

## 1 Literature

## 2 Why regression analysis?

2.1 It is a medicine against alternative facts

2.2 Simpson's paradox

2.3 Correlation does not imply causation

## 3 OLS estimation method

## 4 Caveats of OLS (outliers are bad)

## 5 Example

5.1 How to execute a regression analysis

5.2 First look at data

5.3 Interpretation of the results

5.4 Regression Diagnostics

5.5 Measures of fit

5.6 The miracle of CONTROL VARIABLES in multiple regressions

5.7 When do we need (more) control variables

## 6 Take away messages

# Regression Analysis

Stephan.Huber@hs-fresenius.de (mailto:Stephan.Huber@hs-fresenius.de)

Hochschule Fresenius: Data Science

---

Compiled at 09 May, 2022

Word count: 1618

---

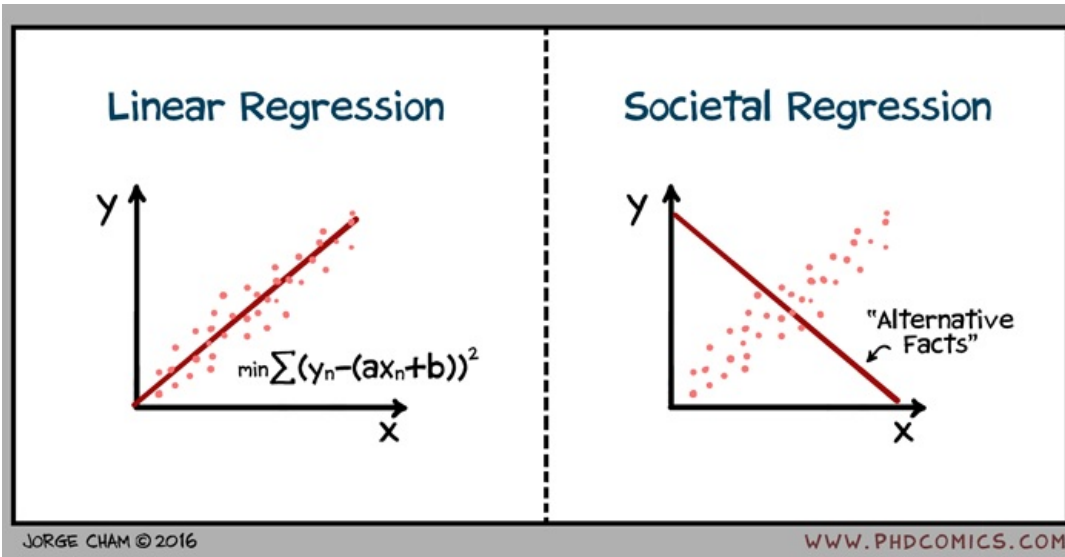
## 1 Literature

Regression analysis is covered by almost all econometric and statistical textbooks. Some are more formal some are more illustrative and intuitive. Here is my selection with a focus on R:

- Book: Applied Statistics with R (<https://davidalpiaz.github.io/appliedstats/>)
- Book: Introduction to Econometrics with R (<https://www.econometrics-with-r.org/>)
- A more formal approach from my Alma Mater: slides ([https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/medien/zeitreihenoekonometrie/eoe\\_eng\\_2020\\_08\\_chapter\\_all.pdf](https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/medien/zeitreihenoekonometrie/eoe_eng_2020_08_chapter_all.pdf)) and handout ([https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/medien/methoden-der-oekonometrie/methoden\\_oekonometrie\\_handout\\_2018\\_11\\_29.pdf](https://www.uni-regensburg.de/wirtschaftswissenschaften/vwl-tschernig/medien/methoden-der-oekonometrie/methoden_oekonometrie_handout_2018_11_29.pdf))
- This presentation is available on my github account (<https://github.com/hubchev/courses>)

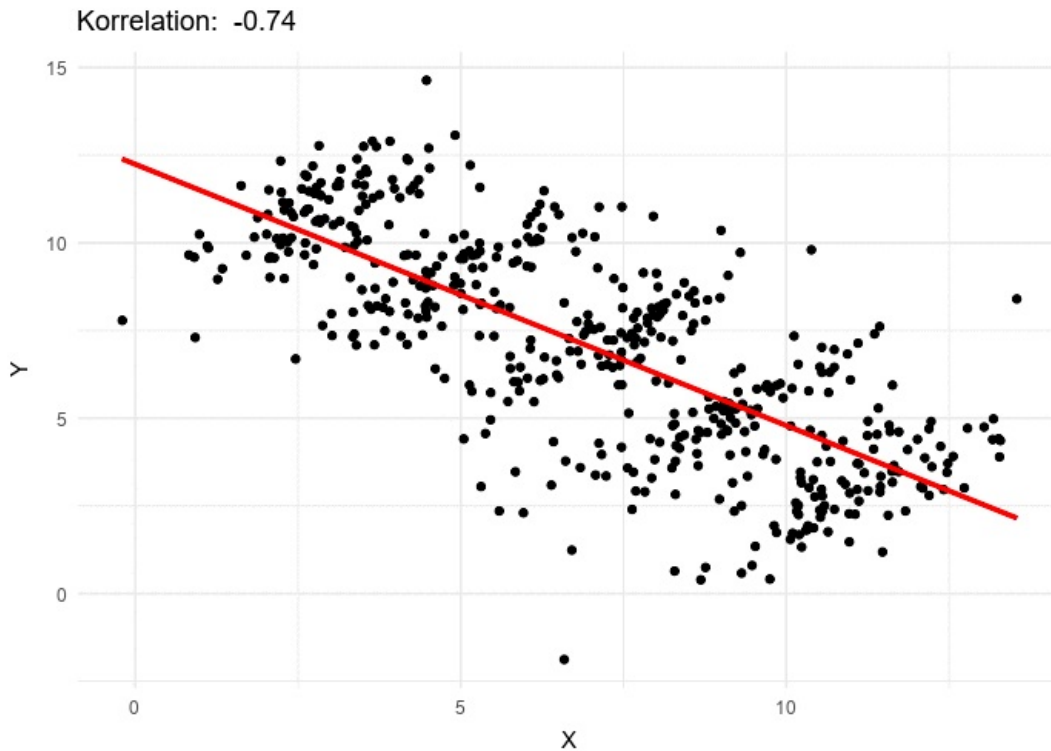
## 2 Why regression analysis?

