Analysis of Quantitative data
Linear regression

Anne Segonds-Pichon
v2020-12

Association between 2 continuous variables

One variable X and One variable Y

One predictor

Correlation

Signal-to-noise ratio

$$\frac{\boxed{\text{Similarity}}}{\boxed{\text{Variability}}} = \frac{\text{Signal}}{\text{Noise}}$$

$$\frac{\text{Signal}}{\text{Noise}} = \text{statistical significance}$$

$$\frac{\text{Signal}}{\text{Noise}} = \text{no statistical significance}$$

Signal-to-noise ratio and Correlation

$$\frac{\text{Similarity}}{\text{Variability}} = \frac{\text{Signal}}{\text{Noise}}$$

- Signal is **similarity** of behaviour between variable x and variable y.
- **Coefficient of correlation:** $\mathbf{r} = \frac{\text{similarity}}{\text{variability}} = \frac{\text{Signal}}{\text{Noise}}$

$$\mathbf{r} = \frac{\text{similarity}}{\text{variability}} = \frac{\overbrace{COV_{xy}}^{\text{covariance}}}{\underbrace{SD_x SD_y}_{\text{Standard Deviation}}} = \frac{\overbrace{\sum (x_i - \bar{x})(y_i - \bar{y})}^{\text{covariance}}}{(n-1)\underbrace{SD_x SD_y}_{\text{Standard Deviation}}}$$

Correlation

- Most widely-used correlation coefficient:
  - **Pearson product-moment correlation coefficient “r”**
    - The **magnitude** and the **direction** of the relation between 2 variables
    - It is designed to range in value between **-1 and +1**
    - **-0.6** < r > **+0.6** : exciting

Coefficient (+ve or –ve)	Strength of the relationship
0.0 to 0.2	Negligible
0.2 to 0.4	Weak
0.4 to 0.7	Moderate
0.7 to 0.9	Strong
0.9 to 1.0	Very strong

  - **Coefficient of determination “r^2”**
    - It gives the proportion of variance in Y that can be explained by X (in percentage).
      - It helps with the interpretation of r
      - It’s basically the **effect size**

Correlation

$p = 0.0002$

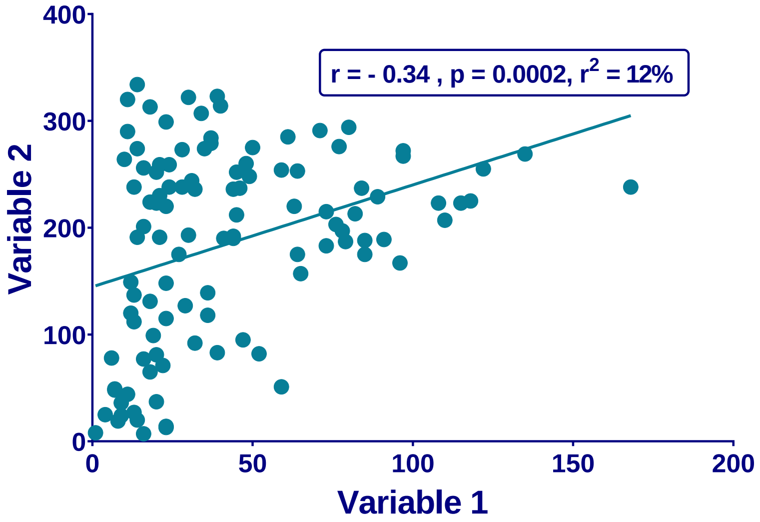
$r = -0.34$

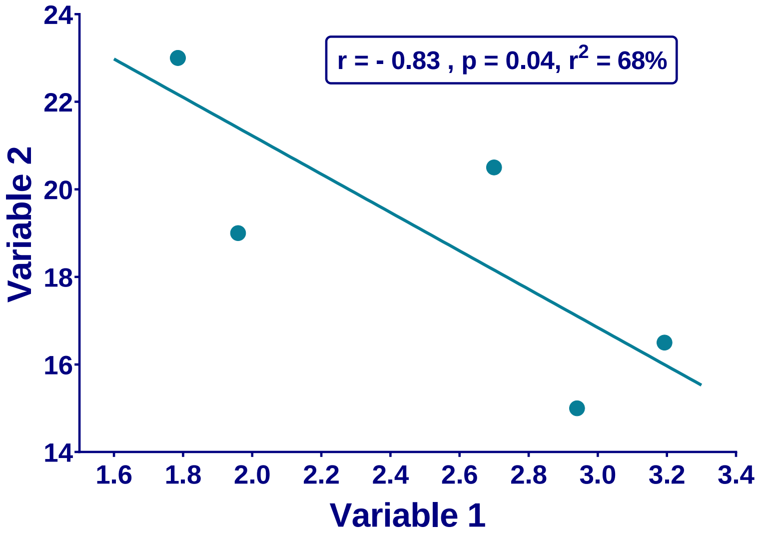
$r^2 = 0.12$

$p = 0.04$

$r = -0.83$

$r^2 = 0.68$





Power!!

Correlation
Assumptions

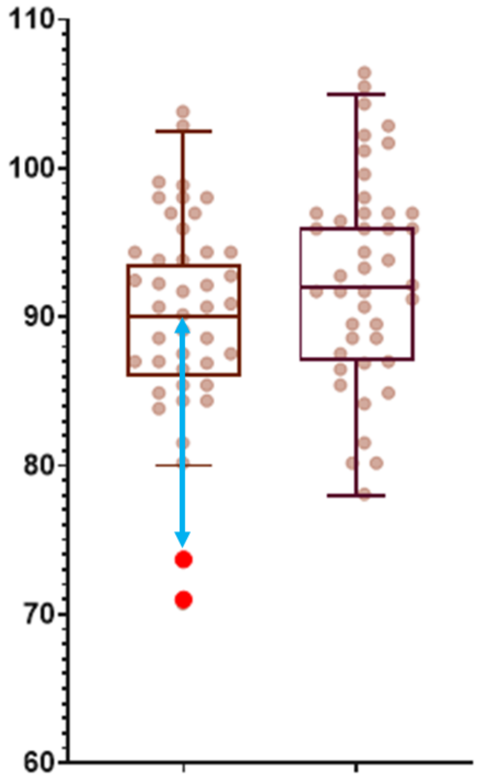
- Assumptions for correlation
  - Regression and linear Model (lm)

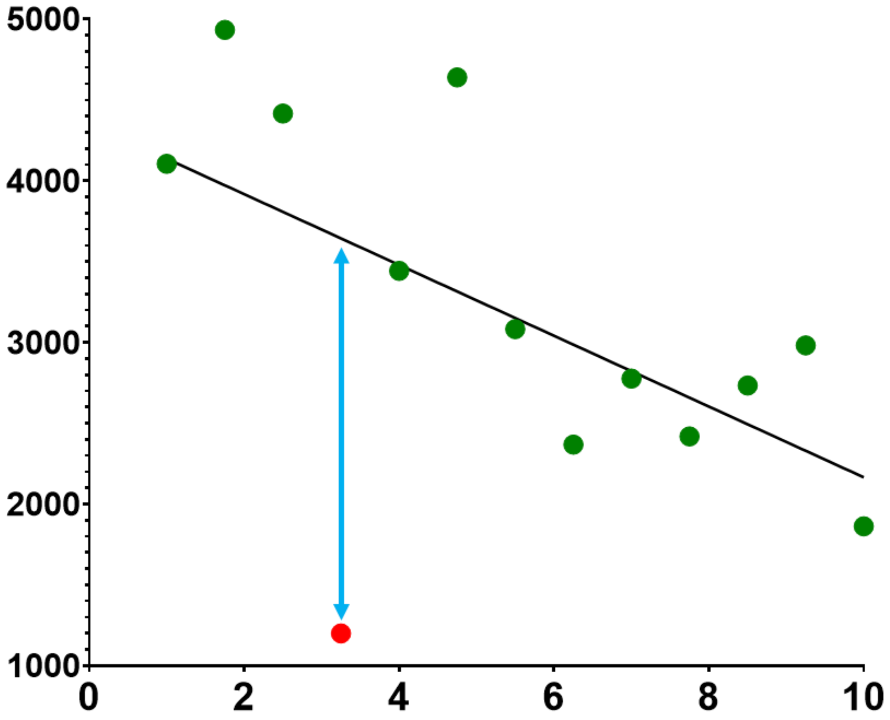
- **Linearity**: The relationship between X and the mean of Y is linear.
- **Homoscedasticity**: The variance of residual is the same for any value of X.
- **Independence:** Observations are independent of each other.
- **Normality:** For any fixed value of X, Y is normally distributed.

Correlation

Outliers and High leverage points

- **Outliers**: the observed value for the point is very different from that predicted by the regression model.





Correlation
Outliers and High leverage points

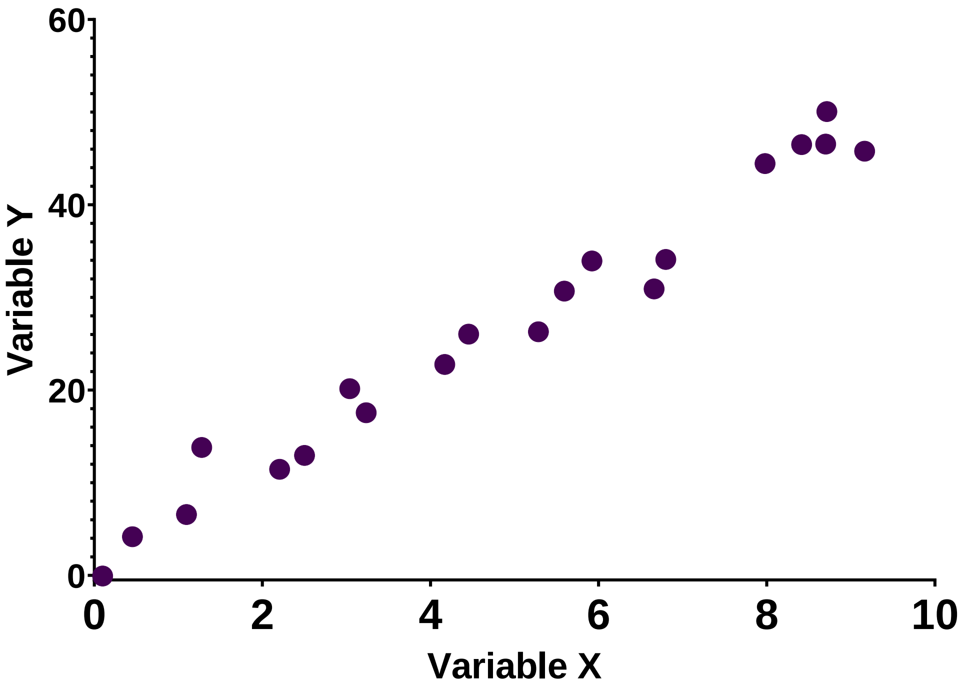
- **Leverage points**: A leverage point is defined as an observation that has a value of x that is far away from the mean of x.

