Lecture 4 - Regression 1

Andrew Stewart

Andrew.Stewart@manchester.ac.uk

 @ajstewart_lang

Session	Topic	Lecturer
1	Introduction, Open Science, and Power	Andrew Stewart
2	Introduction to R	Andrew Stewart
3	Data Wrangling and Visualisation	Andrew Stewart
4	General Linear Model - Regression	Andrew Stewart
5	General Linear Model - Regression	Andrew Stewart
6	General Linear Model - ANOVA	Andrew Stewart
7	General Linear Model - ANOVA	Andrew Stewart
8	General Linear Model - ANOVA	Andrew Stewart
9	Signal Detection Theory	Ellen Poliakoff
10	Signal Detection Theory	Ellen Poliakoff
11	Revision Session	Andrew Stewart

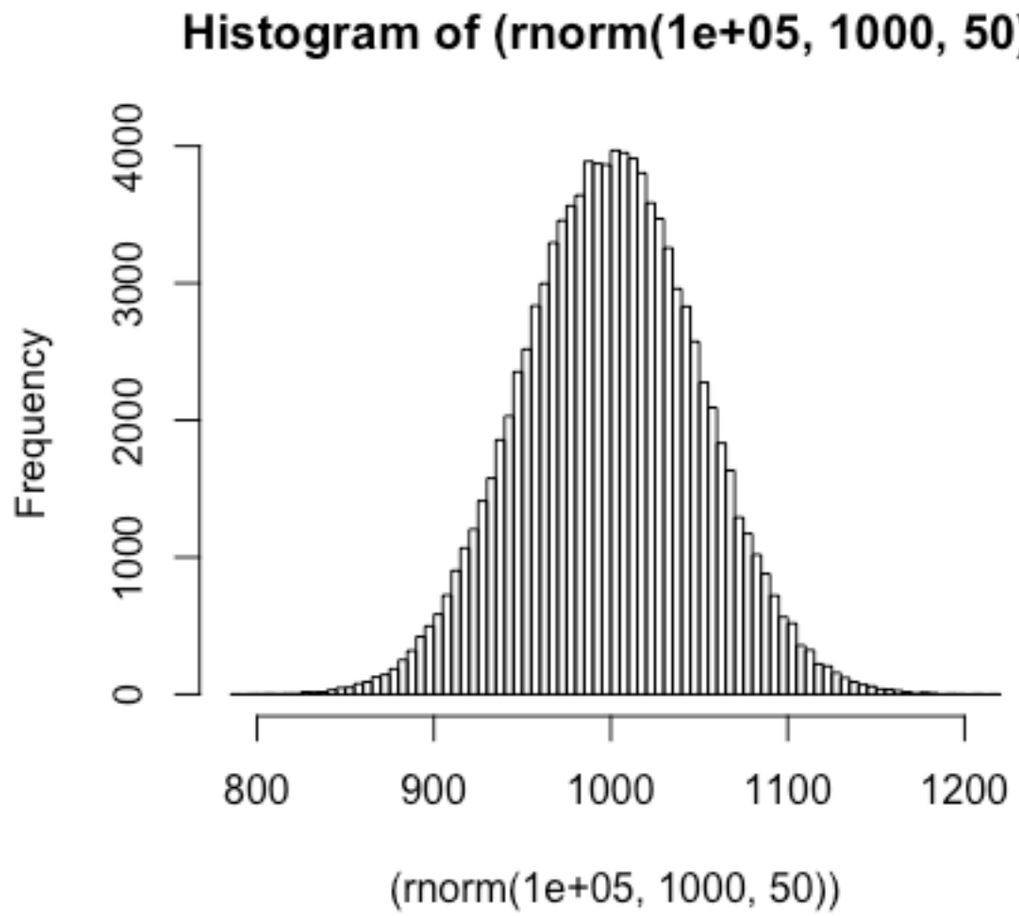
Semester 1 Assignments

ANOVA– Due around the end of November

Signal Detection Analysis – Due around mid-January

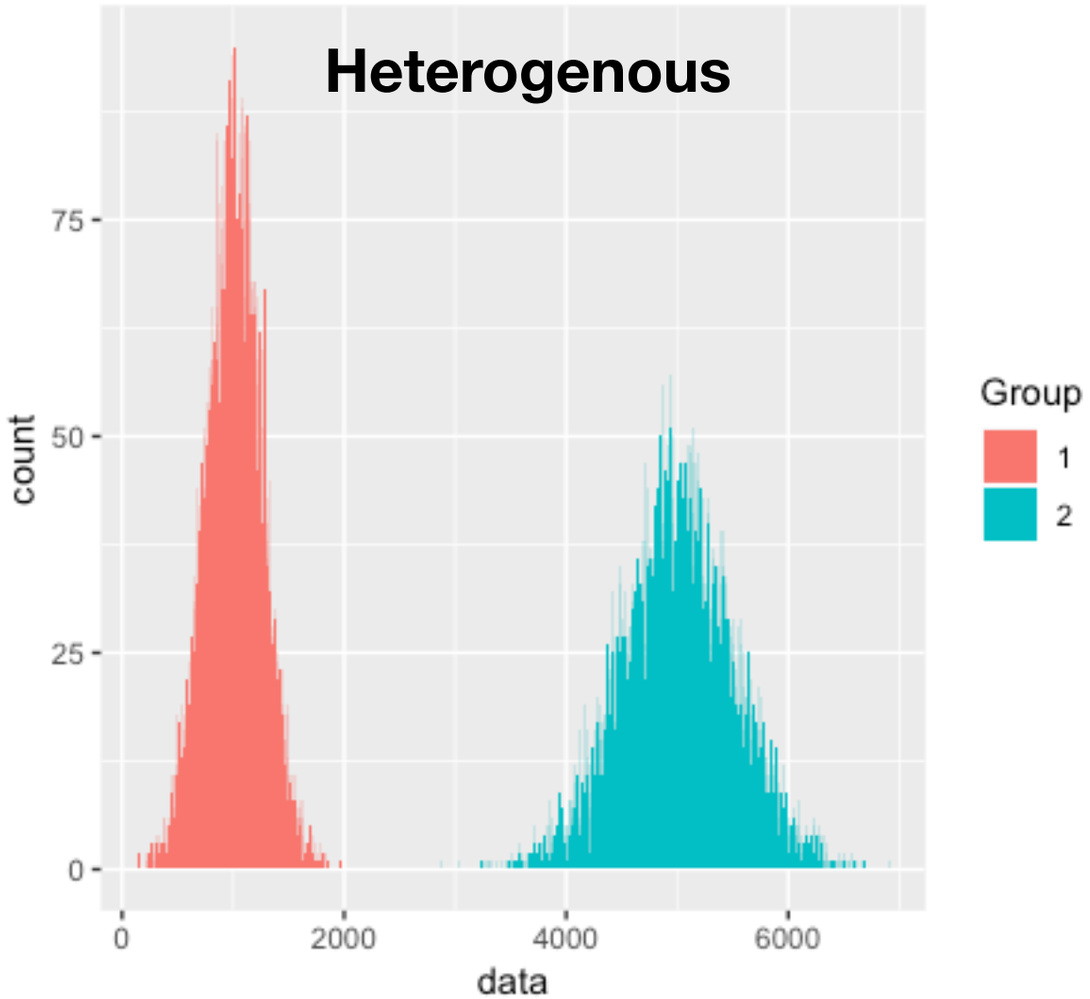
A Brief Reminder of the Assumptions of Parametric Tests

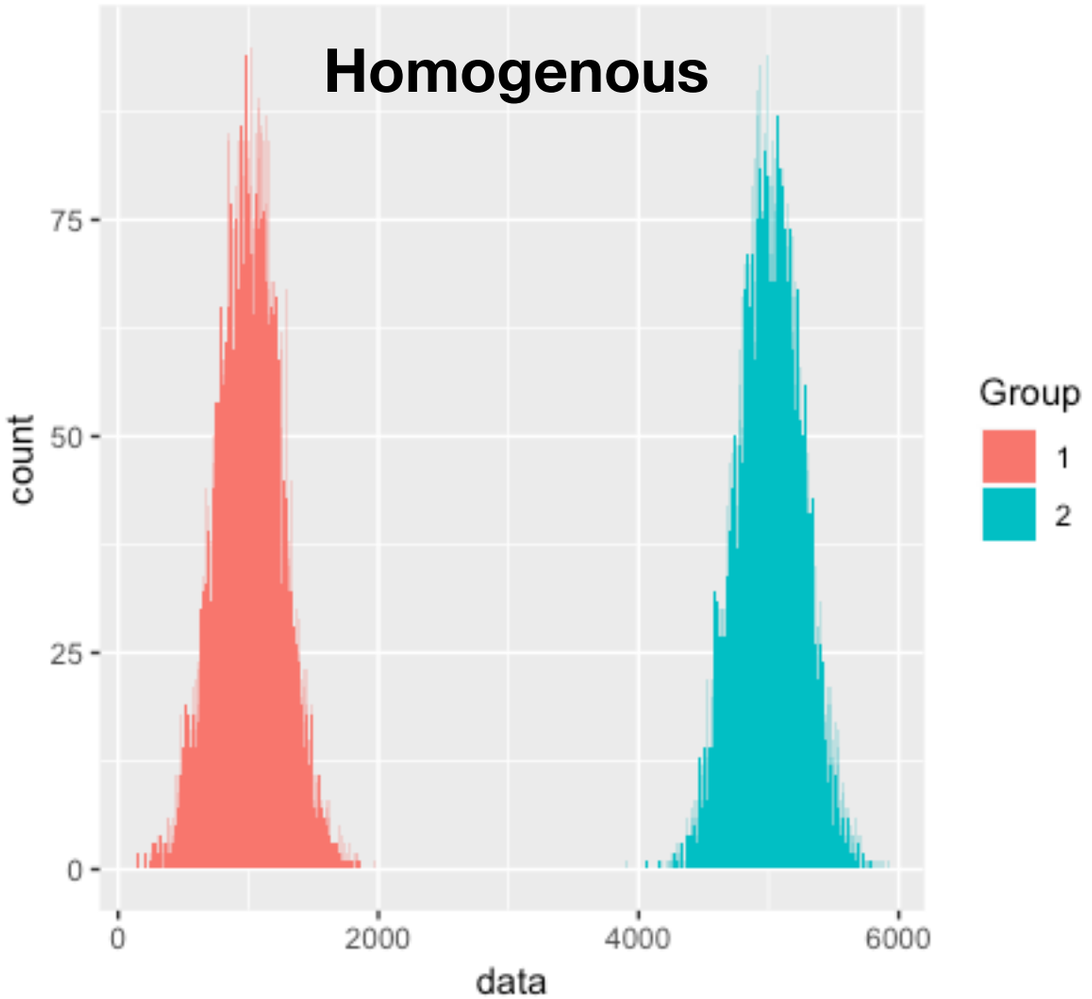
- Assumption 1 - the data are conditionally normally distributed - in practical terms, this means the *residuals* need to be normally distributed (although t-tests require the data to be normal).