### 2.1 It is a medicine against alternative facts

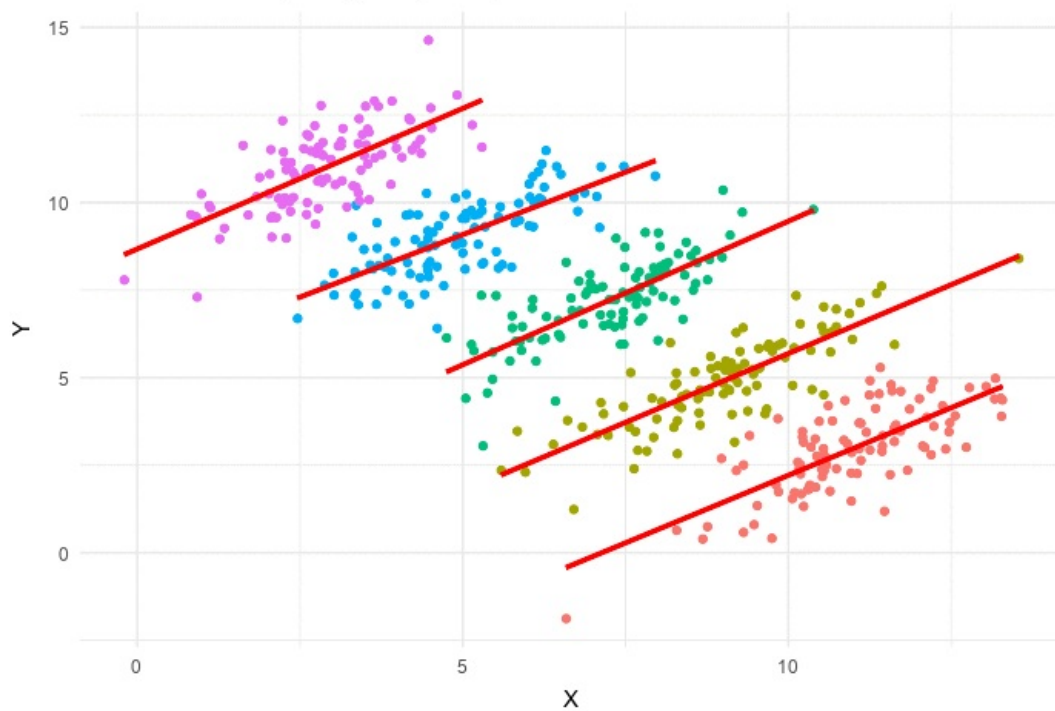


- Regressions allow us to **draw insights** from data,
- to analyze and **interpret** the strength of relationships and
- to reduce the likeliness of **causal fallacy**.

## 2.2 Simpsons paradox



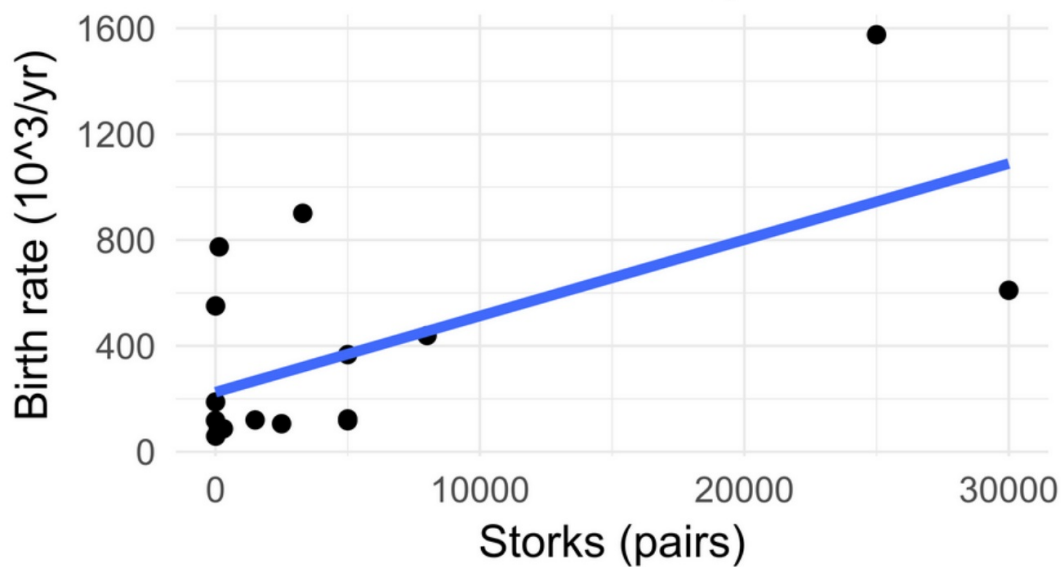
Korrelation: 0.74, 0.82, 0.75, 0.72, 0.69



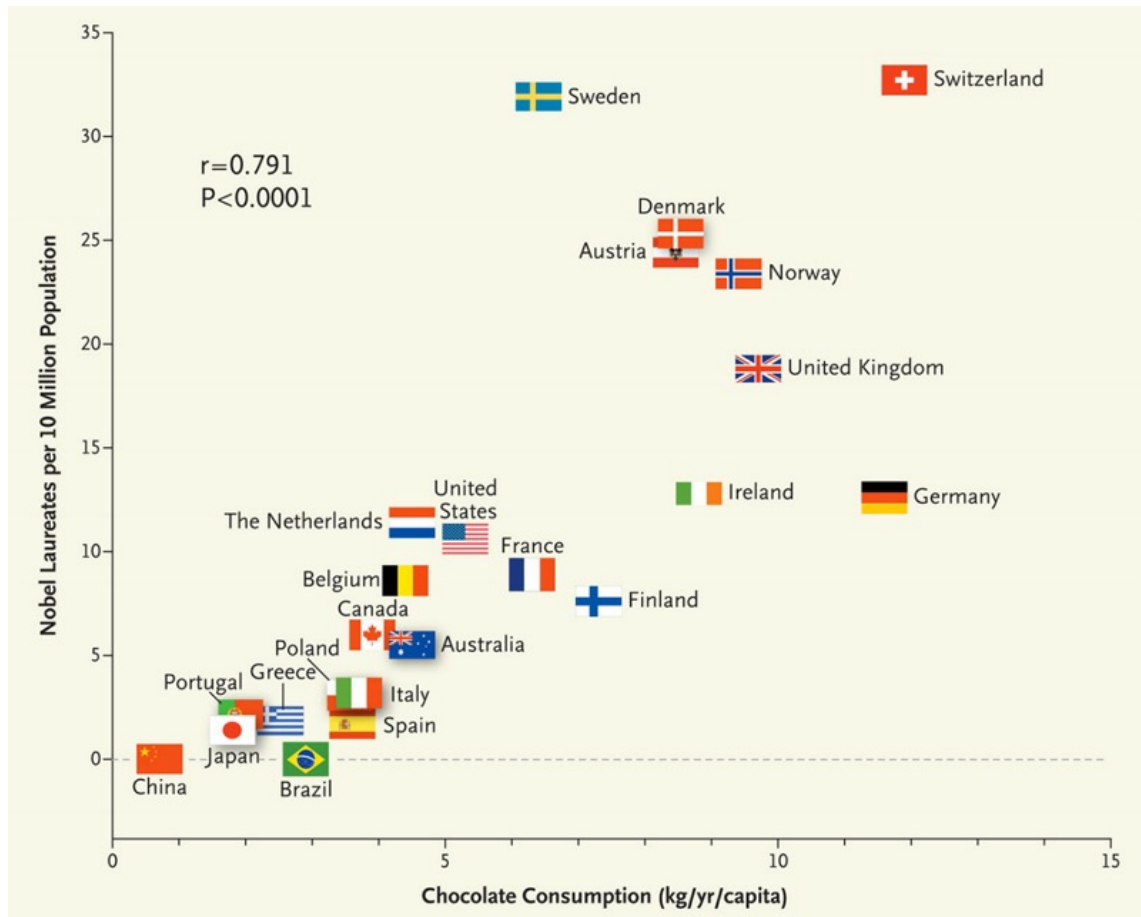
## 2.3 Correlation does not imply causation

### Stork pairs and birth rate

Correlation coefficient = 0.62 , p-value = 0.0079



Matthews, R. (2000), Storks Deliver Babies ( $p = 0.008$ ). Teaching Statistics, 22: 36–38.

