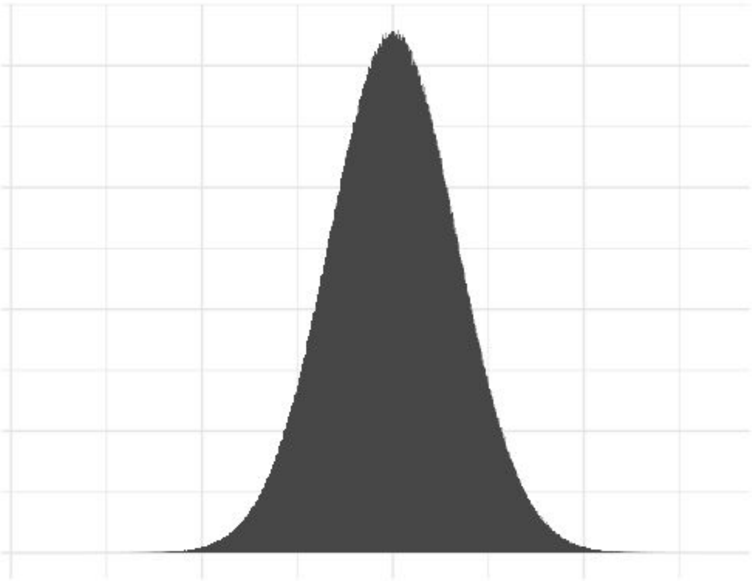
The General Linear Model – Regression Part 1

Dr Andrew J. Stewart

E: drandrewjstewart@gmail.com
T: @ajstewart_lang
G: ajstewartlang

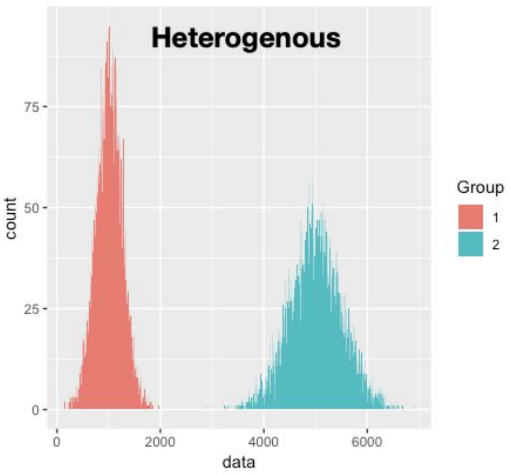
Assumptions of Parametric Statistics

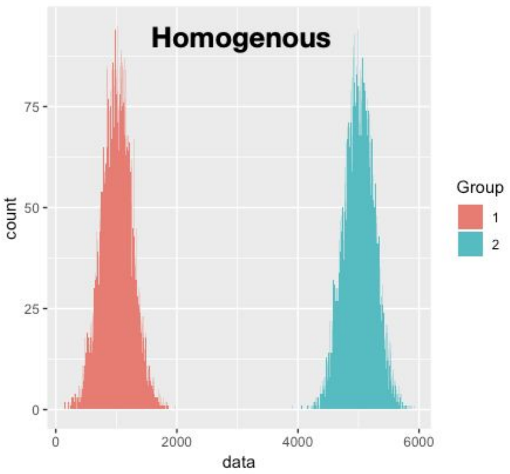
Assumption 1 - the data are conditionally normally distributed - in practical terms, this means the residuals need to be normally distributed (although t-tests require the data to be normal).



Assumptions of Parametric Statistics

Assumption 2 – Homogeneity of variance – the variances should not change systematically throughout the data. In designs where you test several groups of participants this means that the variances of each group should be equivalent.





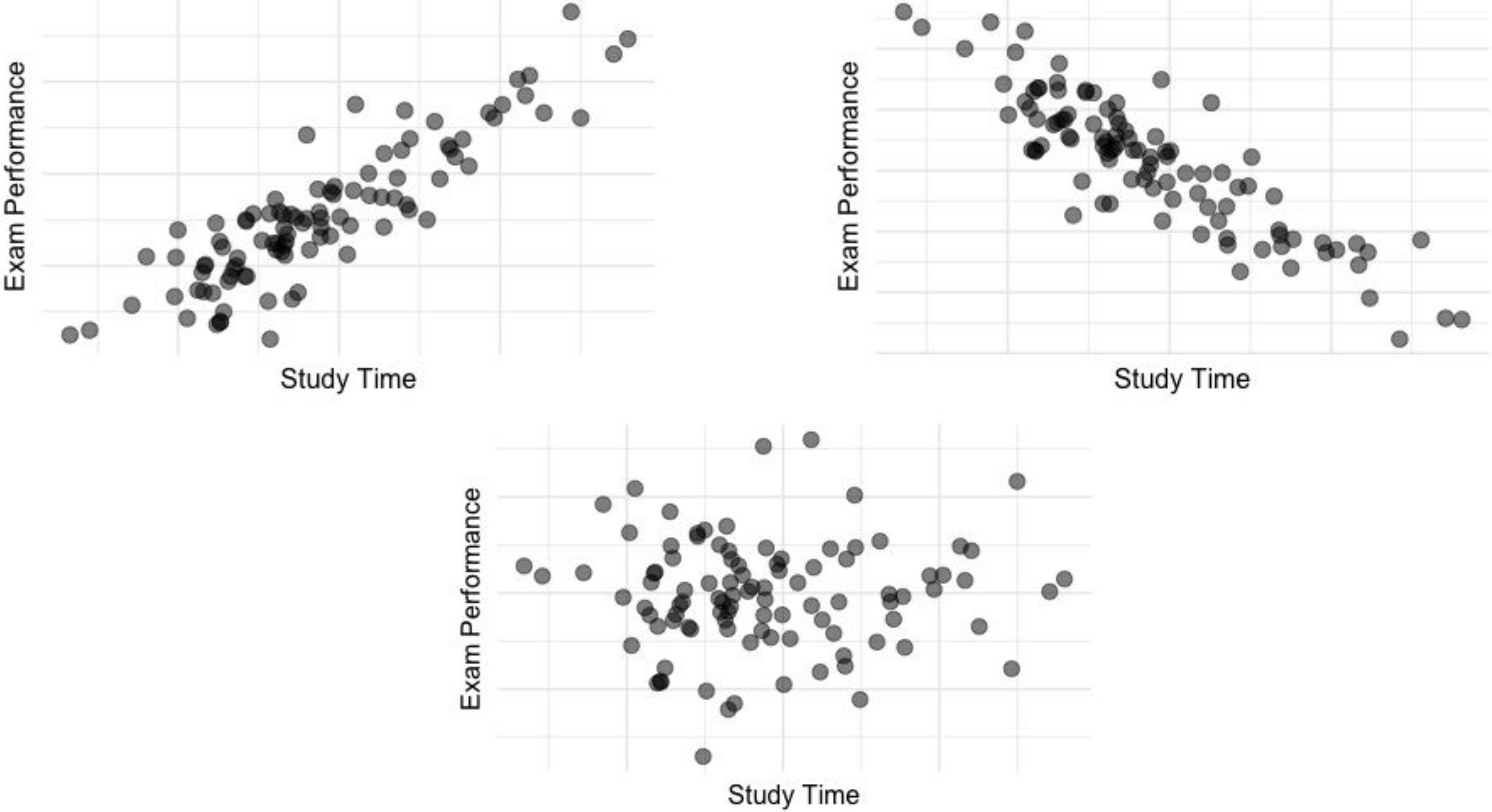
Assumptions of Parametric Statistics

Assumption 3 – Interval data – data should be measured on an interval scale. In other words, the distance between two adjacent points should be the same as the distance between any other two adjacent points. R can’t tell you this – you need to determine it by yourself. Reaction time is a good example of interval data.