- 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.

- Levene's Test for equality of variance. If it is non-significant, then it means that the variances are equivalent (i.e., we have homogeneity of variance).





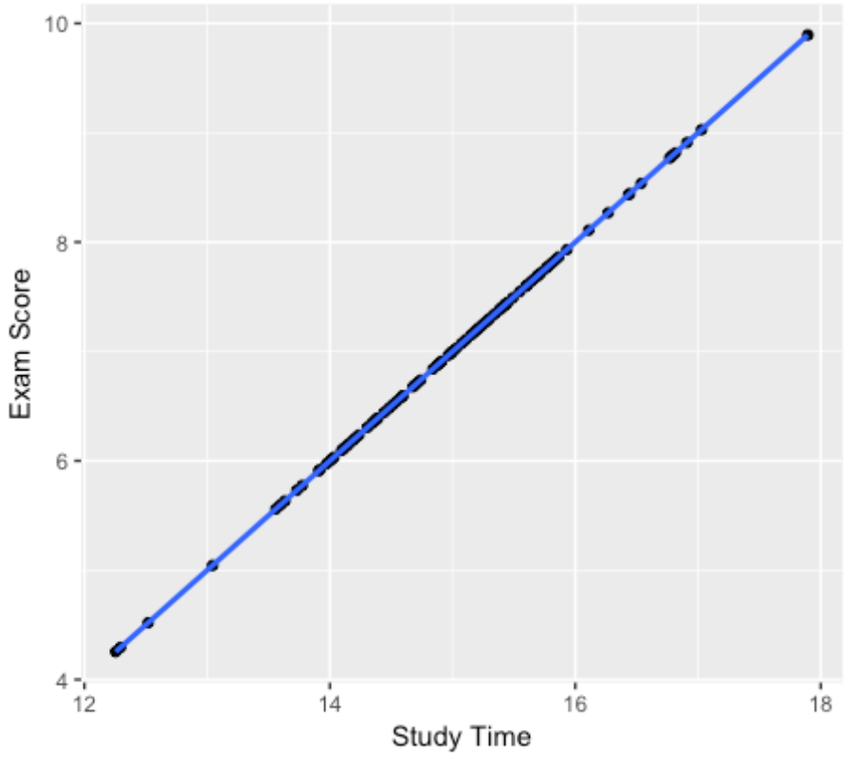
- Assumption 3 – Interval data – data should be measured on an interval scale. In other words, the distance between two adjacent points should be the same as the distance between any other two adjacent points. R can't tell you this – you need to determine it by yourself. Reaction time is a good example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

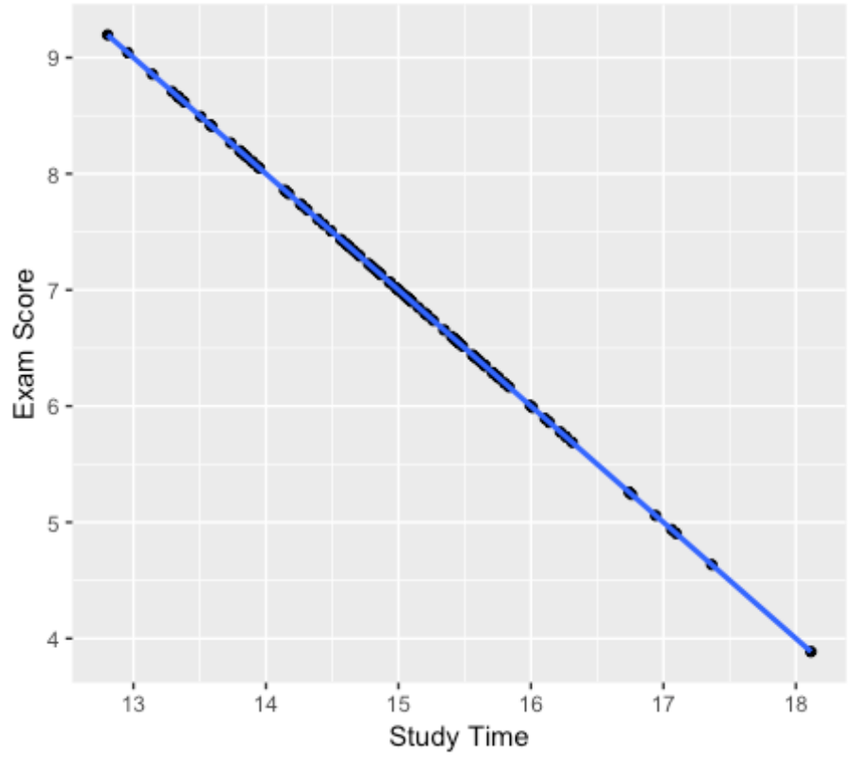
- Assumption 4 – Independence. The data from one participant does not affect the data from another (i.e., they are independent). In repeated measures designs, we expect the scores in the experimental conditions to be independent between participants.

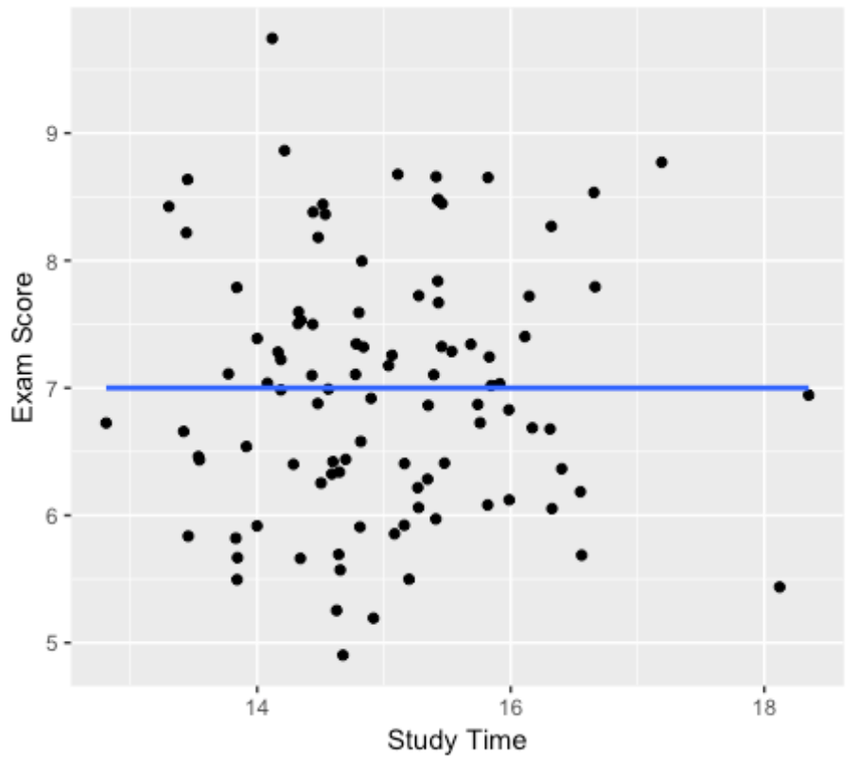
Remember correlation?

- Sometimes we want to know whether there's a relationship between two variables –
    - time spent studying statistics and performance in exam.

- They could be *positively* correlated, *negatively* correlated or *uncorrelated*.