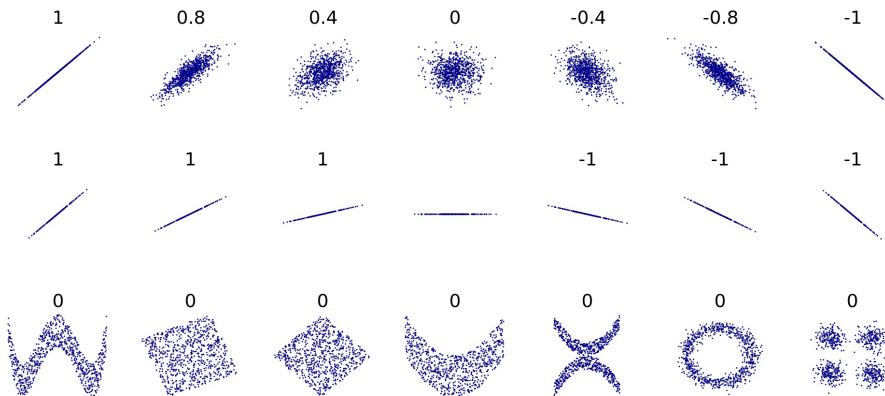


<http://www.nejm.org/doi/full/10.1056/NEJMon1211064> (<http://www.nejm.org/doi/full/10.1056/NEJMon1211064>)

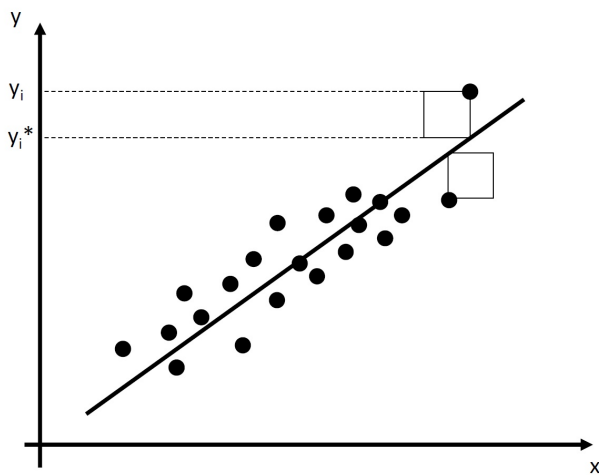
\_Sometimes called spurious correlation\*\_

### 2.3.1 Correlation is often useless

- Correlation reflects the of a linear relationship (top row), it tells us nothing about the slope (strength) of that relationship (middle). Thus, we cannot interpret them economically.
- Moreover, many aspects of nonlinear relationships (bottom) remain unexplained. The figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of  $\backslash(Y\backslash)$  is zero.



- Linear regression** is a predictive modeling techniques that aims to find a mathematical equation for a variable  $\backslash(y\backslash)$  as a function of one (simple linear model) or more variables (multiple linear regression),  $\backslash(x\backslash)$ .
- The method to is called the **ordinary least squared (OLS) method** as it minimizes the sum of the squared differences of all  $\backslash(y\_i\backslash)$  and  $\backslash(y\_i^{*}\backslash)$  as sketched below.



The simple linear regression model is  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$  where

- the index  $i$  runs over the observations,  $i=1, \dots, n$
- $y_i$  is the , the regressand
- $x_i$  is the , the regressor
- $\beta_0$  is the or constant
- $\beta_1$  is the slope of regression line
- $\epsilon_i$  is the or the residual.

### 3 OLS estimation method

- minimize the squared residuals by choosing the estimated coefficients  $\hat{\beta}_0$  and  $\hat{\beta}_1$

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- Minimizing the function requires to calculate the first order conditions with respect to  $\hat{\beta}_0$  and  $\hat{\beta}_1$  and set them zero (see exercises)
- The estimators are:  $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

In the multiple regression model the OLS is derived similarly but we skip the derivation. The  $\hat{\beta}$  coefficient vector can be expressed as

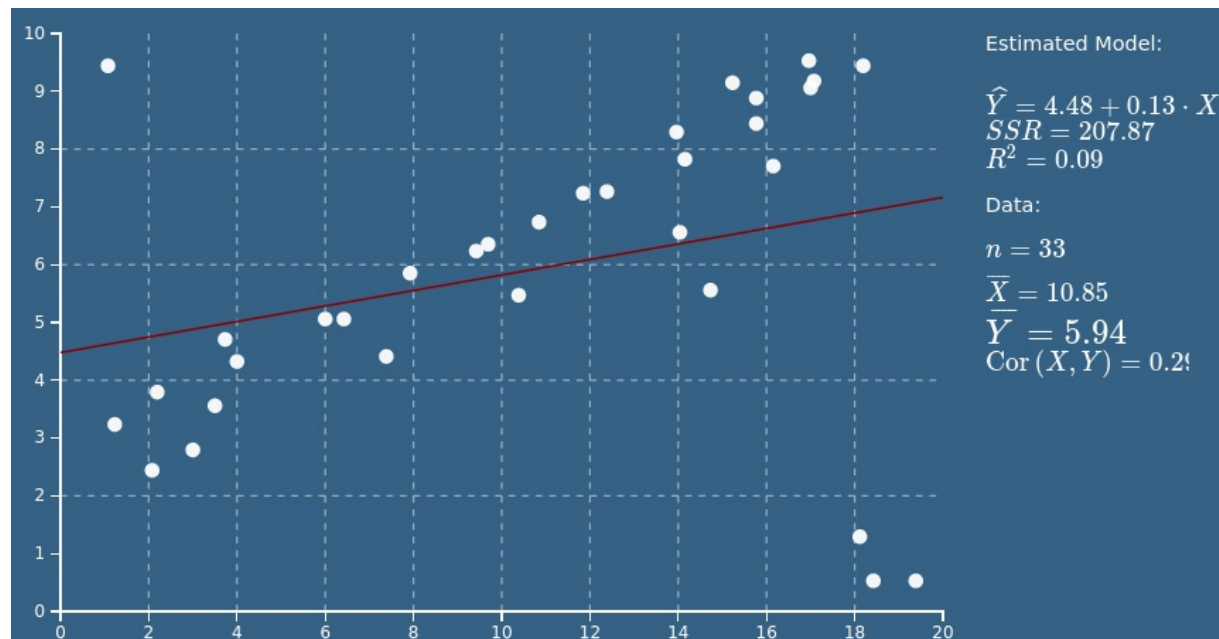
$$\hat{\beta} = (X'X)^{-1}X'y$$



Here you find the derivation of the OLS estimator for the multiple regression model: [ch. 9.1 Applied Statistics with R \(https://daviddalpiazz.github.io/appliedstats/multiple-linear-regression.html\)](https://daviddalpiazz.github.io/appliedstats/multiple-linear-regression.html)

## 4 Caveats of OLS (outliers are bad)

On this website (<https://www.econometrics-with-r.org/4-2-estimating-the-coefficients-of-the-linear-regression-model.html>) you find an interactive application. Play around with it and discuss possible caveats of the OLS method.