Assumption 4 – Independence. The data from one participant does not affect the data from another (i.e., they are independent). In repeated measures designs, we expect the scores in the experimental conditions to be independent between participants.

Remember Correlation?

- Sometimes we’re interested in the possible relationship between two variables such as:
  - Time spent studying and performance on an exam
- Perhaps there’s:
  - A positive correlation between the two where more time spent studying correlates with better exam performance
  - A negative correlation between the two where more time spent study correlates with worse exam performance
  - No correlation between the two variables where time spent studying doesn’t correlate with exam performance


Exam Performance
Study Time
Exam Performance
Study Time
Exam Performance
Study Time

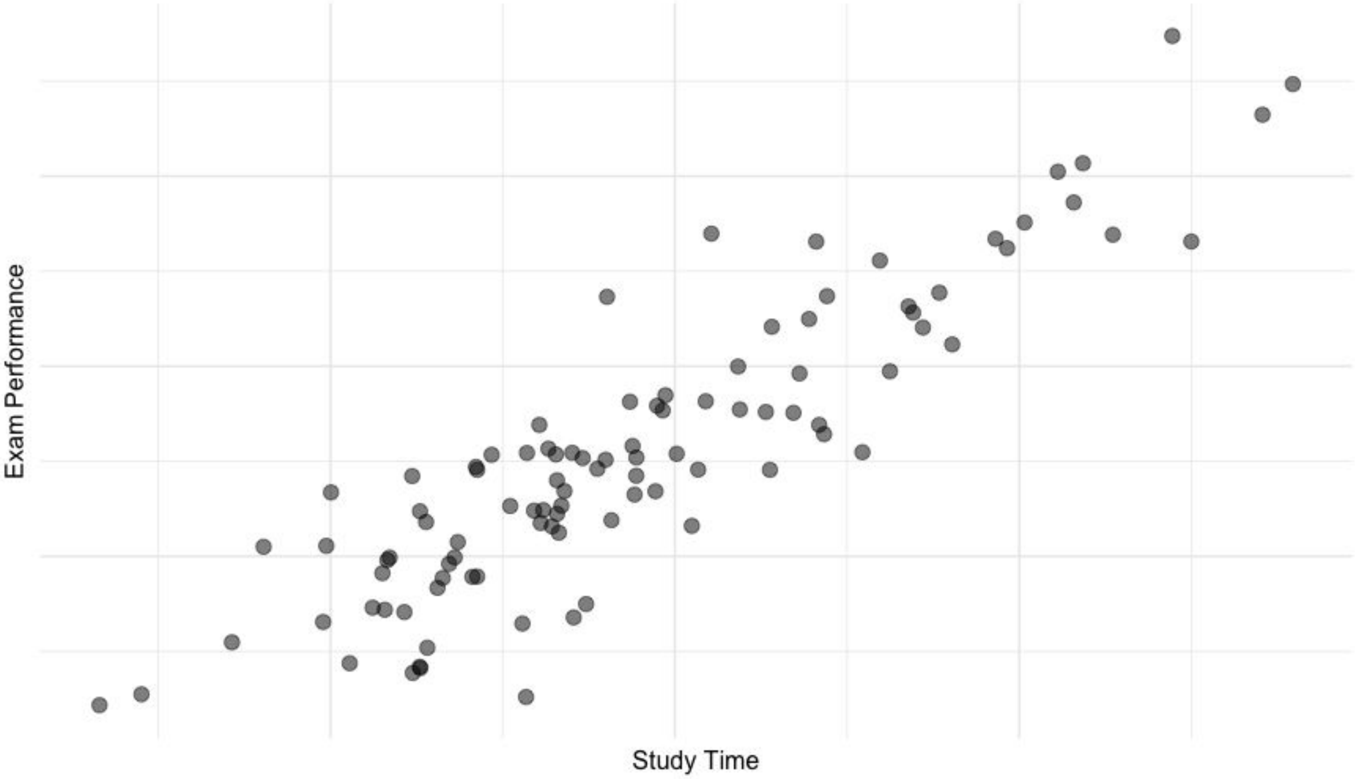
Remember variance?

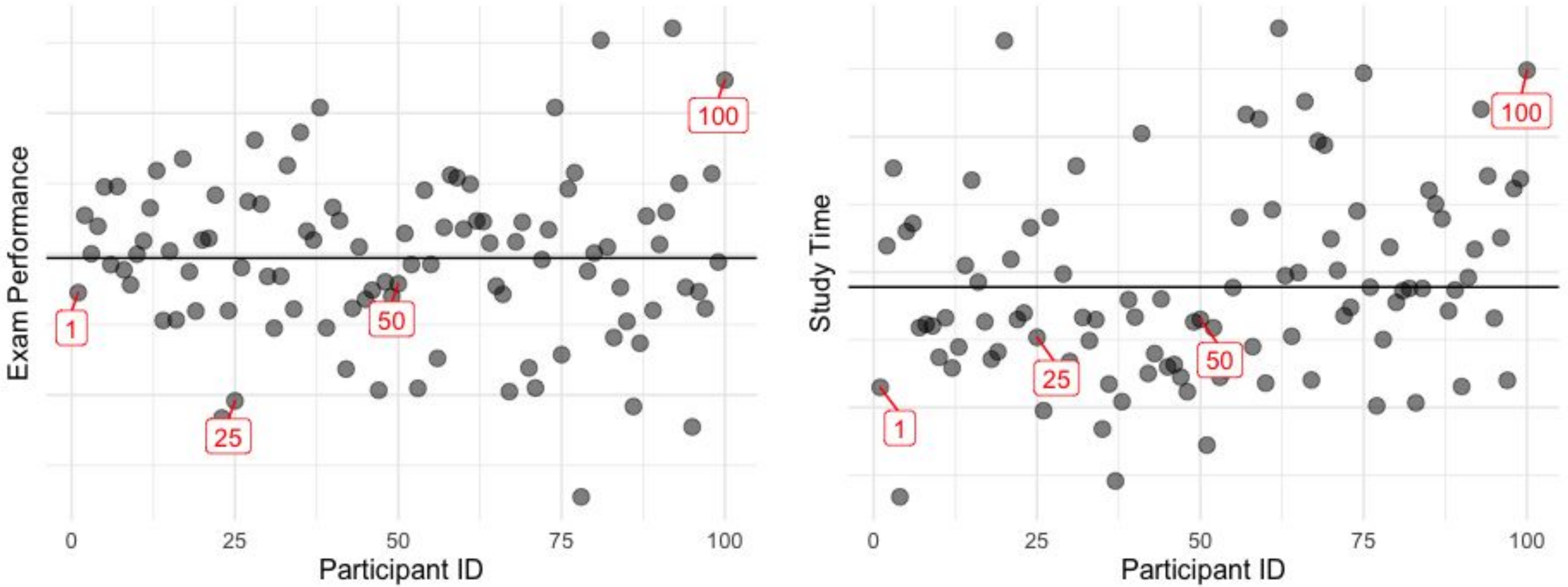
It’s the measure of the amount by which data associated with a variable vary from the mean of that variable…

$$= \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{N - 1}$$

If two variables *covary*, then when one variable deviates from the mean, we expect the other variable to deviate from its mean in a similar way.

