, data = data_blood)

Residuals:
    Min      1Q  Median      3Q     Max
-27.878  -7.548   1.019   9.718  19.035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.5811    19.6060   2.325   0.0384 *
BMI           2.8338     1.1694   2.423   0.0321 *
Age           0.2156     0.3748   0.575   0.5757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.06 on 12 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.5815
F-statistic: 10.73 on 2 and 12 DF,  p-value: 0.00213
```

t-statistics tests the null hypothesis, whether the beta coefficient of the predictor is not significantly different from zero

There is a significant relationship between BMI and blood pressure

The relationship between age and blood pressure is not statistically significant

At least, one of the predictor variables is significantly related to the outcome variable

27

Multiple regression in R

Using only BMI and age, build and interpret a multiple linear regression model

model1 <- lm(Blood.pressure ~ BMI + Age, data = data_blood)
summary(model1)

```
Call:
lm(formula = Blood.pressure ~ BMI + Age, data = data_blood)

Residuals:
    Min      1Q  Median      3Q     Max
-27.878  -7.548   1.019   9.718  19.035

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.5811    19.6060   2.325   0.0384 *
BMI           2.8338     1.1694   2.423   0.0321 *
Age           0.2156     0.3748   0.575   0.5757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.06 on 12 degrees of freedom
Multiple R-squared:  0.6413,    Adjusted R-squared:  0.5815
F-statistic: 10.73 on 2 and 12 DF,  p-value: 0.00213
```

$$y = \alpha + \beta_1 \cdot x_1 + \beta_2 \cdot x_2$$

Blood pressure $= 45.58 + 2.83 \cdot BMI + 0.22 \cdot Age$

The coefficient (β) can be interpreted as the average effect on y of a one unit increase in predictor, holding all other predictors fixed

For example, for a fixed age, increasing the BMI in one unit will result in an increase of 2.83 mmHg in the blood pressure (on average)

28

Multiple regression in R

Using all independent variables, build and interpret a multiple linear regression model

model2 <- lm(Blood.pressure ~ BMI + Age + Sex, data = data_blood)
summary(model2)

```
Call:
lm(formula = Blood.pressure ~ BMI + Age + Sex, data = data_blood)

Residuals:
     Min       1Q   Median       3Q      Max
-21.6172  -2.7685   0.2052   4.6232  18.5268

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.87921   16.28453   4.107  0.00174 **
BMI          2.03747    0.91944   2.216  0.04871 *
Age          0.02967    0.28944   0.103  0.92020
Sex2        21.67923    6.85944   3.160  0.00907 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.63 on 11 degrees of freedom
Multiple R-squared:  0.812,	Adjusted R-squared:  0.7607
F-statistic: 15.84 on 3 and 11 DF,  p-value: 0.0002634
```

There is a significant relationship between BMI and blood pressure

The relationship between age and blood pressure is not statistically significant

There is a significant relationship between BMI and sex

29

Multiple regression in R

We found age is not significant for the model. This means that, for a fixed BMI and sex, changes in age will not significantly affect people's blood pressure. Therefore, one can argue we can remove age from the model

model3 <-lm(Blood.pressure ~ BMI + Sex, data = data_blood)
summary(model3)

```
Call:
lm(formula = Blood.pressure ~ BMI + Sex, data = data_blood)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7759  -2.5372   0.2463   4.7444  18.6447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.8299    15.5919   4.286  0.00106 **
BMI           2.1015     0.6462   3.252  0.00693 **
Sex2         21.8222     6.4333   3.392  0.00535 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.18 on 12 degrees of freedom
Multiple R-squared:  0.8118,	Adjusted R-squared:  0.7805
F-statistic: 25.89 on 2 and 12 DF,  p-value: 4.439e-05
```

Now we only have predictors which are statistically significant

30

Multiple regression in R

Sex:
1 = Female
2 = Male

Reference category: female

$$Blood\ pressure = 66.83 + 2.10 \cdot BMI + 21.82 \cdot Sex(Male)$$