- Outliers and leverage points have the potential to be **Influential observations**:
  - Change the slope of the line. Thus, have a large influence on the fit of the model.

- One method to find influential points is to compare the fit of the model **with** and **without** the dodgy observation.

Correlation

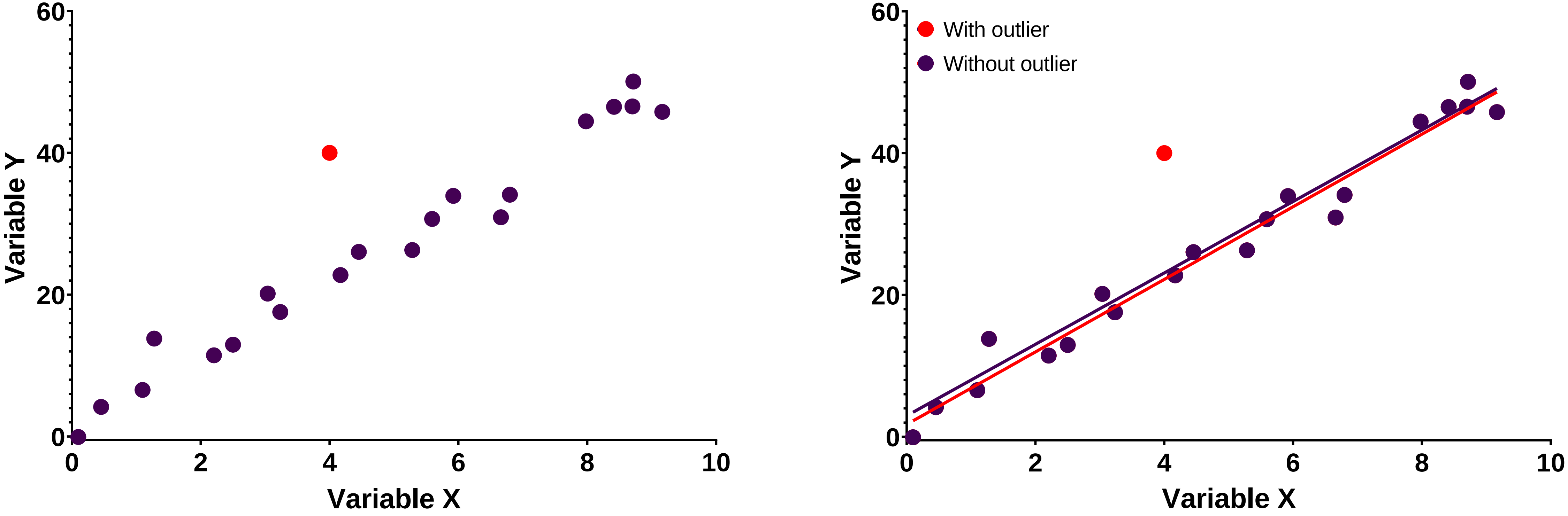
Outliers and High leverage points



All good

Correlation

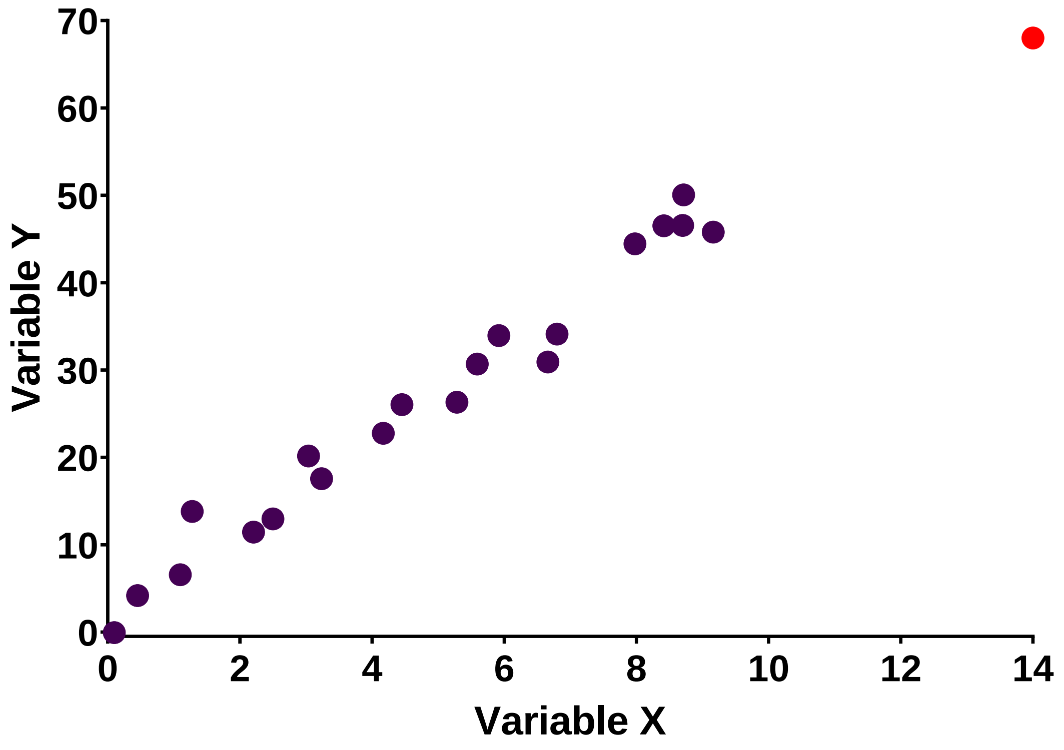
Outliers and High leverage points

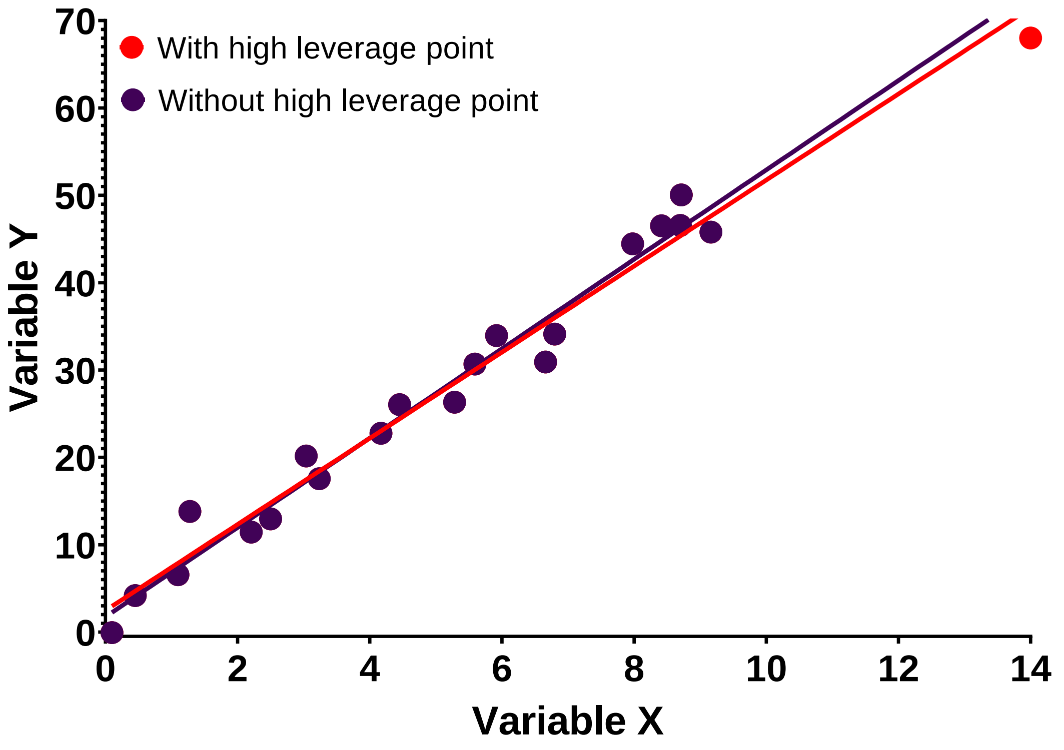


Outlier but not influential value

Correlation

Outliers and High leverage points

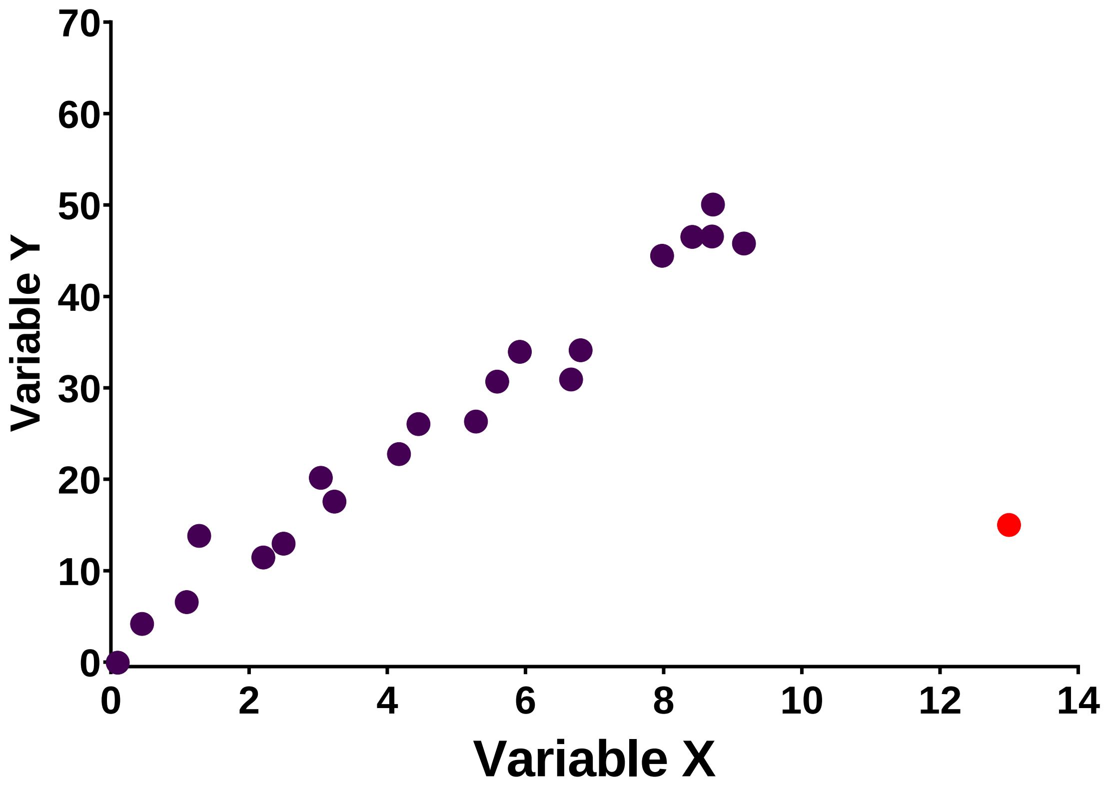