Exam Score
Study Time
10
8
6
4
12
14
16
18


Exam Score
Study Time
9
8
7
6
5
4
13
14
15
16
17
18


Exam Score
Study Time
9
8
7
6
5
14
16
18

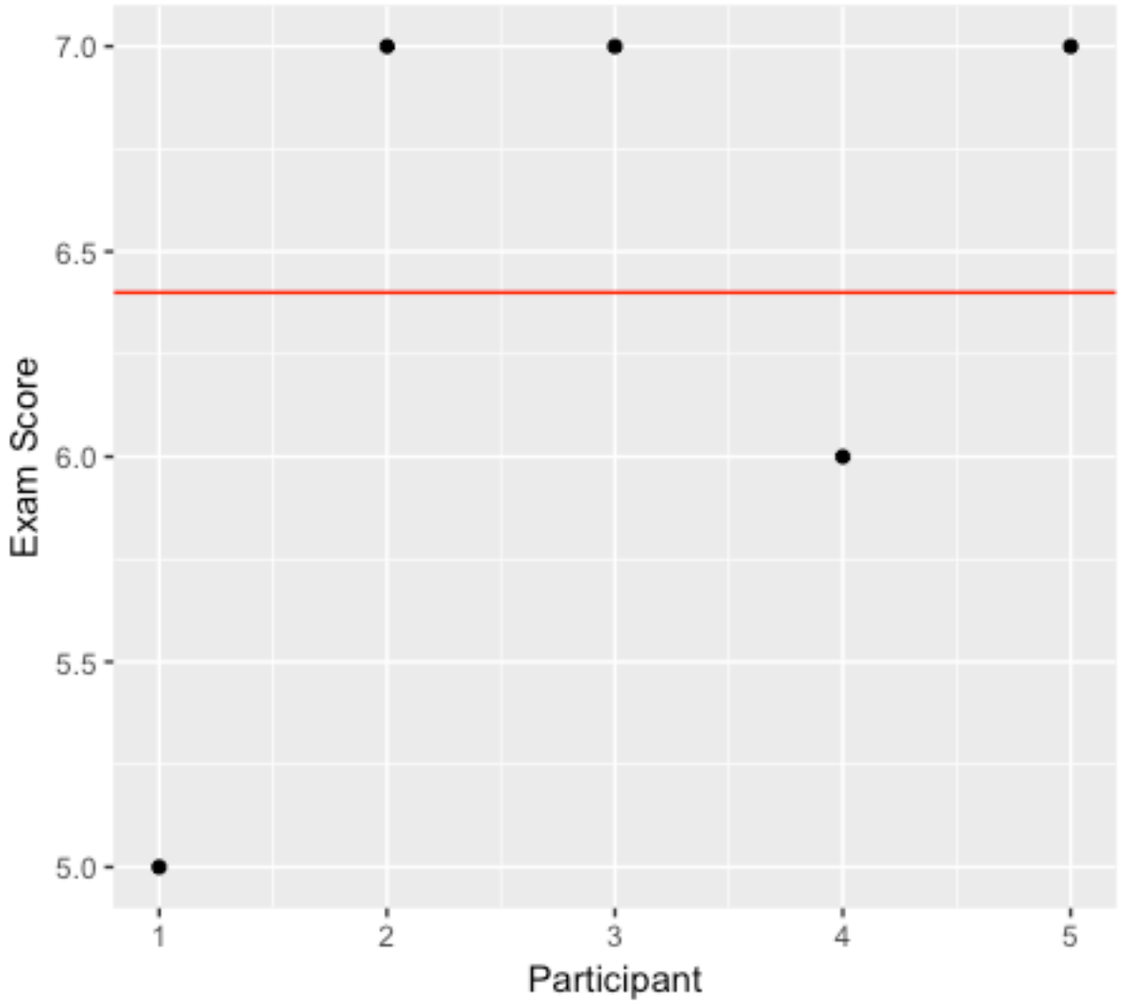
Covariance

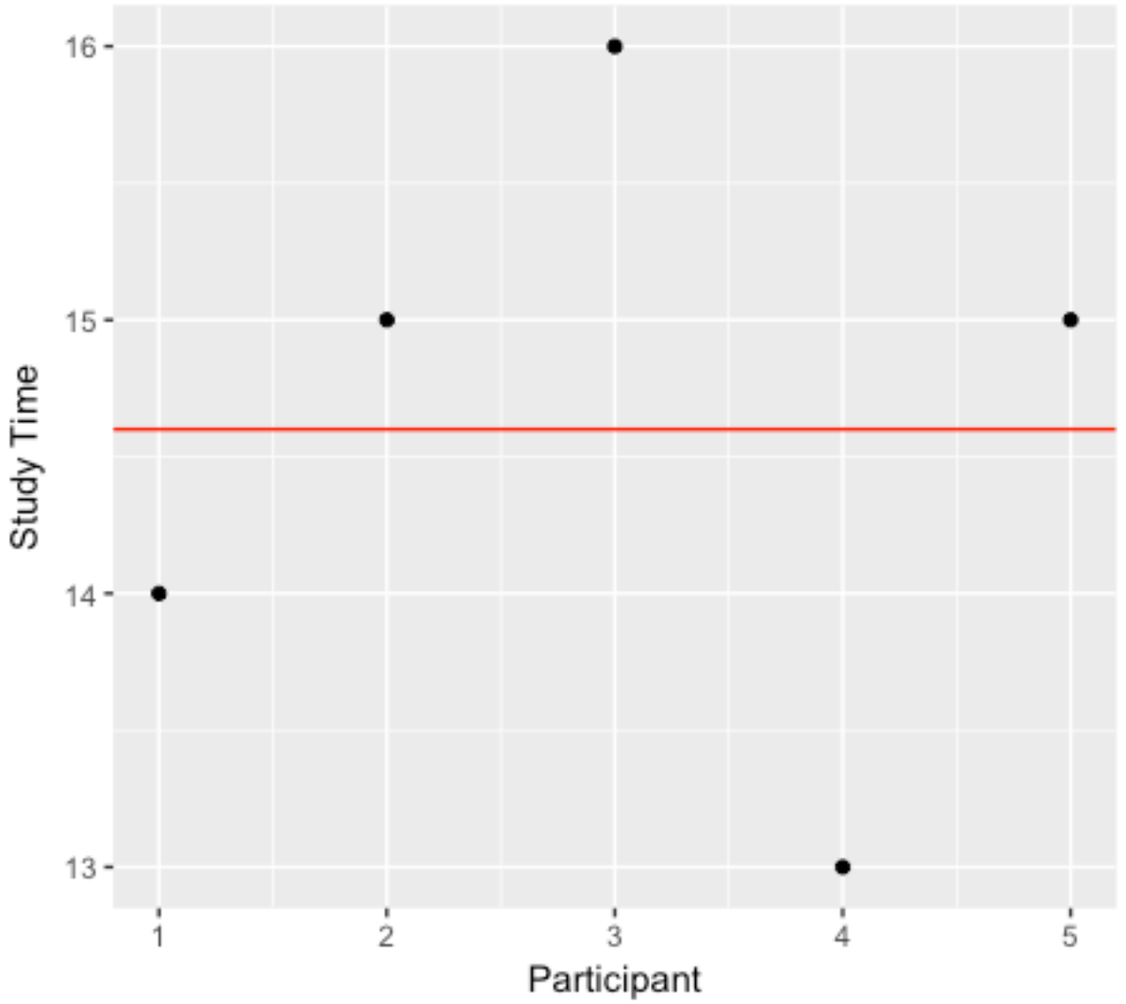
- Remember variance?

- It's the measure of the average 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.





The red 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 variable co-vary.

- For participants 2, 3 and 5, their scores on each variable are all below the respective means for each variable, for participants 1 and 4 their scores are all above the respective means for each variable.

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

$$\mathrm{cov}(x,y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- For our example:

Participant	Study Time (X)	Exam Score (Y)	Mean X	Mean Y	X - Mean X	Y - Mean Y	(X - Mean X) * (Y - Mean Y)
1	14	5	14.6	6.4	-0.6	-1.4	0.84
2	15	7	14.6	6.4	0.4	0.6	0.24
3	16	7	14.6	6.4	1.4	0.6	0.84
4	13	5	14.6	6.4	-1.6	-0.4	0.64
5	15	7	14.6	6.4	0.4	0.6	0.24
							Σ = 2.8

Cov (x,y) = 2.8/N-1 = 2.8/4 = 0.7

- 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:

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

- 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.

- SD of Study Time (X) = 1.140175
- SD of Exam Score (Y) = 0.8944272

$$\text{Pearson's R} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N - 1 s_x s_y}$$

$$\text{Pearson's R} = \frac{2.8}{4 \text{ X } 1.14 \text{ X } 0.89}$$

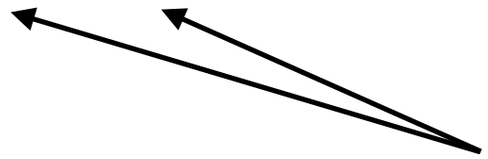
$$= 0.69$$

- In R, you need to install the “Hmisc” package first, and then load it:

```
> library(Hmisc) # Needed for correlation
```

- Our data frame is called “covary” and looks like this:

Participant	Study Time	Exam Score	Mean_Exam_Score	Mean_Study_Time
1	14	5	6.4	14.6
2	15	7	6.4	14.6
3	16	7	6.4	14.6
4	13	6	6.4	14.6
5	15	7	6.4	14.6



- To calculate Pearson’s R for these two variables we type:

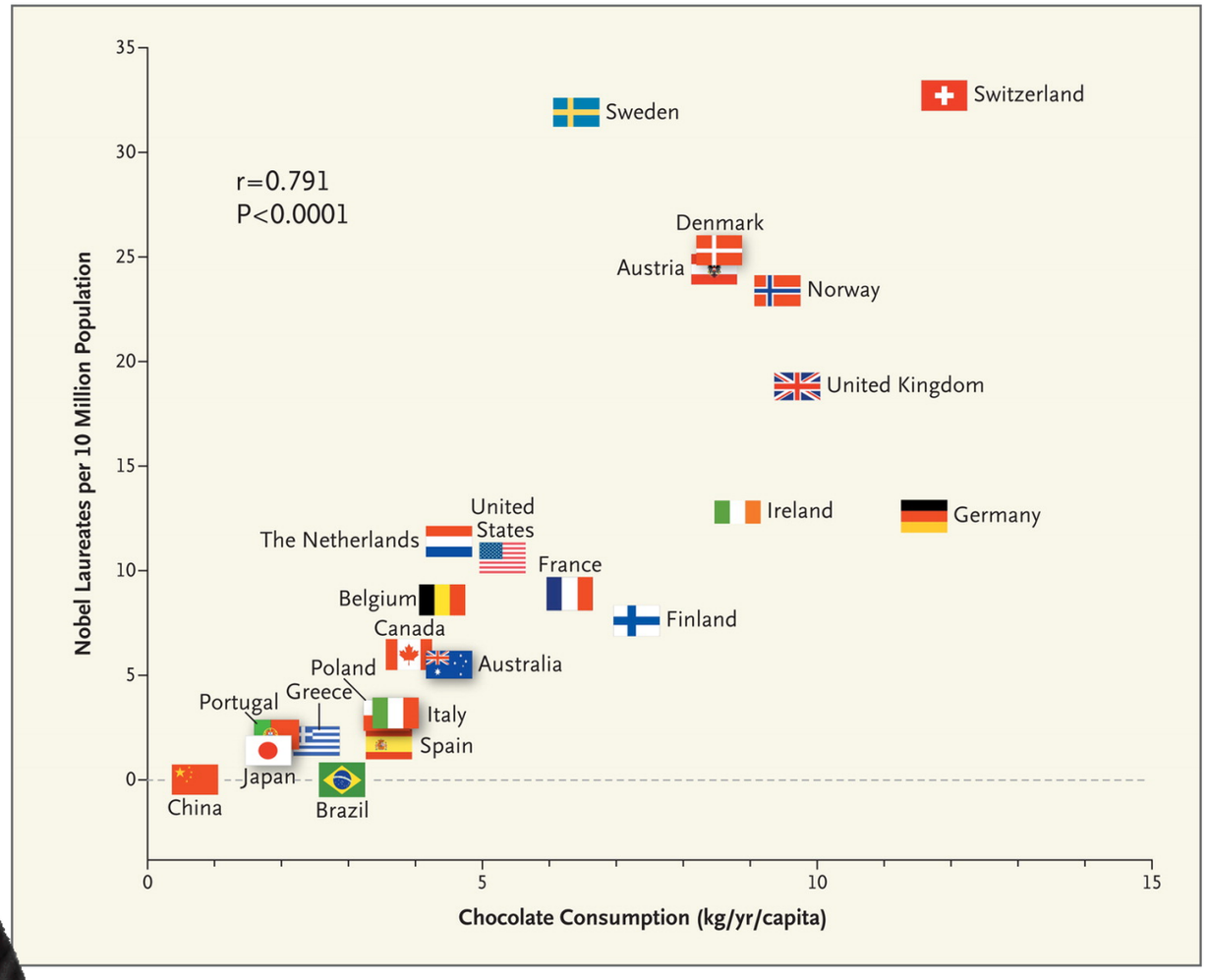
```
> rcorr(covary$`Study Time`, covary$`Exam Score`)
```

```
> rcorr (covary$`Study Time`, covary$`Exam Score`)
     x    y
x 1.00 0.69
y 0.69 1.00

n= 5


P
  x      y
x        0.2006
y 0.2006
```

- The Pearson’s r value is 0.69 - but it is not significant as p = 0.20

Correlation is *not* Causation

There is a high correlation (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