## 5 Example

In the statistic course of WS 2020, I asked 23 students about their weight, height, sex, and number of siblings:

```
library("haven")
classdata <- read.csv("https://raw.githubusercontent.com/hubchev/courses/main/dta/classdata.csv")

head(classdata)
```

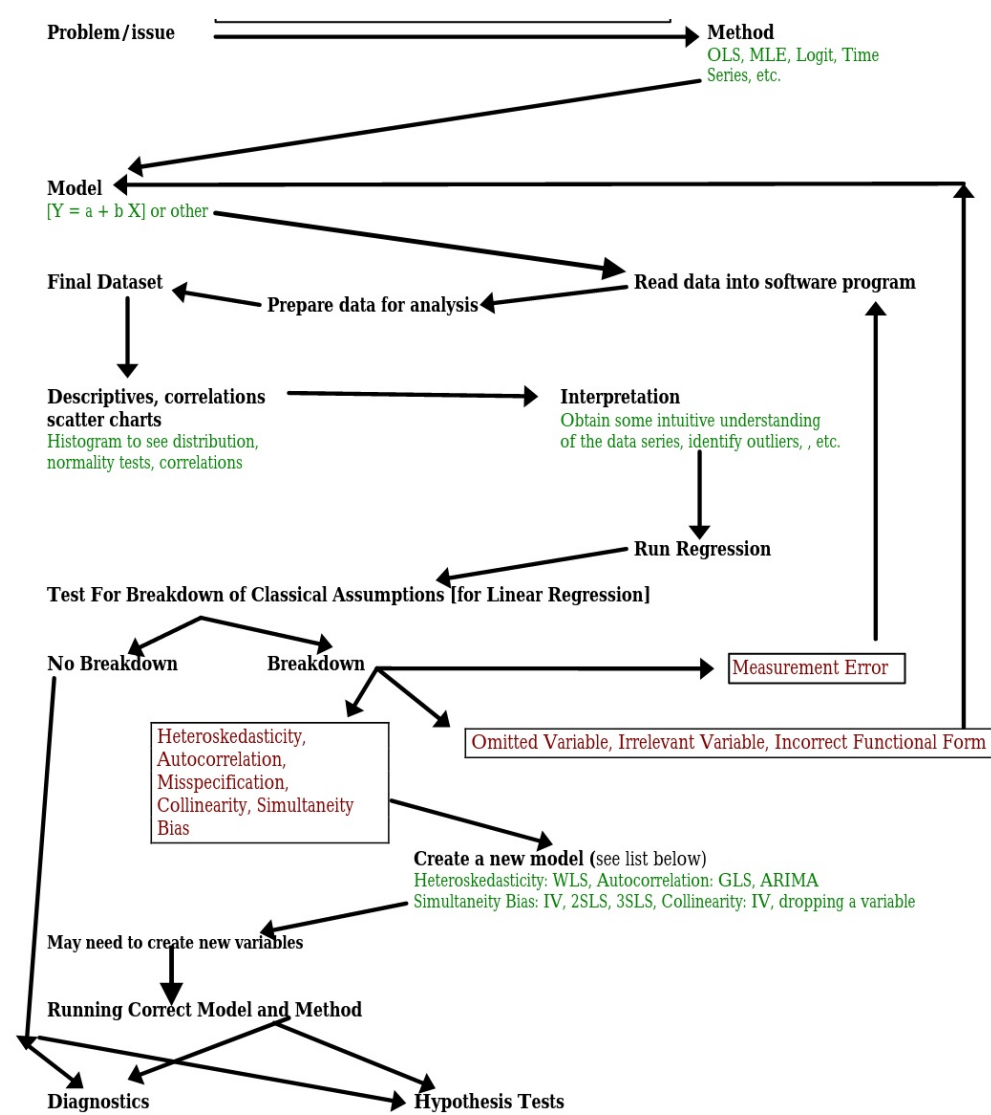
```
##   id sex weight height siblings row
## 1  1  w    53   156         1    g
## 2  2  w    73   170         1    g
## 3  3  m    68   169         1    g
## 4  4  w    67   166         1    g
## 5  5  w    65   175         1    g
## 6  6  w    48   161         0    g
```

```
summary(classdata)
```

```
##      id          sex      weight      height
## Min.   : 1.0      Length:23      Min.   :48.00      Min.   :156.0
## 1st Qu.: 6.5      Class :character 1st Qu.:64.50      1st Qu.:168.0
## Median :12.0      Mode  :character  Median :70.00      Median :175.0
## Mean   :12.0                                Mean   :70.61      Mean   :173.7
## 3rd Qu.:17.5                                3rd Qu.:81.00      3rd Qu.:180.0
## Max.   :23.0                                Max.   :90.00      Max.   :194.0
##      siblings      row
## Min.   :0.000      Length:23
## 1st Qu.:1.000      Class :character
## Median :1.000      Mode  :character
## Mean   :1.391
## 3rd Qu.:2.000
## Max.   :4.000
```

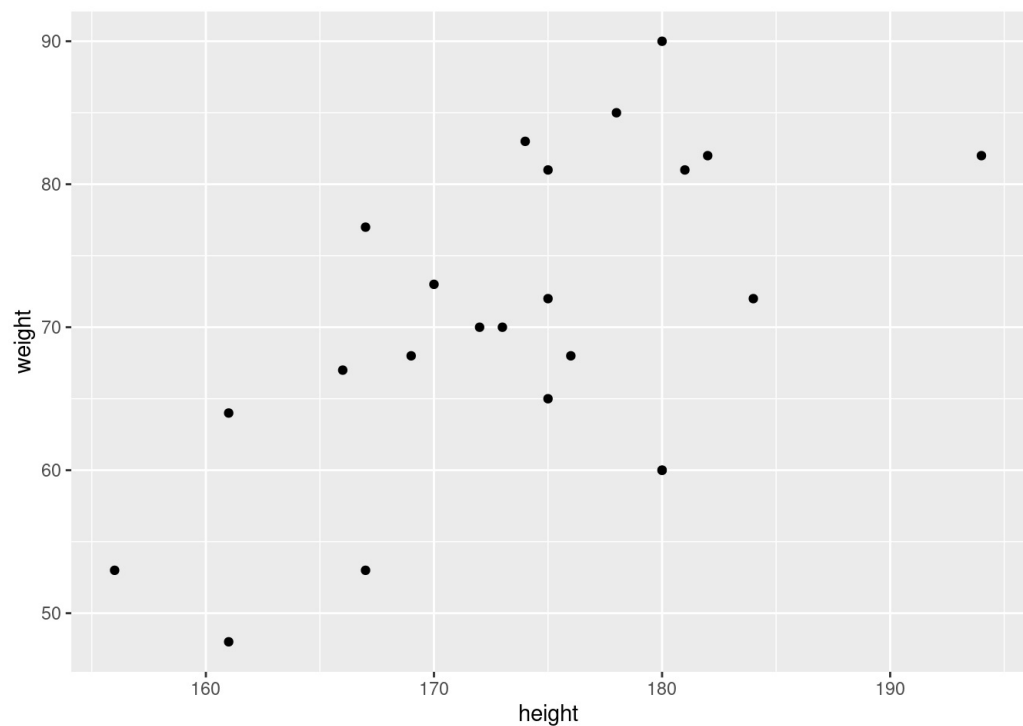
### 5.1 How to execute a regression analysis

1. get known to the data
2. build a theory on how the variables may be related
3. derive a estimated equation from your theory
4. estimate
5. evaluate your empirics
6. go back to 2. and improve your theory
7. interpret your results