Let’s take a close look at the data in this panel:





The horizontal lines represent the mean for each variable - if a participant is below the mean on one variable, notice that they are also below the mean for the other variable - this suggests the two variables co-vary.

For participants 1, 25 and 50, their scores on each variable are all below the respective mean for each variable, for participant 100 their score is above the respective mean for each variable.

To formalise this, we can calculate the combined differences…..

$$= \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{N - 1}$$

$$=\frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{N-1}$$

ID	Study	Score	Mean_Study	Mean_Score	Study - Mean_Study	Score - Mean_Score	(Study - Mean_Study) * (Score - Mean_Score)
1	192	77	199	79.6	-7	-2.6	18.2
2	202	81	199	79.6	3	1.4	4.2
3	208	82	199	79.6	9	2.4	21.6
4	183	75	199	79.6	-16	-4.6	73.6
...	...	...	...	...	...	...	...

Sum = 1435.072

So covariance = (1435.072/99) = 14.49568

Now, one problem with covariance as we’ve calculated it is that the score we end up with depends on the measurement scales associated with our variables.

In other words, the covariance value isn’t standardised.

We can divide any value by the standard deviation and that will give us the distance from the mean in standard deviation units….

We can divide our covariance value by the standard deviations of our two variables (actually standard deviation of x multiplied by standard deviation of y) – in other words:

$$= \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{N - 1 s_x s_y}$$

This is called the Pearson product-moment correlation coefficient and ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation) with 0 meaning on correlation at all.

The Standard Deviation of our Study variable is 7.09566
The Standard Deviation of our Score variable is 2.277481

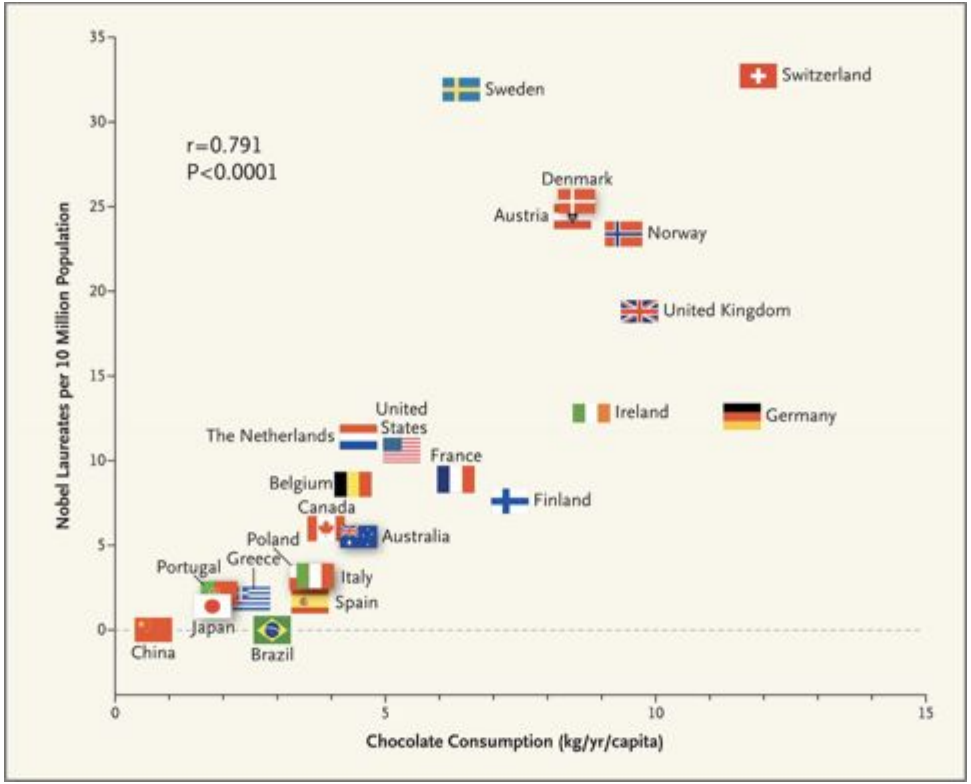
So, we can divide our covariance result (14.49568) by (7.09566) X (2.277481)
Which gives us 0.8969966 - let’s round it to 0.9

Which is the Pearson’s R for the correlation between our two variables...

Correlation is not Causation

There is a high correlation (Pearson’s r = 0.791) between chocolate consumption in a country and the number of Nobel Prize winners in that country.

Why do you think this is?



Correlation is **not** Causation

When interpreting correlation data one common pitfall is to assume that the score on one variable causes a particular score on the other. This is **wrong**!

Very often, common sense might suggest causation – e.g., time spent studying improves exam score.

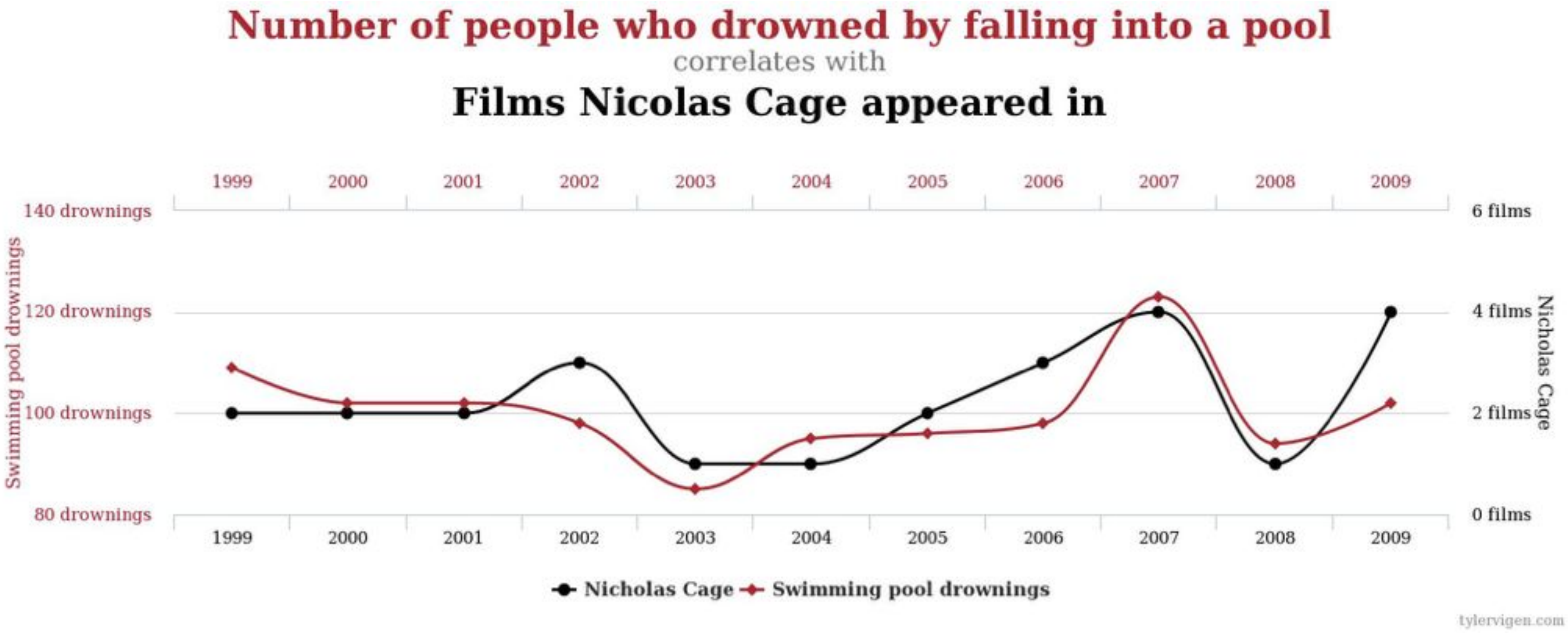
But you cannot make any claim about causation from a correlation.

