UNIVERSITY OF LINCOLN

Correlation and regression

PSY9219M - Research Methods and Skills

Dr Matt Craddock

13/11/2018

Null Hypothesis Significance Testing (NHST)

Think back to our previous questions:

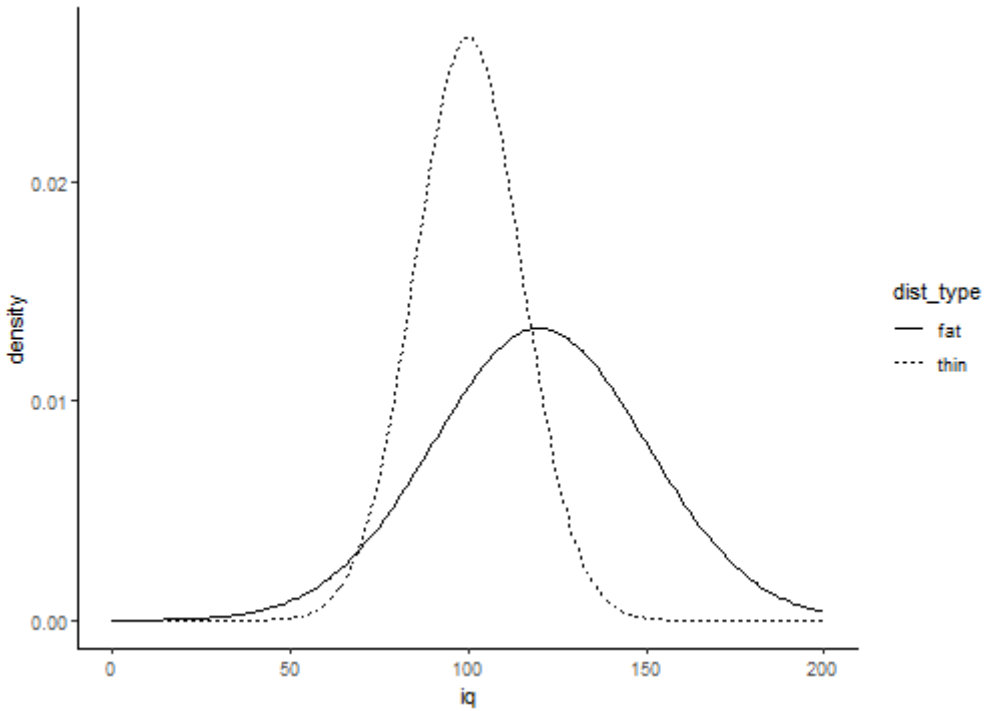
1. Do men and women differ in terms of their fear of crime?
2. Are people who have been a victim of crime more fearful of crime?

The basis of NHST is to phrase these questions as:

If there is only one population, how likely is it that our two samples have values this different from each other?

We answer this question with *test statistics* and *p-values*.

The normal distribution



Performing *t*-tests in R

The tilde (~) symbol in R usually means "modelled by"

`FoC ~ victim_crime` means `FoC modelled by victim_crime`.

`data = crime` tells R to look in the `crime` data frame for the data.

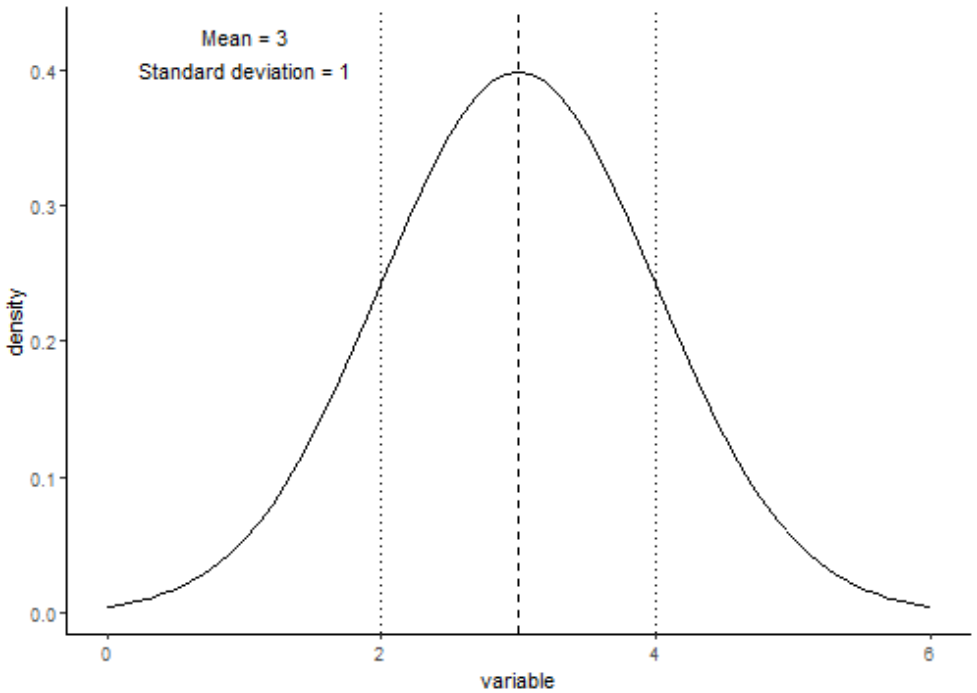
`paired = FALSE` tells R that this is an *independent samples* test.

```
t.test(FoC ~ victim_crime,
       data = crime,
       paired = FALSE)
```

```
##
##     Welch Two Sample t-test
##
## data:  FoC by victim_crime
## t = 0.45309, df = 197.48, p-value = 0.651
## alternative hypothesis: true difference in means i
## 95 percent confidence interval:
##  -0.1873001  0.2990388
## sample estimates:
##  mean in group no mean in group yes
##          2.463636           2.407767
```

