Correlation and Linear Regressions

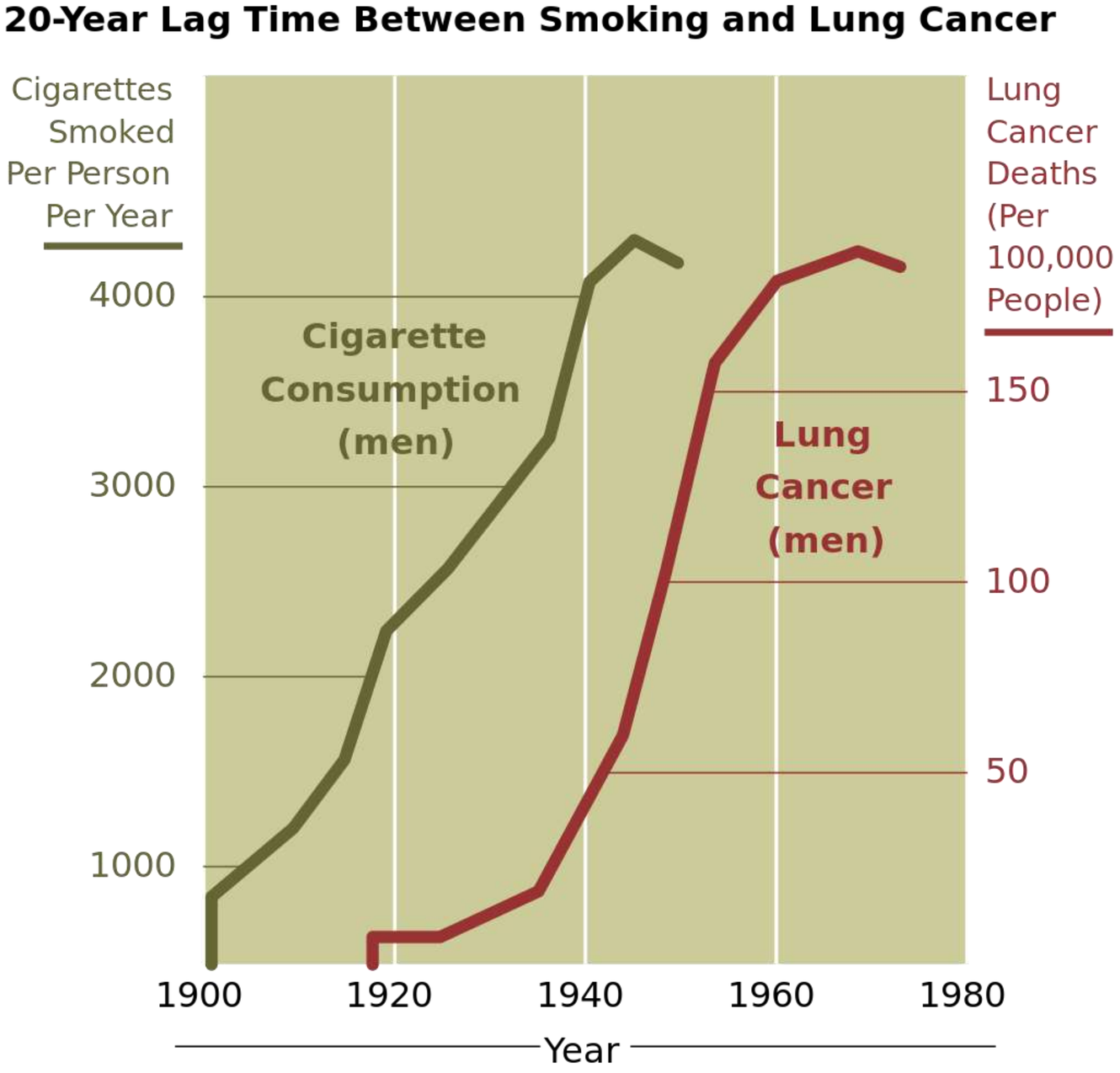
ADS2 - Week 2.6

Lecturer: Zhaoyuan Fang & Samano Hugo

March 18th, 2024

浙江大学爱丁堡大学联合学院

ZJU-UoE INSTITUTE



https://en.m.wikipedia.org/wiki/File:Cancer_smoking_lung_cancer_correlation_from_NIH.svg

Learning Objectives

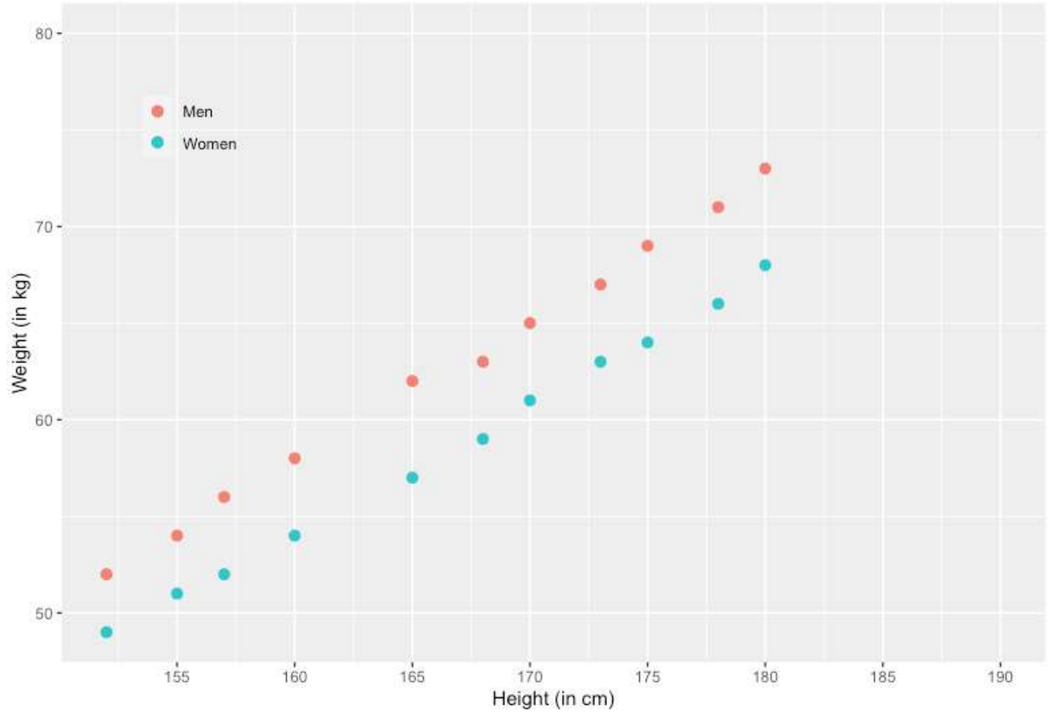
After this lecture you will be able to:

- know the concepts of **correlation** and **linear regressions**
- explore different datasets to **apply** these concepts
- use the relevant **R functions** to perform data analyses

Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
163	56	60
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73



If height increases, weight increases

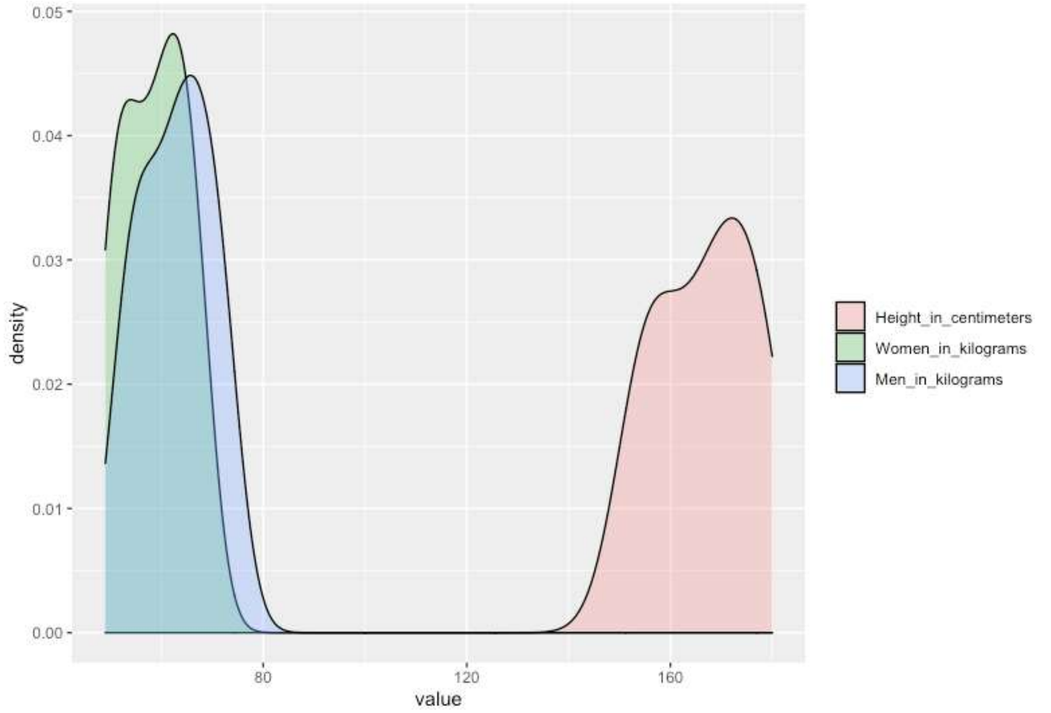
If height decreases, weight decreases

Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
163	56	60
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73

Are height and weight values varying in a similar way?



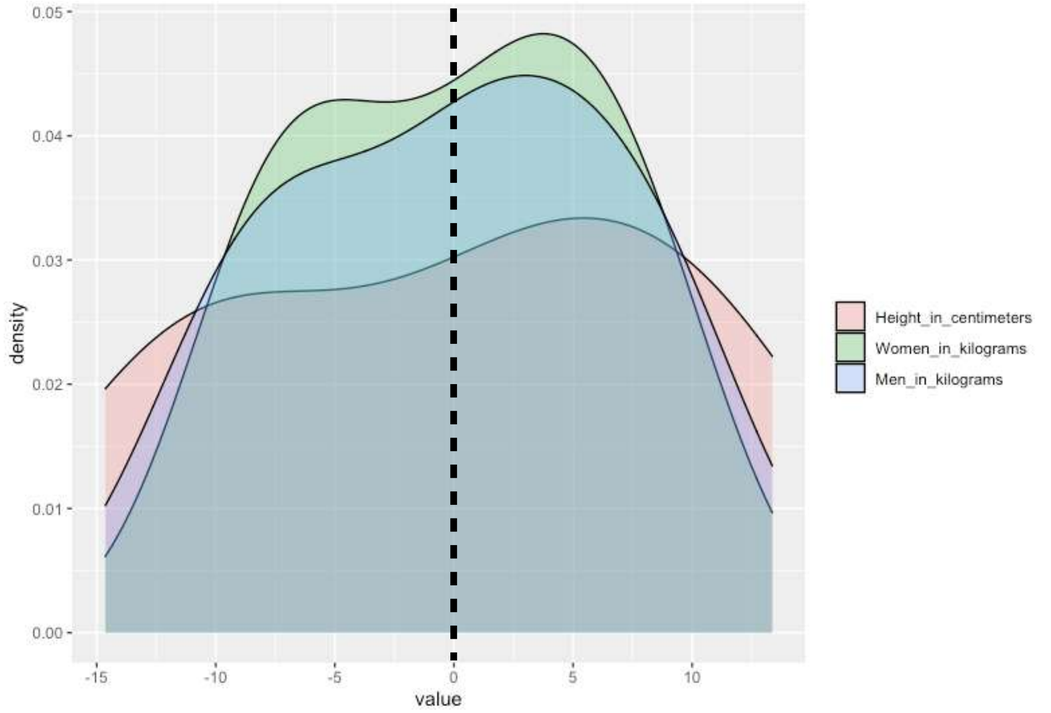
Distributions are not centred

Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
163	56	60
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73
166.64	58.55	62.73	Mean

Are height and weight values varying in a similar way?



Centred distributions

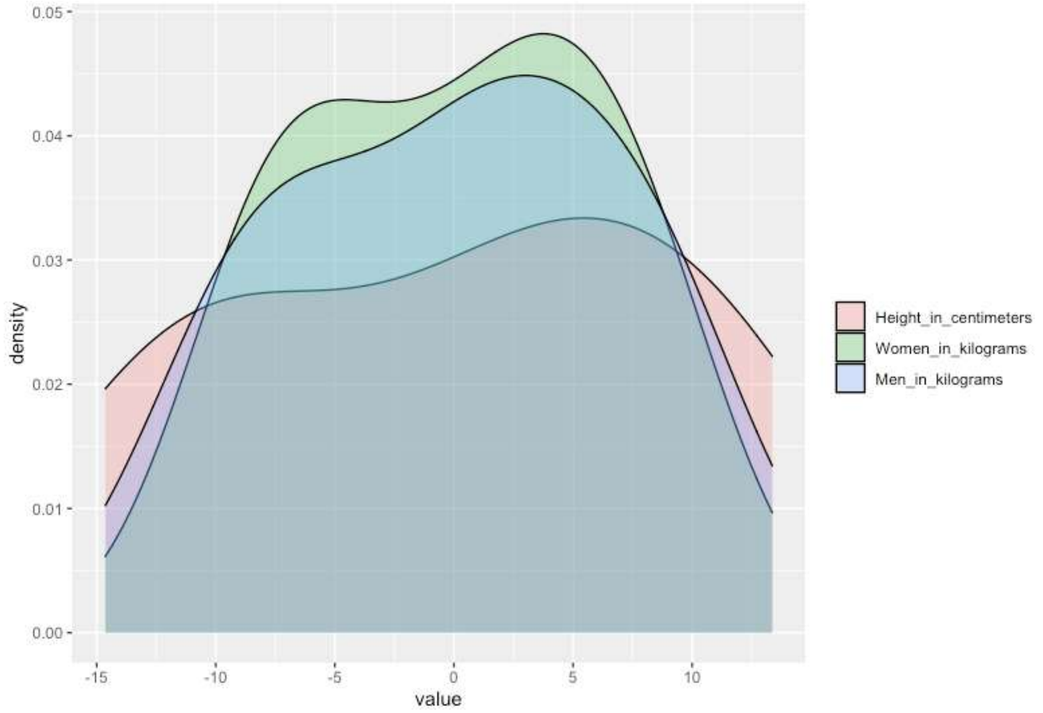
$$(X - \hat{X})$$

Height and Weight Variance and Covariance

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
163	56	60
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73
166.64	58.55	62.73	Mean

We can calculate the variance and covariance



Centred distributions

$$(X - \hat{X})$$

Height and Weight Variance and Covariance

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
163	56	60
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73
166.64	58.55	62.73	Mean

We can calculate the variance and covariance

Variance of Height (S_x^2) :

$$S_x^2 = \frac{\Sigma(X-\hat{X})^2}{n-1} = 92.05$$

Variance of Weight (S_y^2) :

$$S_y^2 = \frac{\Sigma(Y-\hat{Y})^2}{n-1} = 41.47$$

Covariance of height and weight ($Cov(x,y)$) :

$$Cov(x,y) = \frac{\Sigma(X-\hat{X})(Y-\hat{Y})}{n-1} = 61.72$$

Sample Correlation Coefficient

$$S_x^2 = \frac{\Sigma(X-\hat{X})^2}{n-1} = 92.05$$

$$S_y^2 = \frac{\Sigma(Y-\hat{Y})^2}{n-1} = 41.47$$

$$Cov(x,y) = \frac{\Sigma(X-\hat{X})(Y-\hat{Y})}{n-1} = 61.72$$

Sample Correlation Coefficient

$$S_x^2 = \frac{\Sigma(X-\hat{X})^2}{n-1} = 92.05$$

$$S_y^2 = \frac{\Sigma(Y-\hat{Y})^2}{n-1} = 41.47$$

$$Cov(x,y) = \frac{\Sigma(X-\hat{X})(Y-\hat{Y})}{n-1} = 61.72$$

Given these three values, we define the sample correlation coefficient (**r**) as:

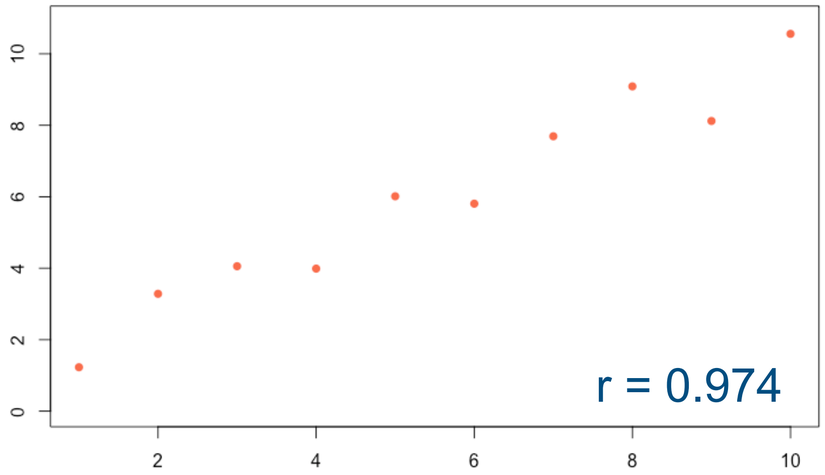
$$r = \frac{Cov(x,y)}{\sqrt{S_x^2 S_y^2}} = 0.9989$$

Sample correlation coefficient ranges from -1 to +1

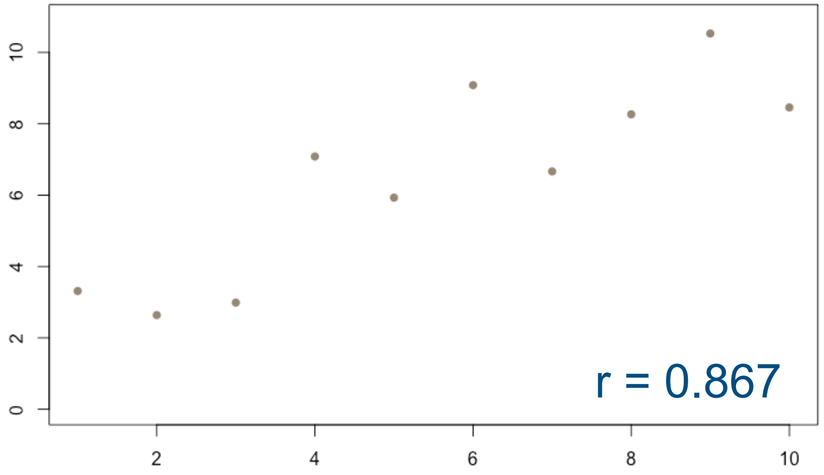
The coefficient quantifies the **direction** and **strength** of the linear association between the two variables

Sample Correlation Coefficient

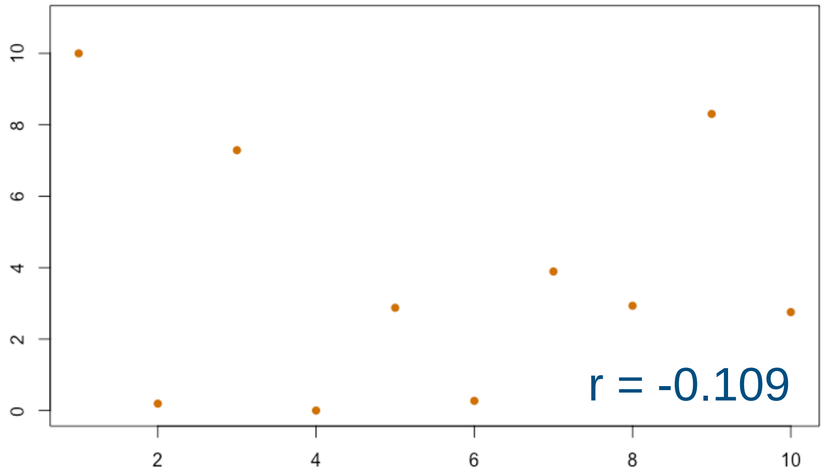
Positive Strong



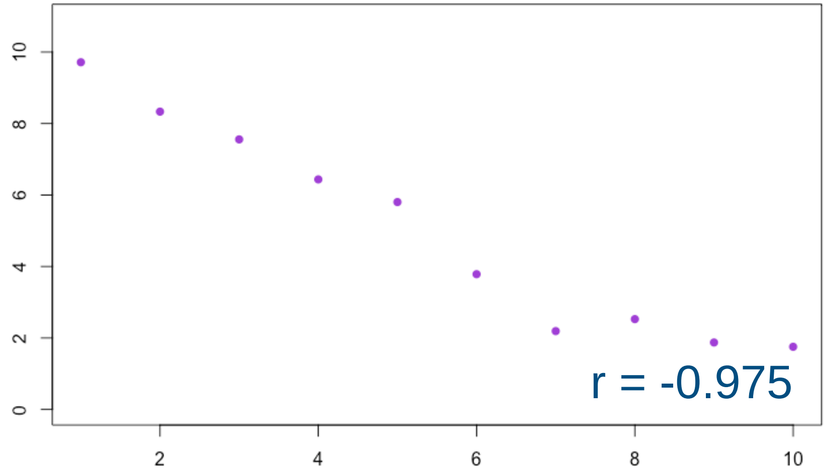
Positive Weak



No Correlation



Negative Strong

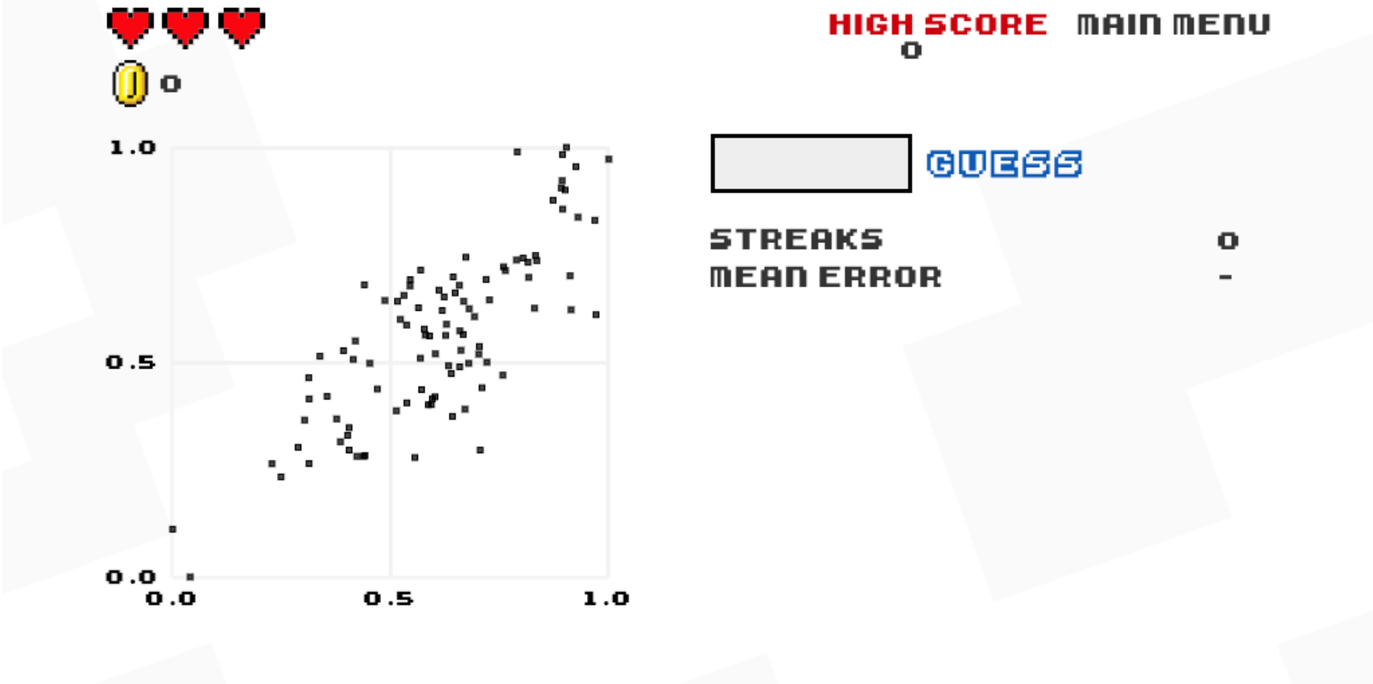


Guess the Correlation Coefficient

http://guessthecorrelation.com

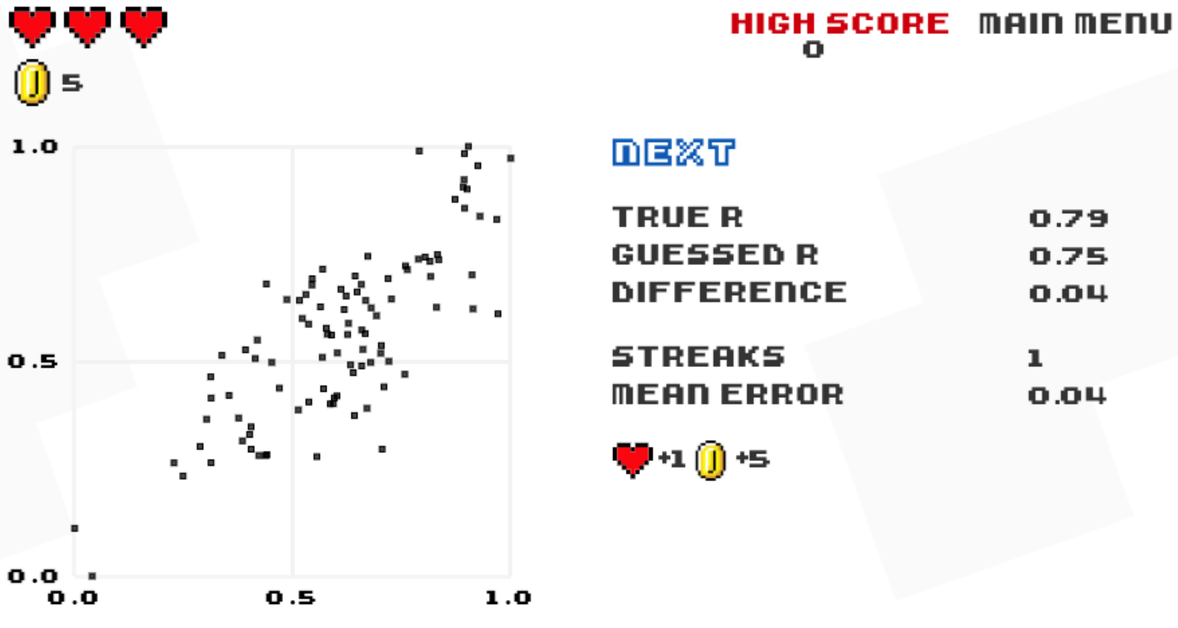
Guess the Correlation Coefficient

http://guessthecorrelation.com



Guess the Correlation Coefficient

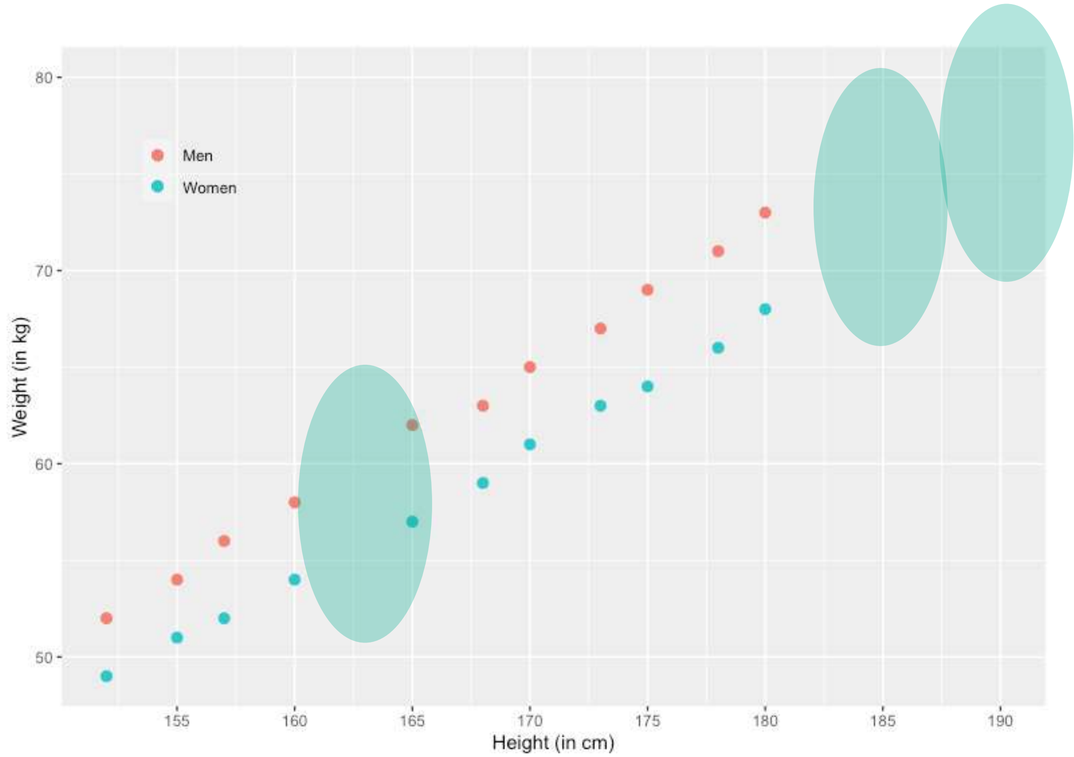
http://guessthecorrelation.com



Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73



$r_m = 0.9986$
$r_w = 0.9989$

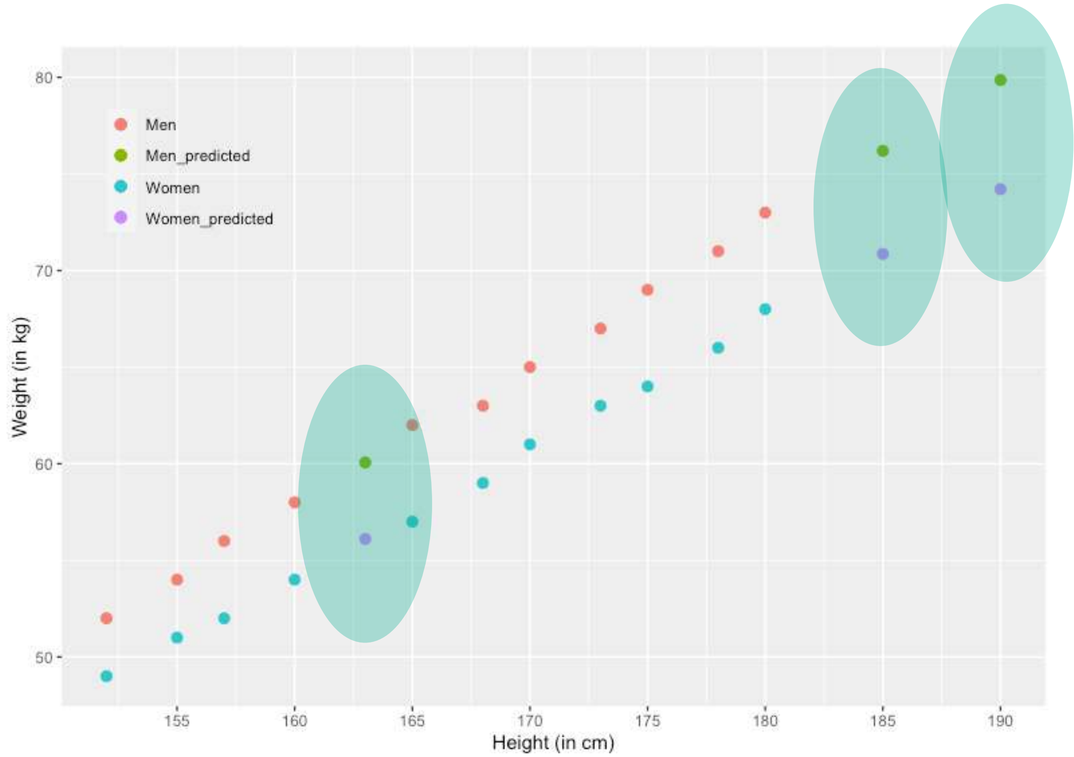
Can we predict the weight corresponding to 163 cm?

How about at 185 cm and 190 cm?

Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73



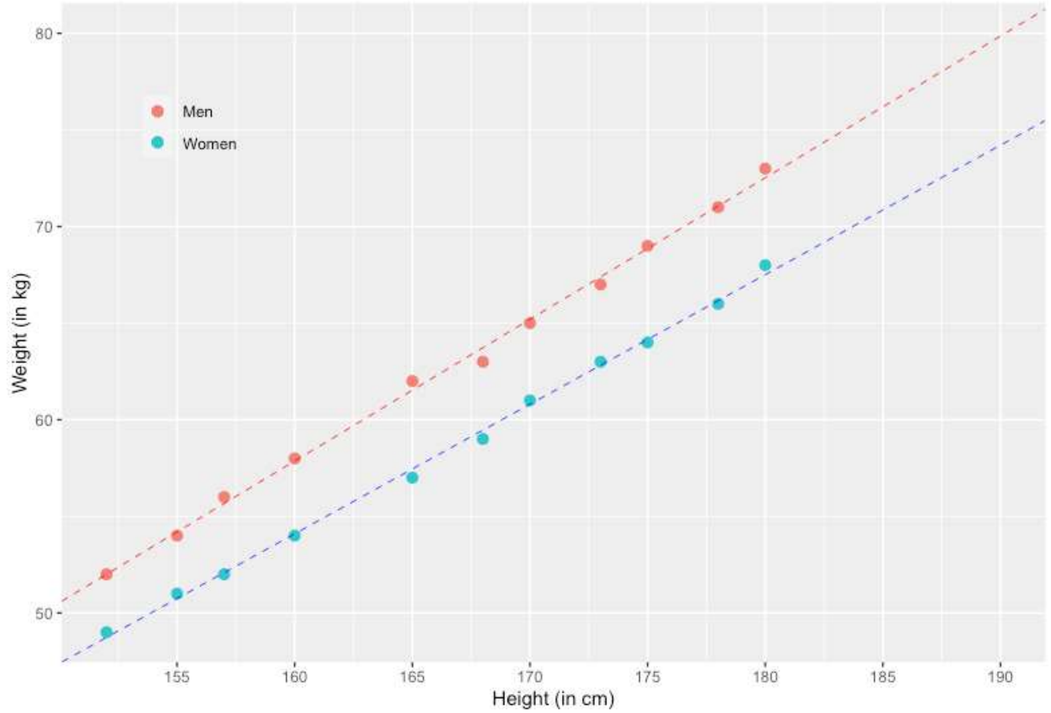
Can we predict the weight corresponding to 163 cm?

How about at 185 cm and 190 cm?

Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73



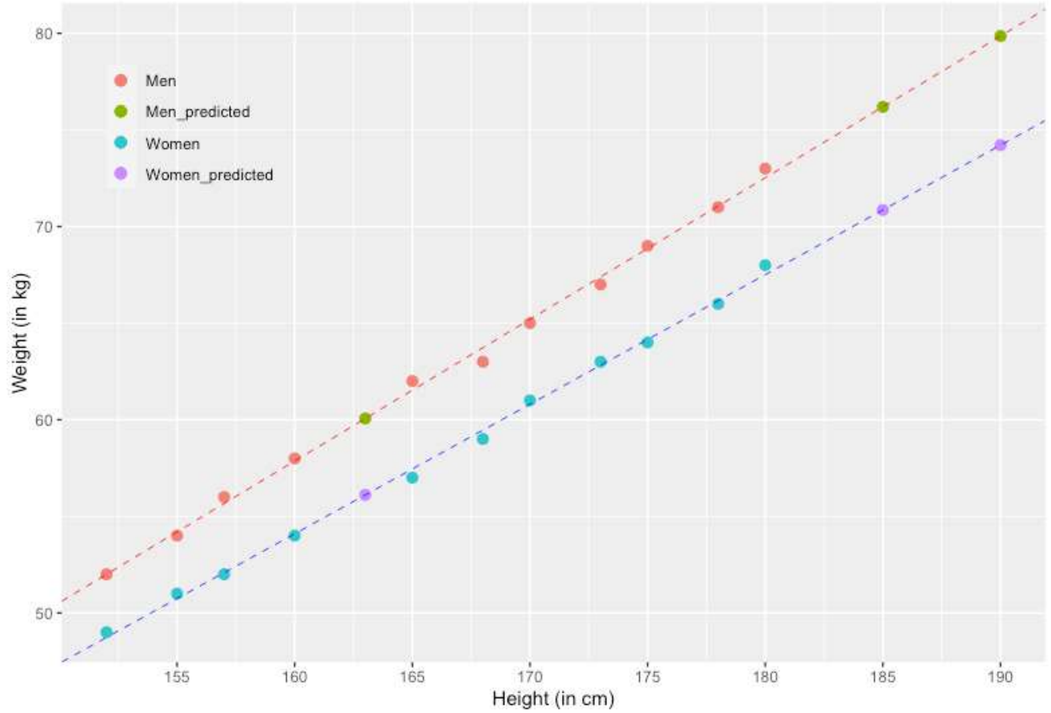
Can we predict the weight corresponding to 163 cm?

How about at 185 cm and 190 cm?

Height and Weight Correlation

Ideal height and weight

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73



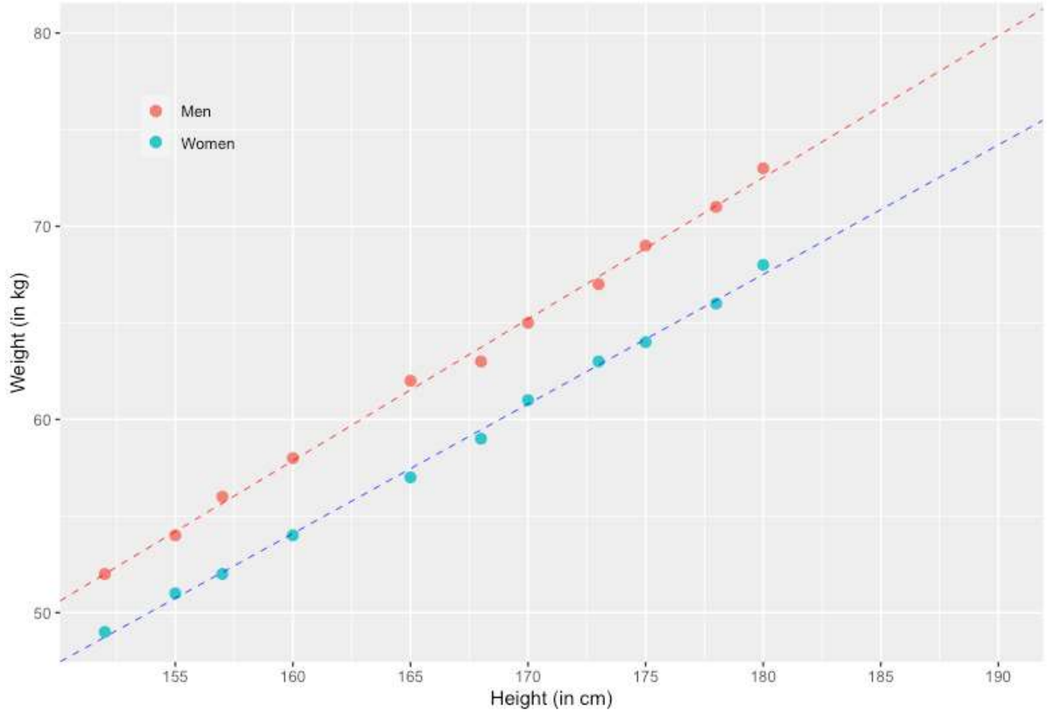
Can we predict the weight corresponding to 163 cm?

How about at 185 cm and 190 cm?

Linear Regression

A **simple linear regression** describes the association between an independent variable and a dependent one

Dependent variable
(outcome)
Y



Independent variable
(predictor)
X

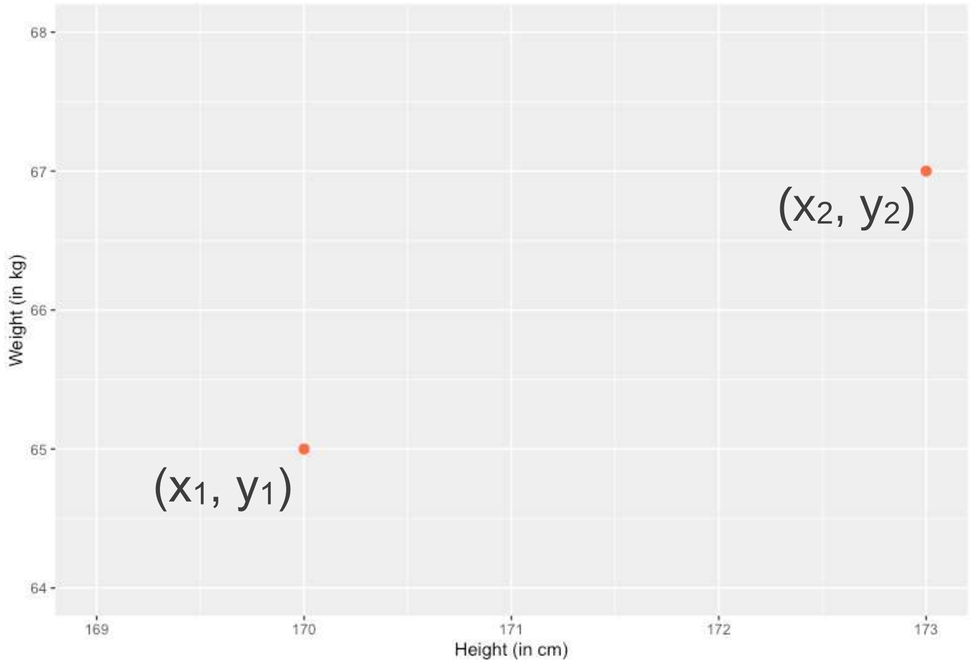
The regression will follow the expression:

$$Y = b_0 + b_1 \cdot X$$

intercept

slope

Linear Regression



Given (x_1, y_1) and (x_2, y_2) we fit it in a straight line:

$$\hat{Y} = b_0 + b_1 \cdot X$$

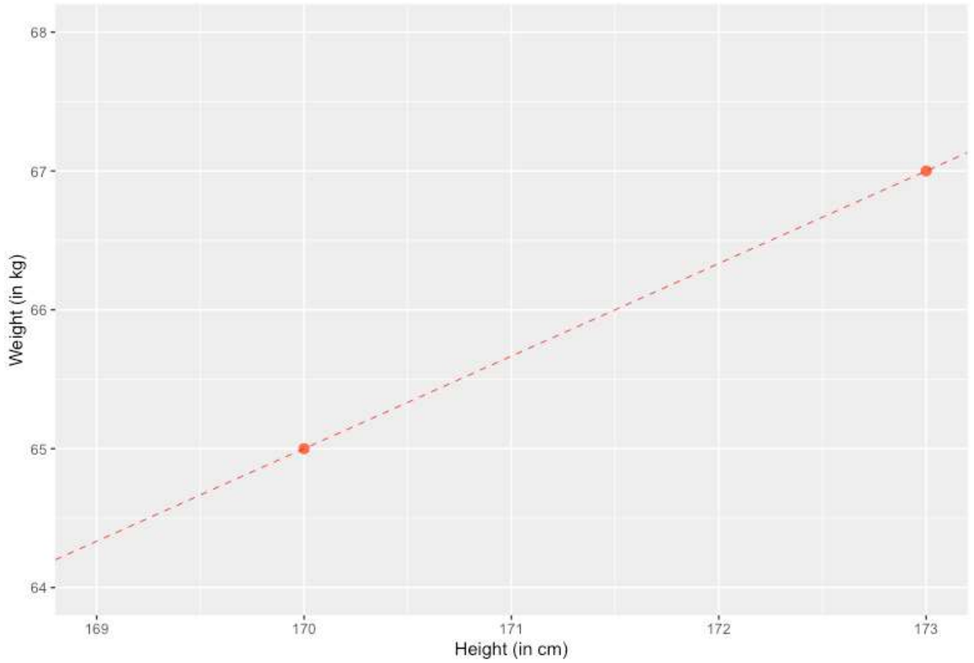
$$y_1 = b_0 + b_1 \cdot x_1$$
$$y_2 = b_0 + b_1 \cdot x_2$$

$$b_0 = y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1$$

$$b_1 = \frac{y_2 - y_1}{x_2 - x_1}$$

$\hat{Y}$ indicates a predicted or expected value

Linear Regression



Given (x_1, y_1) and (x_2, y_2) we fit it in a straight line:

$$\hat{Y} = b_0 + b_1 \cdot X$$

$$y_1 = b_0 + b_1 \cdot x_1$$
$$y_2 = b_0 + b_1 \cdot x_2$$

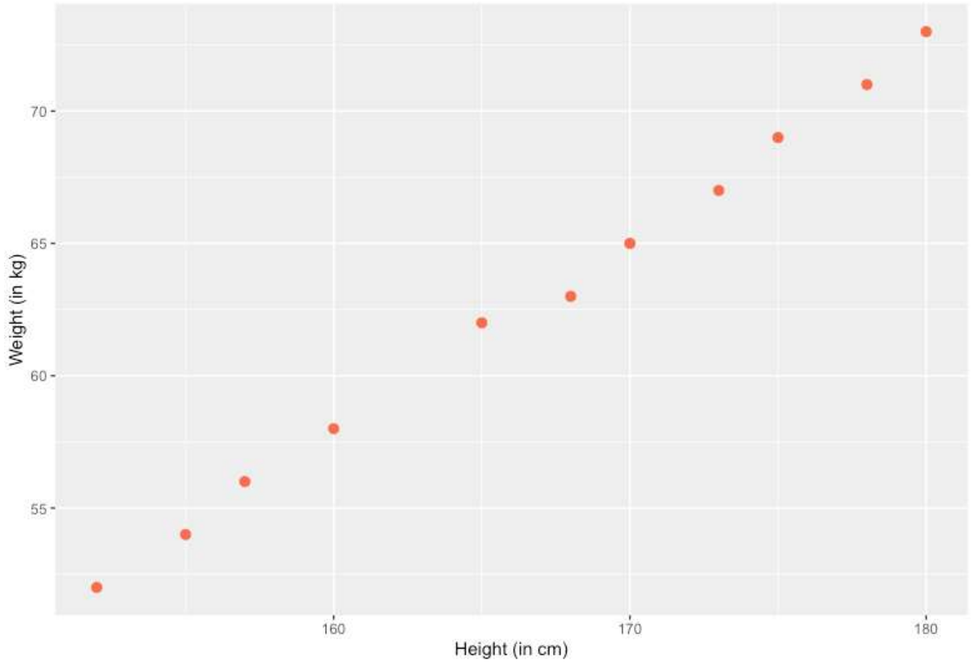
$$b_0 = y_1 - \frac{y_2 - y_1}{x_2 - x_1} x_1$$

$$b_1 = \frac{y_2 - y_1}{x_2 - x_1}$$

$\hat{Y}$ indicates a predicted or expected value

Linear Regression

What if we have more than 2 points?