There may be other variables that we don’t know about – maybe being interested in an academic subject results in more engagement in general with it and it is this that influences time spent studying and exam performance.

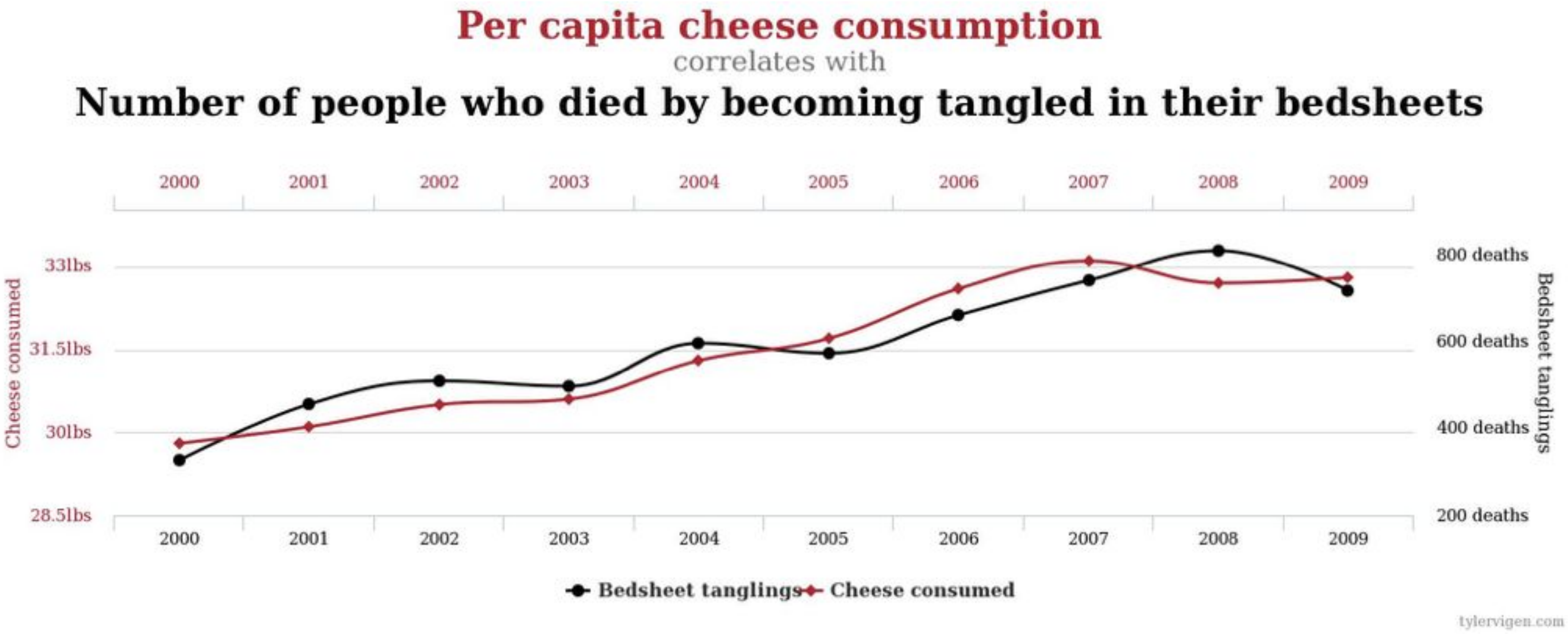
Additionally, spurious correlations can be found all over the place…

Correlation is **not** Causation

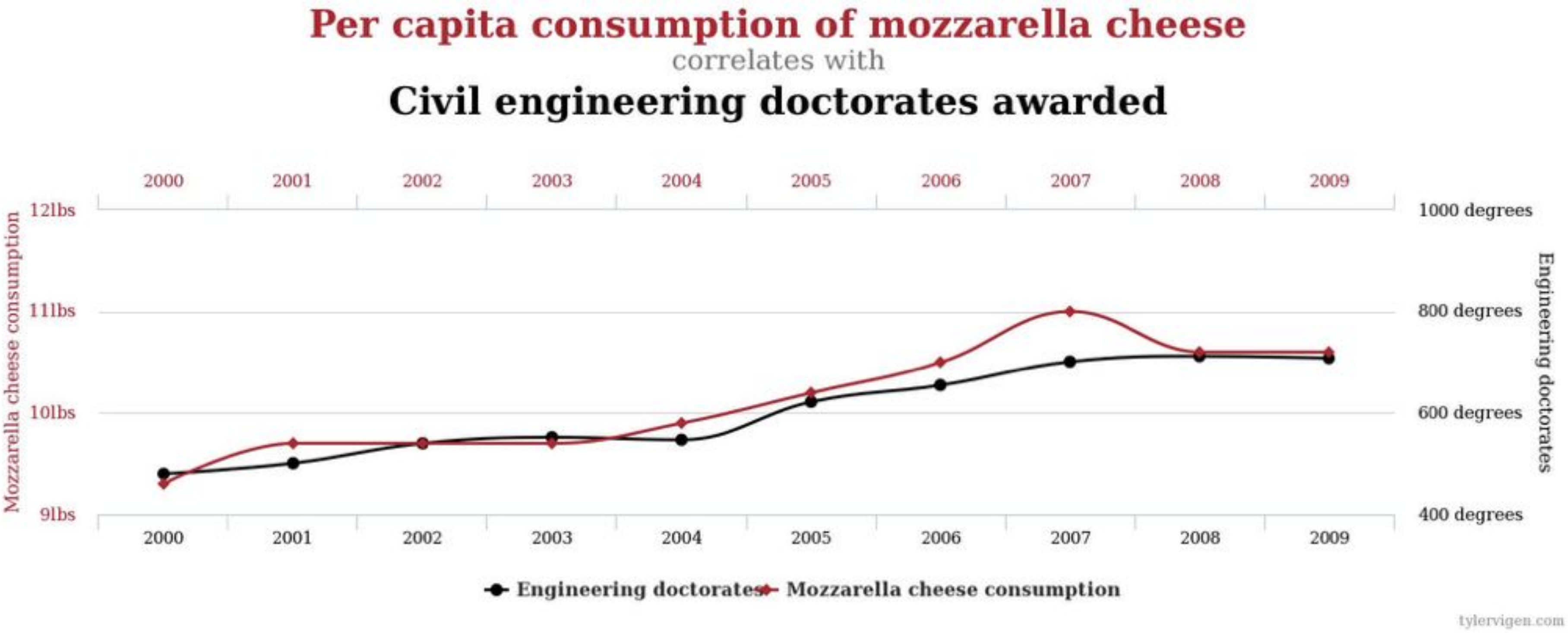
https://www.tylervigen.com/spurious-correlations



Correlation is **not** Causation



Correlation is **not** Causation



R^2 – How much variance in one variable can be explained by the other?

Square Pearson’s r to get R^2.