The mean and the variance

The mean and the variance



The mean - μ - is a measure of *central tendency*.

The standard deviation - σ - is a measure of the *spread* of the data - i.e.

The mean and the variance

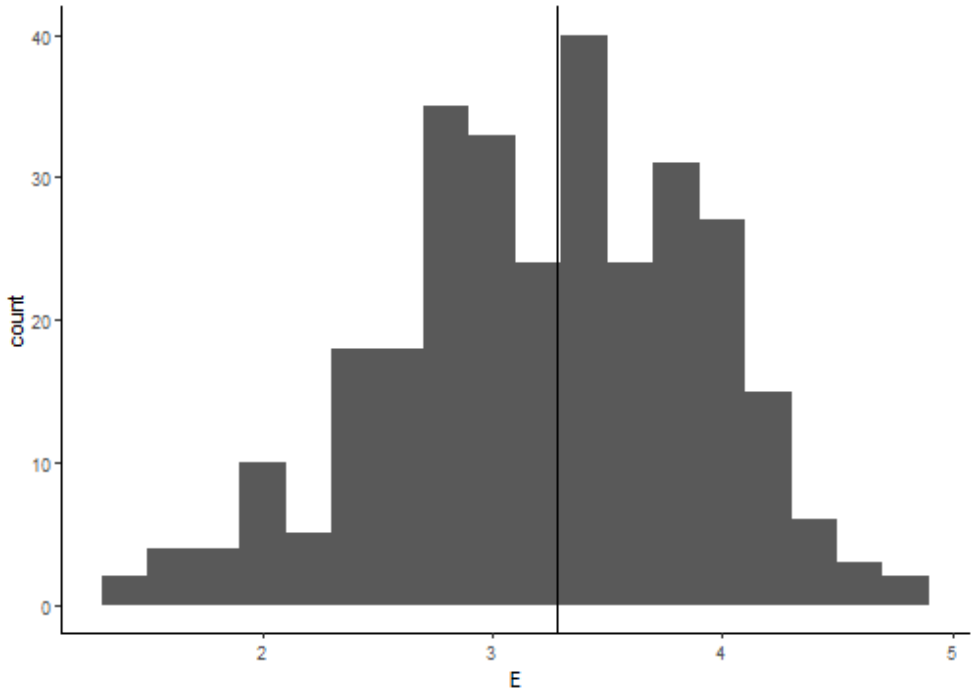
```
ggplot(crime, aes(x = E)) +
  geom_histogram(binwidth = 0.2) +
  geom_vline(aes(xintercept = mean(E))) +
  theme_classic()
```

```
mean(crime$E)
```

```
## [1] 3.279402
```

```
sd(crime$E)
```

```
## [1] 0.6735782
```



The mean and the variance

As you can see, the values don't lie directly on the mean, but are spread around it.