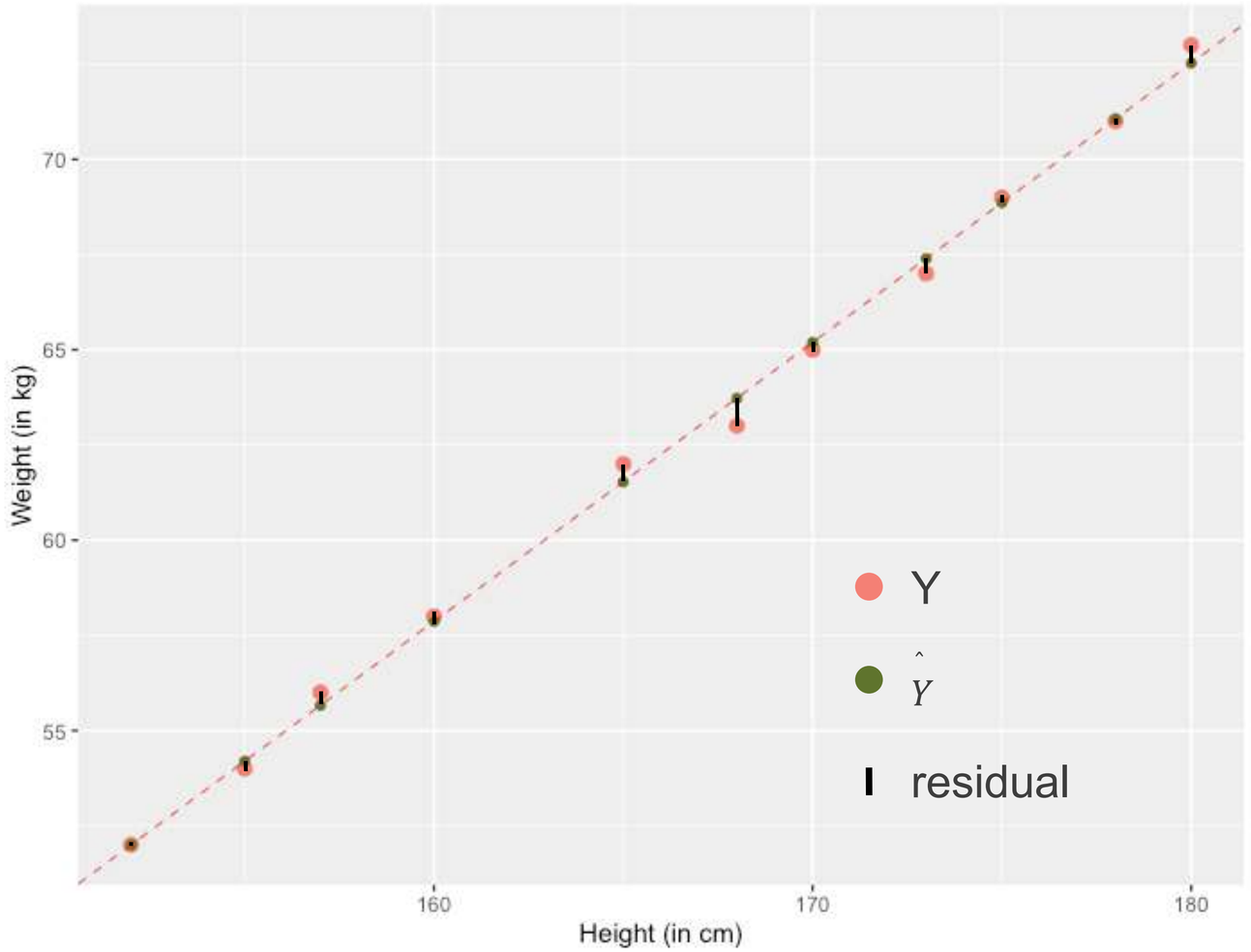
We estimate b_0 and b_1 by minimising the sum of the squared differences between the observed and the predicted values of the outcome

i.e. the minimisation of $\Sigma(Y - \hat{Y})^2$

The differences obtained by $Y - \hat{Y}$ are called residuals (or residual errors)

$\hat{Y}$ indicates a predicted or expected value

Linear Regression: Residuals



$$\hat{Y} = b_0 + b_1 \bullet X$$

The linear regression minimises the sum of the residuals

b_0 and b_1 Estimation

Using least squares we can estimate b_0 and b_1

$$b_1 = r\frac{S_y}{S_x}$$

and

$$b_0 = \hat{Y} - b_1\hat{X}$$

Where r is the sample correlation coefficient,

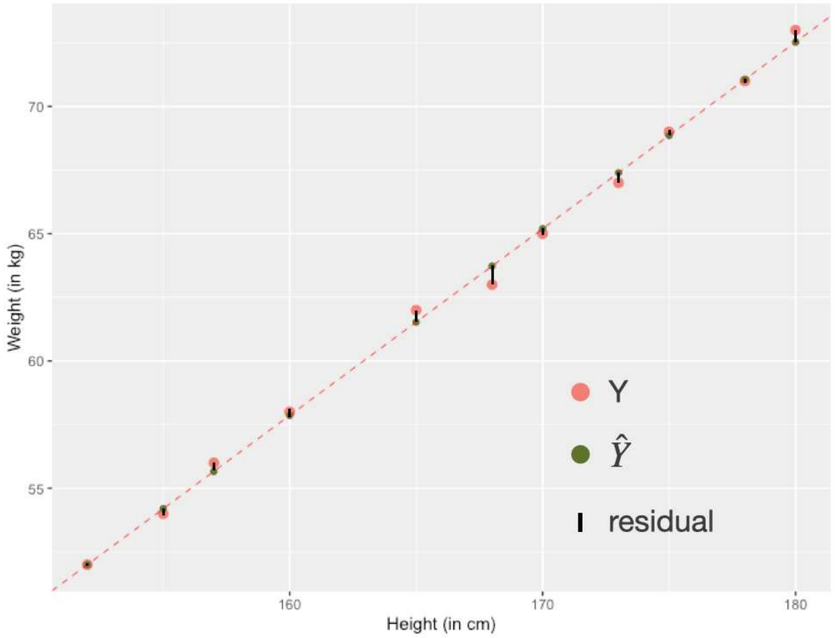
S_y is the square root of the variance of the dependent variable

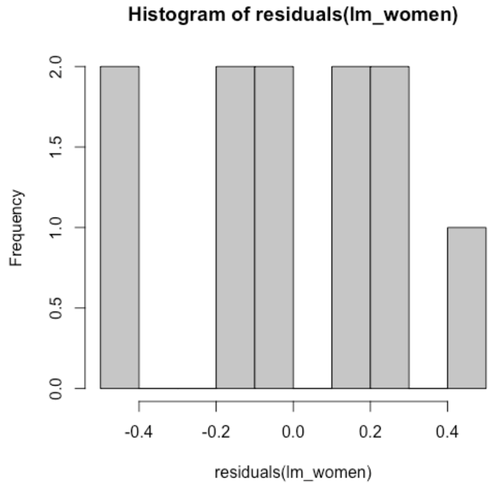
and S_x is the square root of the variance of the independent variable

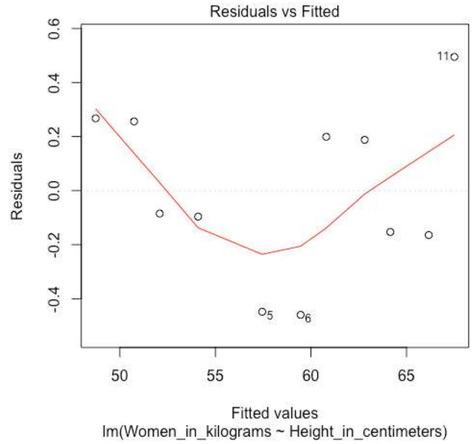
Linear Regression: Residuals

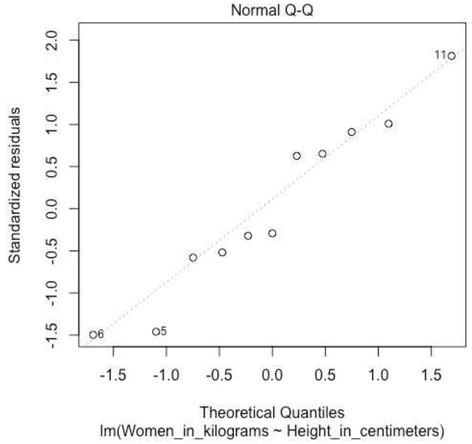
Ordinary least squares regression relies on some assumptions:

- The residuals are normally distributed and homoscedastic
- The errors are independent
- The relationships are linear



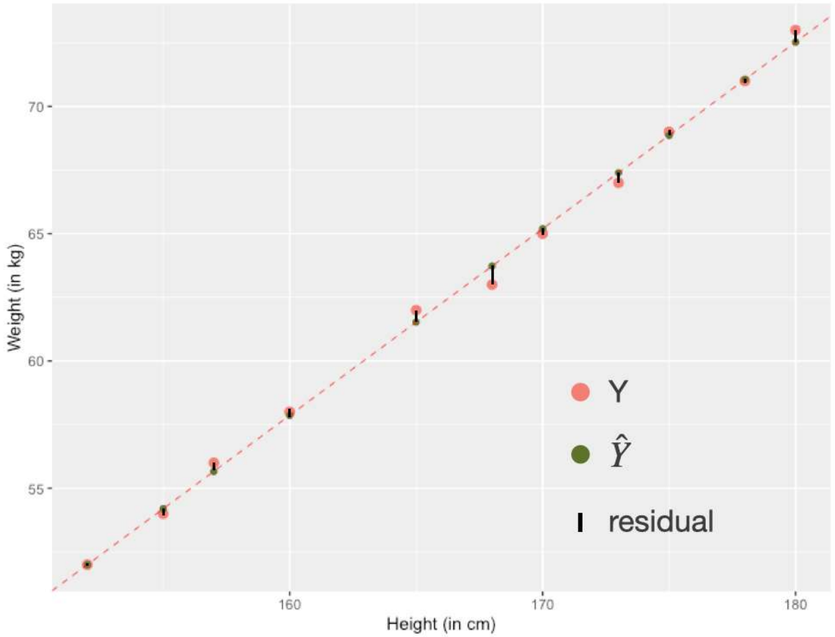




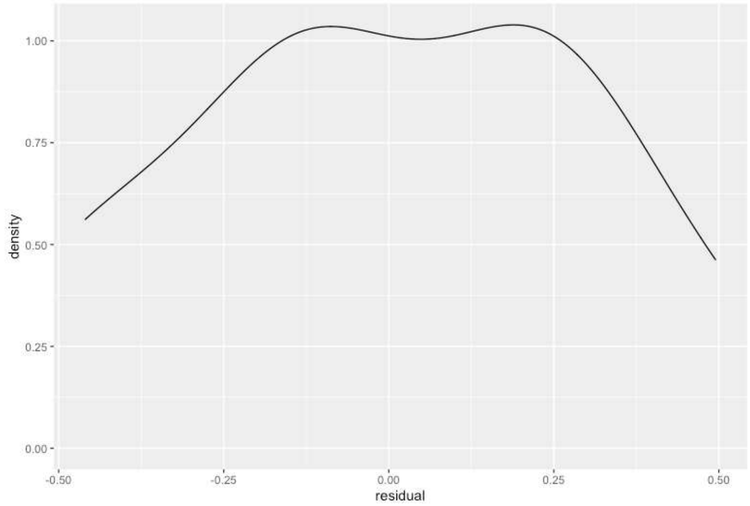


Goodness of Fit

In a simple linear regression with a perfect fit $\Sigma(Y - \hat{Y})^2 = 0$



Therefore, some linear regressions are better than others



Residual Standard Error = $\sqrt{\frac{\Sigma residuals^2}{\#Observations - \#Coeffcients}}$

$\hat{Y}$: predicted value

$\overline{Y}$: mean of y values

$$R^2 = 1 - \frac{\Sigma residuals^2}{\Sigma(Y - \overline{Y})^2}$$

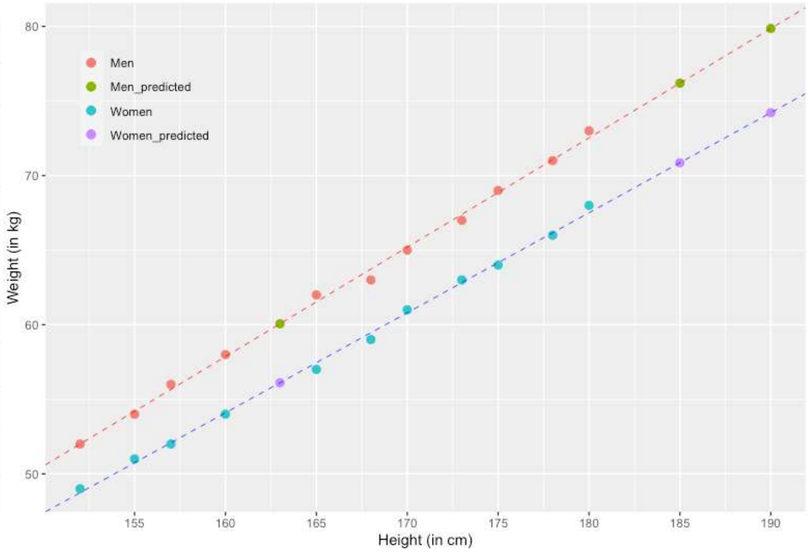
R^2

$$R^2 = 1 - \frac{\Sigma residuals^2}{\Sigma(Y - \overline{Y})^2}$$

R^2 or R-squared represents the proportion of the variance of the dependent variable that it is explained by the independent variable in the regression model

Height and Weight Correlation

Height (cm)	Women (kg)	Men (kg)
152	49	52
155	51	54
157	52	56
160	54	58
165	57	62
168	59	63
170	61	65
173	63	67
175	64	69
178	66	71
180	68	73



$$\text{Residual Standard Error} = \sqrt{\frac{\Sigma residuals^2}{\#Observations - \#Coeffcients}}$$

$$\text{(Multiple) } R^2 = 1 - \frac{\Sigma residuals^2}{\Sigma(Y - \overline{Y})^2}$$

```
>head(handw, 3)
```

	Height_in_centimeters	Women_in_kilograms	Men_in_kilograms
1	152	49	52
2	155	51	54
3	157	52	56

```
>cor(handw$Height_in_centimeters, handw$Women_in_kilograms)
[1] 0.9988705
>(lm_women <- lm(Women_in_kilograms ~ Height_in_centimeters,
                 data = handw))

Call:
lm(formula = Women_in_kilograms ~ Height_in_centimeters, data = handw)

Coefficients:
         (Intercept)  Height_in_centimeters
            -53.1763                 0.6705

>predict(lm_women,
        data.frame(Height_in_centimeters = c(163, 185, 190)))
       1        2        3
56.10745 70.85740 74.20966

>residuals(lm_women)
          1           2           3           4           5           6           7
 0.26752913  0.25617223 -0.08473237 -0.09608928 -0.44835078 -0.45970768  0.19938771
          8           9          10          11
 0.18803081 -0.15287379 -0.16423069  0.49486470

>summary(lm_women)

Call:
lm(formula = Women_in_kilograms ~ Height_in_centimeters, data = handw)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45971 -0.15855 -0.08473  0.22778  0.49486

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -53.17628    1.77414  -29.97  2.5e-10 ***
Height_in_centimeters   0.67045    0.01063   63.07  3.2e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3225 on 9 degrees of freedom
Multiple R-squared:  0.9977,    Adjusted R-squared:  0.9975
F-statistic:  3977 on 1 and 9 DF,  p-value: 3.196e-13
```

Significance

Is Your Model the Best Fit for Your Data?

STATISTIC	CRITERION
R-Squared	Higher the better (> 0.70)
Adj R-Squared	Higher the better
F-Statistic	Higher the better
Std. Error	Closer to zero the better
t-statistic	Should be greater 1.96 for p-value to be less than 0.05
AIC	Lower the better
BIC	Lower the better
Mallows cp	Should be close to the number of predictors in model
MAPE (Mean absolute percentage error)	Lower the better
MSE (Mean squared error)	Lower the better
Min_Max Accuracy => mean(min(actual, predicted)/max(actual, predicted))	Higher the better

http://r-statistics.co/Linear-Regression.html

Outliers

The Cook's distance: To identify data points that negatively affect your regression model (influential outliers)

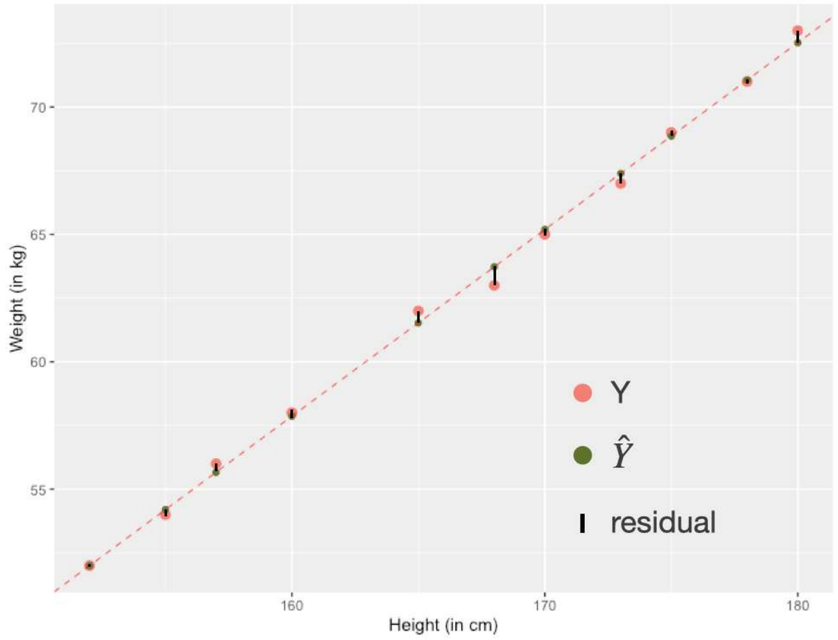
The calculation is similar to follow a *leave-one-out* approach