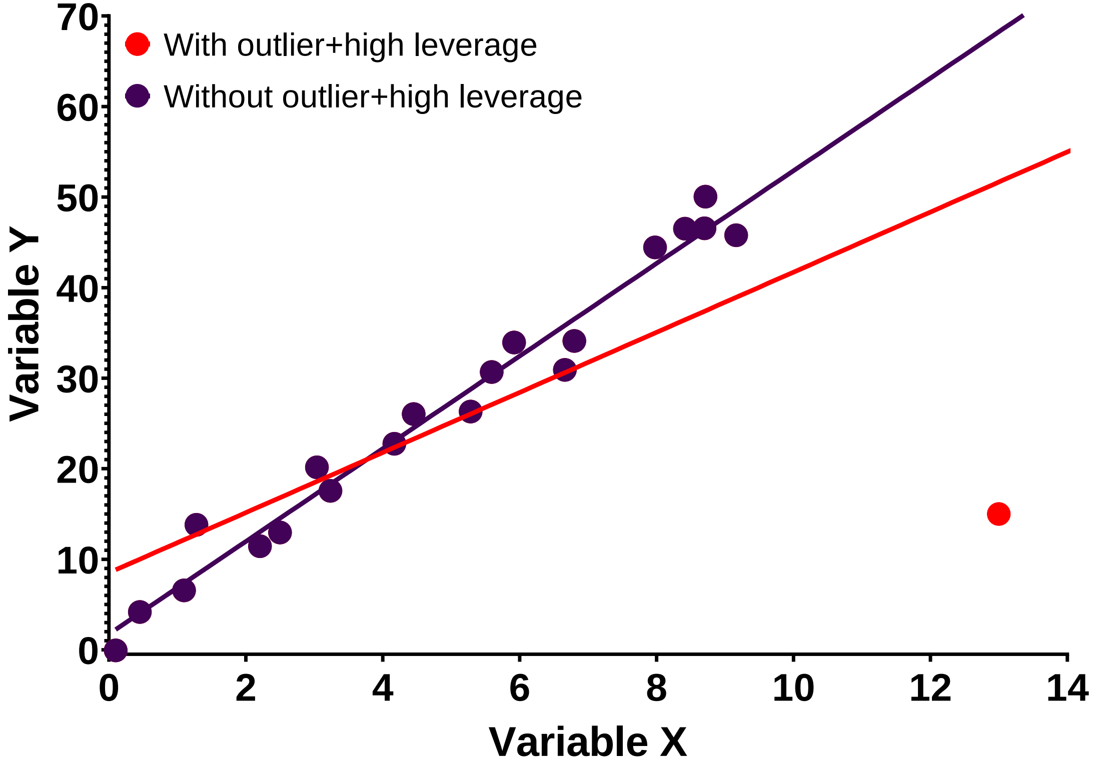


High leverage but not influential value

Correlation

Outliers and High leverage points





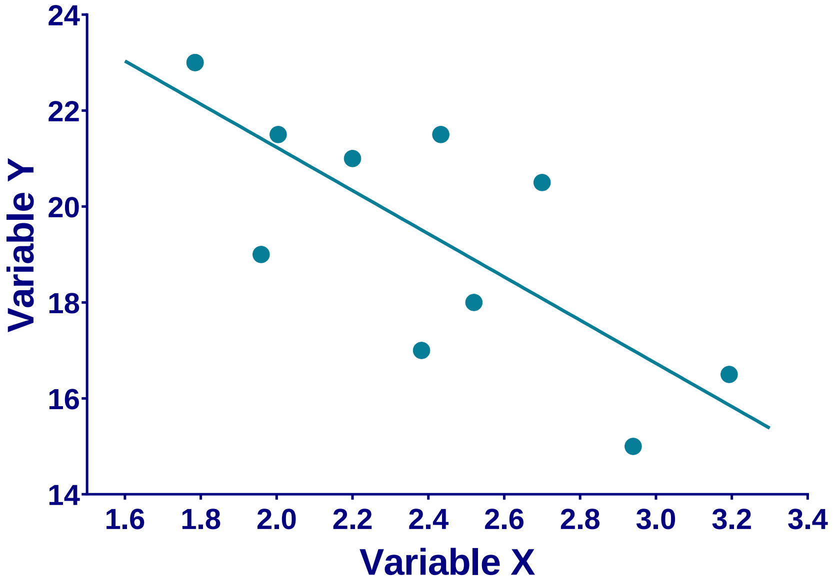
Outlier and High leverage: Influential value

Correlation: Two more things

Thing 1: Pearson correlation is a parametric test

First assumption for parametric test: Normality

Correlation: bivariate Gaussian distribution



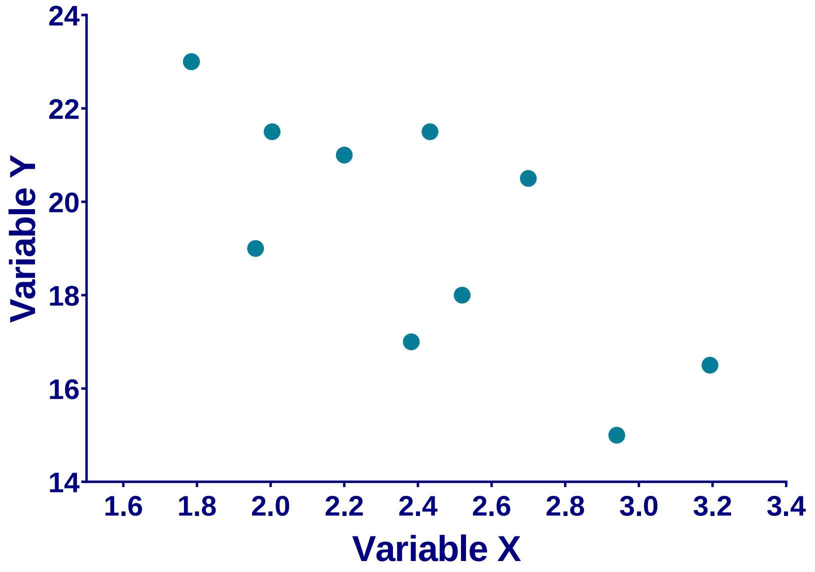
Symmetry-ish of the values on either side of the line of best fit.

Correlation: Two more things

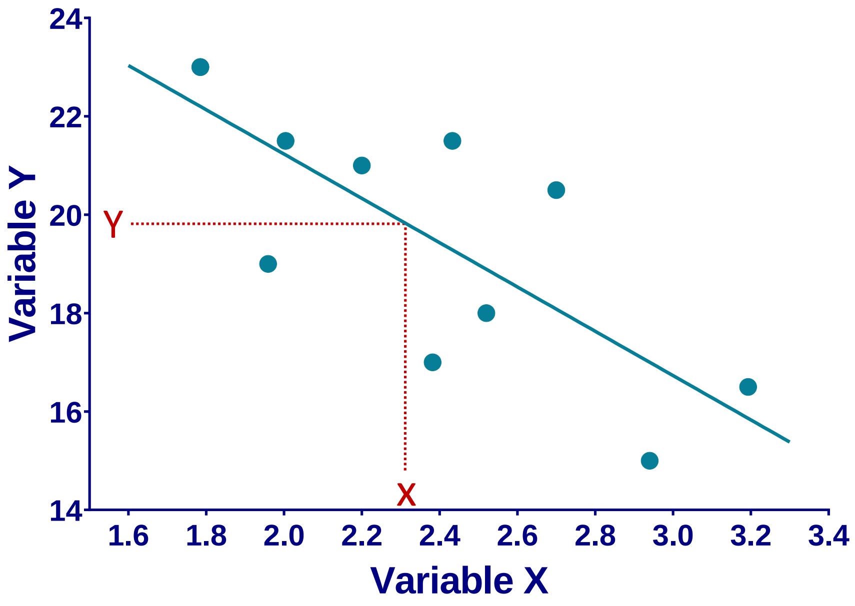
Thing 2: Line of best fit comes from a regression