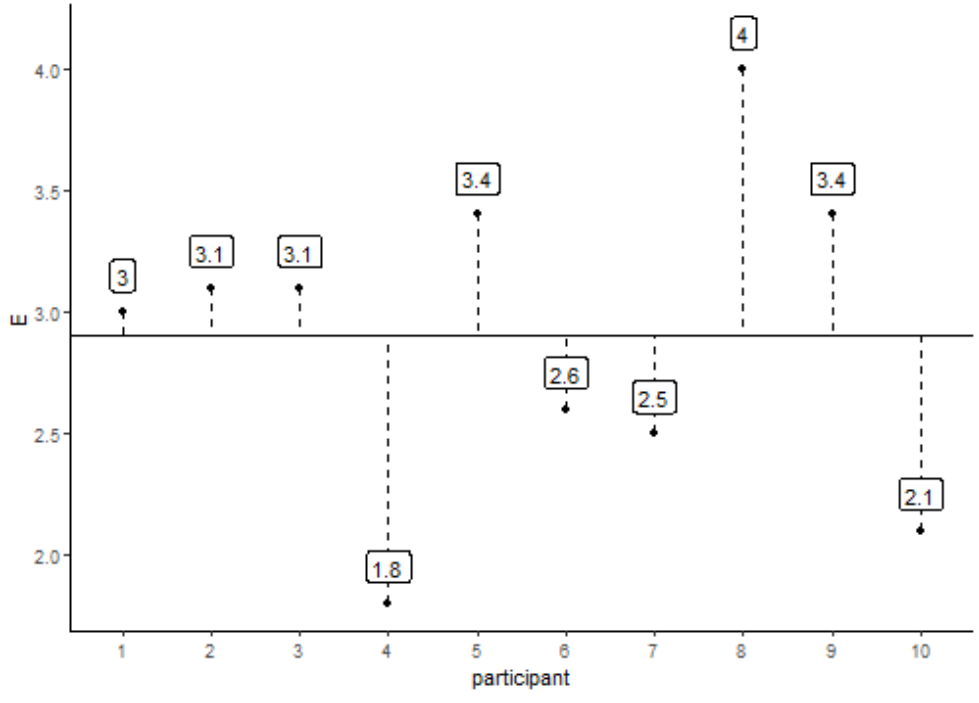
To quantify how much the values vary from the mean, we can calculate the *variance*.

Here's the scary looking formula for the variance:

$$\sigma^2 = \frac{\sum(x - \bar{x})^2}{N - 1}$$

And here's the not-so-scary R function:

```
var(x)
```



The mean and the variance

Let's break down the top part of the formula:

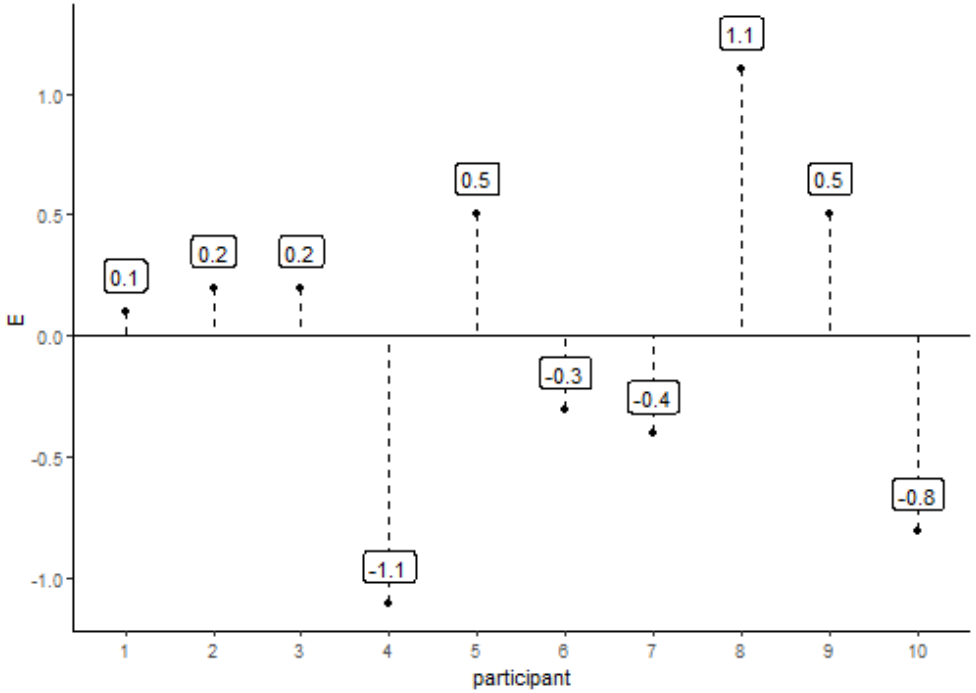
$$\sum (x - \bar{x})^2$$

This is the *sum* of the *squared differences* from the mean.