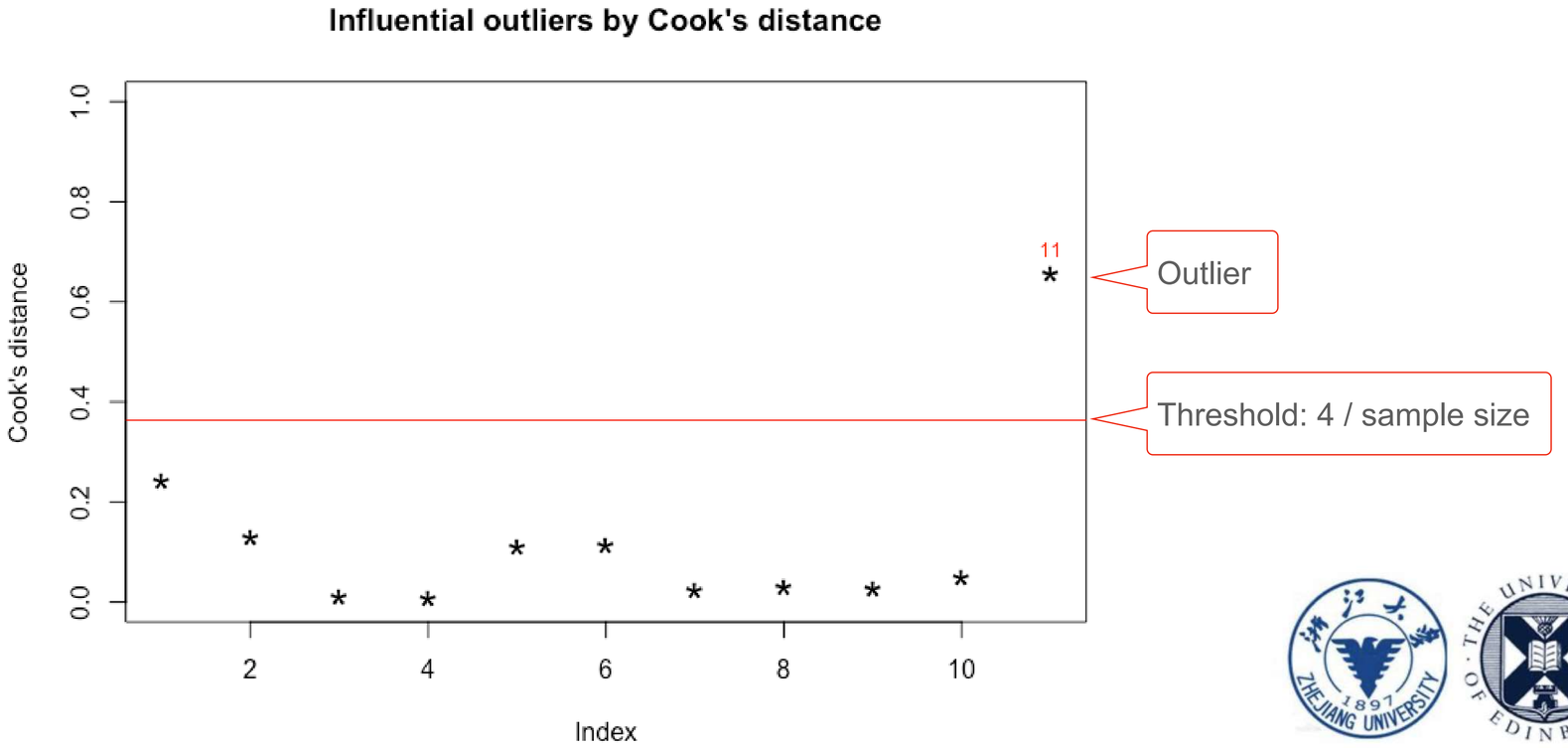
$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{y}_{j(i)})^2}{(p+1)S^2}$$

$\hat{y}_{j(i)}$ fitted response value when excluding i

p number of regression coefficients

S^2 mean squared error of the regression model





World Bank Dataset

```
>head(WBd <- read.csv("WDI.tsv", sep = "\t"))
```

Gross domestic product per capita based on purchased power parity

Total population

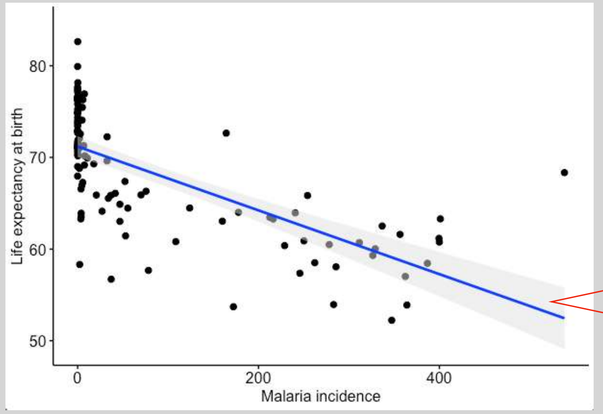
Fertility rate (Births per woman)

Life expectancy at birth

Individuals using internet

	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

```
>ggpubr::ggscatter(WBd, x = "malaria_incidence", y = "lifeexpectancy", add = "reg.line",
                   add.params = list(color = "blue", fill = "lightgray"), conf.int = TRUE) +
  labs(x = "Malaria incidence", y = "Life expectancy at birth", colour = "")
```

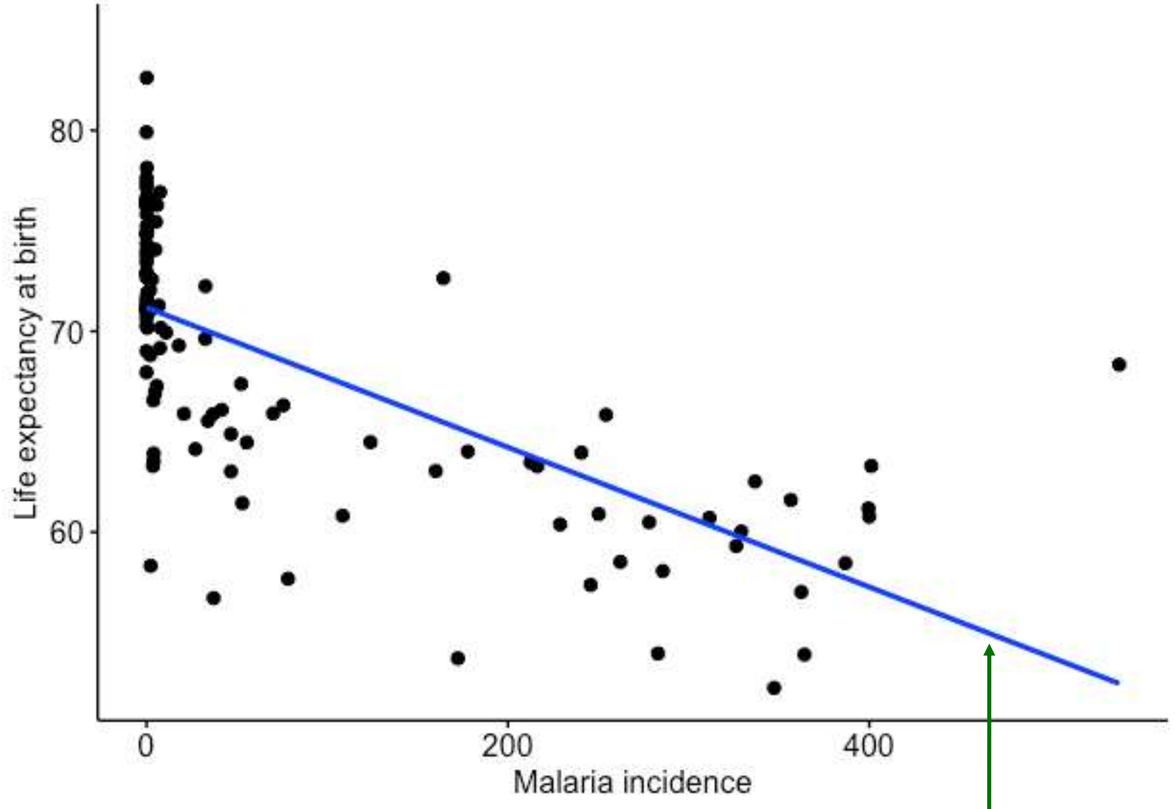


95% confidence interval

```
>summary(lm(lifeexpectancy ~ malaria_incidence, WBd))
```

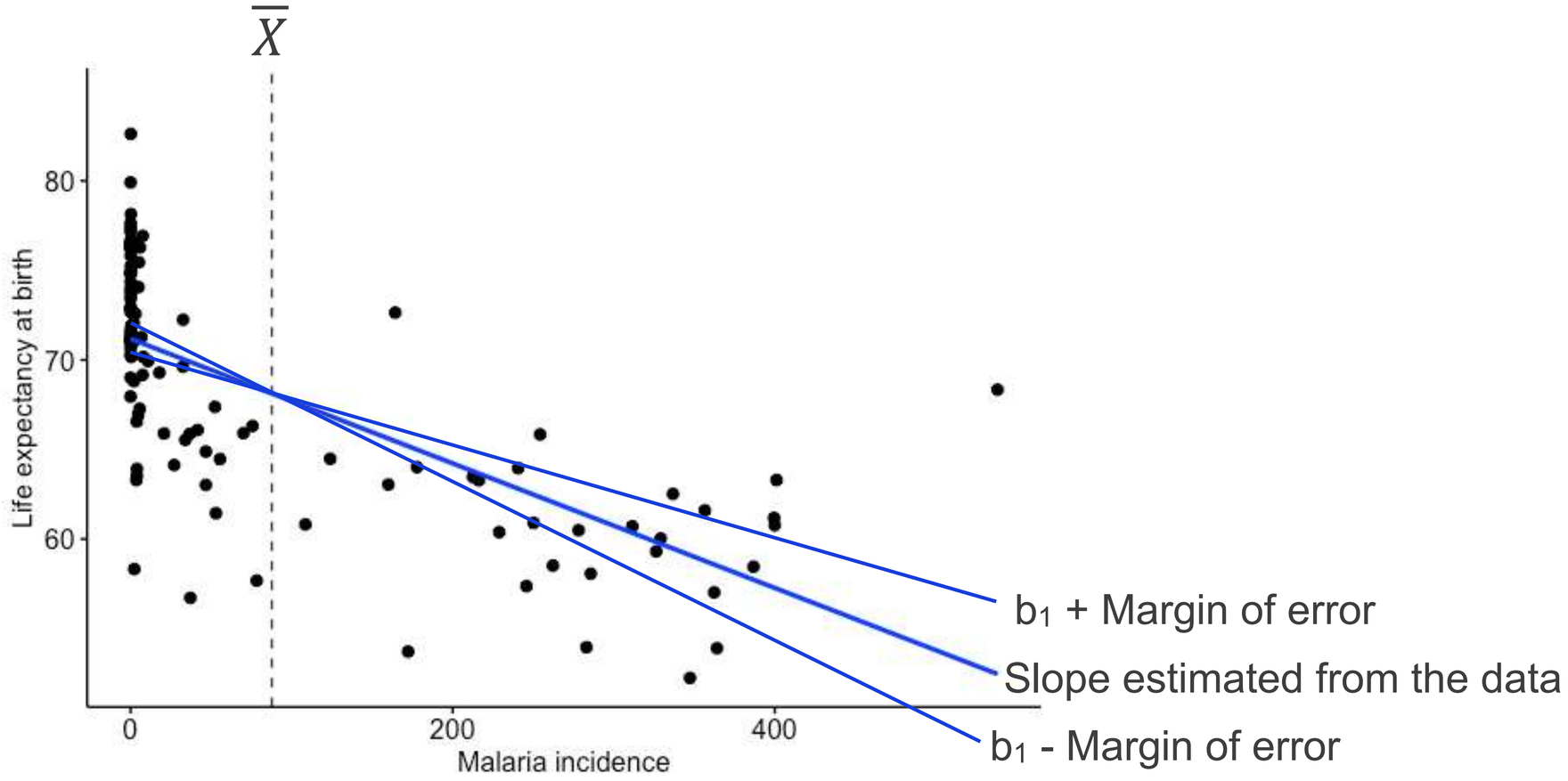
```
Residuals:
     Min       1Q   Median       3Q      Max
-13.1803  -3.0089  -0.0516   3.5594  15.8892

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.18129    0.57398 124.013  < 2e-16 ***
malaria_incidence -0.03479    0.00359  -9.692 3.02e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.965 on 105 degrees of freedom
  (110 observations deleted due to missingness)
Multiple R-squared:  0.4722,    Adjusted R-squared:  0.4671
F-statistic: 93.93 on 1 and 105 DF,  p-value: 3.02e-16
```