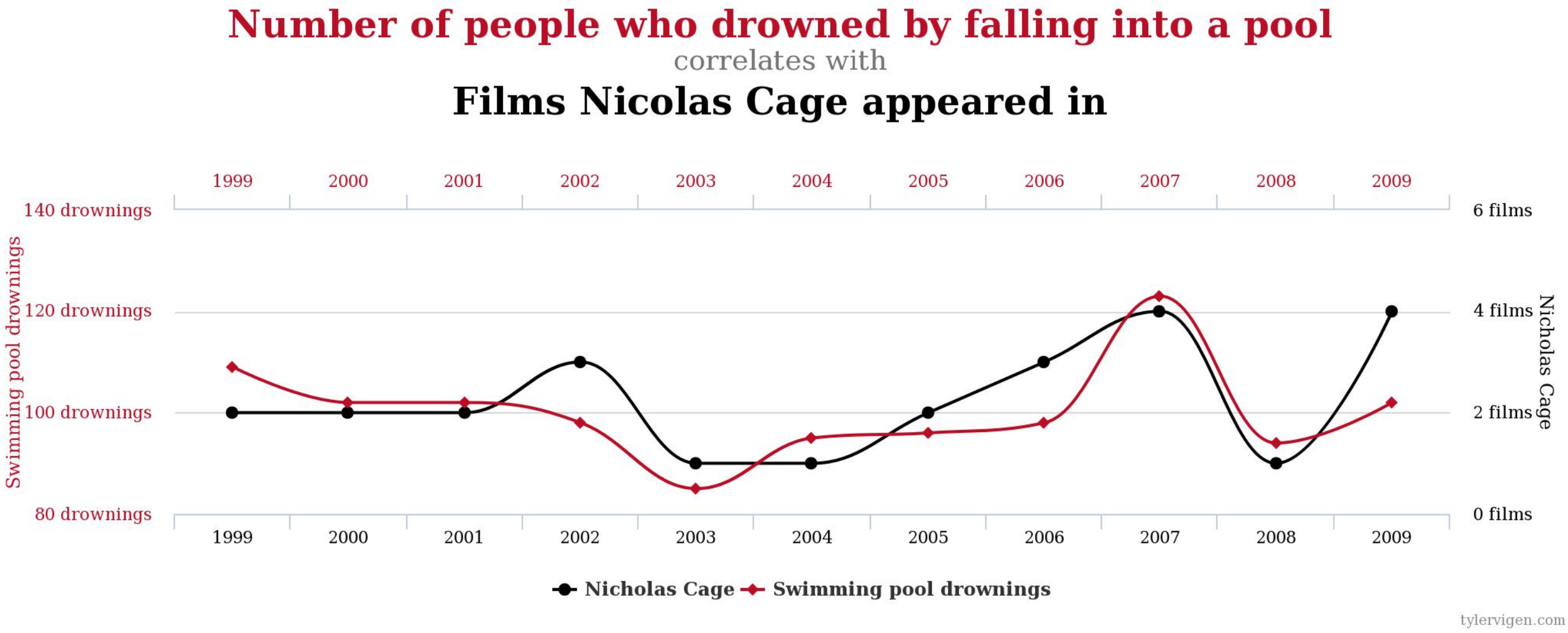
- When interpreting correlation data one common pitfall is to assume that the score on one variable *causes* a particular score on the other. This is wrong!

- Very often, common sense would suggest causation – e.g., time spent studying improves exam score. Again, you cannot make any claim about causation from correlation.

- There may be a third variable that we don't know about – in this case, maybe a positive attitude to studying.

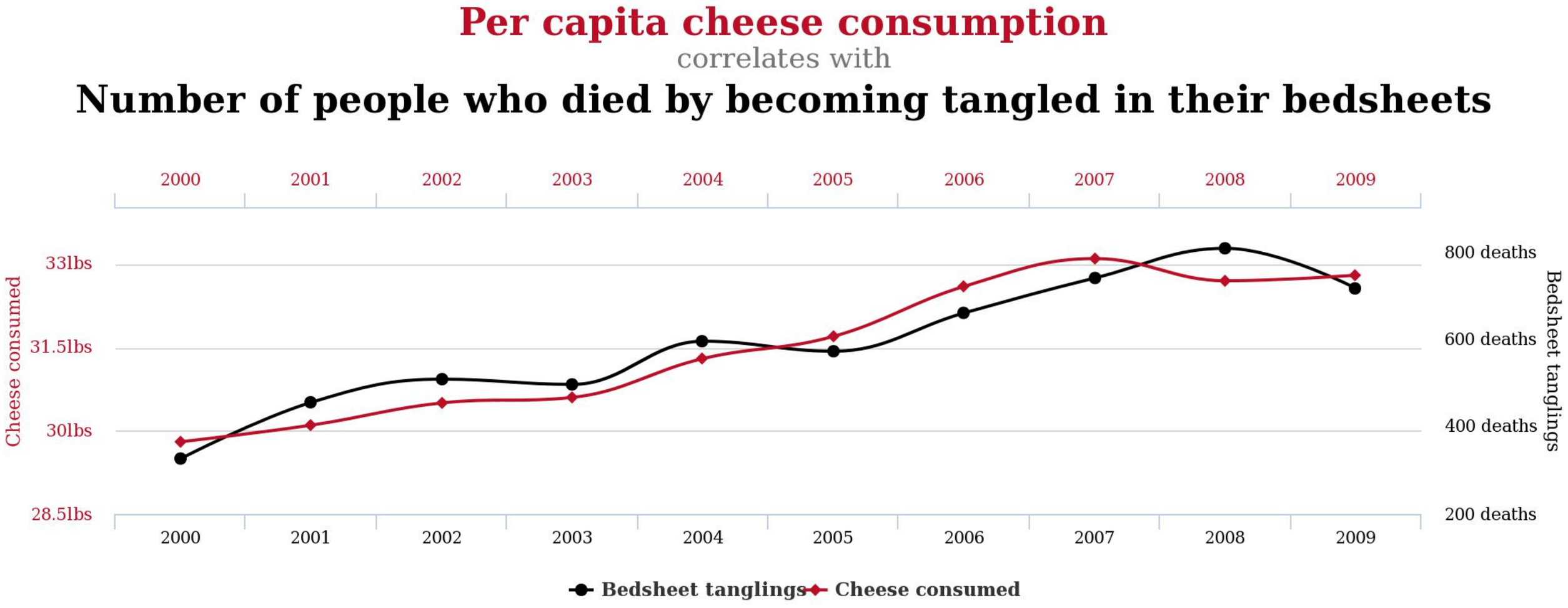
- Additionally, spurious correlations can be found all over the place…

Correlation is *not* Causation



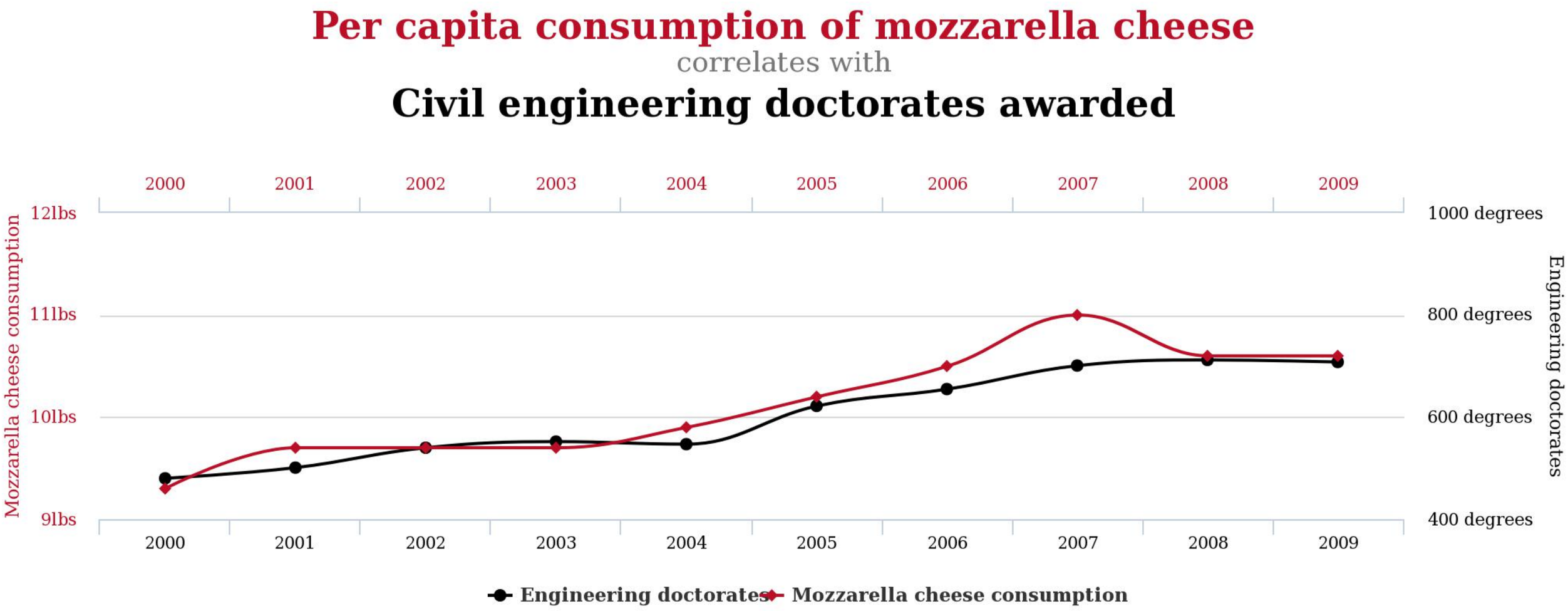
tylervigen.com

Correlation is *not* Causation



tylervigen.com

Correlation is *not* Causation



tylervigen.com

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

- Simply square Pearson's r to get r square.

- If we multiple 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 = 0.4761 as r = 0.69

- This means that about 48% of the variance in exam score is explained by time spent studying. It may not be statistically significant, but you might think it is still meaningful.

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…

- 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.