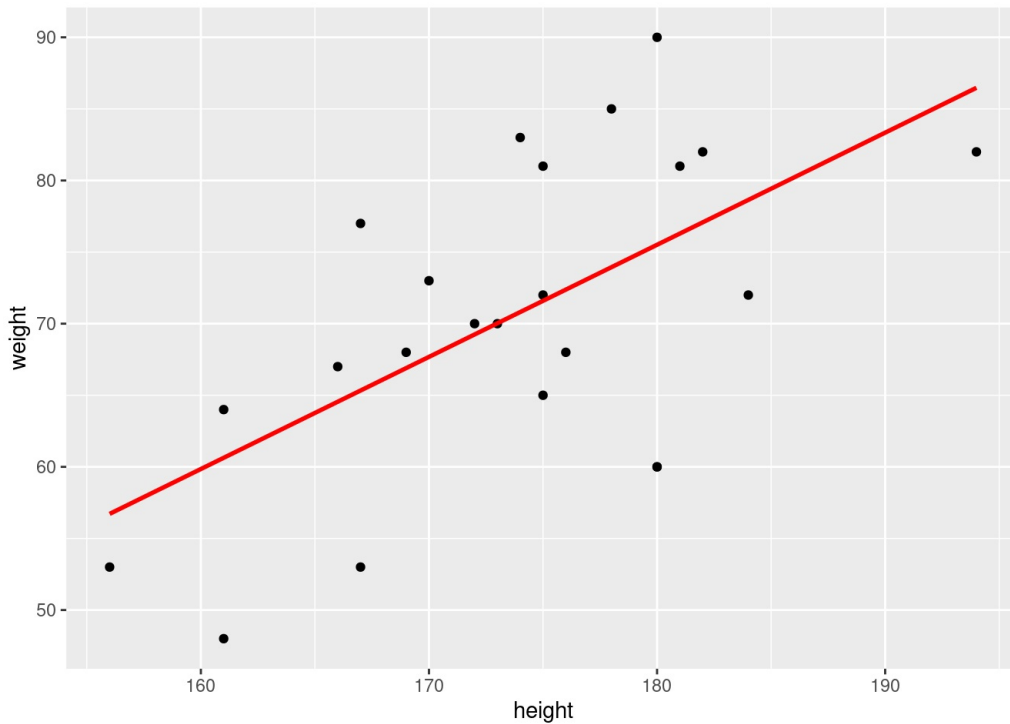
## 5.2 First look at data

```
library("ggplot2")
ggplot(classdata, aes(x=height, y=weight)) + geom_point()
```



include a regression line:

```
ggplot(classdata, aes(x=height, y=weight)) +  
  geom_point() +  
  stat_smooth(formula=y~x, method="lm", se=FALSE, colour="red", linetype=1)
```



distinguish male/female by including a separate constant:

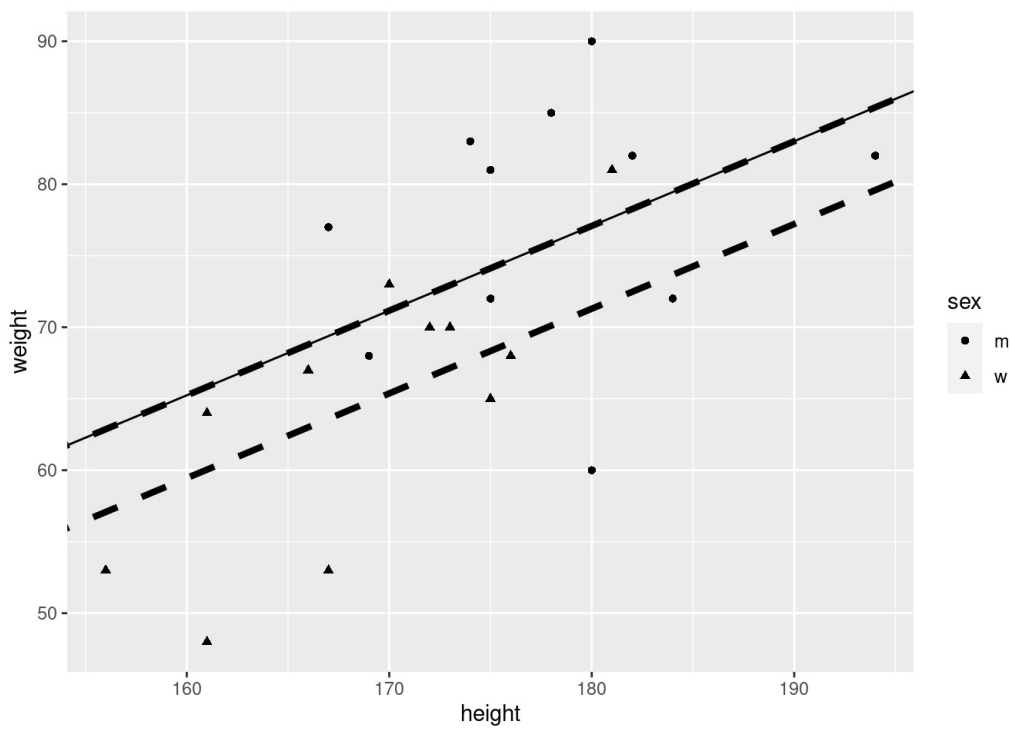
```
## baseline regression model  
model <- lm(weight ~ height + sex , data = classdata )  
show(model)
```

```
##  
## Call:  
## lm(formula = weight ~ height + sex, data = classdata)  
##  
## Coefficients:  
## (Intercept)      height          sexw  
##   -29.5297       0.5923       -5.7894
```

```
interm <- model$coefficients[1]  
slope <- model$coefficients[2]  
interw <- model$coefficients[1]+model$coefficients[3]
```

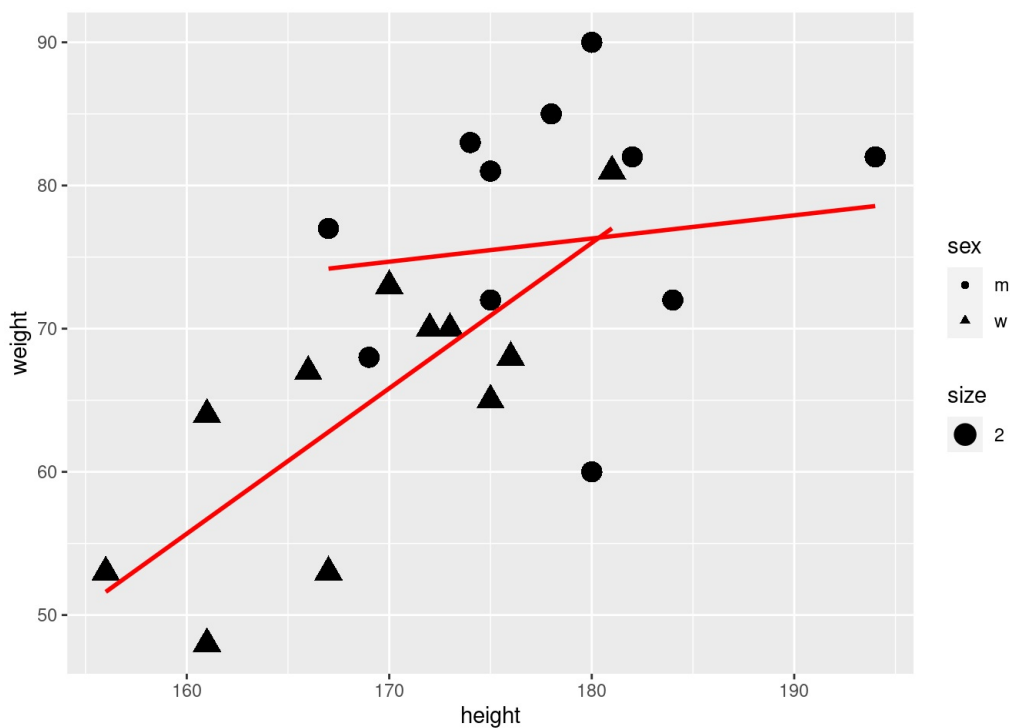
```
ggplot(classdata, aes(x=height, y=weight, shape = sex)) +  
  geom_point() +  
  geom_abline(slope = slope, intercept = interw, linetype = 2, size=1.5)+  
  geom_abline(slope = slope, intercept = interm, linetype = 2, size=1.5) +  
  geom_abline(slope = coef(model)[[2]], intercept = coef(model)[[1]])
```