Confidence Interval

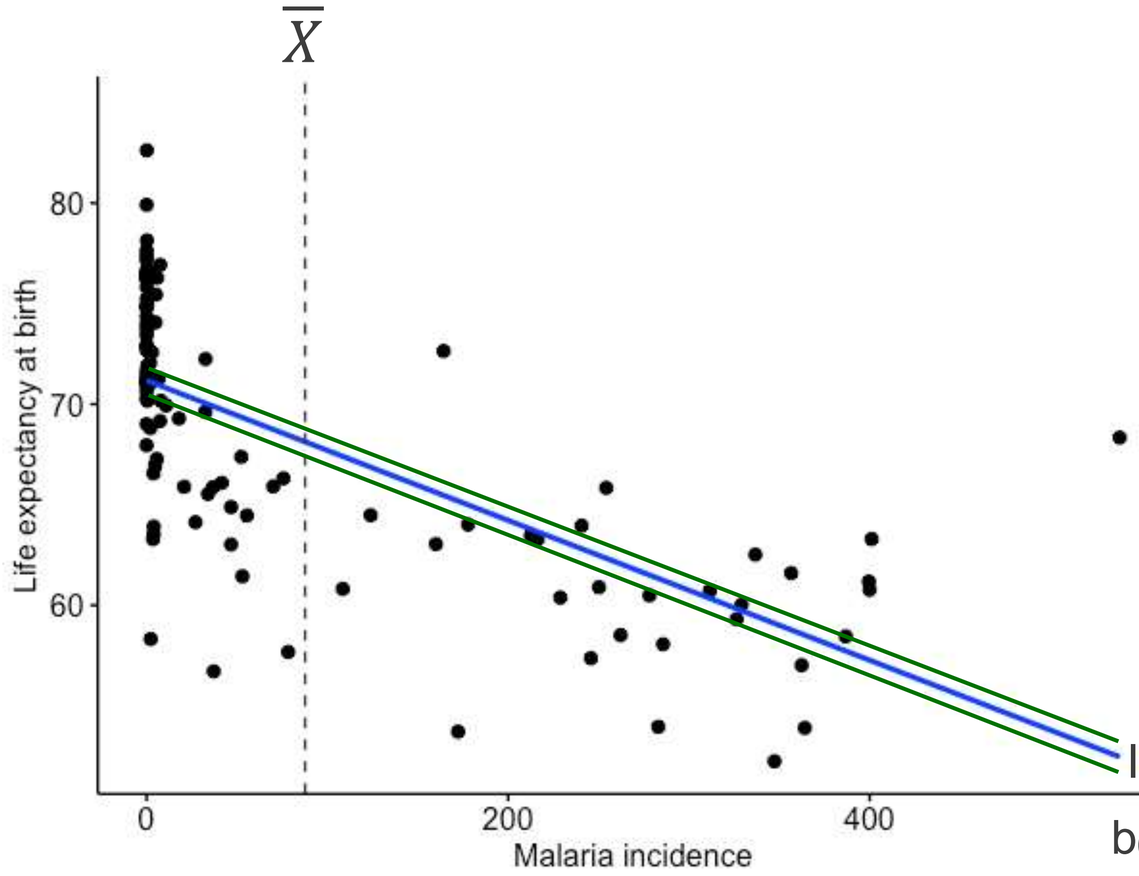


$$Lifeexpectancy = 71.18 - 0.03 \cdot Malariaincidence$$

Confidence Interval



Confidence Interval

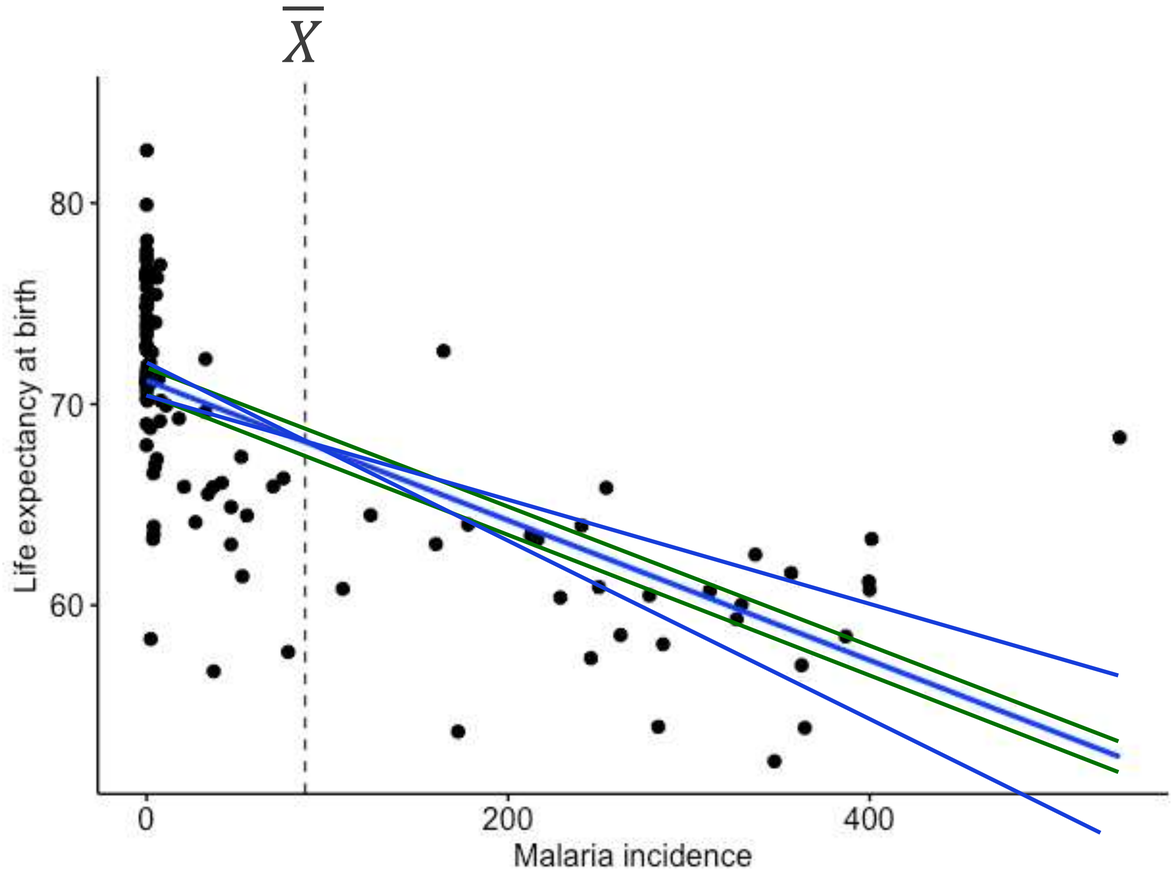


b_0 + Margin of error

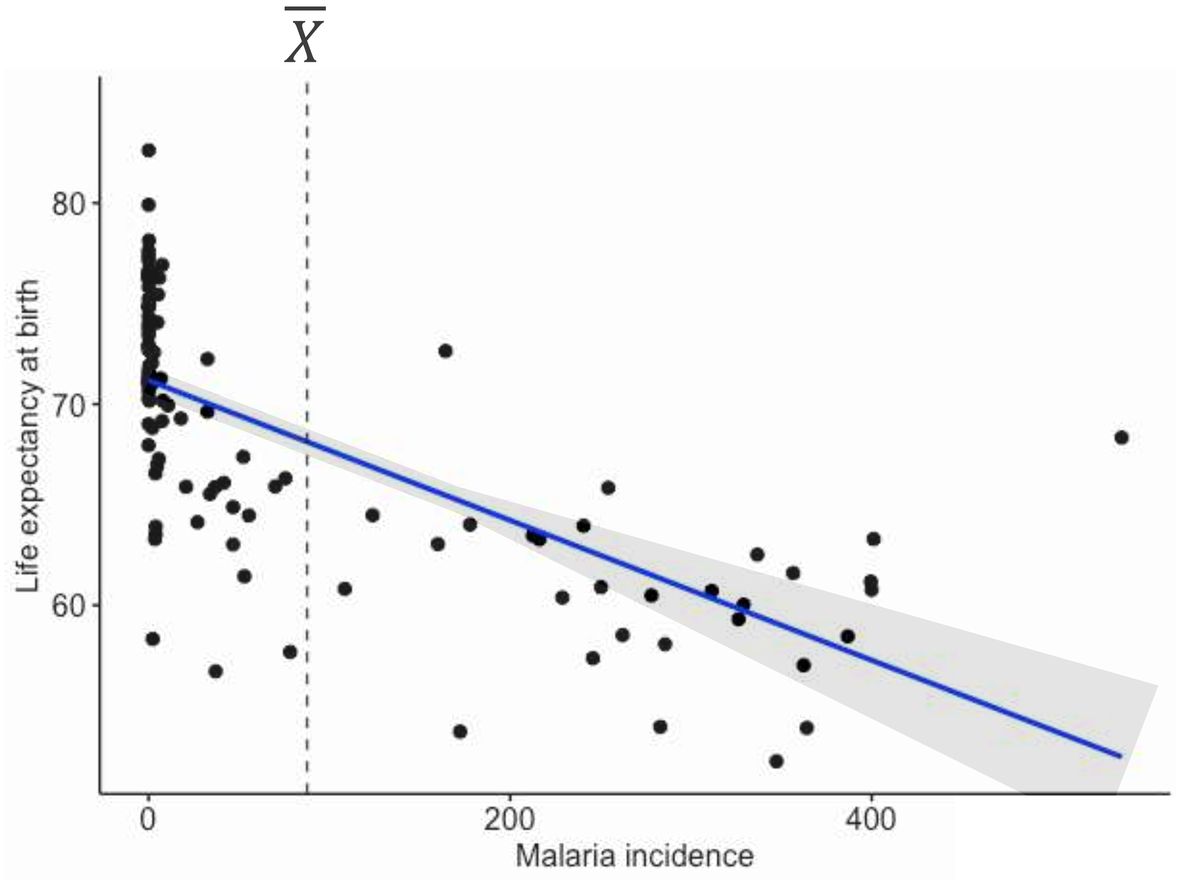
Intercept estimated from the data

b_0 - Margin of error

Confidence Interval



Confidence Interval



World Bank Dataset

```
>head(WBd <- read.csv("WDI.tsv", sep = "\t"))
```

Gross domestic product per capita based on purchased power parity

Total population

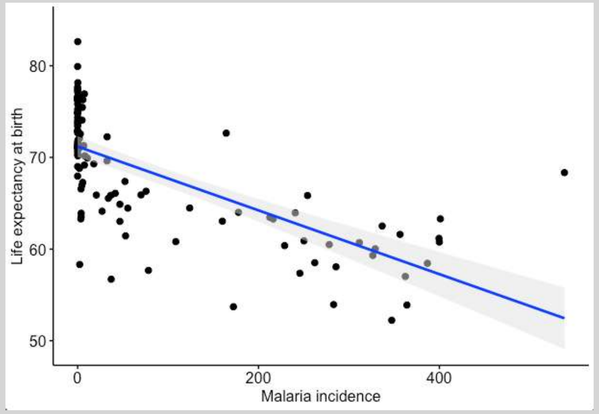
Fertility rate (Births per woman)

Life expectancy at birth

Individuals using internet

	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

```
>ggpubr::ggscatter(WBd, x = "malaria_incidence", y = "lifeexpectancy", add = "reg.line",
                   add.params = list(color = "blue", fill = "lightgray"), conf.int = TRUE) +
  labs(x = "Malaria incidence", y = "Life expectancy at birth", colour = "")
```




```
>summary(lm(lifeexpectancy ~ malaria_incidence, WBd))
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-13.1803  -3.0089  -0.0516   3.5594  15.8892

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       71.18129    0.57398 124.013  < 2e-16 ***
malaria_incidence -0.03479    0.00359  -9.692 3.02e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.965 on 105 degrees of freedom
  (110 observations deleted due to missingness)
Multiple R-squared:  0.4722,    Adjusted R-squared:  0.4671
F-statistic: 93.93 on 1 and 105 DF,  p-value: 3.02e-16
```

World Bank Dataset

```
>head(WBd <- read.csv("WDI.tsv", sep = "\t"))
```

Gross domestic product per capita based on purchased power parity

Total population

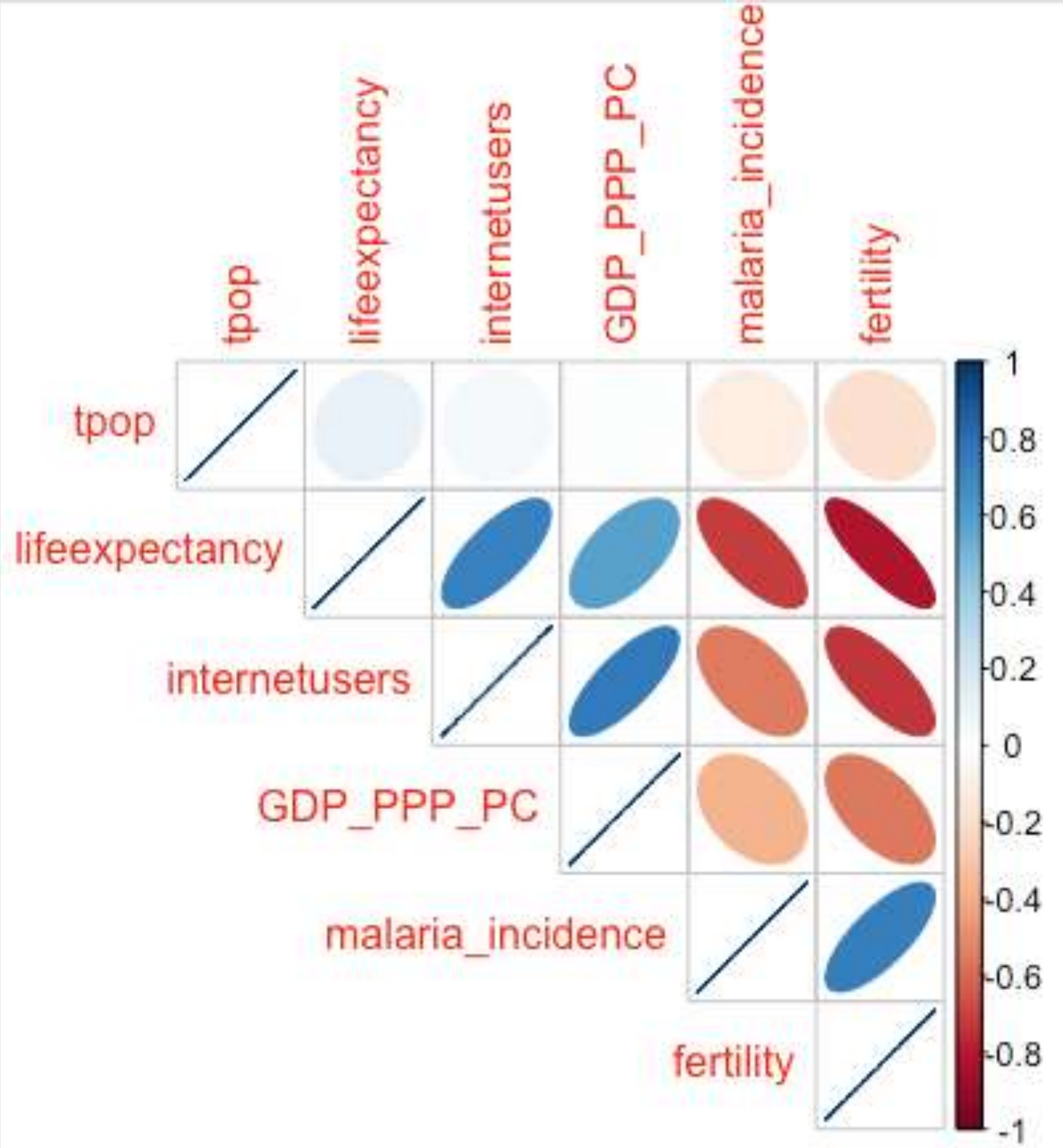
Fertility rate (Births per woman)

Life expectancy at birth

Individuals using internet

	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

```
># Removing rows with NAs
>withNAs <- apply(WBd[ , 3:8], 1, function(x){
  sum(is.na(x))
>})
>corrplot::corrplot(cor(WBd[withNAs == 0, 3:8]), order = "AOE",
  method = "ellipse", type = "upper")
```



World Bank Dataset

```
>head(WBd <- read.csv("WDI.tsv", sep = "\t"))
```

Gross domestic product per capita based on purchased power parity | Total population | Fertility rate (Births per woman) | Life expectancy at birth | Individuals using internet

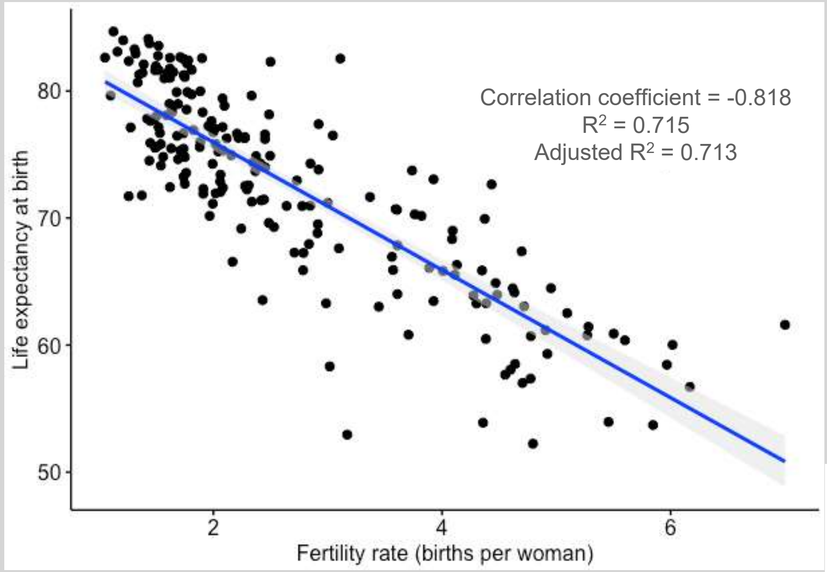
	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

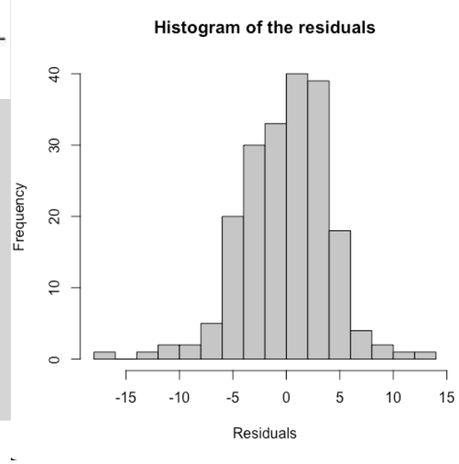
```
># Removing rows with NAs
>withNAs <- apply(WBd[ , 3:8], 1, function(x){
  sum(is.na(x))
>})
>corrplot::corrplot(cor(WBd[withNAs == 0, 3:8]), order = "AOE",
  method = "ellipse", type = "upper", tl.pos = "d")
>corrplot::corrplot(cor(WBd[withNAs == 0, 3:8]), order = "AOE",
  add = TRUE, method = "number", type = "lower", diag = FALSE,
  tl.pos = "n", cl.pos = "n")
```

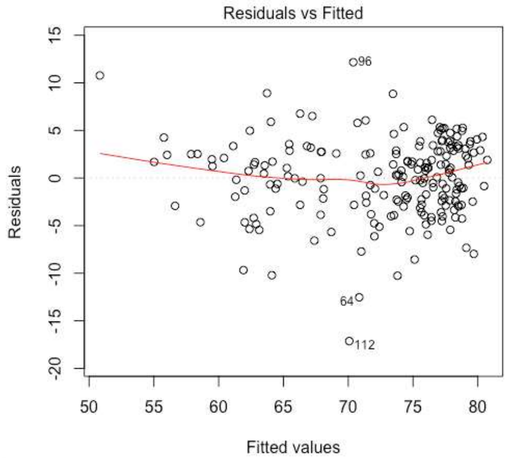
World Bank Dataset

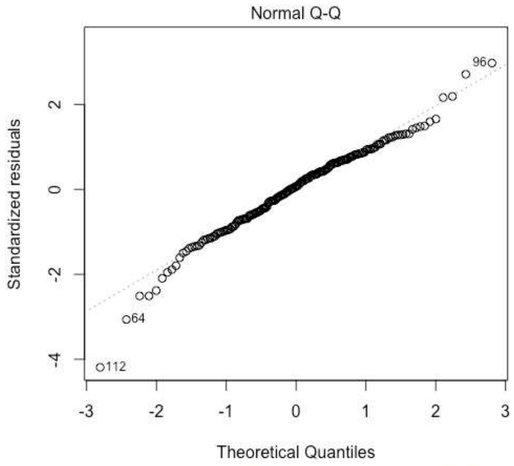
	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

```
> # Fertility vs life expectancy
```