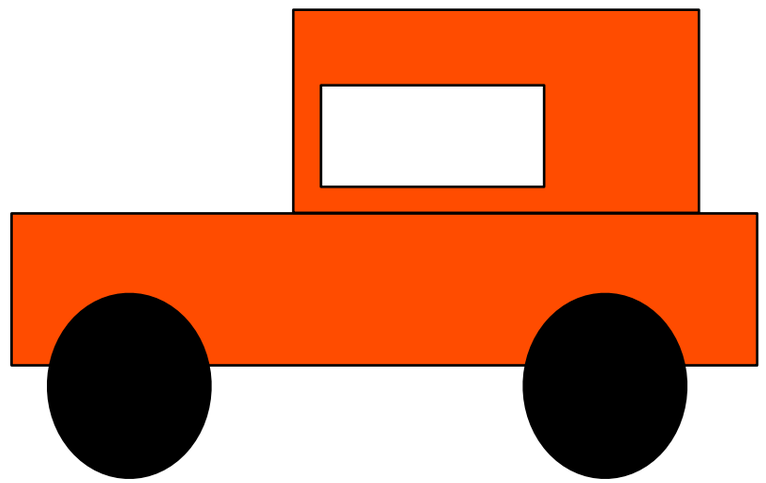
Statistical Models

- Most of what we do in applying statistics in Psychology is model building. We build a statistical model and test whether it is a good fit for our data - in other words, whether it describes 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)…

Real data

=

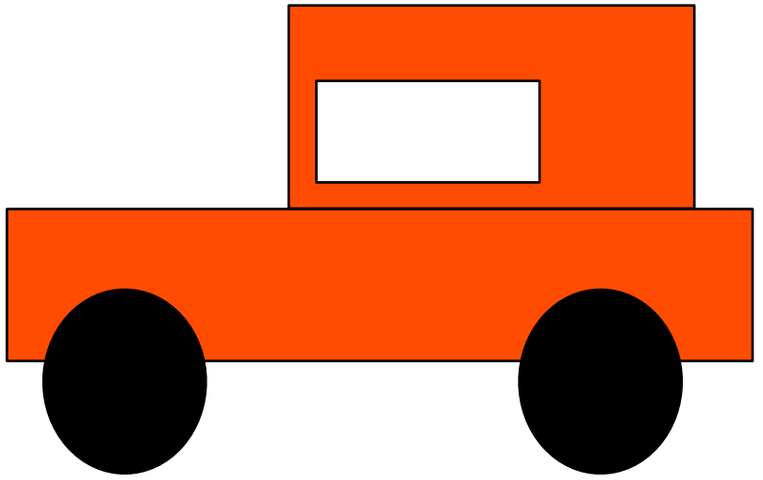
Model 1

+ Error

Real data

=

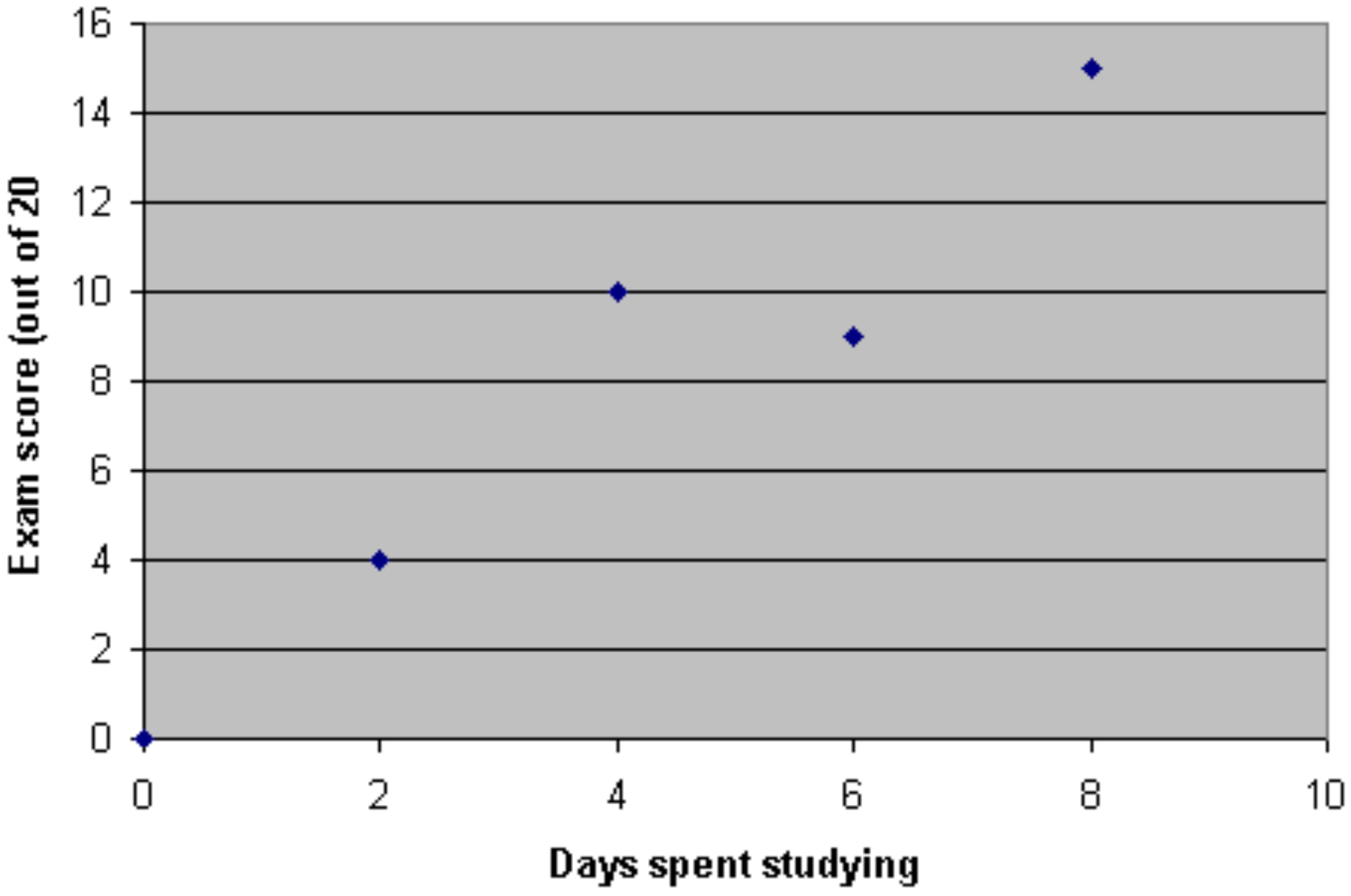
Model 2



+ 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) ?

Plotting a straight line

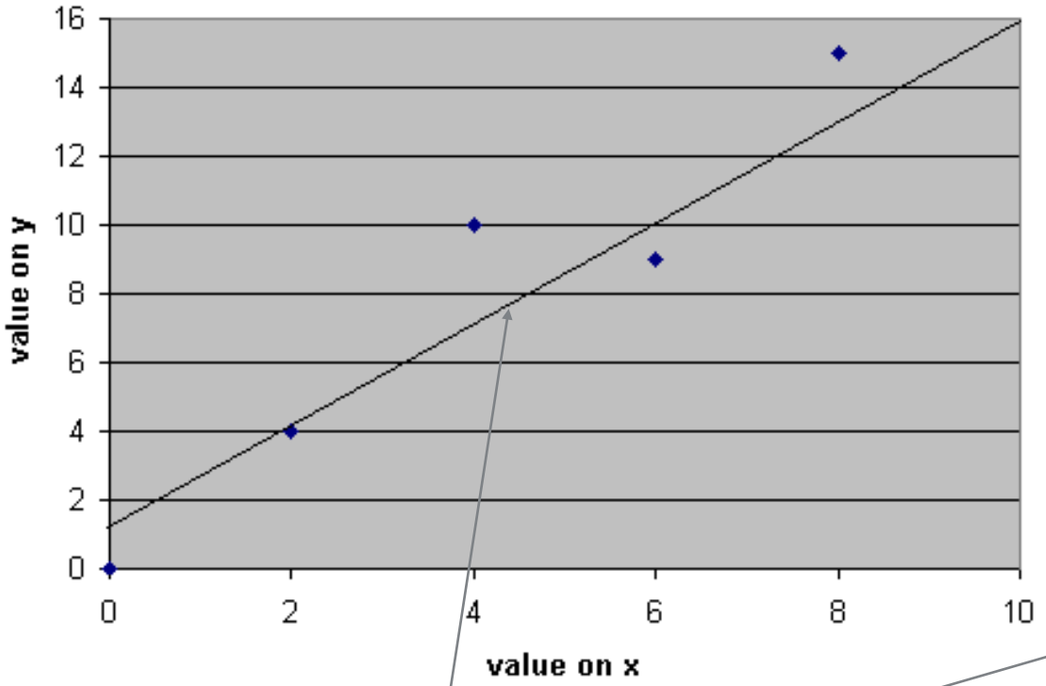
- For any data plots such as on the previous slide, when we have one predictor (x) we could plot many straight lines.

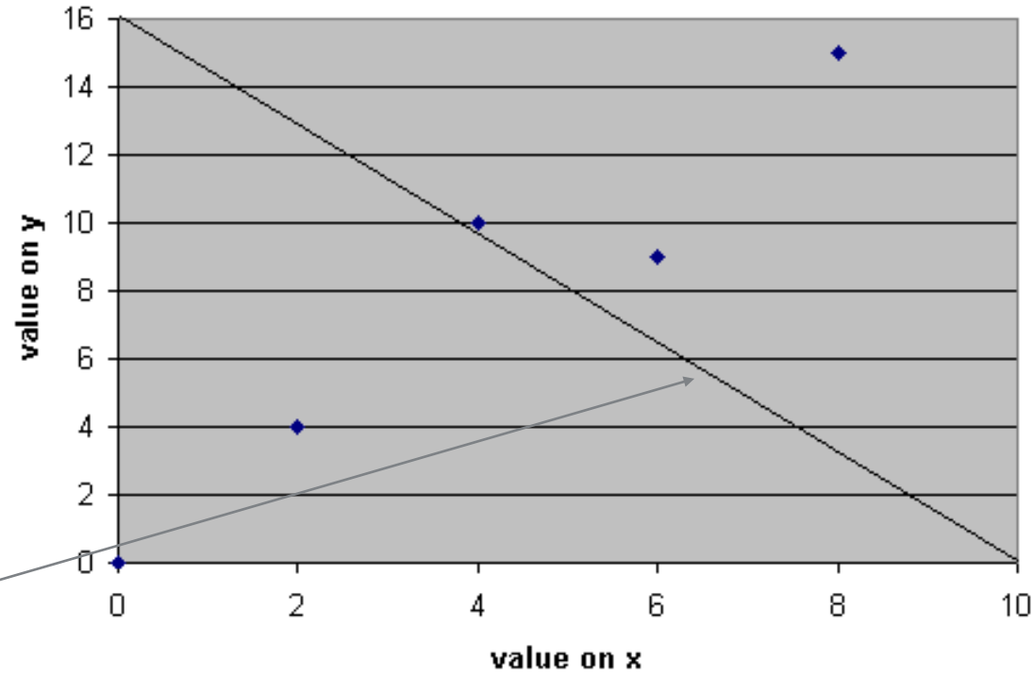
$$y = \beta x_i + \beta_o + residual_i$$