If we multiply this value by 100, that will be the % of variance explained in one variable by the other.

For our example on time spent studying and exam score, r squared = 0.81 as r = 0.9

This means that about 81% of the variance in exam score is explained by time spent studying.

Regression

Regression is where we want to predict the value of one variable (called our Outcome variable) on the basis of the value of one or more Predictor variables.

Simple regression is when we have one predictor, multiple regression is when we have more than one…

One of the most commonly used regression type is OLS (ordinary least squares) which works by minimising the distance (deviation) between the observed data and the linear model.

Statistical Models

Most of what we do in applying statistics to particular domains involves model building.

We build a statistical model and test whether it is a good fit for our data - in other words, whether it captures our data well.

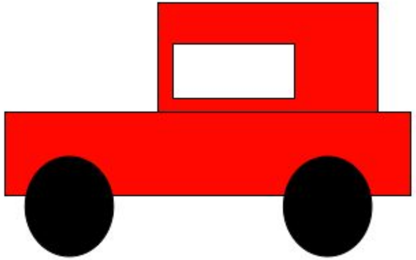
All models are an approximation of reality, and some are better than others.

Or to paraphrase the statistician George Box,

“all models are wrong but some are useful.”

Real data

Model 1



Model 2

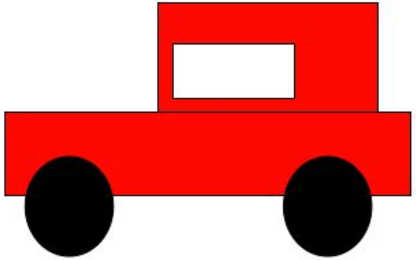
So how do we tell if a particular statistical model is a good fit to our data?

We can look at the extent to which our data deviate from a particular model (where deviation = error).

=

+ Error

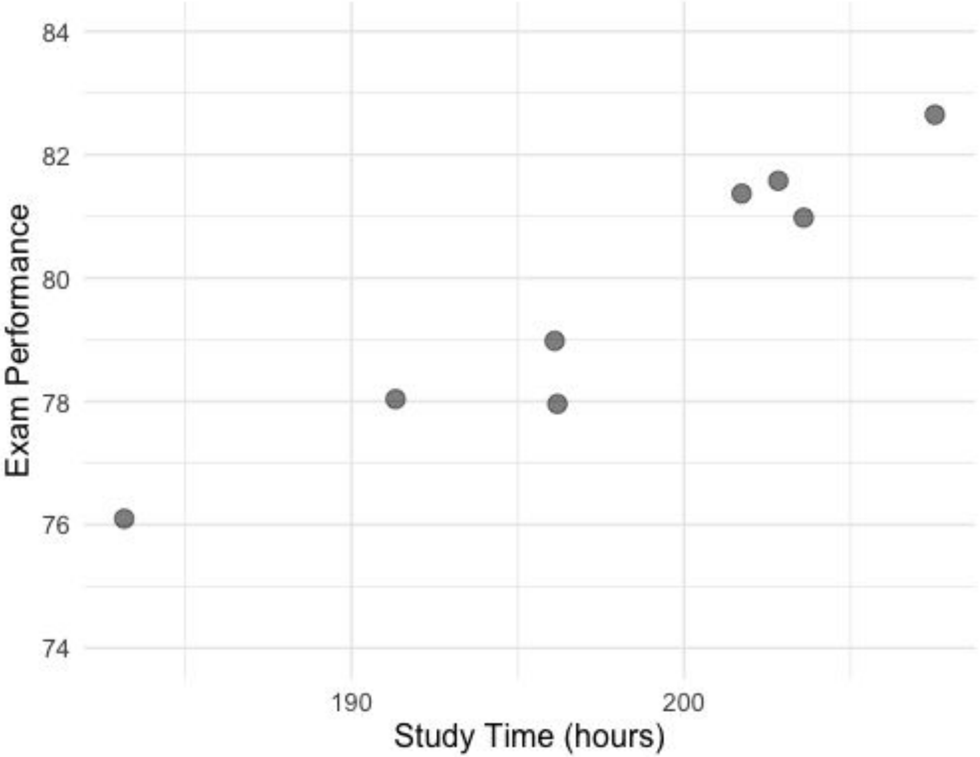
=



+ Error

We want to select the model which has the smallest error (aka model residuals).

Regression



We can plot data on exam performance and days spent studying.

Wouldn’t it be helpful if we could draw a straight line such that if we know the value on one axis (x say), we could predict the value on the other (y say) ?

Determining the best line

For any line, we can calculate what’s known as the Ordinary Least Squares.

The Ordinary Least Squares (OLS) method in regression provides us with a line that results in the least differences between the values predicted by the line and the data themselves.

Plotting a Regression Line

With OLS regression, we can plot a straight line that minimises the residuals (i.e., the error).

β_0 = intercept (when x=0)
β = gradient of the line
$Residual_i$ = difference between predicted score and actual score for participant *i*

$$y = \beta_0 + \beta x_i + Residual_i$$

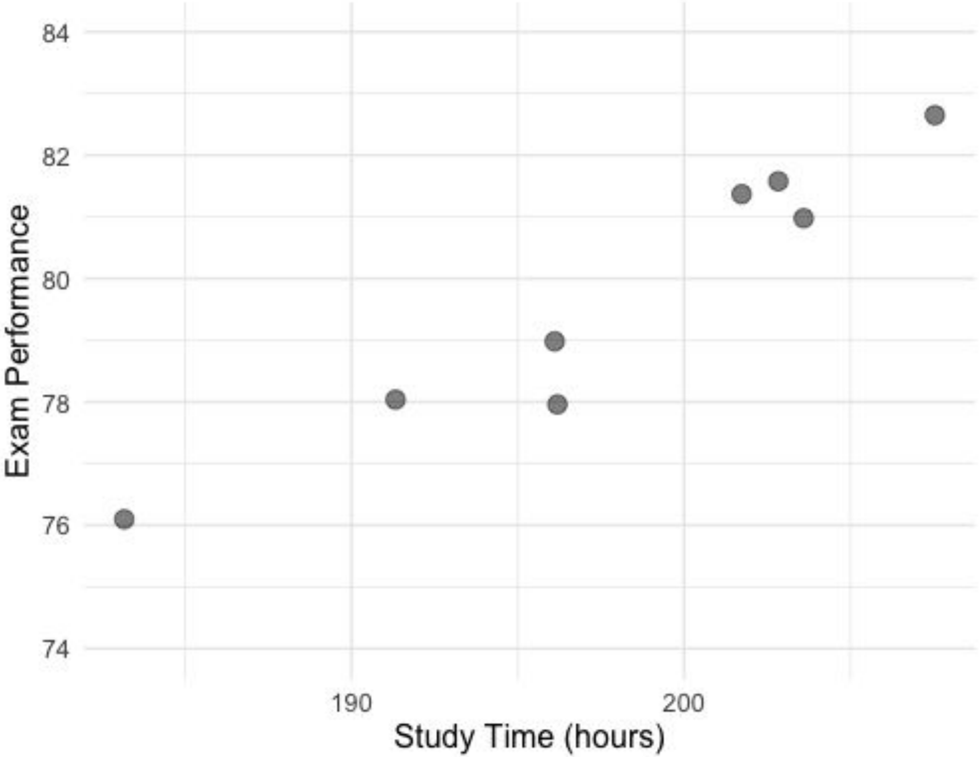
How do we determine how good a fit our regression line is?