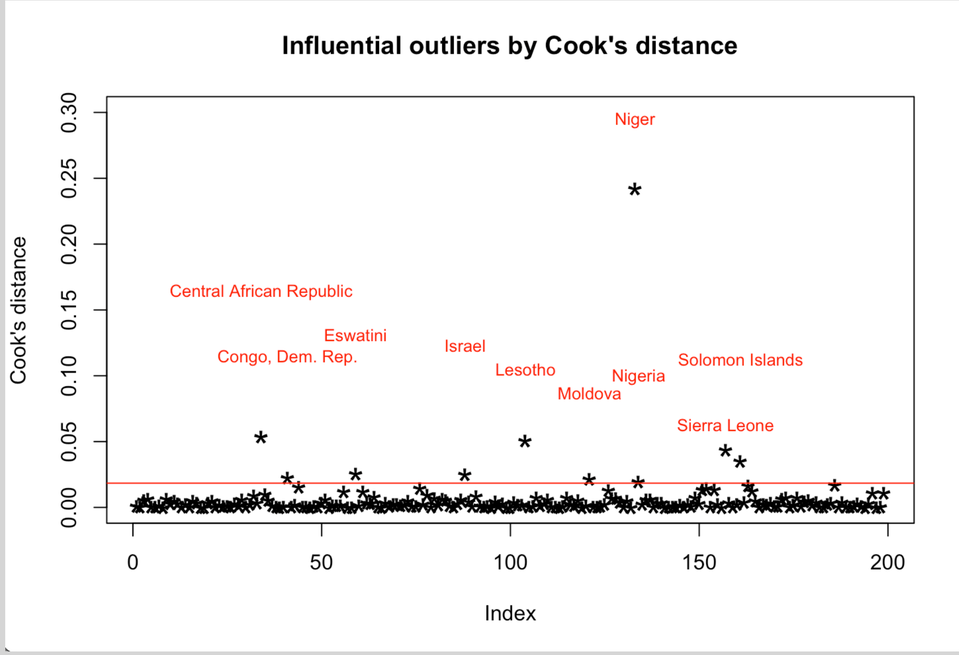


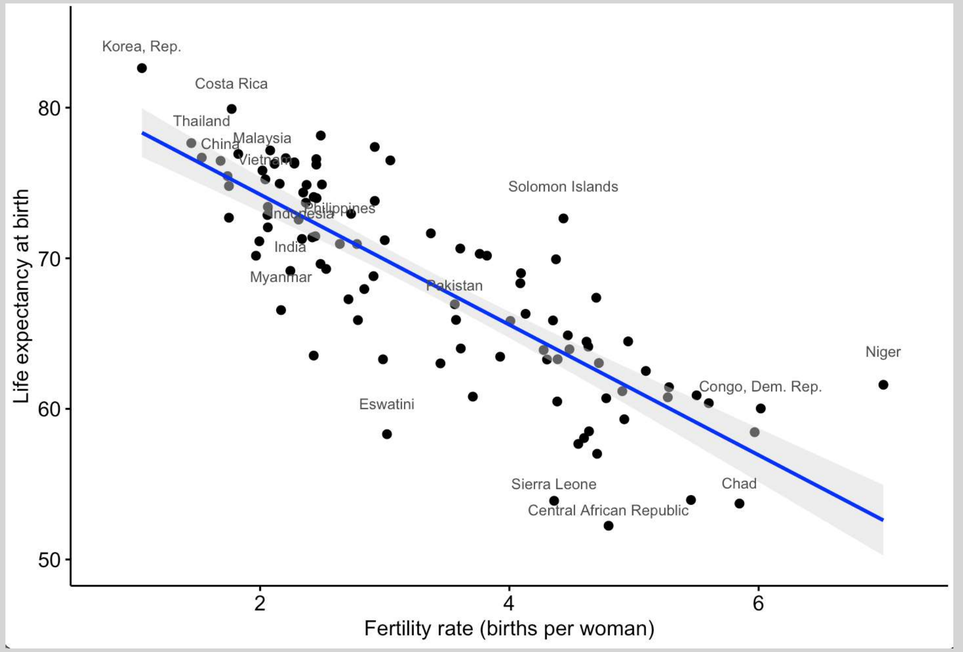




Outliers in Life Expectancy vs Fertility Rate

```
>cooksd <- cooks.distance(lm(lifeexpectancy ~ fertility, WBd))
>sample_size <- nrow(WBd)
# Not NAs
>not_nas <- which(!is.na(WBd$lifeexpectancy) & !is.na(WBd$fertility))
# Plot Cook's distance
>plot(cooksd, pch = "*", cex = 2, main = "Influential outliers by Cook's distance",
      ylim = c(0, 0.3), ylab = "Cook's distance")
>abline(h = 4/sample_size, col = "red")  # add cutoff line
>text(x = 1:length(cooksd), y = cooksd + 0.05, col = "red", cex = 0.8,
      labels = ifelse(cooksd > 4 / sample_size, names(cooksd), ""))  # add labels
```





Life Expectancy as the Outcome

Fertility rate and Life expectancy correlate

Life expectancy correlates with Wealth (GDP PC)

Fertility rate correlates with Poverty (1 / GDP PC)

Is Life expectancy a function of both Fertility and Wealth?

$$LifeExpectancy = f(Fertility, GDP) = b_0 + b_1 Fertility + b_2 GDP$$

	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

Multiple Linear Regression

```
>lm_le_fe_GDP <- lm(lifeexpectancy ~ fertility + GDP_PPP_PC, WBd)
>summary(lm_le_fe_GDP)

# Is malaria incidence a predictor of Life expectancy?
```

```
Call:
lm(formula = lifeexpectancy ~ fertility + GDP_PPP_PC, data = WBd)

Residuals:
     Min       1Q   Median       3Q      Max
-15.6187  -2.0670   0.2578   2.4264   9.9080

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.5741798  0.9620719  83.751  < 2e-16 ***
fertility   -3.8801390  0.2606233 -14.888  < 2e-16 ***
GDP_PPP_PC   0.0001062  0.0000151   7.028 4.18e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.665 on 180 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.7692,    Adjusted R-squared:  0.7666
F-statistic:   300 on 2 and 180 DF,  p-value: < 2.2e-16
```

LE_Fe: Adj R^2 = 0.713

LE_Fe_GDP: Adj R^2 = 0.767

Life Expectancy as the Outcome

Fertility rate and Life expectancy correlate

Life expectancy correlates with Wealth (GDP PC)

Fertility rate correlates with Poverty (1 / GDP PC)

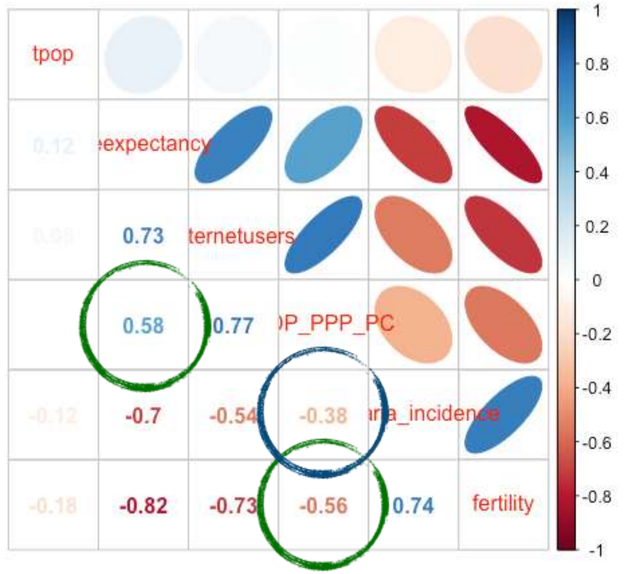
Is Life expectancy a function of Fertility, Wealth and Malaria?

$$LifeExpectancy = f(Fertility, GDP, Malariaincidence)$$

```
>lm_le_fer_GDP_Mal <-
                      lm(lifeexpectancy ~ fertility + GDP_PPP_PC + malaria_incidence, WBd)
>summary(lm_le_fer_GDP_Mal)
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.778e+01  1.693e+00  45.943  < 2e-16 ***
fertility         -2.881e+00  4.837e-01  -5.957 4.23e-08 ***
GDP_PPP_PC         1.267e-04  4.322e-05   2.931  0.00423 **
malaria_incidence -1.105e-02  4.077e-03  -2.710  0.00797 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.706 on 96 degrees of freedom
  (117 observations deleted due to missingness)
Multiple R-squared:  0.7137,    Adjusted R-squared:  0.7047
F-statistic: 79.75 on 3 and 96 DF,  p-value: < 2.2e-16
```

	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

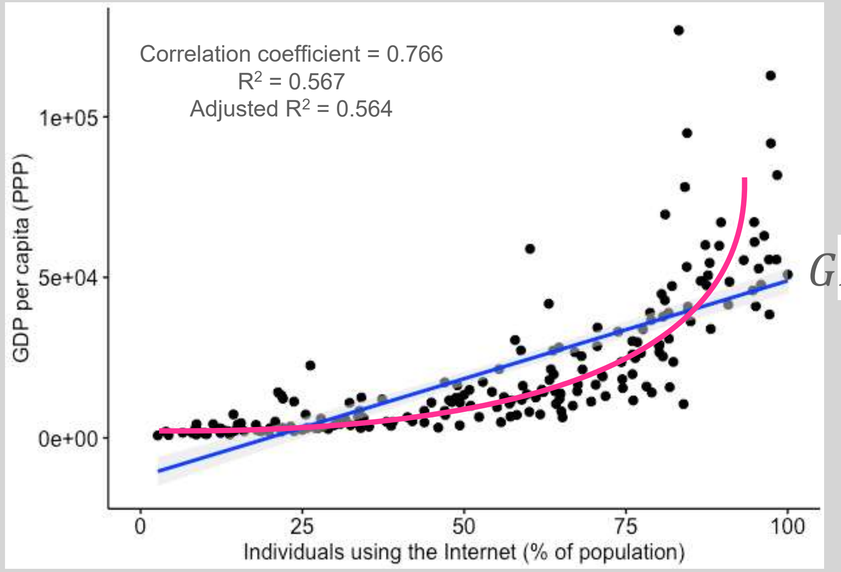


LE_Fe: Adj R^2 = 0.713

LE_Fe_GDP: Adj R^2 = 0.767

Internet Users vs GDP per Capita (PPP)

```
>ggpubr::ggscatter(WBd, x = "internetusers", y = "GDP_PPP_PC",
  add = "reg.line",
  add.params = list(color = "blue", fill = "lightgray"),
  conf.int = TRUE) +
  labs(x = "Individuals using the Internet (% of population)",
       y = "GDP per capita (PPP)", colour = "")
```



$$GDP = a \cdot e^{r \cdot InternetUsers}$$

	Country.Name	Country.Code	GDP_PPP_PC	malaria_incidence	tpop	fertility	lifeexpectancy	internetusers
1	Afghanistan	AFG	2058.384	27.07227	36296400	4.633	64.130	13.5000
2	Albania	ALB	13037.010	NA	2873457	1.638	78.333	71.8470
3	Algeria	DZA	11737.409	0.00000	41389198	3.045	76.499	47.6911
4	American Samoa	ASM	NA	NA	55620	NA	NA	NA
5	Andorra	AND	NA	NA	77001	NA	NA	91.5675
6	Angola	AGO	7310.902	228.90894	29816748	5.600	60.379	14.3391

Internet Users vs GDP per Capita (PPP)

$$GDP =$$

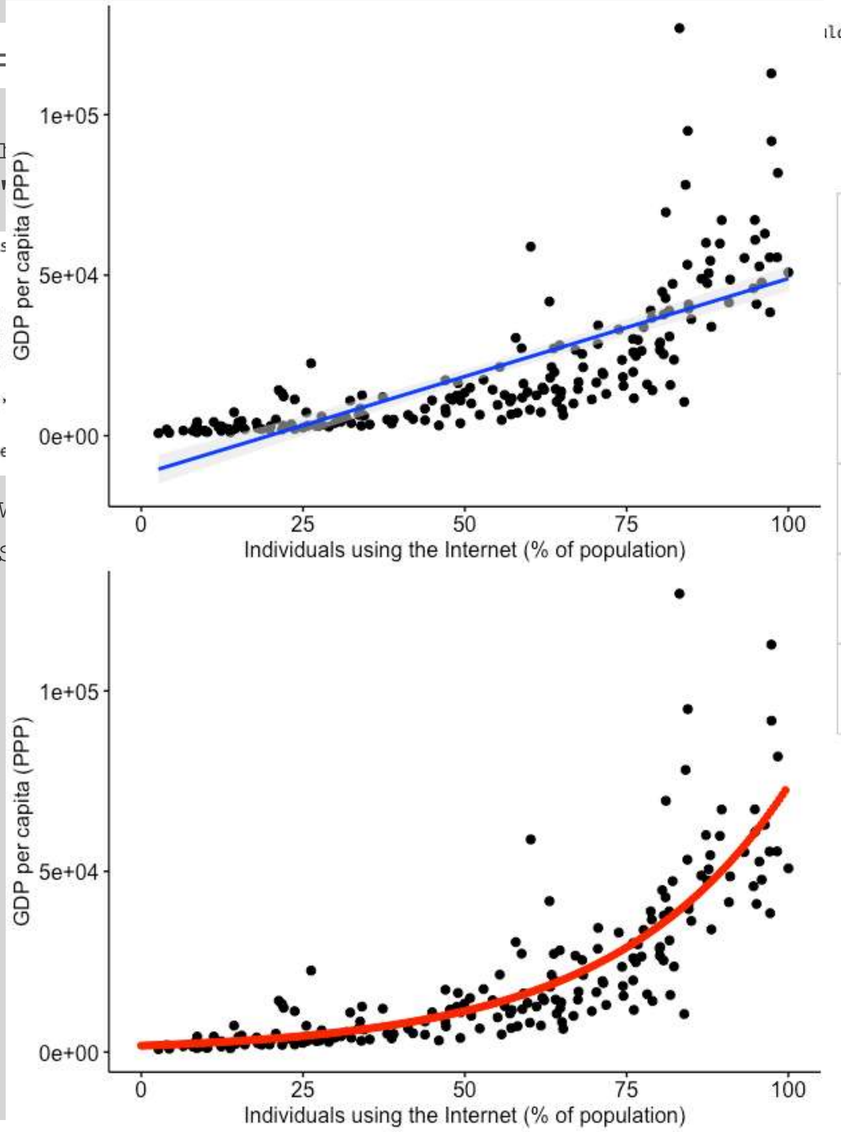
```
>nls_GDP <- nls(GDP_PPP_PC ~ exp(r*internetuser
                start = list(r = 0.02), alg = '
>summary(nls_GDP)
```

```
Formula: GDP_PPP_PC ~ exp(r * internetusers

Parameters:
      Estimate Std. Error t value Pr(>|t|)
r    3.742e-02  2.523e-03  14.834  < 2e-16
.lin 1.748e+03  3.828e+02   4.566 8.96e-06
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'

Residual standard error: 12480 on 188 degre
```

```
>df <- data.frame(x = head(seq(0, 100, 100/nrow
>df$y <- predict(nls_GDP, data.frame(internetus
>ggpubr::ggscatter(WBd, x = "internetusers",
                        y = "GDP_PPP_PC") +
  labs(x = "Individuals using the Internet
            (% of population)",
       y = "GDP per capita (PPP)") +
  geom_point(aes(x = df$x, y = df$y),
             col = "red")
```

Fitting an exponential curve is not a linear regression

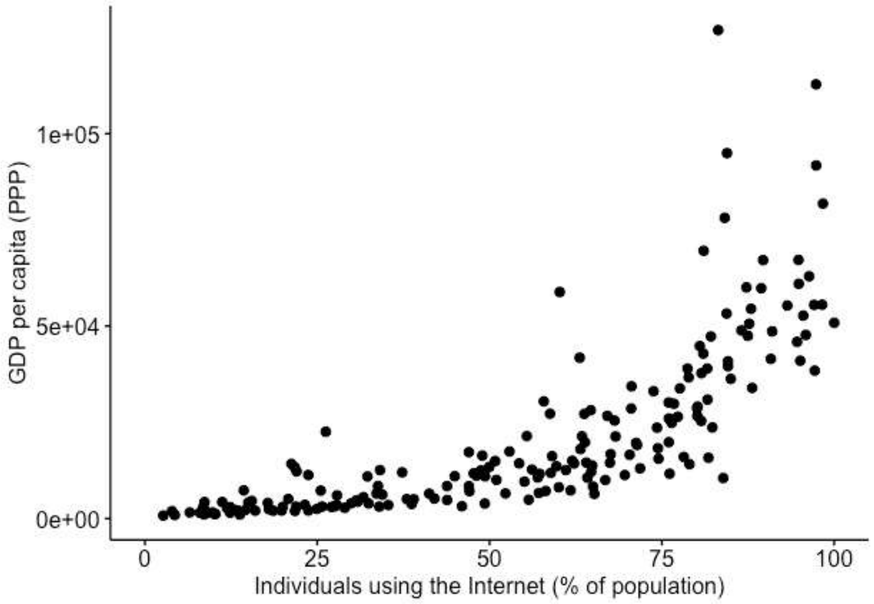
But there is a trick!