Correlation: nature and strength of the association

Regression: nature and strength of the association <u>and</u> **prediction**



Correlation = Association

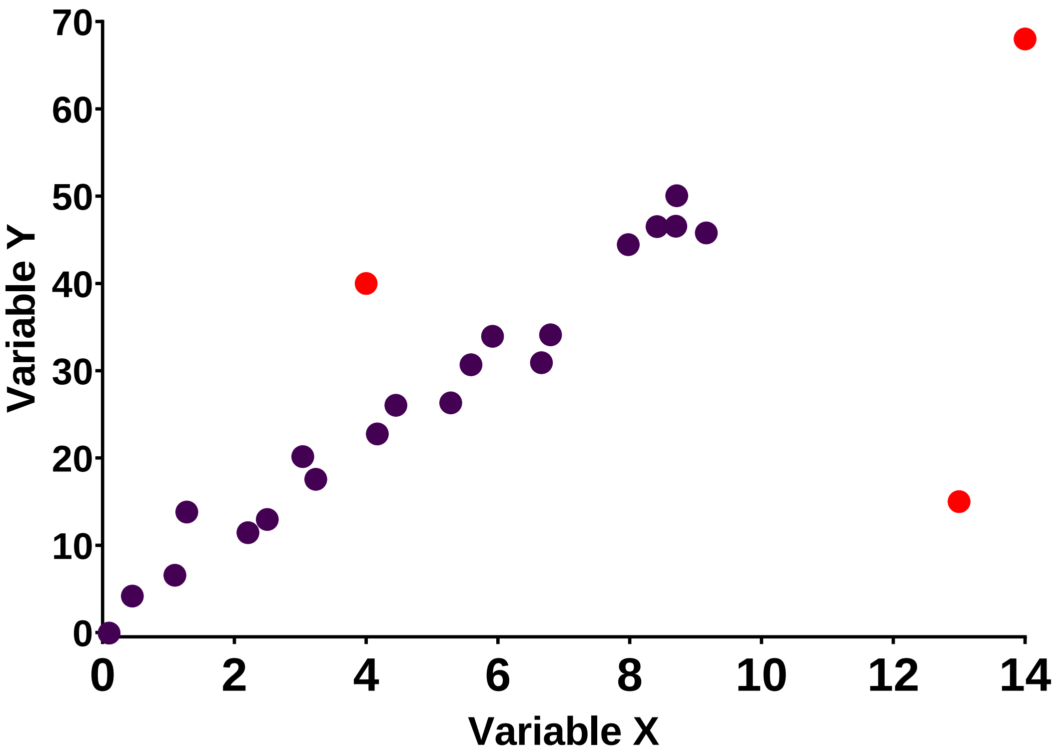


Regression = Prediction

Y = A + B*X

Correlation: correlation.csv

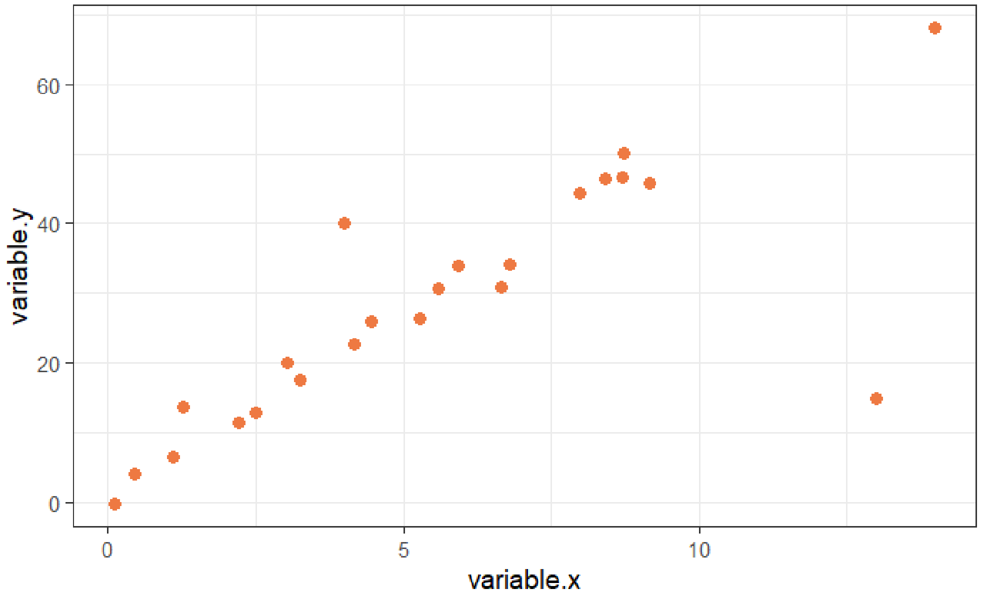
- **Questions**:
  - What is the nature and the strength of the relationship between X and Y?
  - Are there any dodgy points?



Correlation: correlation.csv

- **Question**: are there any dodgy points?

```
read_csv("correlation.csv") -> correlation

correlation %>%
  ggplot(aes(variable.x, variable.y, colour=Gender)) +
  geom_point(size=3, colour="sienna2")
```



ID <dbl>	variable.x <dbl>	variable.y <dbl>
1	0.10000	-0.0716
2	0.45401	4.1673
3	1.09765	6.5703
4	1.27936	13.8150
5	2.20611	11.4501
6	2.50064	12.9554
7	3.04030	20.1575
8	3.23583	17.5633
9	4.45308	26.0317
10	4.16990	22.7573

1-10 of 23 rows

Correlation: correlation.csv

- For the lines of best-fit: 3 new functions:

```
lm(y~x, data=) -> fit
coefficients(fit) -> cf.fit
geom_abline(intercept=cf.fit[1], slope=cf.fit[2])
```

(vector of 2 values)

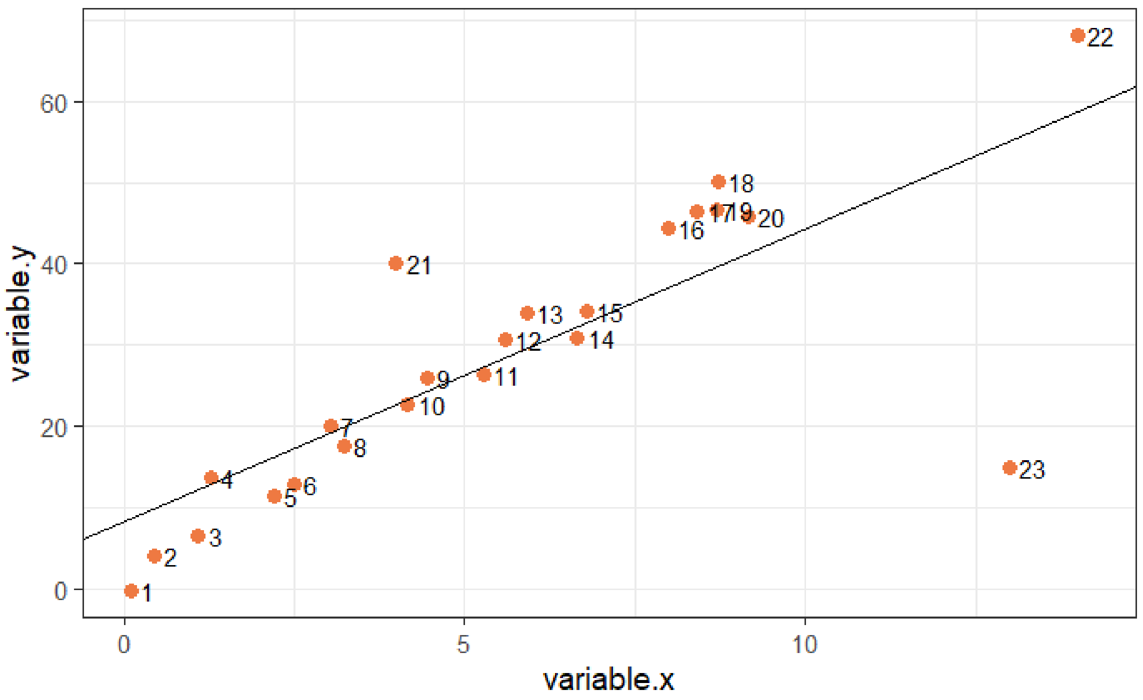
```
lm(variable.y ~ variable.x, data=correlation)-> fit.correlation
coefficients(fit.correlation) -> coef.correlation
coef.correlation
```

```
(Intercept)  variable.x
   8.379803    3.588814
```

intercept slope

Correlation: correlation.csv

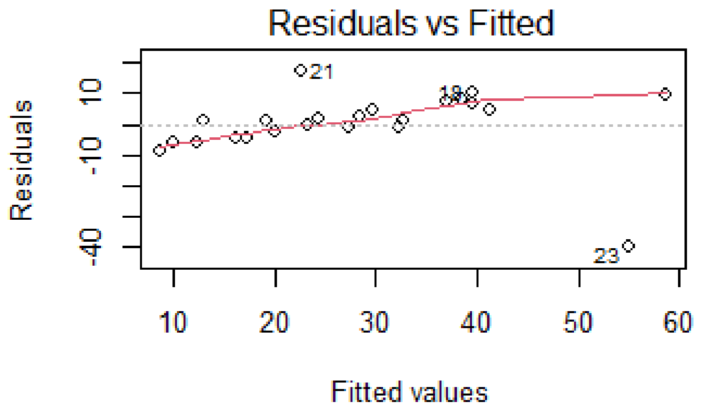
```
correlation %>%
  ggplot(aes(variable.x, variable.y, label = ID)) +
  geom_point(size=3, colour="sienna2") +
  geom_abline(intercept = coef.correlation[1], slope = coef.correlation[2])+
  geom_text(hjust = 0, nudge_x = 0.15)
```