β = gradient of the line

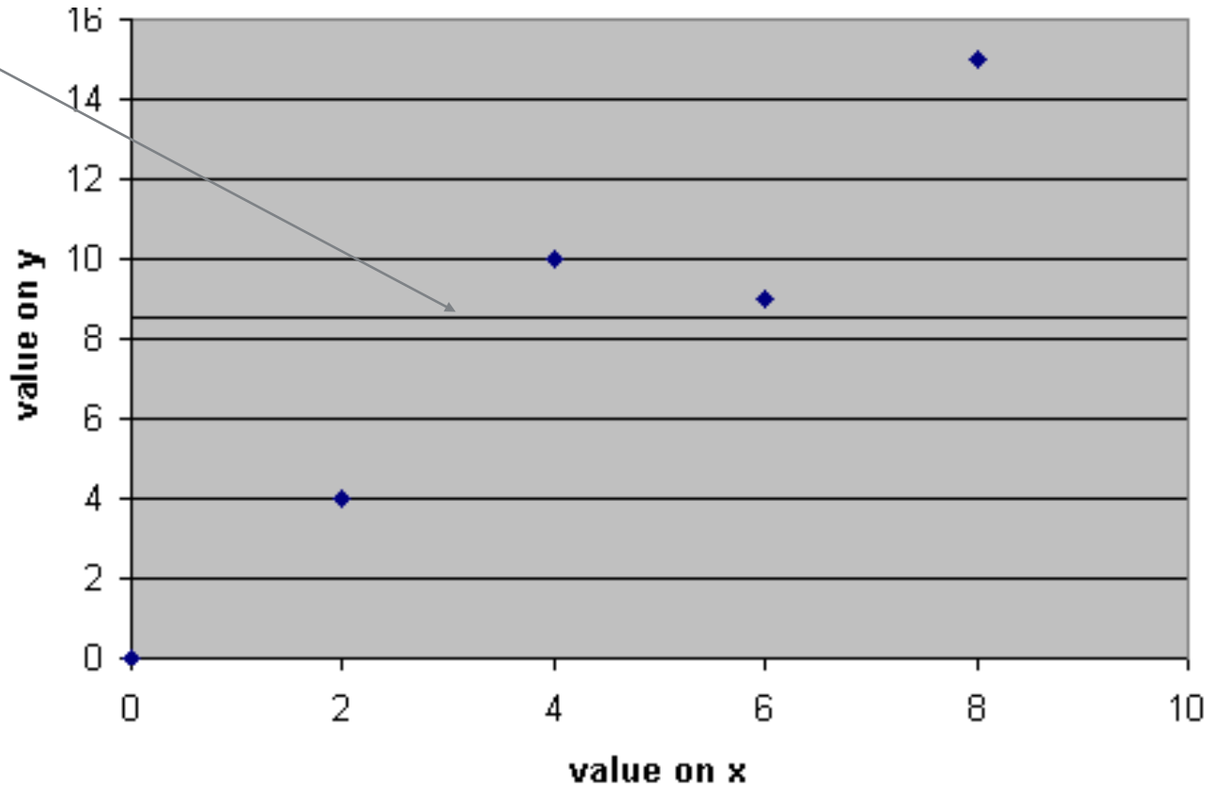
β_o = intercept (when x=0)

$residual_i$ = difference between predicted score and actual score for participant i





Which line seems the best fit ?

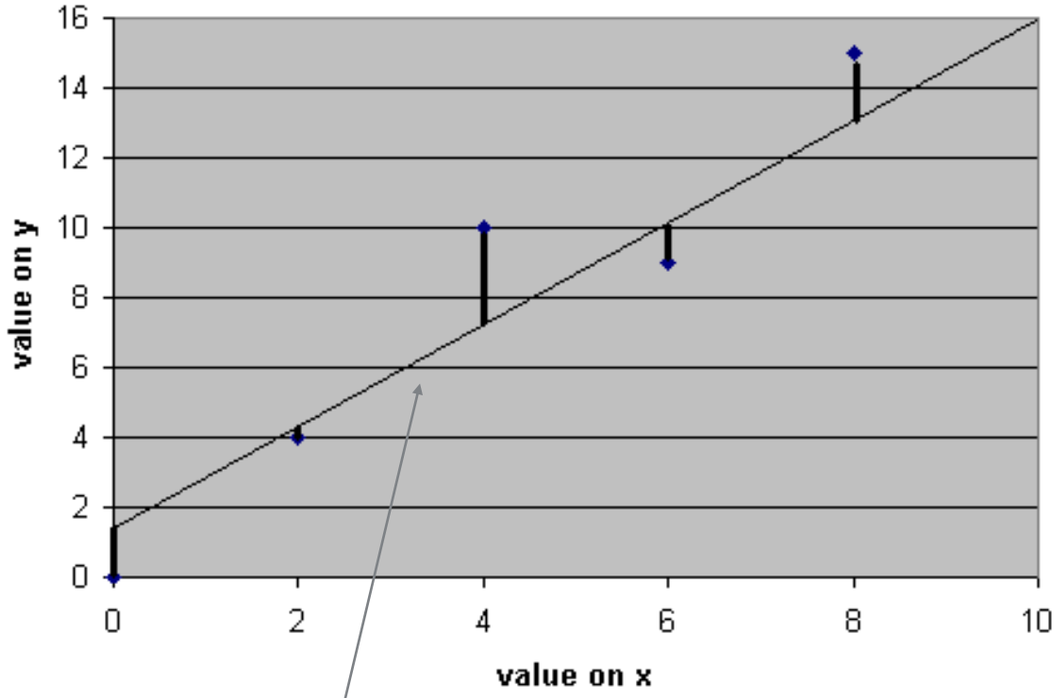


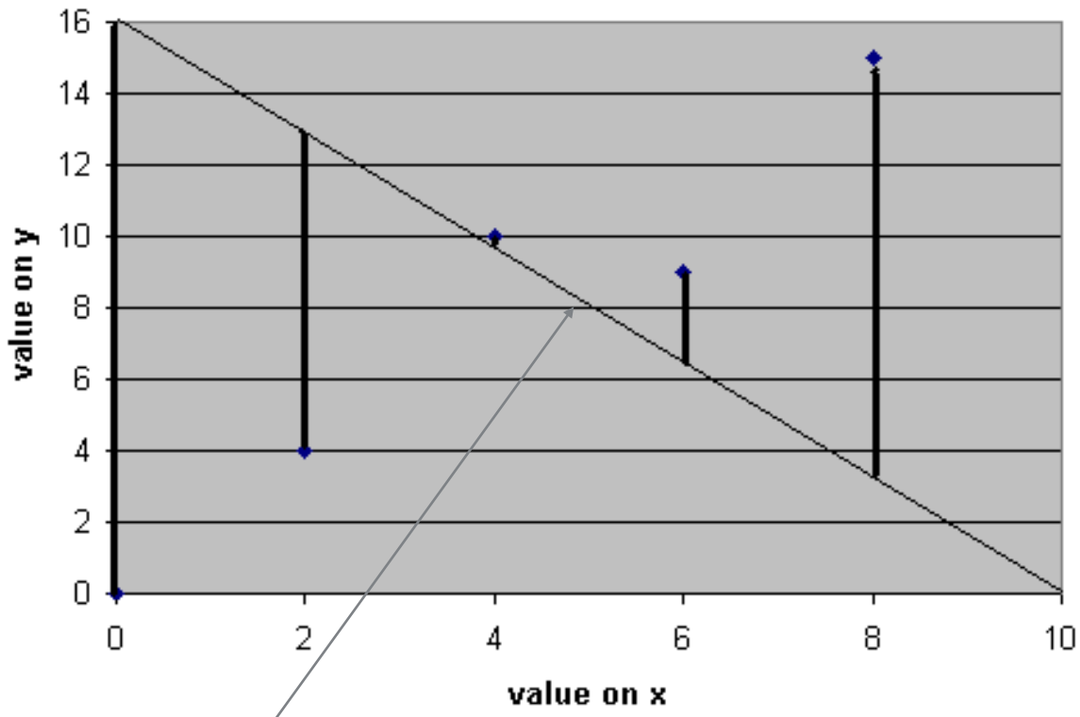
Determining the best line

- For any line, we can calculate what's known as the Least Squares.

- The Least Squares 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….

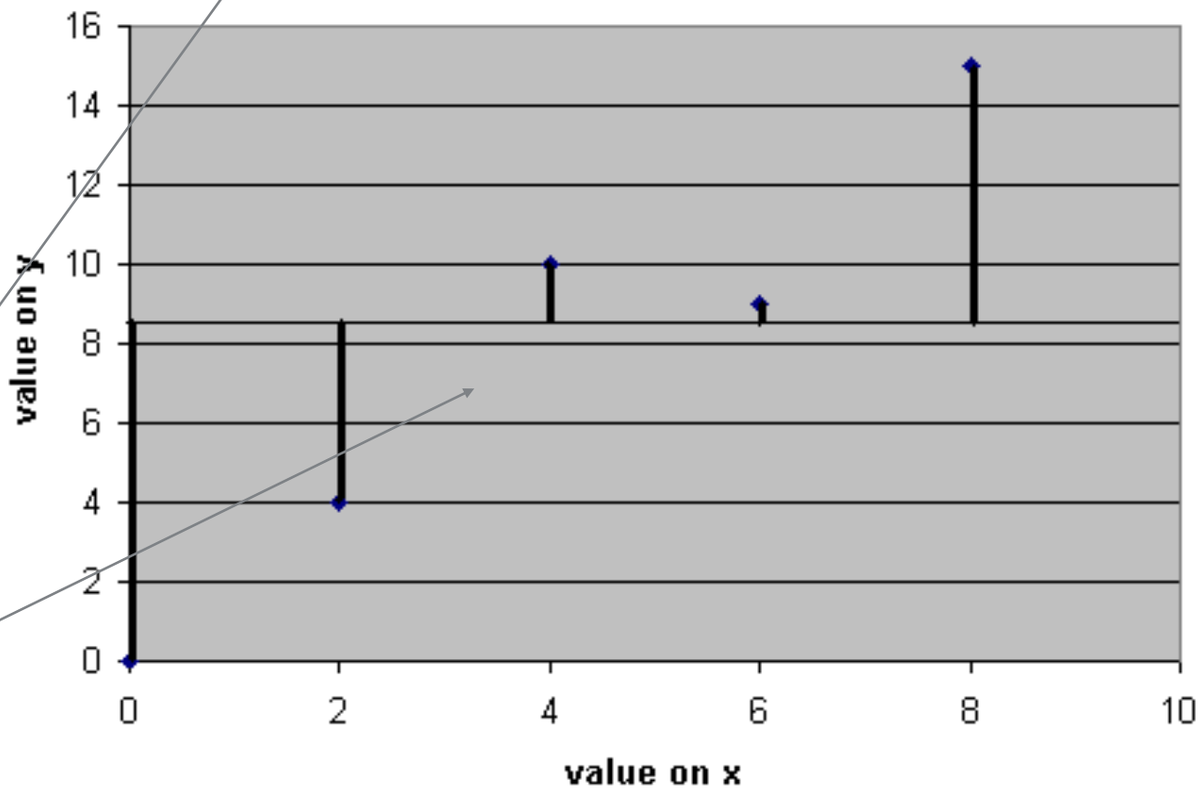
- So, for the three possible lines we just looked at….





We can see that this line seems to be the best fit as it leads to the least error between the predicted data (the line) and our observed data (the points).

These two lines aren't much good as they lead to a lot of error between predicted and observed data.