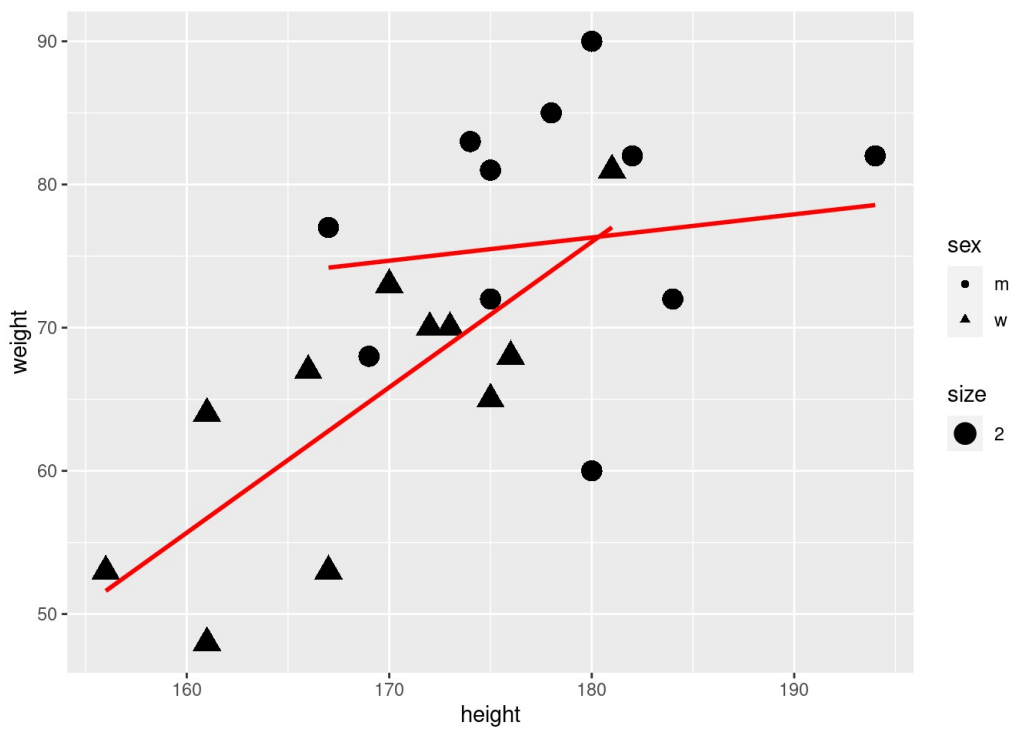
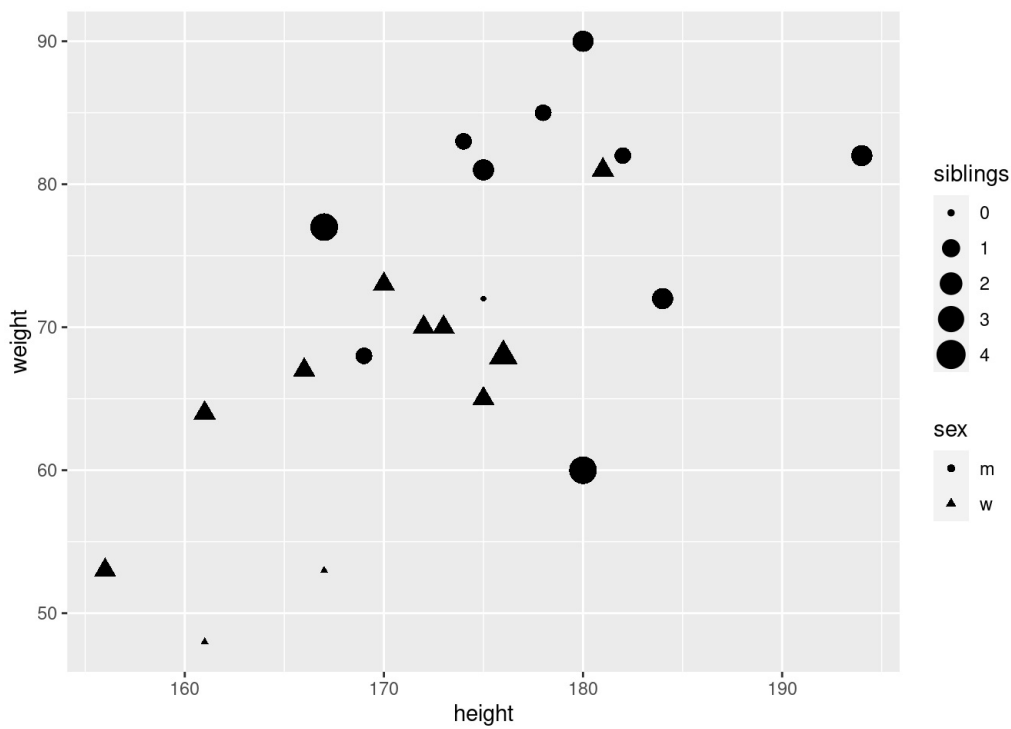
does not look to good, maybe we should introduce also different slopes for m/w

```
ggplot(classdata, aes(x=height, y=weight, shape = sex)) +
  geom_point( aes(size = 2)) +
  stat_smooth(formula = y ~ x, method = "lm",
    se = FALSE, colour = "red", linetype = 1)
```



Can we use other available variables: siblings?

```
ggplot(classdata, aes(x=height, y=weight, shape = sex)) +
  geom_point( aes(size = siblings))
```



Let us look at regression output:

```
m1 <- lm(weight ~ height , data = classdata )
m2 <- lm(weight ~ height + sex , data = classdata )
m3 <- lm(weight ~ height + sex + height * sex , data = classdata )
m4 <- lm(weight ~ height + sex + height * sex + siblings , data = classdata )
m5 <- lm(weight ~ height + sex + height * sex , data = subset(classdata, siblings < 4 ))
```

```
library(sjPlot)
```

```
## Registered S3 method overwritten by 'parameters':
##   method                from
##   format.parameters_distribution datawizard
```

```
## #refugeeswelcome
```

```
tab_model(m1, m2, m3, m4, m5,
          p.style = "stars",
          p.threshold = c(0.2, 0.1, 0.05),
          show.ci = FALSE,
          show.se = FALSE)
```

	weight	weight	weight	weight
Predictors	Estimates	Estimates	Estimates	Estimates
(Intercept)	-65.44 *	-29.53	47.14	50.27
height	0.78 ***	0.59 ***	0.16	0.16
sex [w]		-5.79	-153.96 **	-161.92 **
height * sex [w]			0.85 *	0.89 *
siblings				-1.16
Observations	23	23	23	23
R <sup>2</sup> / R <sup>2</sup> adjusted	0.363 / 0.333	0.412 / 0.354	0.487 / 0.407	0.496 / 0.385

•  $p < 0.2$  \*\*  $p < 0.1$  \*\*\*  $p < 0.05$

excluding outliers with four siblings:

	weight	weight
Predictors	Estimates	Estimates
(Intercept)	47.14	27.69
height	0.16	0.28
sex [w]	-153.96 **	-134.51 *
height * sex [w]	0.85 *	0.74 *
Observations	23	21
R <sup>2</sup> / R <sup>2</sup> adjusted	0.487 / 0.407	0.572 / 0.497

•  $p < 0.2$  \*\*  $p < 0.1$  \*\*\*  $p < 0.05$

## 5.3 Interpretation of the results

- We can make predictions about the impact of height on male and female
- As both, the intercept and the slope differs for male and female we should interpret the regressions separately:
- One centimeter more for **MEN** is *on average* and *ceteris paribus* related with 0.16 kg more weight.
- One centimeter more for **WOMEN** is *on average* and *ceteris paribus* related with 1.01 kg more weight.