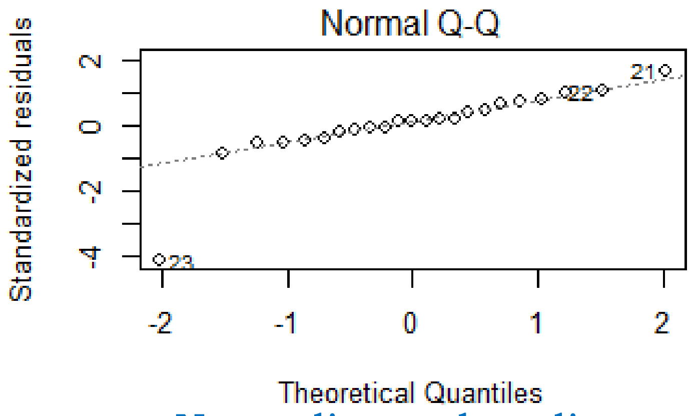
Correlation: correlation.csv

Assumptions, outliers and influential cases

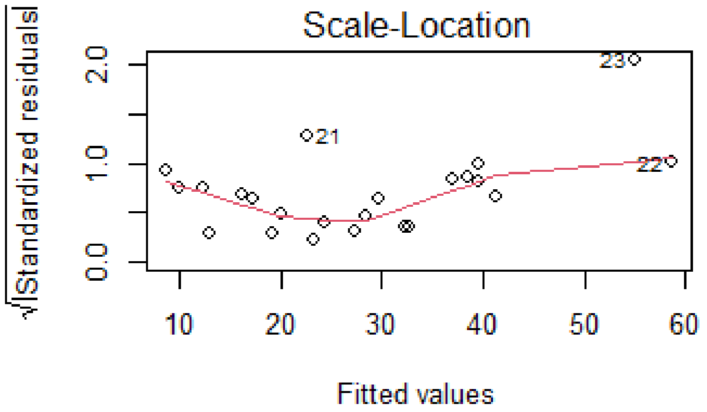
```
par(mfrow=c(2,2))
plot(fit.correlation)
```



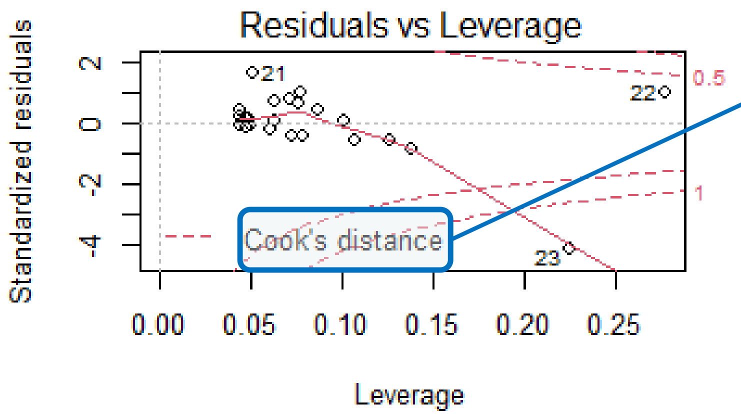
Linearity, homoscedasticity and outlier



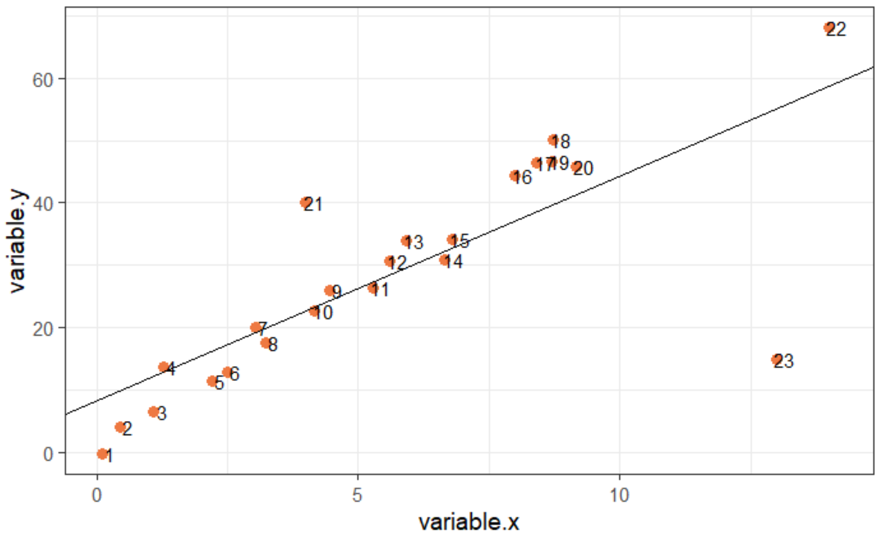
Normality and outlier



Homoscedasticity



Influential cases

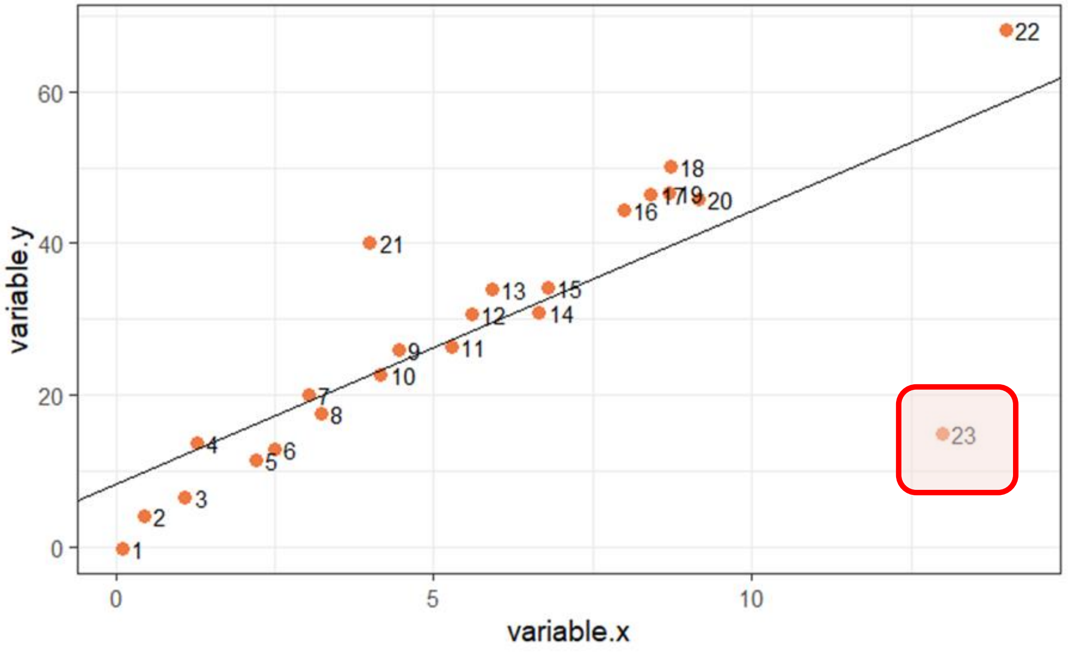


`cooks.distance()`

The **Cook's distance** is a combination of each observation's leverage and residual values ; the higher the leverage and residuals, the higher the Cook's distance (influential observation).

- It summarizes how much all the values in the regression model change when the ith observation is removed.
- Consensus: cut-off point =1 (0.5).

Correlation: correlation.csv




```
summary(fit.correlation)
```

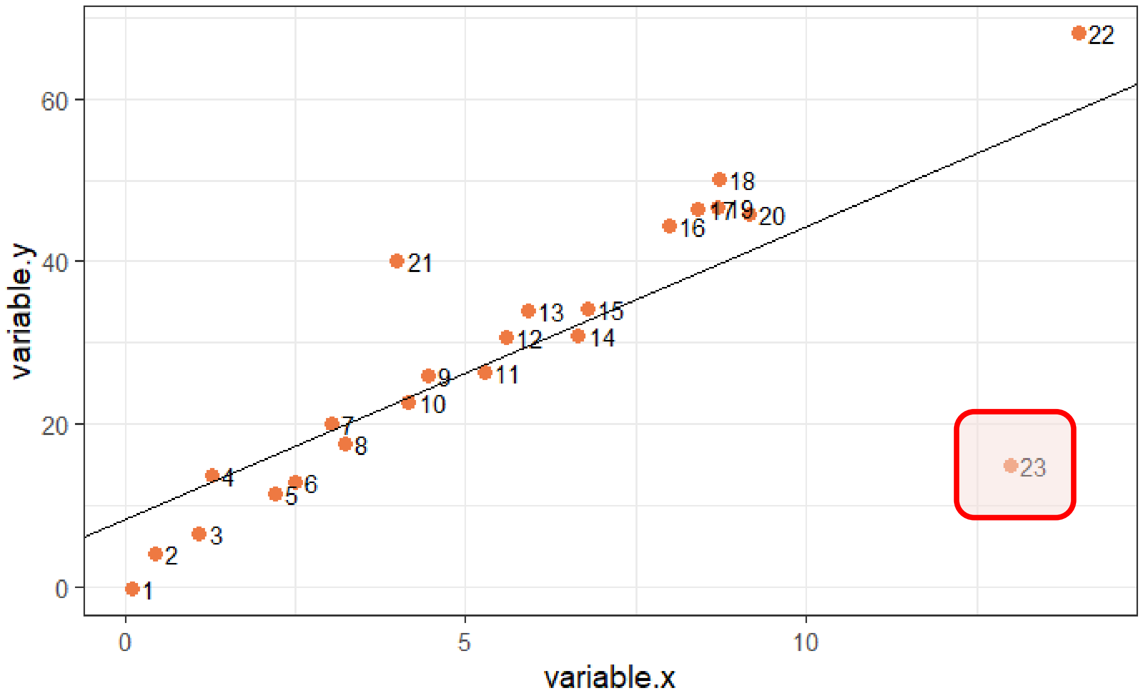
Line of best fit: Y=8.38 + 3.59*X

```
Call:
lm(formula = variable.y ~ variable.x, data = correlation)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.034  -3.414   0.867   5.723  17.265 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   8.3798     4.1195   2.034   0.0548 .  
variable.x    3.5888     0.6225   5.765 1.01e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.93 on 21 degrees of freedom
Multiple R-squared:  0.6128,	Adjusted R-squared:  0.5943 
F-statistic: 33.23 on 1 and 21 DF,  p-value: 1.01e-05
```

```
correlation %>%
  cor_test(variable.x, variable.y)
```

var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>	method <chr>
variable.x	variable.y	0.78	5.764871	1.01e-05	0.5471597	0.9034793	Pearson