We could work out by how much each observed value differs from the mean of y.

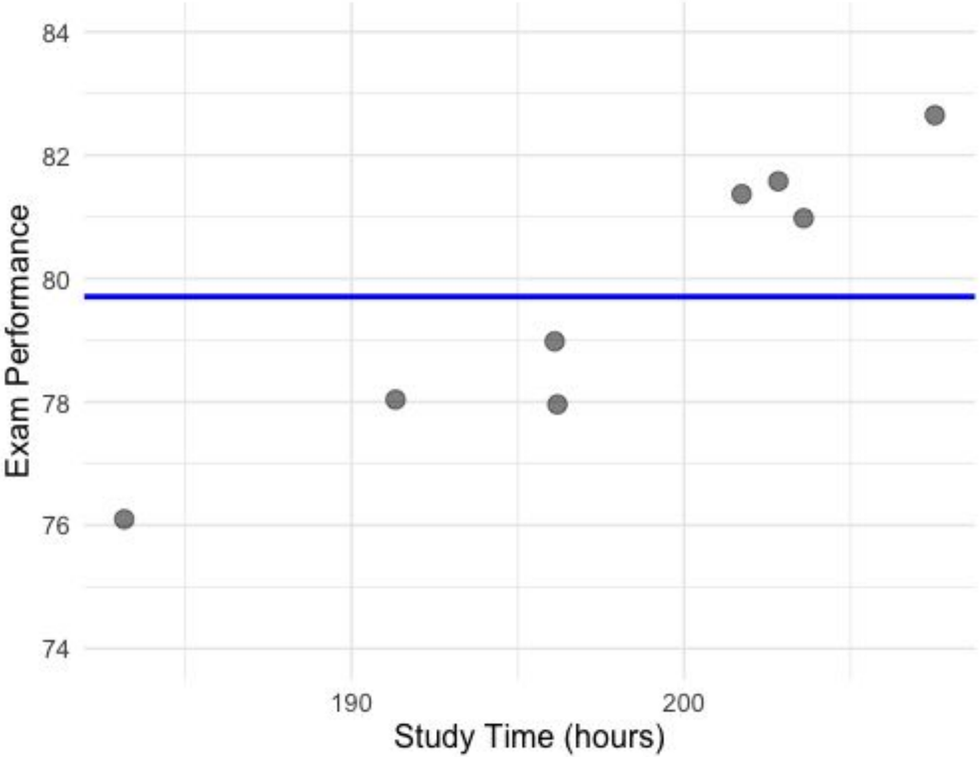
We could work out by how much each observed value differs from the regression line.

We could work out by how much the mean value of y differs from the regression line (for different values of x).

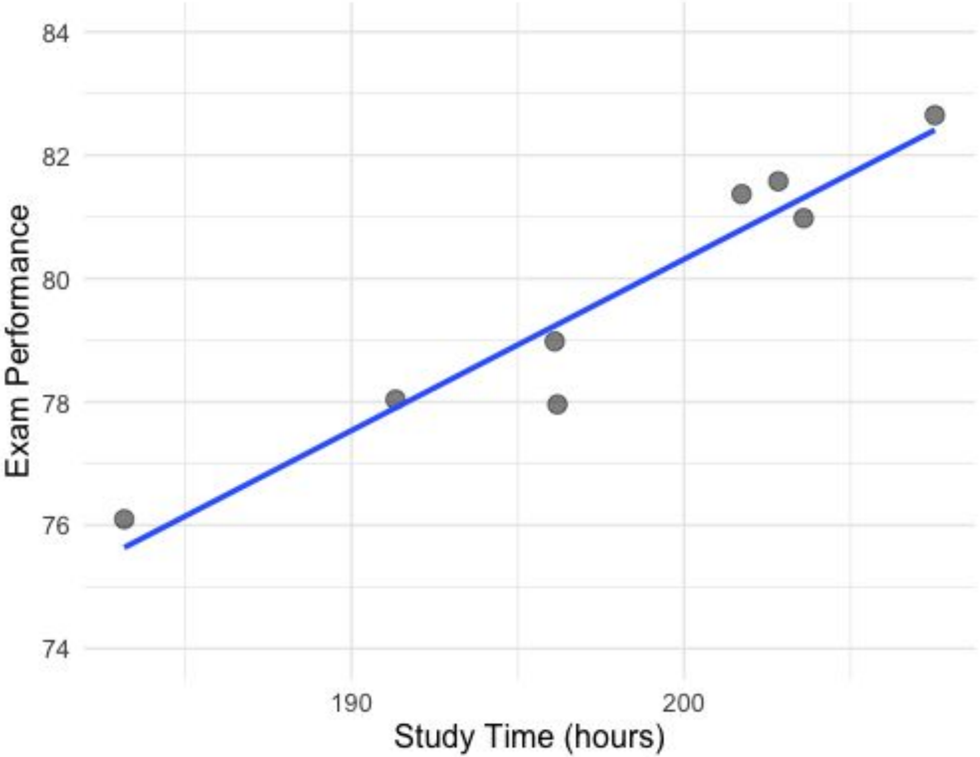
Our data



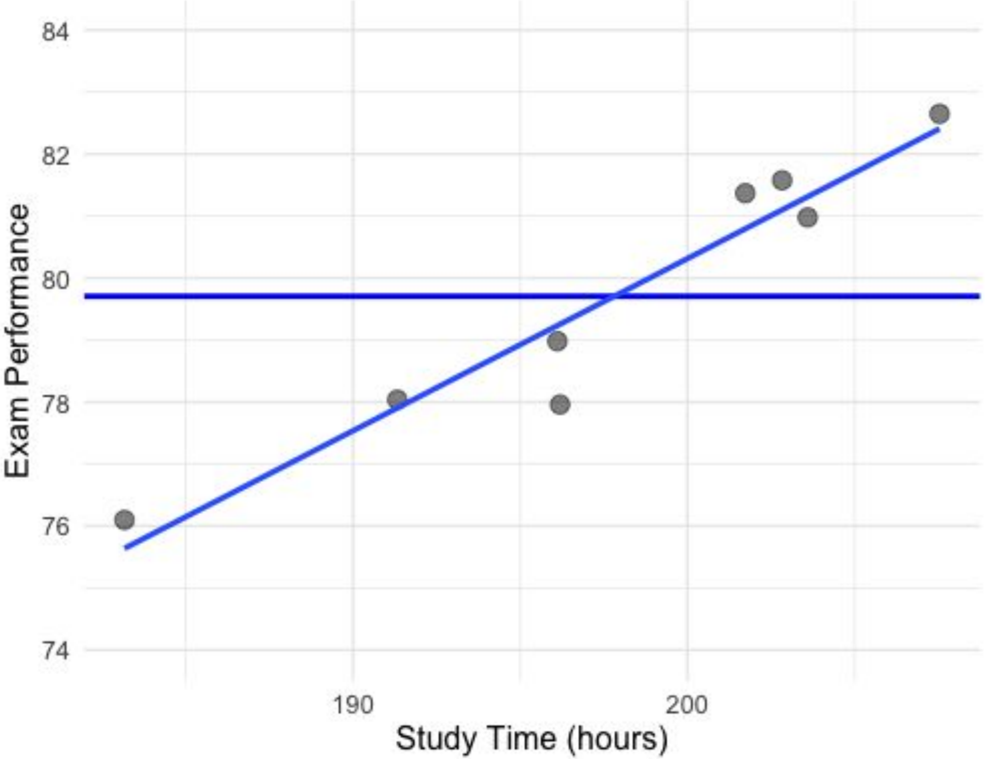
The mean as a model of our data (SS_T)



The regression line as a model of our data (SS_R)



Comparing the mean and regression lines as models of our data (SS_M)



Is our regression model any good?