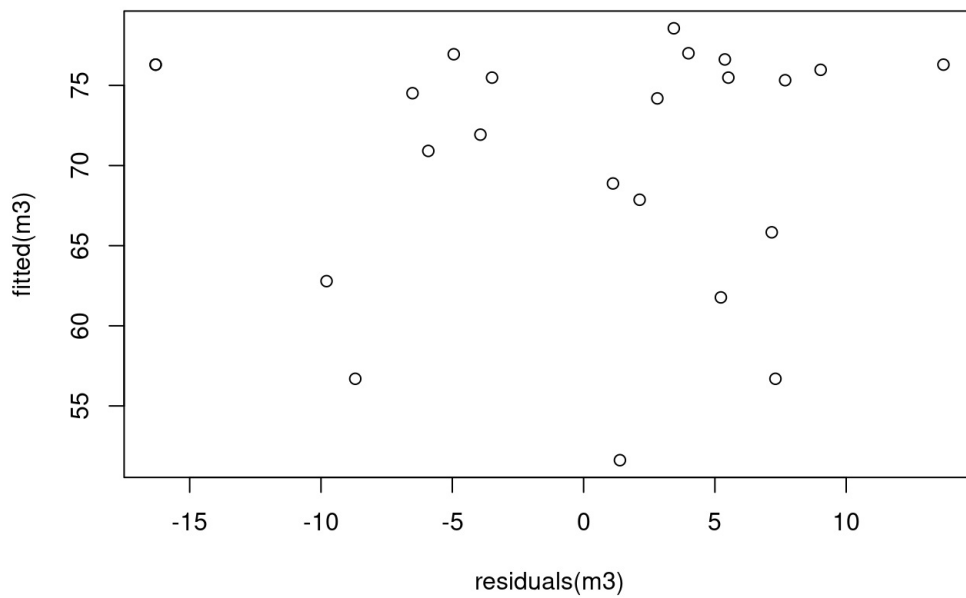
## 5.4 Regression Diagnostics

Linear Regression makes several assumptions about the data, the model assumes that:

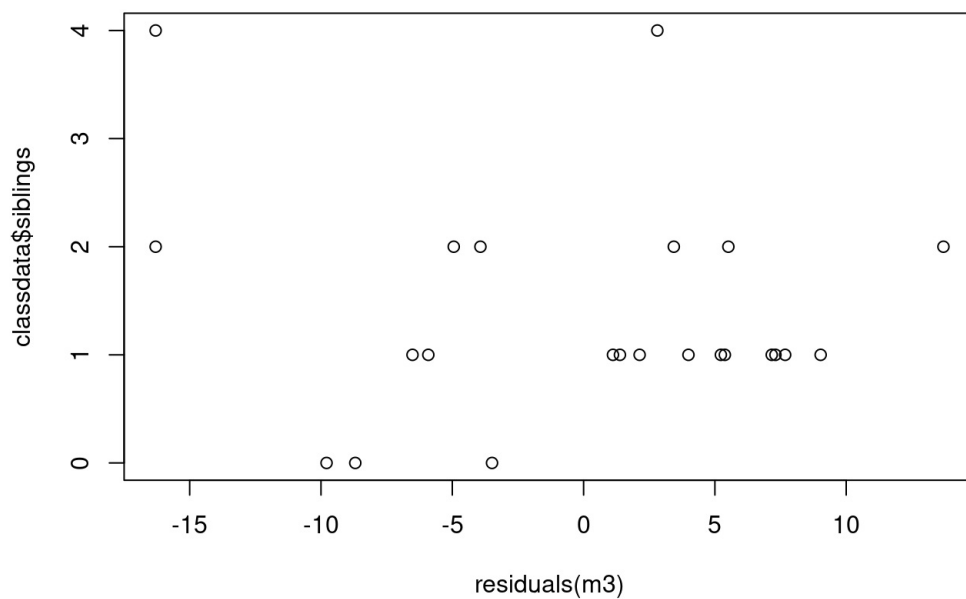
- The relationship between the predictor (x) and the dependent variable (y) has linear relationship.
- The residuals are assumed to have a constant variance.
- The residual errors are assumed to be normally distributed.
- Error terms are independent and have zero mean.

More on regression Diagnostics can be found Applied Statistics with R: 13 Model Diagnostics (<https://daviddalpiatz.github.io/appliedstats/model-diagnostics.html#r-markdown-6>)

```
plot(residuals(m3), fitted(m3))
```



```
plot(residuals(m3), classdata$siblings)
```

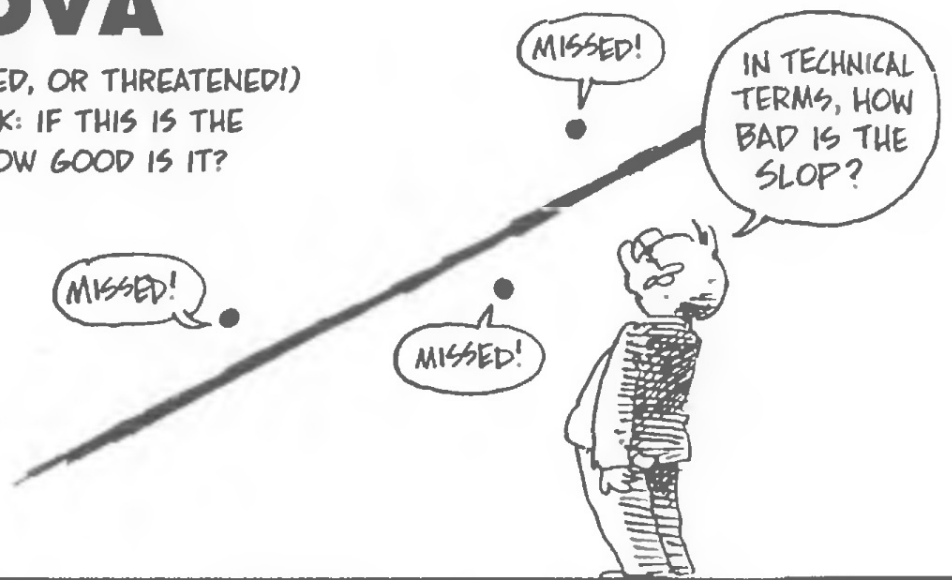


## 5.5 Measures of fit

### 5.5.1 R squared

# ANOVA

(AS PROMISED, OR THREATENED!)  
NOW WE ASK: IF THIS IS THE  
BEST FIT, HOW GOOD IS IT?

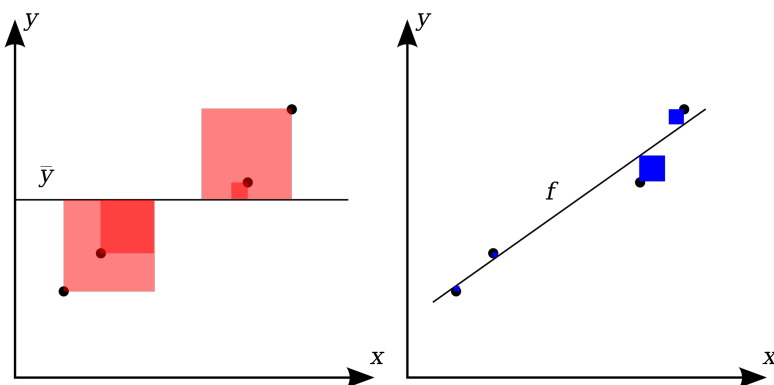


$R^2$  is the fraction of the sample variance of  $(Y_i)$  that is explained by  $(X_i)$ . It can be written as the ratio of the explained sum of squares (ESS) to the total sum of squares (TSS):  $ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$ ,  $TSS = \sum_{i=1}^n (Y_i - \bar{Y})^2$ ,  $R^2 = \frac{ESS}{TSS}$ .