Correlation: correlation.csv

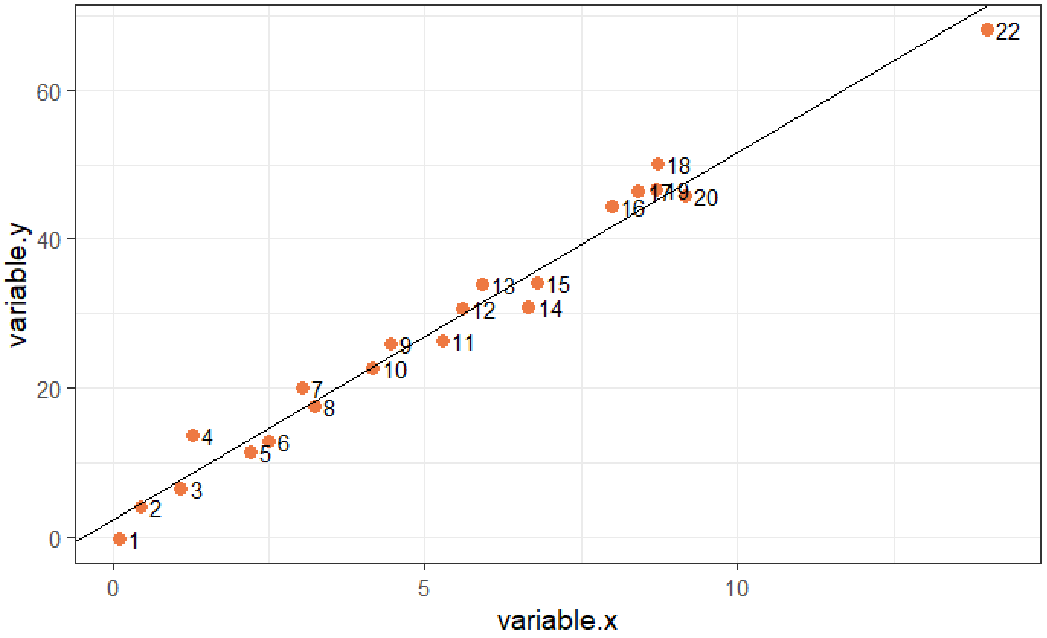


Have a go: Remove ID 23, then re-run the model and plot the graph again. Then decide what you want to do with ID 21 and 22.

```
correlation %>%
  filter(ID != 23) -> correlation.23
```

Correlation: correlation.csv

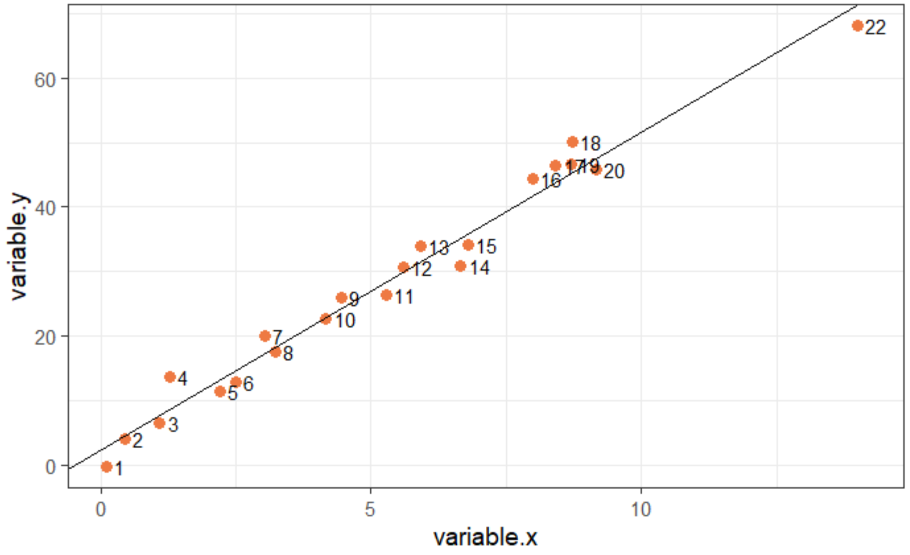
```
correlation %>%
  filter(ID != 23) -> correlation.23

lm(variable.y ~ variable.x, correlation.23) -> fit.correlation.23
summary(fit.correlation.23)
```




```
Call:
lm(formula = variable.y ~ variable.x, data = correlation.23)

Residuals:
   Min     1Q Median     3Q    Max 
-5.049 -2.784 -1.446  1.679 16.915 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.7103     1.8338   2.023   0.0566 .  
variable.x    4.8436     0.2971  16.303 5.13e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.695 on 20 degrees of freedom
Multiple R-squared:  0.93,	Adjusted R-squared:  0.9265 
F-statistic: 265.8 on 1 and 20 DF,  p-value: 5.13e-13
```

Correlation: correlation.csv

```
correlation.23 %>%
  filter(ID != 21) -> correlation.23.21

lm(variable.y ~ variable.x, correlation.23.21) -> fit.correlation.23.21
summary(fit.correlation.23.21)
```




```
Correlation.23.21 %>%
  cor_test(variable.x, variable.y)
```

var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>	method <chr>
variable.x	variable.y	0.99	28.66085	4.23e-17	0.9716067	0.9954718	Pearson

```
Call:
lm(formula = variable.y ~ variable.x, data = correlation.23.21)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.3636 -1.8607 -0.5376  2.2987  5.0434 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.4679     1.0757   2.294   0.0333 *  
variable.x    4.9272     0.1719  28.661   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.709 on 19 degrees of freedom
Multiple R-squared:  0.9774,	Adjusted R-squared:  0.9762 
F-statistic: 821.4 on 1 and 19 DF,  p-value: < 2.2e-16
```

Extra exercise

Correlation: exam.anxiety.csv

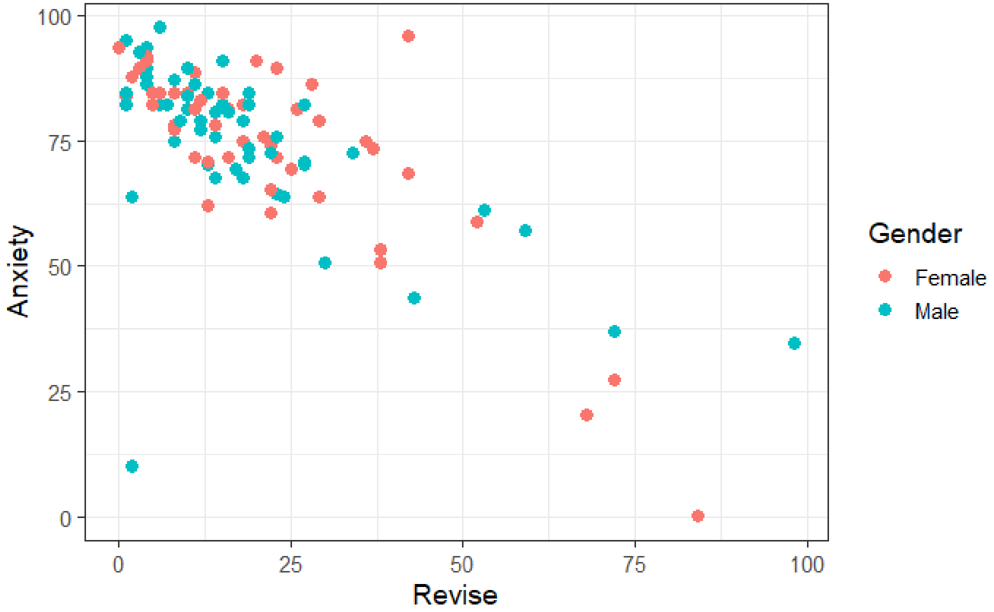
- **Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

- Build a fit for the boys and a fit for the girls
  - `data %>% filter() lm(y~x, data=)`
- Plot the 2 lines of best fit on the same graph
  - `coefficients()  geom_abline()`
- Check the assumptions visually from the data and with the output for models
  - `par(mfrow=c(2,2))  plot(fit.male)`
- Filter out misbehaving values based on the standardised residuals
  - **`rstandard`**`()` **`add_column()`**
- Plot the final (improved!) model
  - **`bind_rows()`**

Correlation: exam.anxiety.csv

- **Question**: Is there a relationship between time spent revising and exam anxiety? And, if yes, are boys and girls different?

```
read_csv("exam.anxiety.csv") -> exam.anxiety

exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender)) + geom_point(size=3)
```



A	B	C	D	E
Code	Revise	Exam	Anxiety	Gender
1	4	40	86.298	Male
2	11	65	88.716	Female
3	27	80	70.178	Male
4	53	80	61.312	Male
5	4	40	89.522	Male
6	22	70	60.506	Female
7	16	20	81.462	Female
8	21	55	75.82	Female
9	25	50	69.372	Female

Correlation: exam anxiety.csv

- Is there a relationship between time spent revising and exam anxiety?

```
exam.anxiety %>%
  filter(Gender=="Female") -> exam.anxiety.female

lm(Anxiety~Revise, data=exam.anxiety.female) -> fit.female

coefficients(fit.female) -> cf.fit.female
```