```
Call:
lm(formula = Blood.pressure ~ BMI + Sex, data = data_blood)

Residuals:
     Min       1Q   Median       3Q      Max
-21.7759  -2.5372   0.2463   4.7444  18.6447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.8299    15.5919   4.286  0.00106 **
BMI           2.1015     0.6462   3.252  0.00693 **
Sex2         21.8222     6.4333   3.392  0.00535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.18 on 12 degrees of freedom
Multiple R-squared:  0.8118,	Adjusted R-squared:  0.7805
F-statistic: 25.89 on 2 and 12 DF,  p-value: 4.439e-05
```

This means that, the blood pressure in men are estimated to be 21.82 mmHg higher than the blood pressure in women

This means that, for a male with BMI = 25, his predicted blood pressure will be:

$$BP = 66.83 + 2.10 \cdot BMI + 21.82 \cdot Sex(Male)$$
$$BP = 66.83 + 2.10 \cdot 25 + 21.82 \cdot 1$$
$$BP = 141.15$$

For a female with BMI = 25, her predicted blood pressure will be:

$$BP = 66.83 + 2.10 \cdot BMI + \overbrace{21.82 \cdot Sex(Male)}^{\textbf{Zero}}$$
$$BP = 66.83 + 2.10 \cdot 25$$
$$BP = 119.33$$

31

Coefficient of determination (R^2)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.87921   16.28453   4.107  0.00174 **
BMI          2.03747    0.91944   2.216  0.04871 *
Age          0.02967    0.28944   0.103  0.92020
Sex2        21.67923    6.85944   3.160  0.00907 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.63 on 11 degrees of freedom
Multiple R-squared:  0.812,	Adjusted R-squared:  0.7607
F-statistic: 15.84 on 3 and 11 DF,  p-value: 0.0002634
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  66.8299    15.5919   4.286  0.00106 **
BMI           2.1015     0.6462   3.252  0.00693 **
Sex2         21.8222     6.4333   3.392  0.00535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.18 on 12 degrees of freedom
Multiple R-squared:  0.8118,	Adjusted R-squared:  0.7805
F-statistic: 25.89 on 2 and 12 DF,  p-value: 4.439e-05
```

Multiple R^2: explains how much of the variation in the dependent variable (y) is accounted for by the variations in the independent variables (x1, x2, x3, ….)

$$R^2 = \frac{Explained\ variation}{Total\ variation}$$

The values are basically the same for both models.

Adjusted R^2: compensates for the addition of new variables. $\bar{R}^2 = 1 - (1 - R^2)\dfrac{n-1}{n-p-1}$

As Adjusted R^2 is higher for the simpler model, some people may decide to remove age from the model

For the exam, you do not need to worry to make these decisions yourself!

32

Content:

- What is a multiple regression?
- Simple linear regression vs. multiple linear regression
- Types of variables
- Multiple regression in R
- **Prediction**

Predicting outcomes

The same way as the class for simple linear regression, the regression line can be used for prediction.

The better the fit of the model, the better the prediction will be.

Predicting outcomes

Sex:
1 = Female
2 = Male

What will be the expected blood pressure of a man with a BMI of 25?

$$Blood\ pressure = 66.83 + 2.10 \cdot \underbrace{BMI}_{25} + 21.82 \cdot \underbrace{Sex(Male)}_{1} \qquad BP = 141.15$$

model3 <-lm(Blood.pressure ~ BMI + Sex, data = data_blood)
predict(model3, data.frame(BMI =25, Sex = "2"))

```
> predict(model3, data.frame(BMI =25, Sex = "2")
```

> 141.1894

35

In-class exercise

A group of marine biologists have developed a model to predict the weight of fish individuals, based on the fish species and their measurements. They have therefore built a model based on the following parameters collected from 159 different fish:

- Species: Name of the fish species (Bream, Parkki, Perch, Pike, Roach, Smelt, and Whitefish)
- Weight: Weight of individual fish in grams
- Length 1: Longitudinal length in cm
- Length 2: Diagonal length in cm
- Length 3: Cross length in cm
- Height: Height in cm
- Width: Width in cm

36

36

In-class exercise

They started the process by including all predictors in the model. The following model output was obtained:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      -918.3321   127.0831  -7.226   2.5e-11 ***
SpeciesParkki     164.7227    75.6995   2.176 0.031152 *  
SpeciesPerch      137.9489   120.3135   1.147 0.253419    
SpeciesPike      -208.4294   135.3064  -1.540 0.125607    
SpeciesRoach      103.0400    91.3084   1.128 0.260954    
SpeciesSmelt      446.0733   119.4303   3.735 0.000268 ***
SpeciesWhitefish   93.8742    96.6580   0.971 0.333045    
Length1           -80.3030    36.2785  -2.214 0.028403 *  
Length2            79.8886    45.7180   1.747 0.082653 .  
Length3            32.5354    29.3002   1.110 0.268633    
Height              5.2510    13.0560   0.402 0.688128    
Width              -0.5154    23.9130  -0.022 0.982832    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.83 on 147 degrees of freedom
Multiple R-squared:  0.9361,	Adjusted R-squared:  0.9313 
F-statistic: 195.7 on 11 and 147 DF,  p-value: < 2.2e-16
```


In-class exercise

Classify the following statements as TRUE or FALSE:

STATEMENT 1: All fish that are Perch, Pike, Roach and Whitefish should be removed from the analysis.

STATEMENT 2: The weight of a Pike fish is estimated to be 208 grams lower than a Perch fish (considering all other variables are fixed).

STATEMENT 3: Increasing the longitudinal length of a fish in 1 cm will result in a decrease of 80.3 grams in the fish’s weight (considering all other variables are fixed).

Content:

- What is a multiple regression?
- Simple linear regression vs. multiple linear regression
- Types of variables
- Multiple regression in R
- Prediction
- **Questions related to the course/exam**

39

39

Exam

- Date of exam: 9th of January.
- Multiple choice exam in ItsLearning.
- Questions will involve concepts' understanding and calculations for problem solving and interpretation of findings.
- It will cover the entire course content.
- 120 minutes.
- Around 20 questions summing 100 points in total.
- You can find in ItsLearning a detailed file with information on what you can (or cannot) bring.

40

40

Opening the data in R

- In class #7, you learned different ways to upload your data in R

Two options

Creating vectors
+
Creating a dataframe

Importing a dataset

Using clickable actions in the environment tab

Using the appropriate lines of code

Opening the data in R

Creating vectors
+
Creating a dataframe

- You can create vectors for each variable and later use the function data.frame()

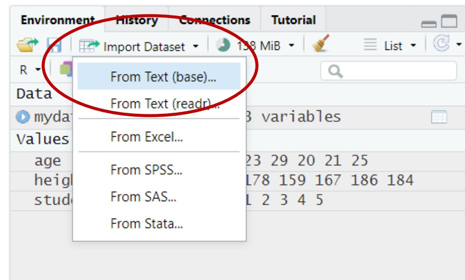
```
# Creating vectors:
student <- c(1, 2, 3, 4, 5)
age <- c(23, 29, 20, 21, 25)
height <- c(178, 159, 167, 186, 184)

#Creating a dataframe
mydata <- data.frame(student, age, height)
```

OR

- We will give you the precise commands to import the data in R (as we did in the last point-giving activity)

Opening the data in R



Obs: Here you may need to select heading = yes

Importing a Txt data file

Using clickable actions in the environment tab

Using the appropriate lines of code

```
> Nicolas_cage_drowning <-
read.table("C:/Users/vbv/Desktop/My_documents/Teaching/Teaching/Statistical_Data_Analysis/2022_2023/Exercises_in_R/Nicolas_cage.txt", header=TRUE)
```

43

43

Questions?

44