If SS_M is large, then the regression model is better than the mean in terms of predicting values of the outcome variable.

If SS_M is small, then the regression model is not much better than the mean in terms of predicting values of the outcome variable.

Is our regression model any good?

We can calculate the proportion of improvement in prediction by looking at the ratio of SS_M to SS_T. Actually, this is R^2:

$$R^2 = \frac{SS_M}{SS_T}$$

And this is the same R^2 that we worked out by squaring the Pearson correlation coefficient.

Is our regression model any good?

We can also assess how good our model is by using the F-test.

The F-test is based on the ratio of the improvement due to the model (SS_M) and the difference between the model and the observed data (SS_R).

Rather than use the Sums of Squares themselves, we use the Mean Squares (MS_M and MS_R) which we get by dividing the Sums of Squares by their respective degrees of freedom.

$$F = \frac{MS_M}{MS_R}$$

Is our regression model any good?

A good model will have large MS_M and a small MS_R

In other words, the improvement of the model compared to the mean will be good.

$$F = \frac{MS_M}{MS_R}$$

The difference between the model and our observed data will be small.

Is our regression model any good?

If MS_M is large and MS_R is small, then F will be large.

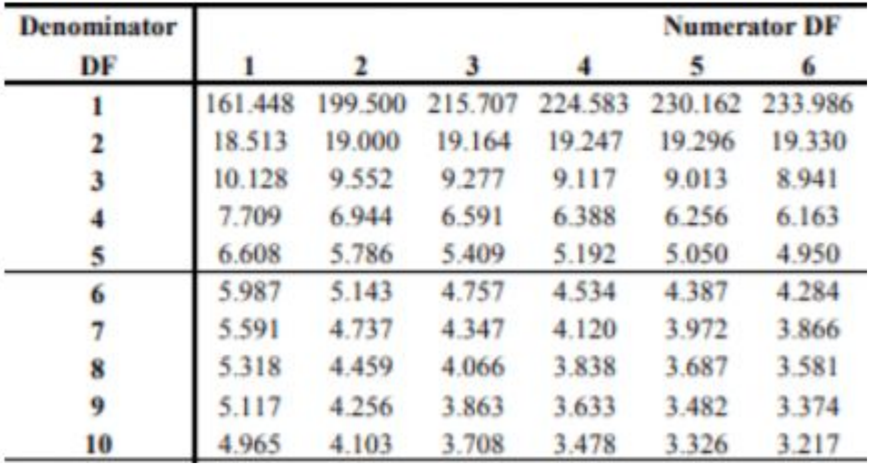
We can determine whether our F value is significant by looking up the critical values on the F table.

For SS_M the degrees of freedom = number of variables in model (in our case 2).

For SS_R the degrees of freedom = number of observations – number of parameters being estimated, including the constant (in our case 8-2 = 6)

Is our regression model any good?

In our example, df numerator = 2, df denominator = 6 for our example. Here is a portion of the F table for a .05 alpha level.



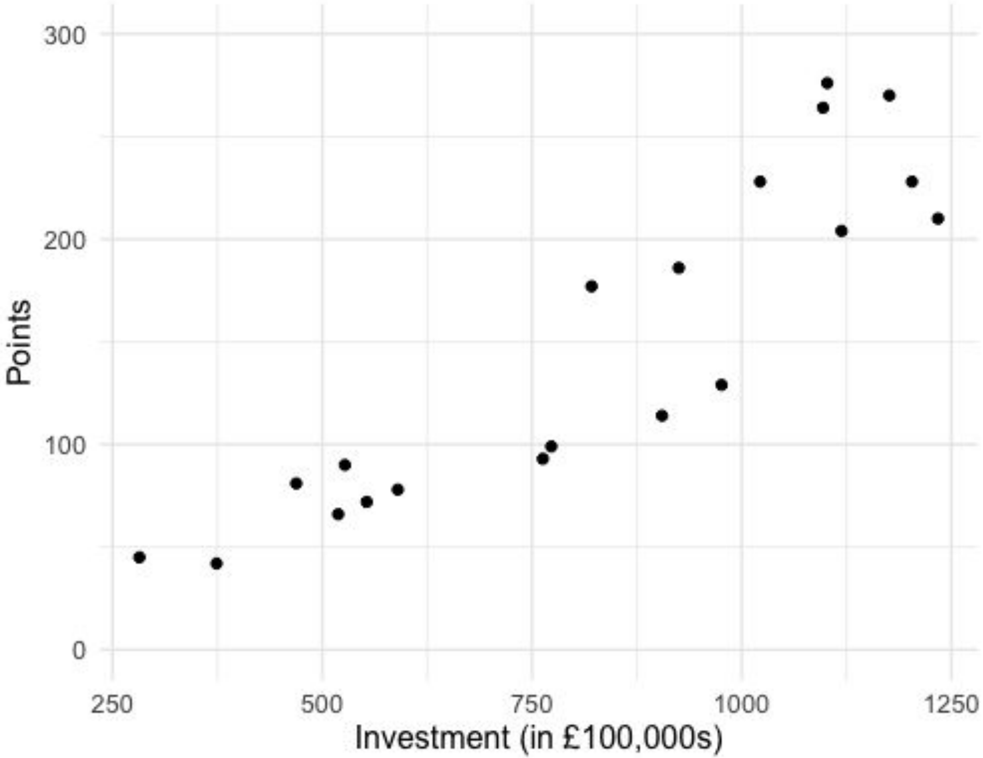
Denominator					Numerator DF	
DF	1	2	3	4	5	6
1	161.448	199.500	215.707	224.583	230.162	233.986
2	18.513	19.000	19.164	19.247	19.296	19.330
3	10.128	9.552	9.277	9.117	9.013	8.941
4	7.709	6.944	6.591	6.388	6.256	6.163
5	6.608	5.786	5.409	5.192	5.050	4.950
6	5.987	5.143	4.757	4.534	4.387	4.284
7	5.591	4.737	4.347	4.120	3.972	3.866
8	5.318	4.459	4.066	3.838	3.687	3.581
9	5.117	4.256	3.863	3.633	3.482	3.374
10	4.965	4.103	3.708	3.478	3.326	3.217

So we would need an F value greater than 5.143 for our result to be significant at $p < 0.05$.

Example

Imagine that you are Formula 1 team director. You’re interested in understanding how the number of points that a team scores is predicted by the amount of money invested in the team. As well as being in charge of F1, you also have a secret interest in statistical analysis. In `dataset1.csv` you will find (for each of the 20 drivers) the amount of money invested in their particular car (in £100,000s) plus the total number of points they were awarded over the season. Work out the simple linear regression equation that captures the relationship between investment (as our predictor) and points awarded (as our outcome).