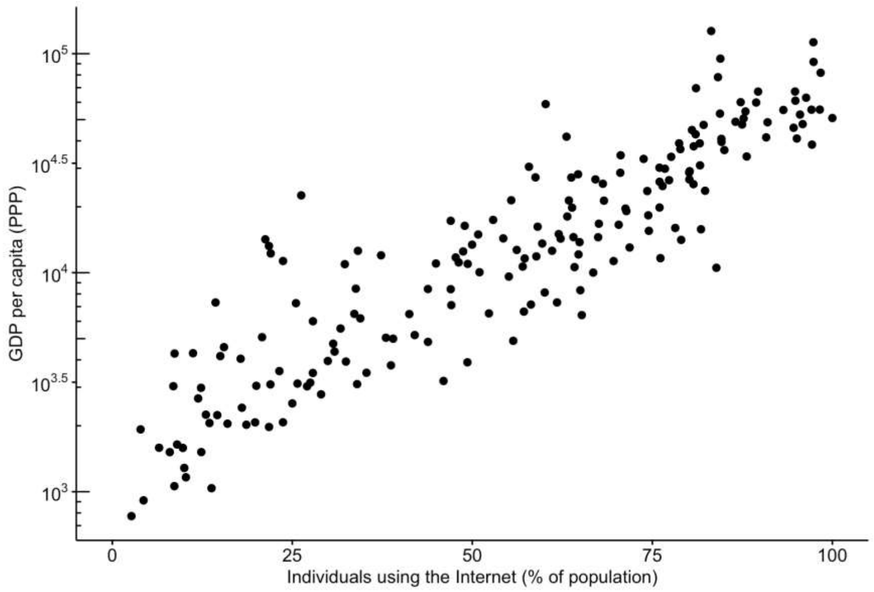


ılaria_incidence	tpop	fertility	lifeexpectancy	internetusers
27.07227	36296400	4.633	64.130	13.5000
NA	2873457	1.638	78.333	71.8470
0.00000	41389198	3.045	76.499	47.6911
NA	55620	NA	NA	NA
NA	77001	NA	NA	91.5675
228.90894	29816748	5.600	60.379	14.3391

Internet Users vs GDP per Capita (PPP)

Axis transformed to Log10 scale

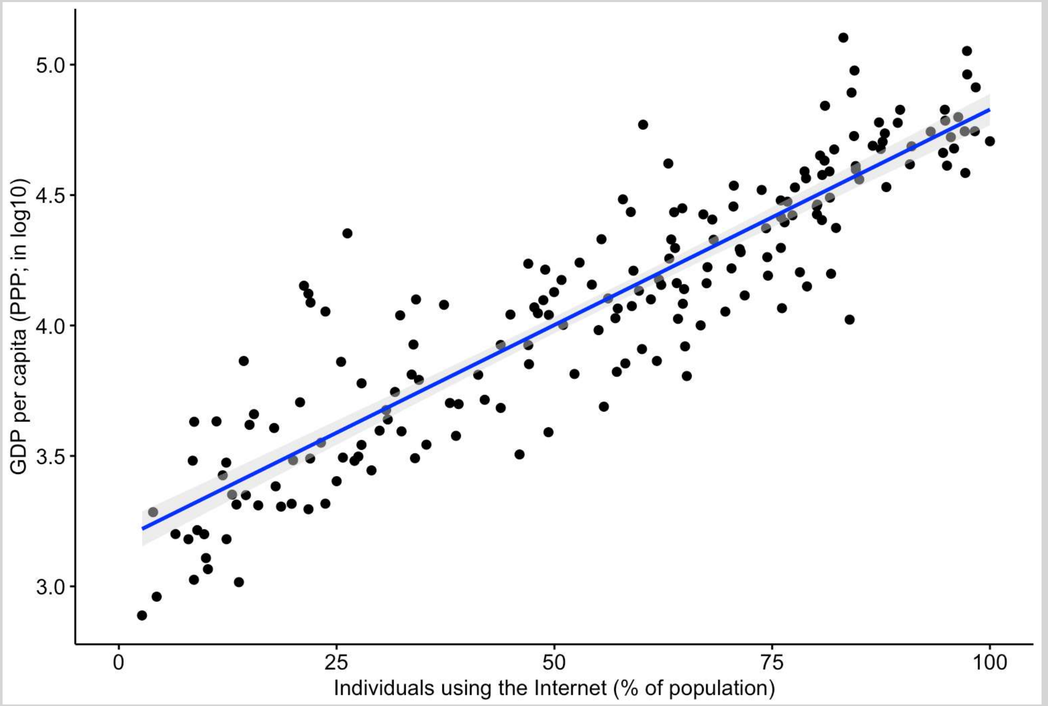




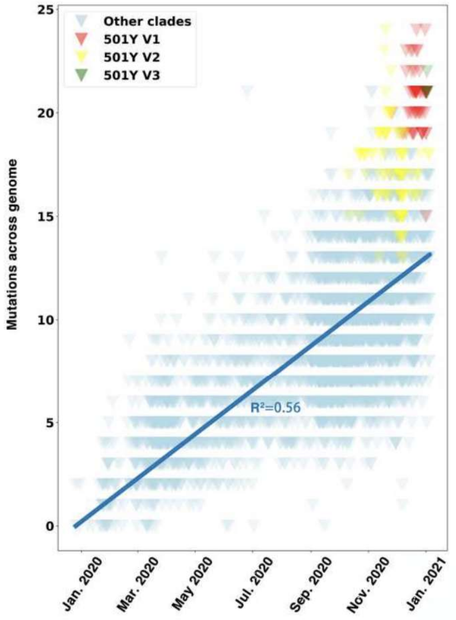
Internet Users vs GDP per Capita (PPP)

```
>WBd$log10GDP <- log10(WBd$GDP_PPP_PC)
>cor(WBd$log10GDP, WBd$internetusers, use = "complete.obs")
[1] 0.9018969
>lm_log10GDP <- lm(log10GDP ~ internetusers, WBd)
```

```
Call:
lm(formula = log10GDP ~ internetusers, data = WBd)

Residuals:
     Min       1Q   Median       3Q      Max
-0.53951 -0.12952 -0.02846  0.11727  0.74365

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.176032   0.035377   89.78   <2e-16 ***
internetusers 0.016520   0.000577   28.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2187 on 188 degrees of freedom
  (27 observations deleted due to missingness)
Multiple R-squared:  0.8134,    Adjusted R-squared:  0.8124
F-statistic: 819.6 on 1 and 188 DF,  p-value: < 2.2e-16
```

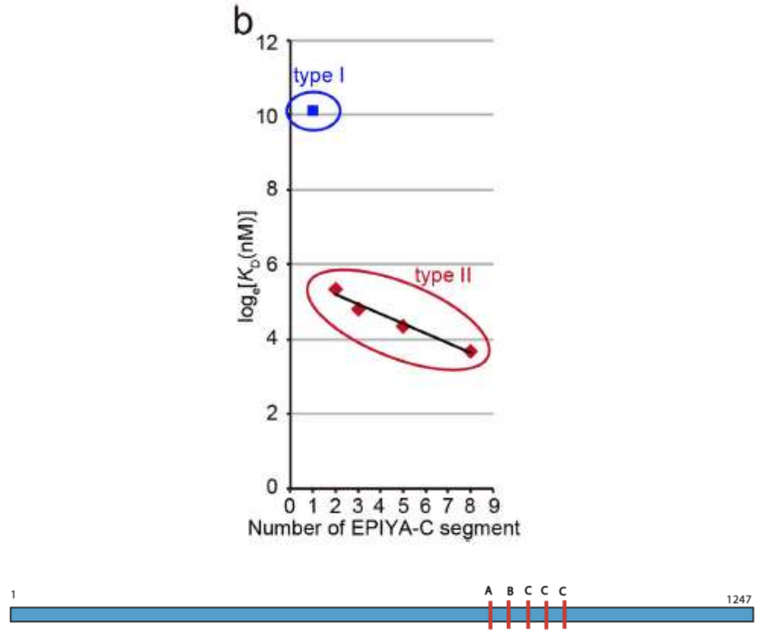


Applications



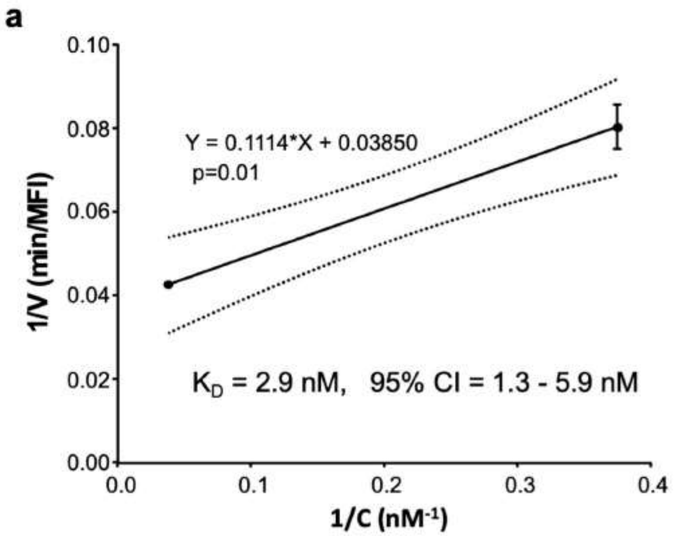
The accumulated mutations in SARS CoV-2 strains compared with early reference strain since January 2021 shows an approximate of **one new mutation** every 29 days.

Wu, A., et al. (2021) One Year of SARS-CoV-2 Evolution. *Cell Host and Microbe*



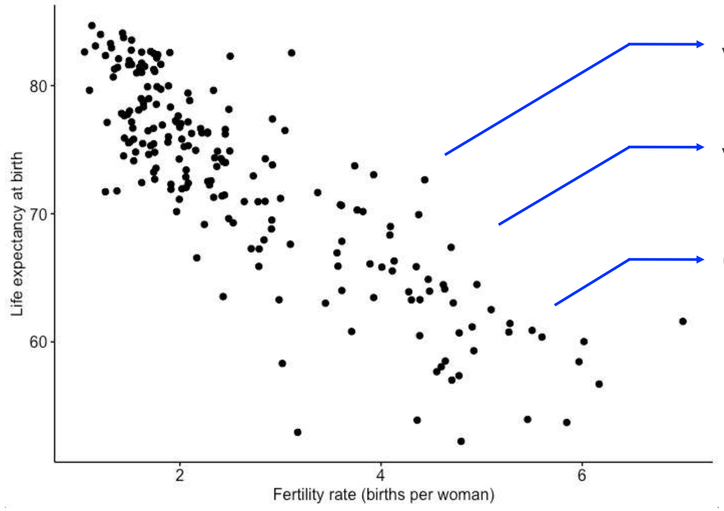
The affinity of the *Helicobacter pylori* CagA protein for the human SHP2/SH2 domain **increases linearly** with the number of repeats of the EPIYA-C motif. Blue: Western type I *H. pylori* strains. Red: Western type II *H. pylori* strains.

Nagase, L., et al. (2015) Dramatic increase in SHP2 binding activity of *Helicobacter pylori* Western CagA by EPIYA-C duplication: its implications in gastric carcinogenesis. *Scientific Reports.*



Kinetics of binding between the recombinant monoclonal antibody mAb7899 and PfGARP-A measured at two antibody concentrations.

Raj, D.K., et al. (2020) Anti-PfGARP activates programmed cell death of parasites and reduces severe malaria. *Nature.*



Variance in x: S_x^2

Variance in y: S_y^2

Covariance: $Cov(x, y)$

Sample Correlation Coefficient: r

[-1, 1]

Summary

Simple Linear Regression

$$\hat{Y} = b_0 + b_1 X$$

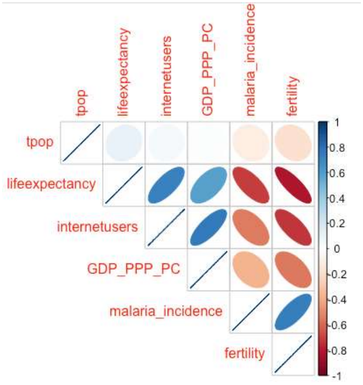
Intercept: b_0

Slope: b_1

Estimated using least squares

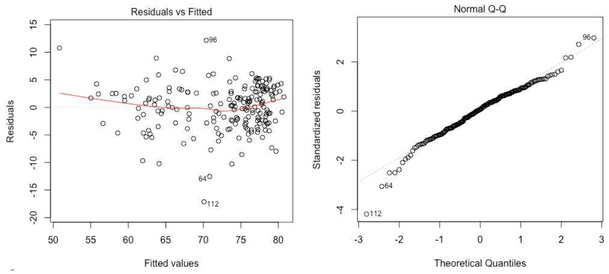
$$\Sigma(Y - \hat{Y})^2$$

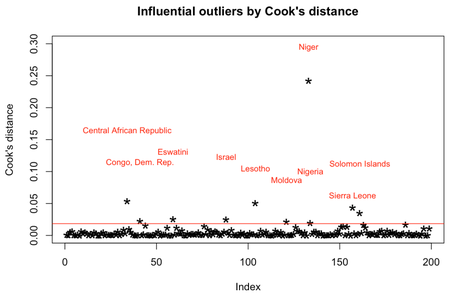
residuals

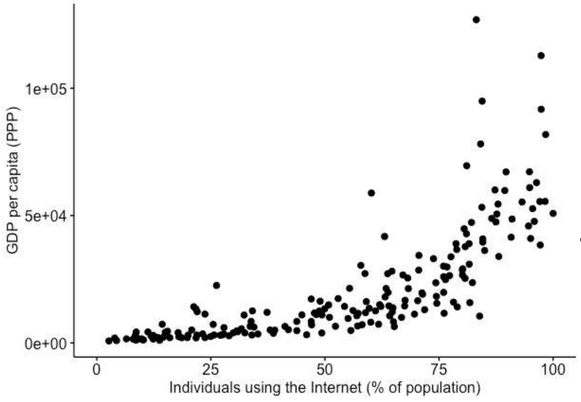


Multiple Linear Regression

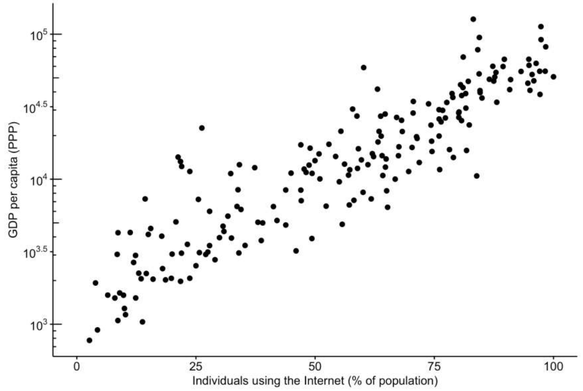
$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots$$







Logarithmic transformation



Goodness of Fit

Residual Standard Error

R^2

F-statistic