Fit for the females

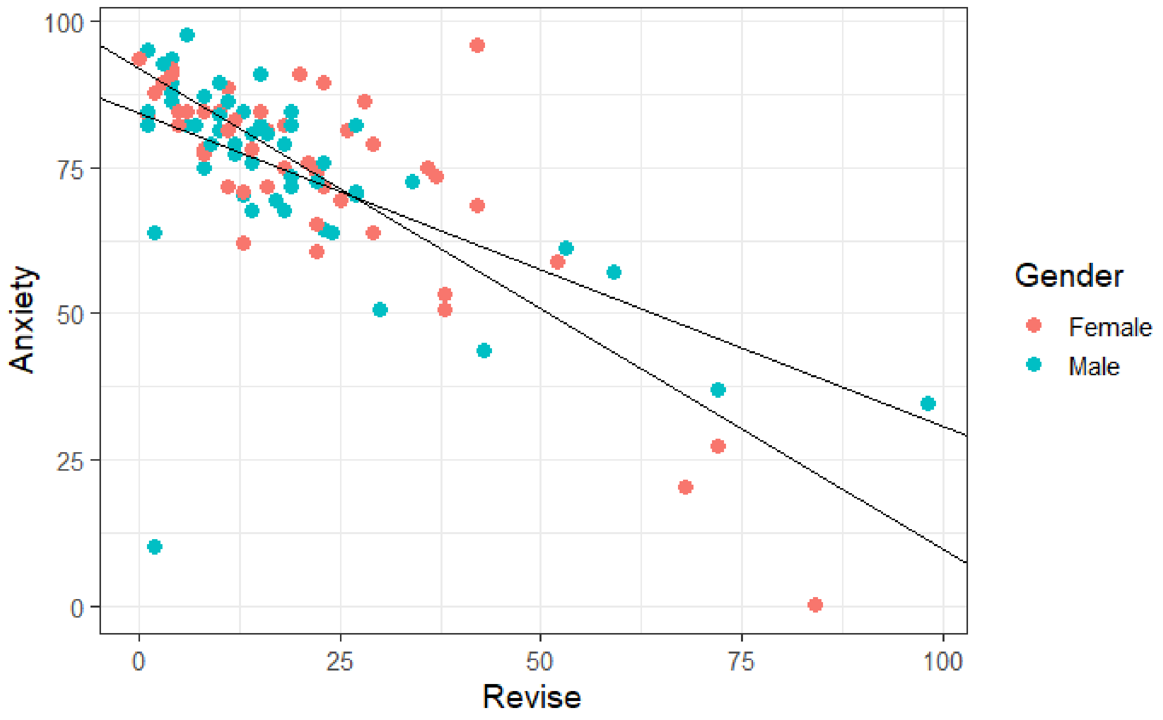
```
exam.anxiety %>%
  filter(Gender=="Male") -> exam.anxiety.male

lm(Anxiety~Revise, data=exam.anxiety.male) -> fit.male

coefficients(fit.male) -> cf.fit.male
```

Fit for the males

Correlation: exam anxiety.csv

- Is there a relationship between time spent revising and exam anxiety?

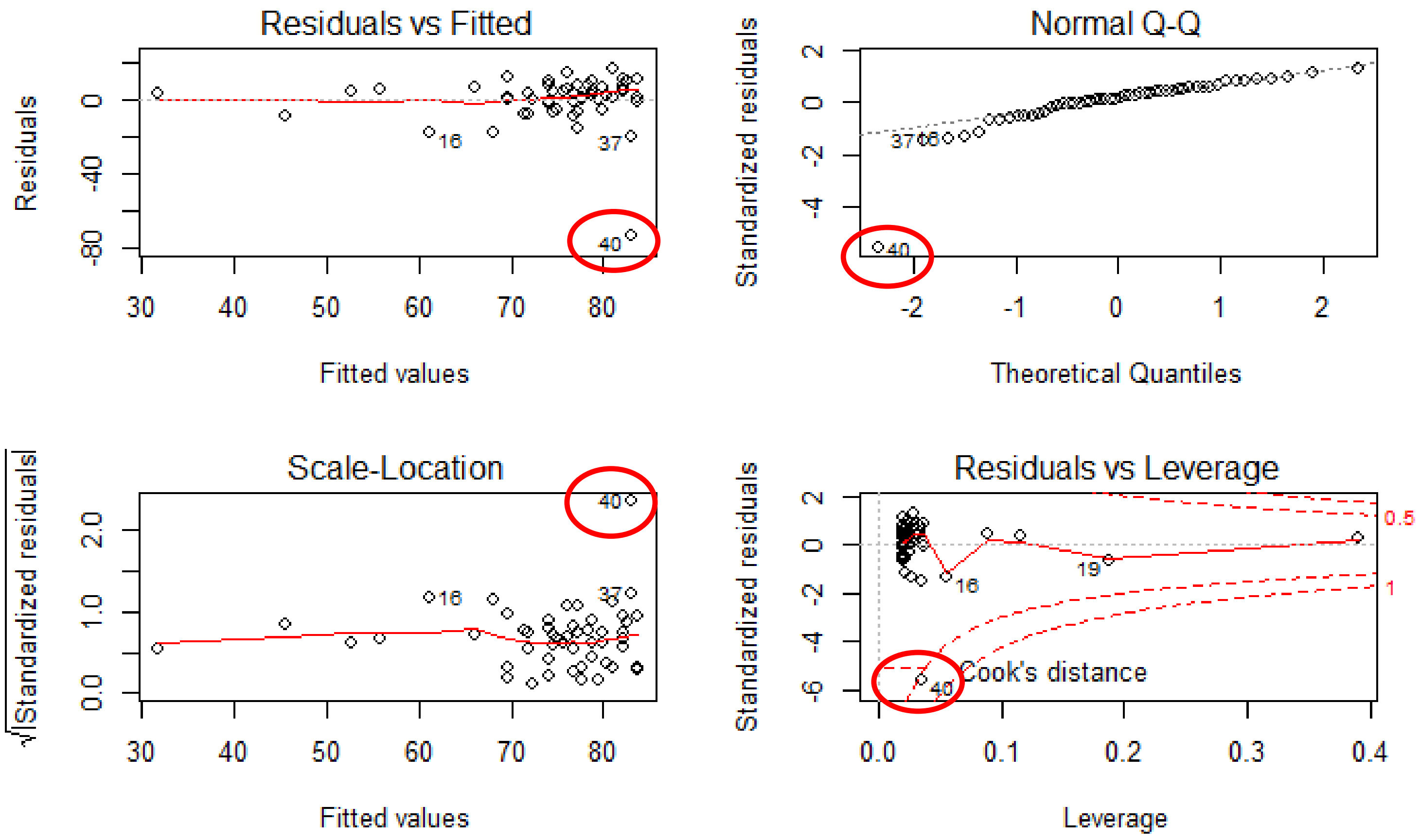
```
exam.anxiety %>%
  ggplot(aes(x=Revise, y=Anxiety, colour=Gender))+
  geom_point(size=3)+
  geom_abline(intercept=cf.fit.male[1], slope=cf.fit.male[2])+
  geom_abline(intercept=cf.fit.female[1], slope=cf.fit.female[2])
```



Correlation: exam anxiety.csv

Assumptions, outliers and influential cases

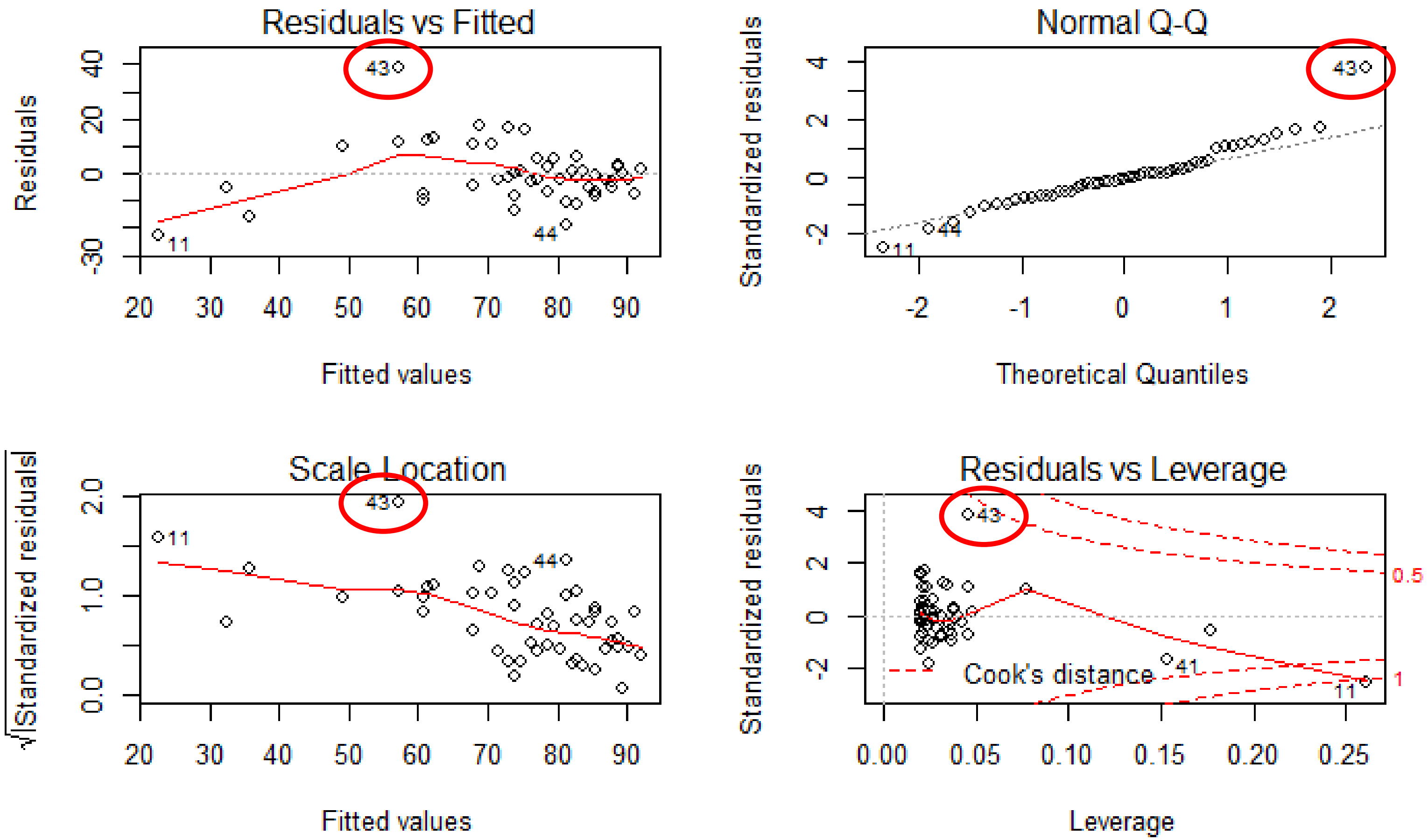
```
par(mfrow=c(2,2))
plot(fit.male)
```