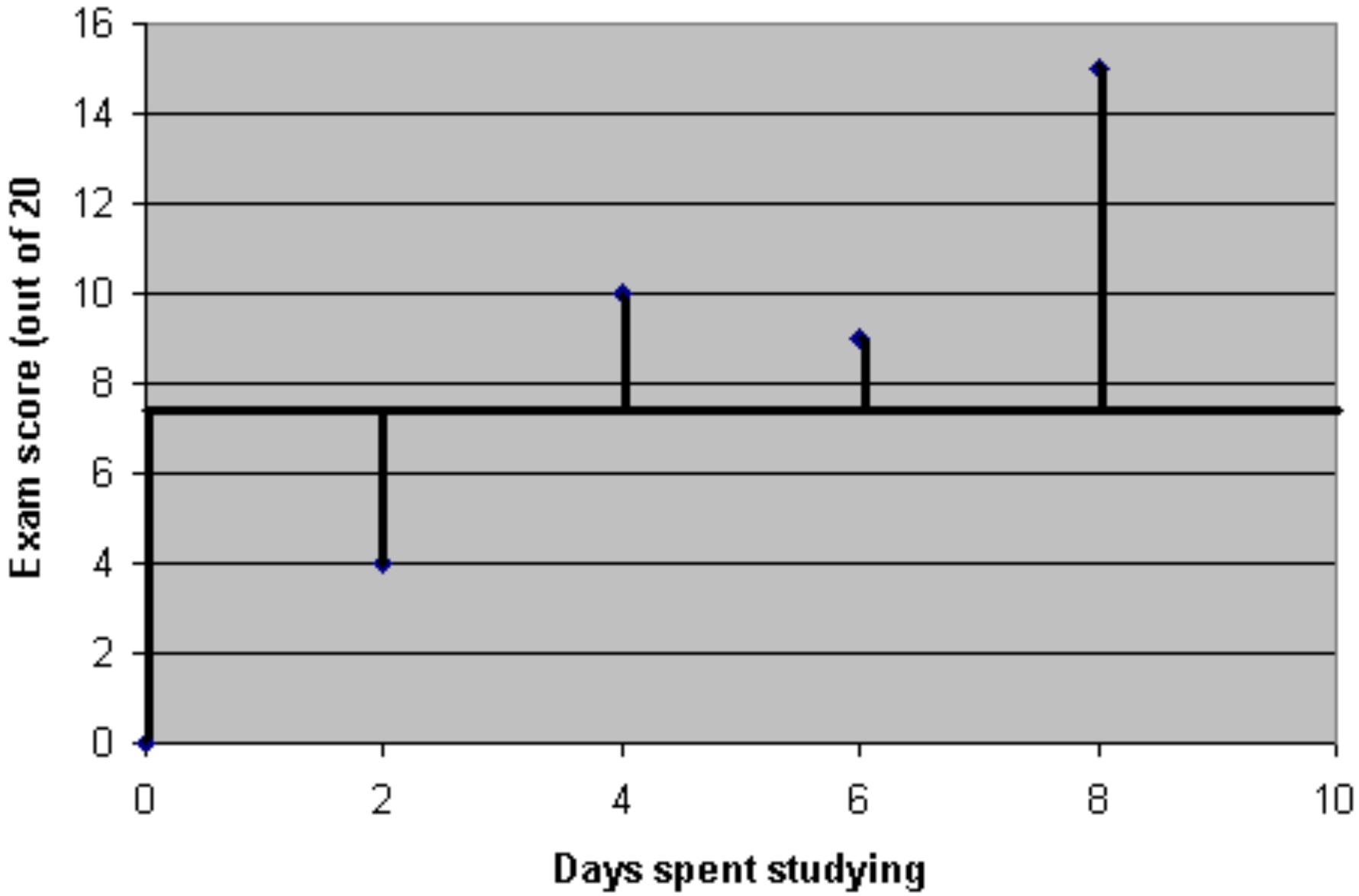
How do we determine how good a fit our line is ?

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

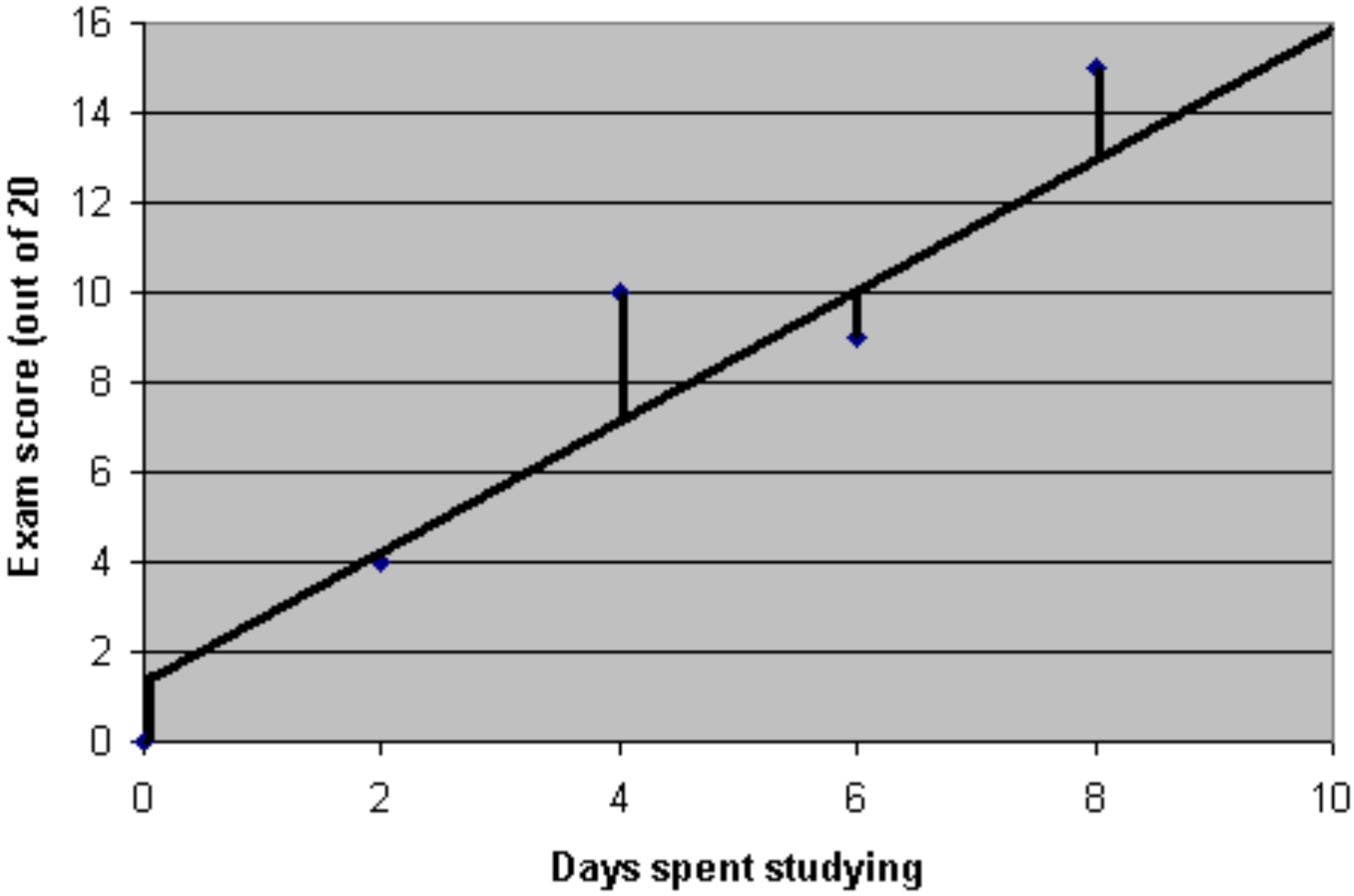
- We could work out by how much each observed value differs from the regression line.

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

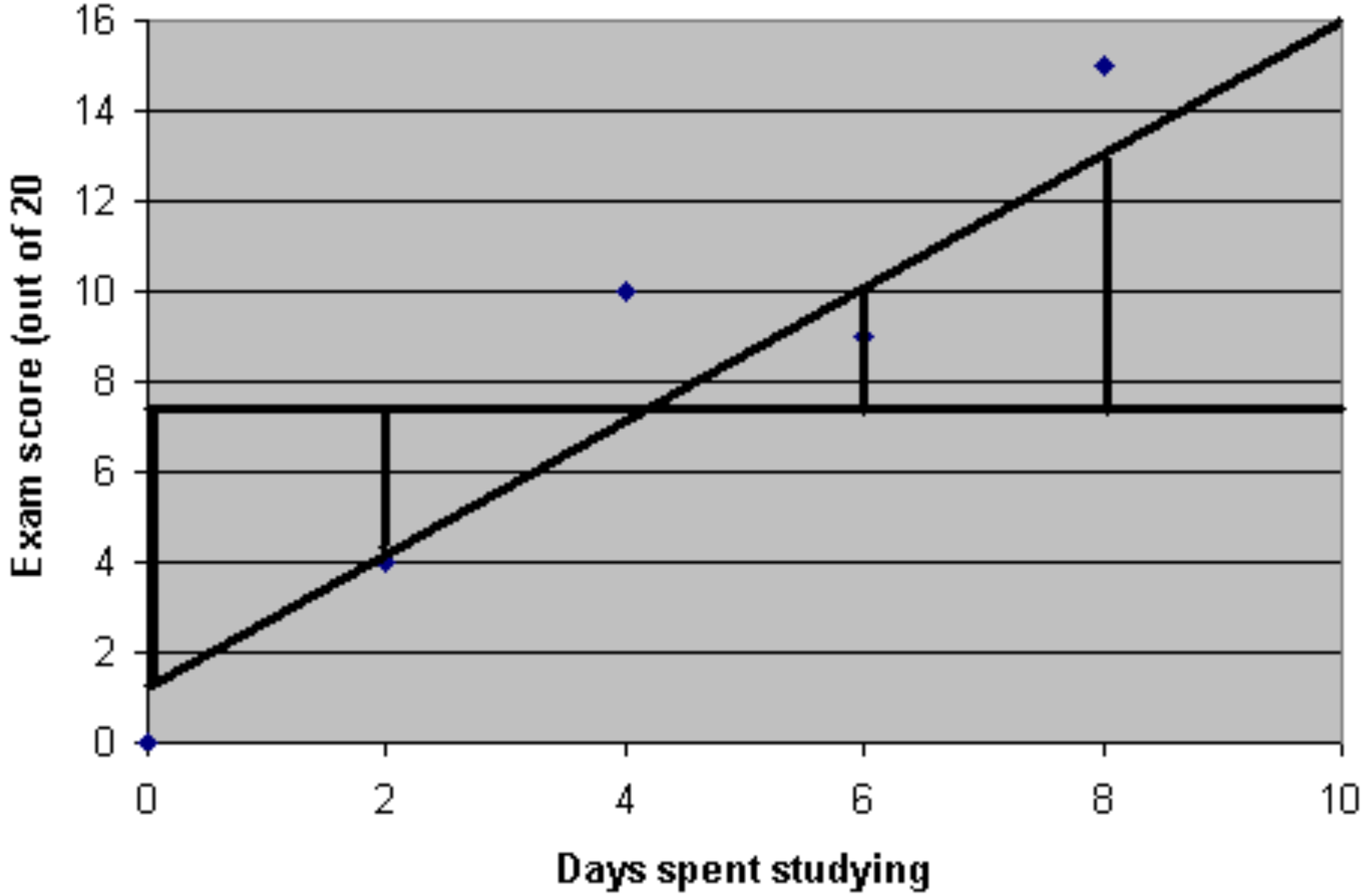
How much each observed value differs from the mean of y (called SS_T):



How much each observed value differs from the regression line (called SS_R):



How much the mean value of Y differs from the regression line for different values of x (called SS_M):



- 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.