Since  $(TSS = ESS + SSR)$  we can also write

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

with  $(SSR = \sum_{i=1}^n \epsilon_i^2)$



$R^2$  lies between 0 and 1. It is easy to see that a perfect fit, i.e., no errors made when fitting the regression line, implies  $(R^2=1)$  since then we have  $(SSR=0)$ . On the contrary, if our estimated regression line does not explain any variation in the  $(Y_i)$ , we have  $(ESS=0)$  and consequently  $(R^2=0)$ .

## 5.5.2 Adjusted R-squared

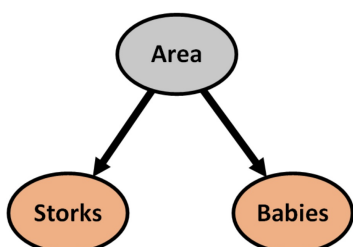
Including more independent variables into an estimated model must decrease SSR. Thus, the R-squared never decreases, when adding new variables even when it just a chance correlation between variables. Having more coefficients to estimate the precision at which we can estimate the effects decrease. Thus, we need to deal with the trade-off. The adjusted  $(R^2)$  can help here:

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}$$

Always use adjusted  $(R^2)$  when you compare specifications with different number of coefficients.

## 5.6 The miracle of CONTROL VARIABLES in multiple regressions

Control variables are usually variables that you are not particularly interested in, but that are related to the dependent variable. You want to remove their effects from the equation. A control variable enters a regression in the same way as an independent variable – the method is the same.

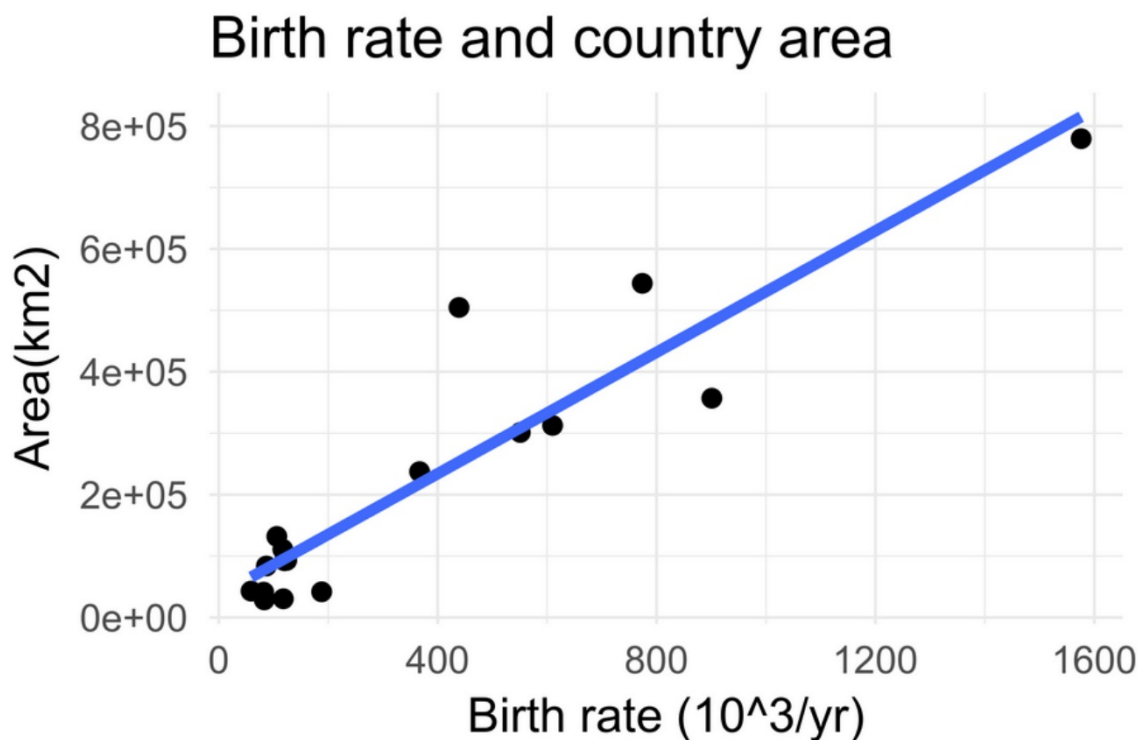


## 5.7 When do we need (more) control variables

From the Gauss-Markov theorem (<https://www.econometrics-with-r.org/5-5-the-gauss-markov-theorem.html>) we know that if the OLS assumptions (<https://www.econometrics-with-r.org/6-4-ols-assumptions-in-multiple-regression.html>) are fulfilled, the OLS estimator is (in the sense of smallest variance) the **best linear conditionally unbiased estimator (BLUE)**.

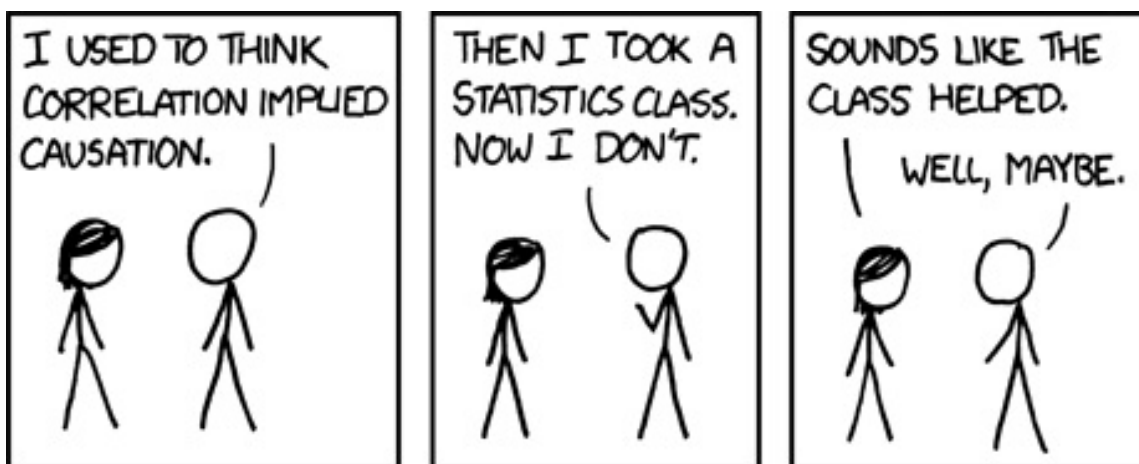
However, OLS estimates can suffer from **omitted variable bias** when the regressor,  $X$ , is correlated with an omitted variable. For omitted variable bias to occur, two conditions must be fulfilled:

1.  $X$  is correlated with the omitted variable.
2. The omitted variable is a determinant of the dependent variable  $Y$ .



## 6 Take away messages

- Regressions rule the world and may kill alternative facts.
- Correlation does not imply causation.
- It is hard to find the true *data generating process*.



```
rmarkdown::render("regress_lecture.Rmd", "all")
```

wkhtmltopdf regress\_lecture.html regress\_lecture.pdf

```
## Loading page (1/2)
## [>] 0%
[=====>] 10%
[=====>] 90%
[=====] 100%
Printing pages (2/2)
## [>]
Done
```