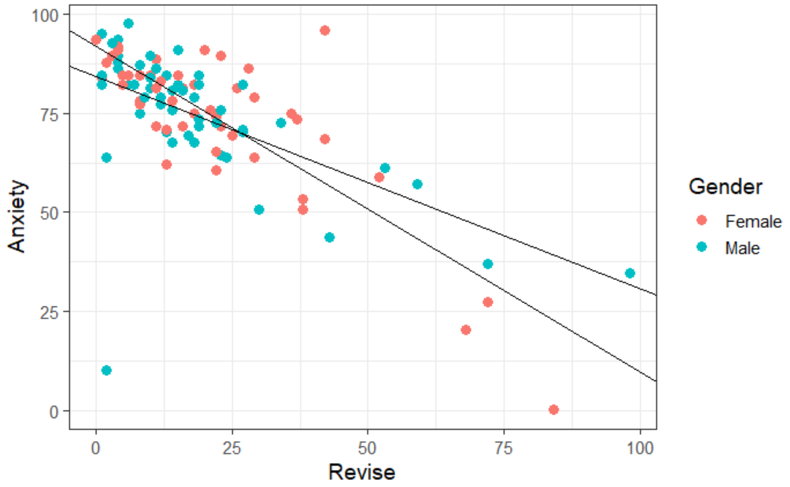
Correlation: exam anxiety.csv

Assumptions, outliers and influential cases

```
plot(fit.female)
```



Correlation: exam anxiety.csv




```
exam.anxiety %>%
  group_by(Gender) %>%
    cor_test(Revise, Anxiety) %>%
      ungroup()
```

Gender <chr>	var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>	method <chr>
Female	Revise	Anxiety	-0.82	-10.079994	1.54e-13	-0.8944820	-0.7054746	Pearson
Male	Revise	Anxiety	-0.60	-5.267088	2.94e-06	-0.7482821	-0.3876660	Pearson

```
summary(fit.male)
```

Anxiety=84.19 - 0.53*Revise

```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.male)

Residuals:
    Min      1Q  Median      3Q     Max
-73.124  -2.900   2.221   6.750  16.600

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.1941     2.6213  32.119  < 2e-16 ***
Revise       -0.5353     0.1016  -5.267 2.94e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.3 on 50 degrees of freedom
Multiple R-squared:  0.3568,	Adjusted R-squared:  0.344
F-statistic: 27.74 on 1 and 50 DF,  p-value: 2.937e-06
```

```
summary(fit.female)
```

Anxiety=91.94 - 0.82*Revise

```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.female)

Residuals:
    Min      1Q  Median      3Q     Max
-22.687  -6.263  -1.204   4.197  38.628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.94181    2.27858   40.35  < 2e-16 ***
Revise      -0.82380    0.08173  -10.08 1.54e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.42 on 49 degrees of freedom
Multiple R-squared:  0.6746,	Adjusted R-squared:  0.668
F-statistic: 101.6 on 1 and 49 DF,  p-value: 1.544e-13
```

Correlation: exam.anxiety.csv

Influential outliers: Boys

```
rstandard(fit.male) -> st.resid.m

exam.anxiety.male %>%
  add_column(st.resid.m) %>%
      filter(abs(st.resid.m)<3) -> exam.anxiety.male.clean

lm(Anxiety~Revise, data=exam.anxiety.male.clean) -> fit.male2

summary(fit.male2)
```

```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.male.clean)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0296  -3.8704   0.5626   6.0786  14.2525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 86.97461    1.64755  52.790  < 2e-16 ***
Revise      -0.60752    0.06326  -9.603 7.59e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.213 on 49 degrees of freedom
Multiple R-squared:  0.653,	Adjusted R-squared:  0.6459
F-statistic: 92.22 on 1 and 49 DF,  p-value: 7.591e-13
```

```
exam.anxiety.male.clean %>%
  cor_test(Revise, Anxiety)
```

var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>
Revise	Anxiety	-0.81	-9.602995	7.59e-13	-0.8863013	-0.6850763

Correlation: exam.anxiety.csv

Influential outliers: Girls

```
rstandard(fit.female) -> st.resid.f

exam.anxiety.female %>%
  add_column(st.resid.f) %>%
  filter(abs(st.resid.f) < 3) -> exam.anxiety.female.clean

lm(Anxiety~Revise, data=exam.anxiety.female.clean) -> fit.female2

summary(fit.female2)
```

```
Call:
lm(formula = Anxiety ~ Revise, data = exam.anxiety.female.clean)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.7518  -5.7069  -0.7782   3.2117  18.5538 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 92.24536    1.93591   47.65   <2e-16 ***
Revise      -0.87504    0.07033  -12.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.849 on 48 degrees of freedom
Multiple R-squared:  0.7633,	Adjusted R-squared:  0.7584 
F-statistic: 154.8 on 1 and 48 DF,  p-value: < 2.2e-16
```

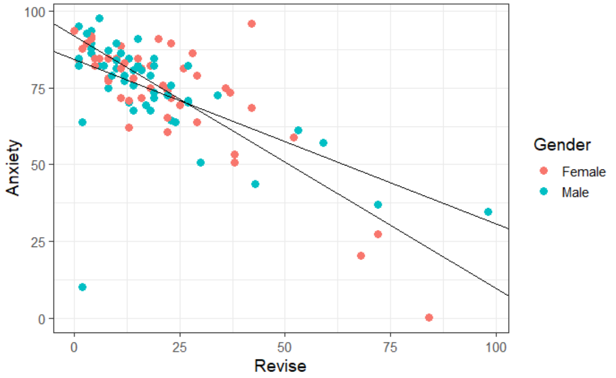
```
exam.anxiety.female.clean %>%
  cor_test(Revise, Anxiety)
```

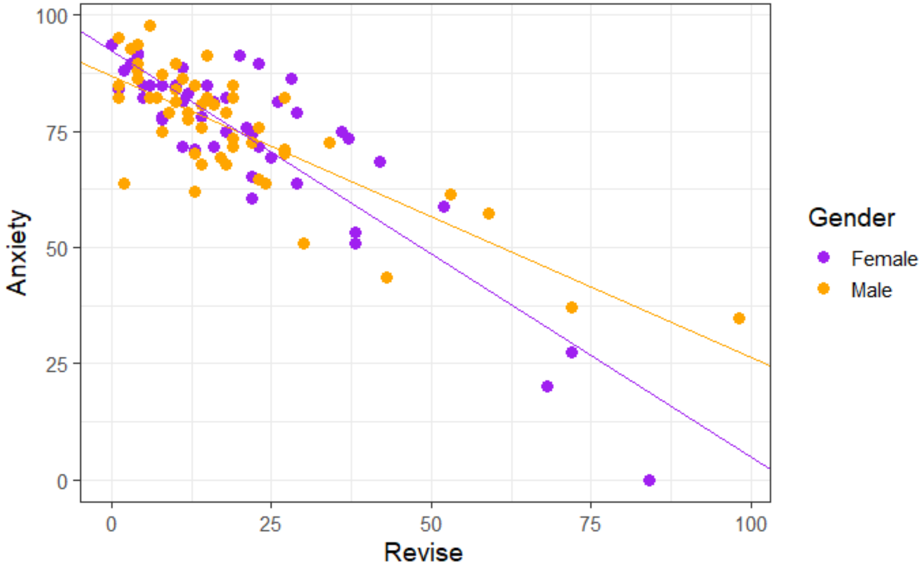
var1 <chr>	var2 <chr>	cor <dbl>	statistic <dbl>	p <dbl>	conf.low <dbl>	conf.high <dbl>
Revise	Anxiety	-0.87	-12.44127	1.25e-16	-0.9266661	-0.7866117

Correlation: exam.anxiety.csv

- **Question**: Is there a relationship between time spent revising and exam anxiety? Yes!

```
bind_rows(exam.anxiety.female.clean, exam.anxiety.male.clean) -> exam.anxiety.clean
coefficients(fit.male2) -> cf.fit.male2
coefficients(fit.female2) -> cf.fit.female2
exam.anxiety.clean %>%
  ggplot(aes(Revise, Anxiety, colour=Gender))+geom_point(size=3)+
  geom_abline(aes(intercept=cf.fit.male2[1], slope=cf.fit.male2[2]), colour="orange")+
  geom_abline(aes(intercept=cf.fit.female2[1], slope=cf.fit.female2[2]), colour="purple")+
  scale_colour_manual(values = c("purple", "orange"))
```





Correlation: exam.anxiety

Influential outliers: Another check

```
exam.anxiety.male %>%
  shapiro_test(st.resid.m)
```

variable <chr>	statistic <dbl>	p <dbl>
st.resid.m	0.6992772	5.05199e-09

```
exam.anxiety.female %>%
  shapiro_test(st.resid.f)
```

variable <chr>	statistic <dbl>	p <dbl>
st.resid.f	0.9442729	0.01828732

```
exam.anxiety.male.clean %>%
  shapiro_test(st.resid.m)
```

variable <chr>	statistic <dbl>	p <dbl>
st.resid.m	0.9539309	0.04607996

```
exam.anxiety.female.clean %>%
  shapiro_test(st.resid.f)
```

variable <chr>	statistic <dbl>	p <dbl>
st.resid.f	0.9767888	0.4258592

Correlation: exam anxiety.csv

- Difference between boys and girls?

```
lm(Anxiety~Revise*Gender, data=exam.anxiety.clean) -> fit.genders

summary(fit.genders)
```

```
Call:
lm(formula = Anxiety ~ Revise * Gender, data =
exam.anxiety.clean)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0296  -5.6022  -0.3294   5.6091  18.5538

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       92.24536    1.86694  49.410  < 2e-16 ***
Revise            -0.87504    0.06783 -12.901  < 2e-16 ***
GenderMale        -5.27075    2.53296  -2.081  0.04008 *
Revise:GenderMale  0.26752    0.09445   2.832  0.00562 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.534 on 97 degrees of freedom
Multiple R-squared:  0.7228,    Adjusted R-squared:  0.7142
F-statistic: 84.32 on 3 and 97 DF,  p-value: < 2.2e-16
```