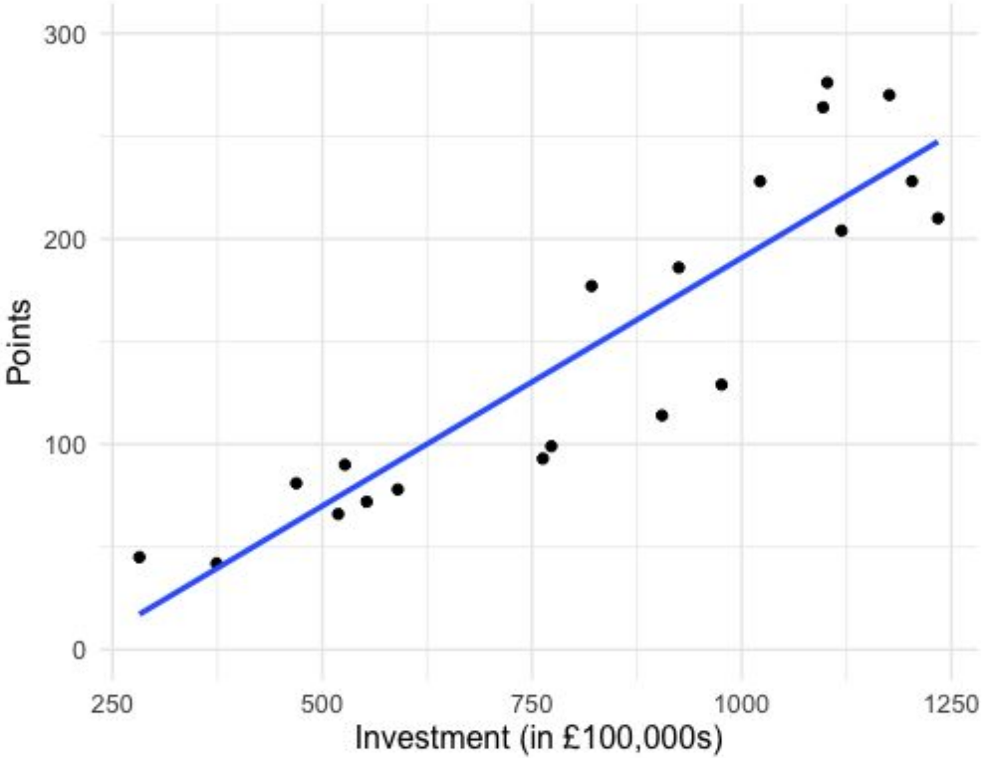
The data



We are going to use the tidyverse and the Hmisc libraries:

```
> library(tidyverse)
> library(Hmisc)
```

Visualising the regression line



```
set.seed(1234)
ggplot(dataset1, aes(x = investment,
                     y = points)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Investment (in £100,000s)",
       y = "Points") +
  theme_minimal() +
  theme(text = element_text(size = 12)) +
  ylim(0,300)
```

Building a simple linear model

Let's build two models using the `lm()` function- one (model0) is just the intercept (the mean of our outcome) predicting the outcome (points) while the second (model1) is a model with investment predicting the outcome (points).

```
model0 <- lm(points ~ 1, data = dataset1)
model1 <- lm(points ~ investment, data = dataset1)
```

We can compare the models to each other and calculated the F-ratio using the `anova()` function.

```
anova(model0, model1)
```

Comparing the two models

```
> anova(model0, model1)
Analysis of Variance Table

Model 1: points ~ 1
Model 2: points ~ investment
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1     19 120827                                  
2     18  22046  1     98781 80.654 4.547e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-ratio comparing our two models is 80.654 indicating our model with our predictor (investment) is a better fit than our model with just the intercept (the mean).

Summary of our regression model

```
> summary(model1)

Call:
lm(formula = points ~ investment, data = dataset1)

Residuals:
    Min      1Q  Median      3Q     Max 
-55.936 -20.840  -2.978  28.212  60.615 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -50.92329   23.44967  -2.172   0.0435 *  
investment    0.24166    0.02691   8.981 4.55e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35 on 18 degrees of freedom
Multiple R-squared:  0.8175,	Adjusted R-squared:  0.8074 
F-statistic: 80.65 on 1 and 18 DF,  p-value: 4.547e-08
```

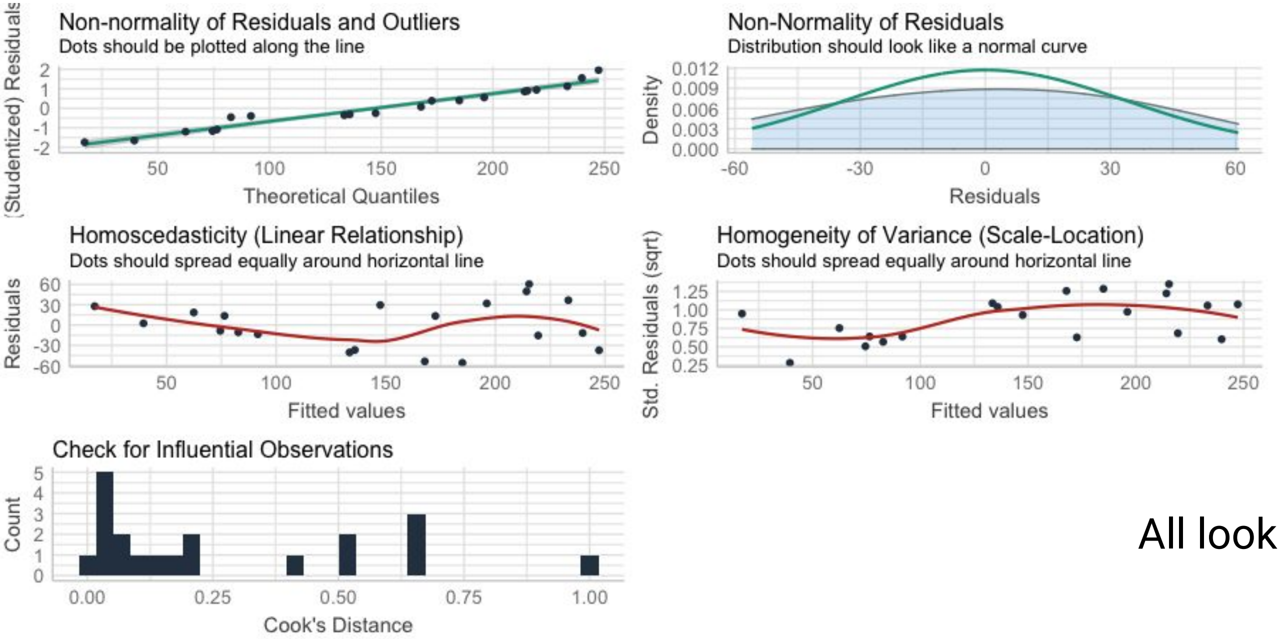
Here we have our parameter estimates.

Here we have the t-test associated with our predictor (investment).

Here are the R-squared and Adjusted R-squared values (which reflects the number of predictors in our model).

Checking our Assumptions

```
> performance::check_model(model1)
```



Non-normality of Residuals and Outliers

Dots should be plotted along the line

Non-Normality of Residuals

Distribution should look like a normal curve

Homoscedasticity (Linear Relationship)

Dots should spread equally around horizontal line

Homogeneity of Variance (Scale-Location)

Dots should spread equally around horizontal line

Check for Influential Observations

All looks generally ok...

What does it mean?

We would conclude from this that the amount of money spent on a driver does indeed predict the number of points they score in a season of F1. Specifically, for every £24,166 spent on them they will score one additional point.

Remember, regression is nothing more than prediction - a simple regression model allows us to predict the value of a variable on the basis of knowing about another variable (and its relationship to that variable).