- We can calculate the proportion of improvement in prediction by looking at the ratio of SS_M to SS_T.

- Actually, this is called R^2 so:

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

And this is the <u>same</u> R^2 that we worked out by squaring the Pearson correlation coefficient…..

- 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).

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

- Rather than use the sums of squares themselves, we use the mean sums of squares (MS_M and MS_R).

- 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.

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

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

- 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 5-2 = 3)

df numerator = 2, df denominator = 3 for our example.

df for numerator

df for denominator

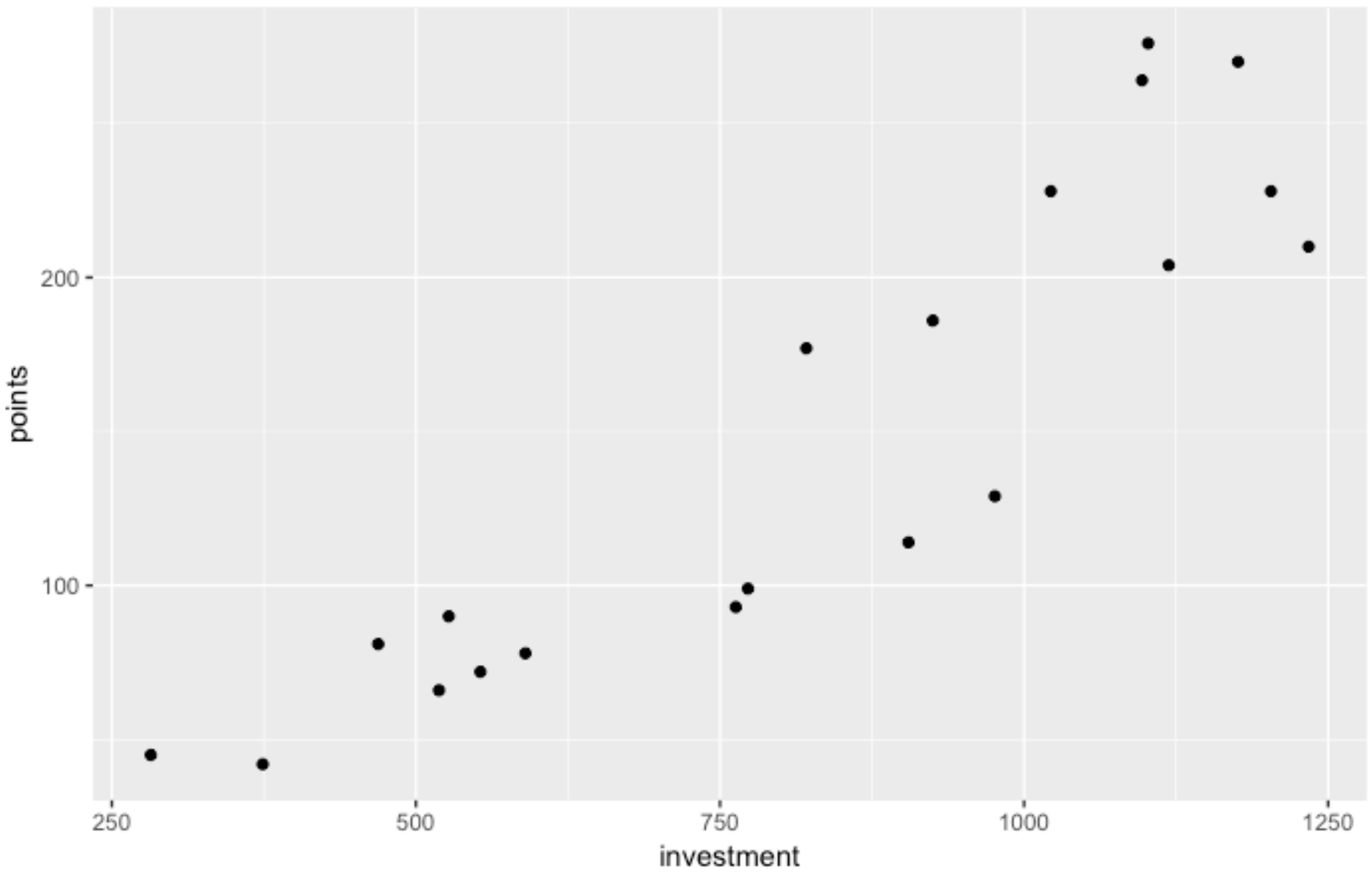
df2/df1	1	2	3	4	5	
1	161.4476	199.5000	215.7073	224.5832	230.1619	
2	18.5128	19.0000	19.1643	19.2468	19.2964	
3	10.1280	9.5521	9.2766	9.1172	9.0135	
4	7.7086	6.9443	6.5914	6.3882	6.2561	
5	6.6079	5.7861	5.4095	5.1922	5.0503	

So we would need an F value of greater than 9.5521 for our result to be significant at $p < 0.05$

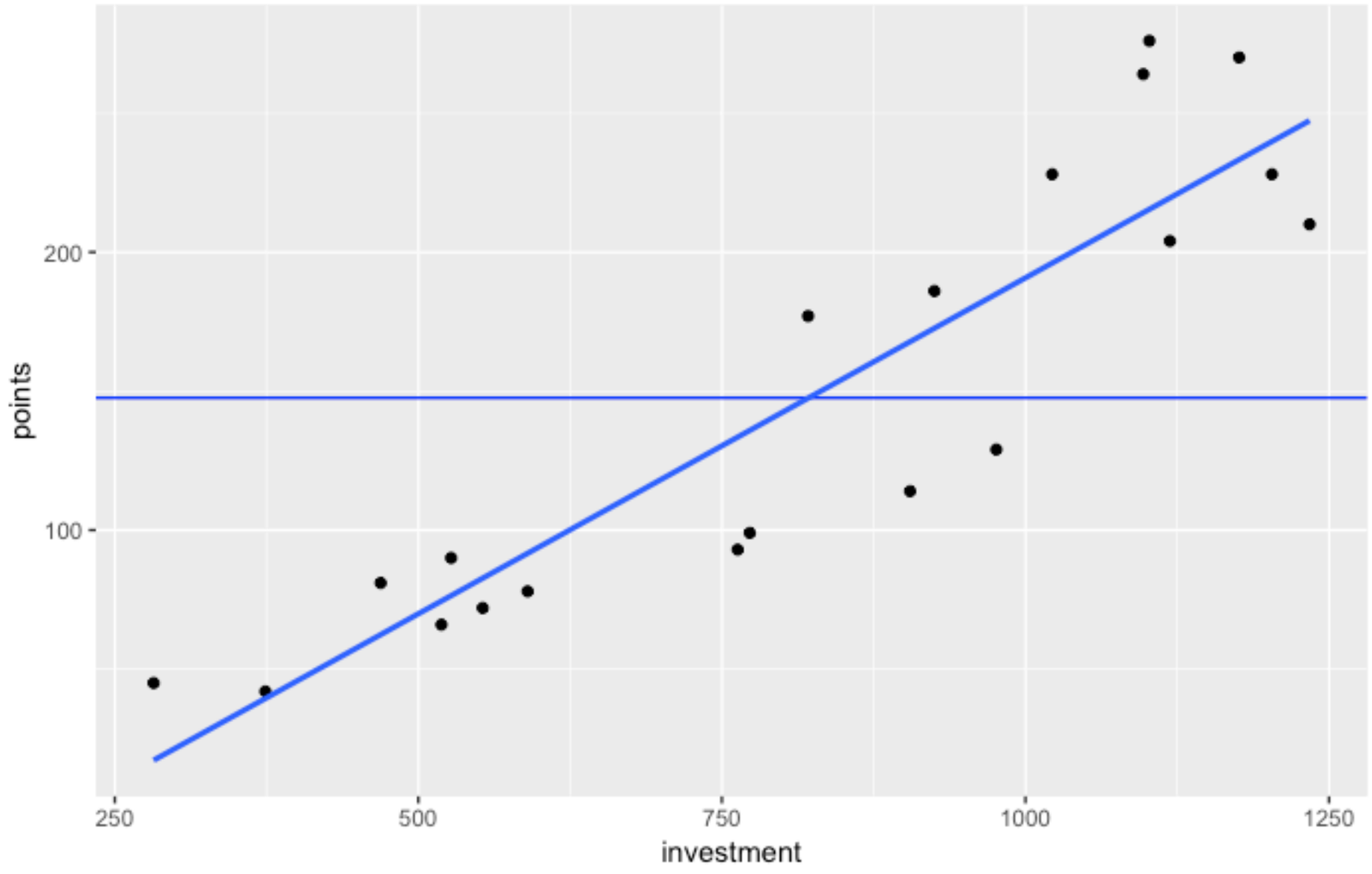
An example

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" 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).

```
> library(ggplot2) # For building ggplots
> library(Hmisc) # Needed for correlation

> #let's do a plot first
> ggplot(dataset1, aes (x=investment, y=points)) + geom_point()
```



```
> # Let's add a regression line and a line of our outcome mean
> ggplot(dataset1, aes(x = investment, y = points)) + geom_point() +
geom_hline(yintercept = mean(dataset1$points), colour = "blue") +
geom_smooth(method = "lm", se = FALSE)

> # Let's calculate Pearson's r
> rcorr(dataset1$investment, dataset1$points)
```



Pearson's r = 0.9, p < .001

Building a simple linear model

```
> # Let's do regression with just the one predictor

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

We have built two models - *model0* is a model with just the intercept (so the mean of our outcome) predicting the outcome (*points*) while *model1* is a model with *investment* predicting the outcome (*points*).

```
> # You can compare the two models to each other

> anova(model0, model1)
```

```
> 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(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).

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 £24,1666 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 future on the basis of knowing about that variable (and it’s relationship to another variable) now…

To the computer cluster…