The first step is to subtract the mean of all values from the values themselves: $x - \bar{x}$

In R code this is:

```
x - mean(x)
```



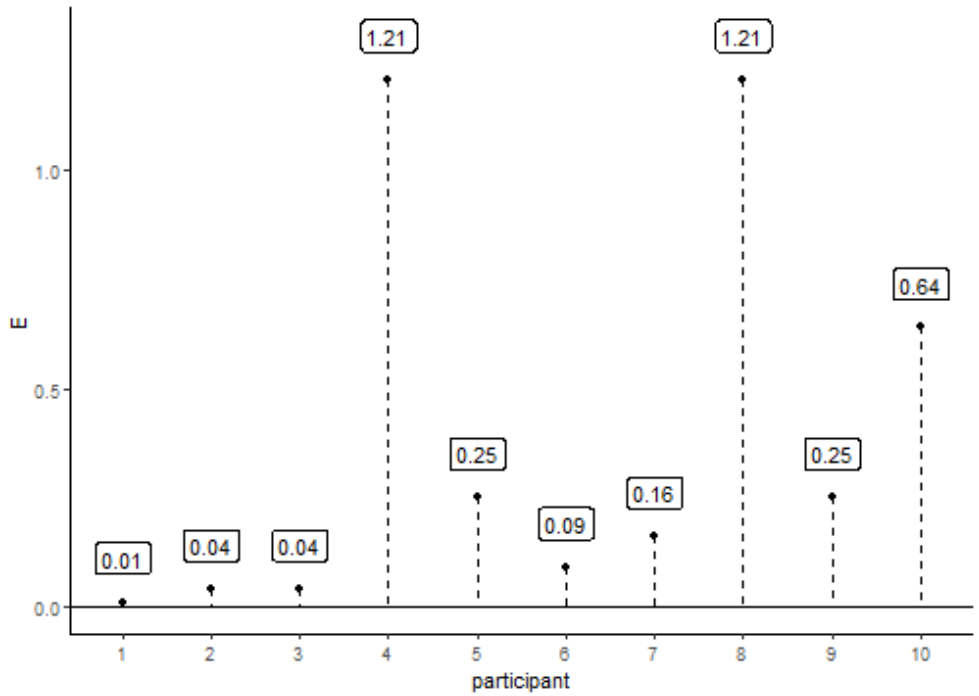
The mean and the variance

The next step is to *square* those differences to get $(x - \bar{x})^2$.

This has a couple of consequences:

1. Negative values become positive.
2. Values that are further away from the mean often get even further away.

This prevents "errors" from cancelling out, and effectively penalises values that are far away from the mean.



The mean and the variance

Next, we add those numbers together:

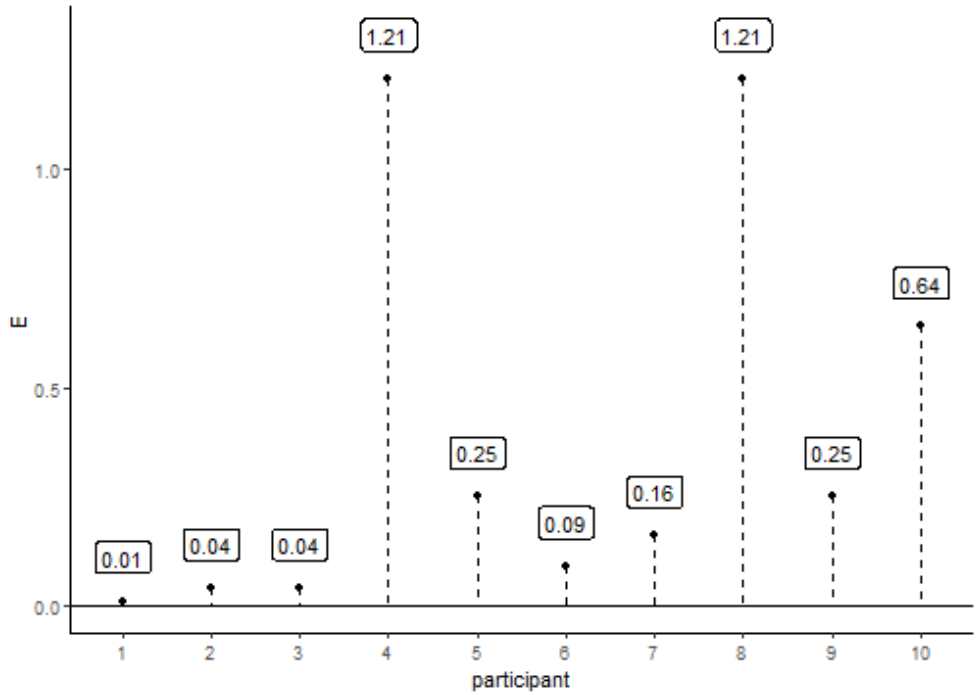
$$\sum (x - \bar{x})^2$$

In R code, this is:

```
sum((x - mean(x))^2)
```

```
sum((crime$E[1:10] - mean(crime$E[1:10]))^2)
```

```
## [1] 3.9
```



The mean and the variance

The final step is to take the average:

$$\frac{\sum (x - \bar{x})^2}{N - 1}$$

In R code, this is:

```
sum((x - mean(x))^2) / (length(x) - 1)
```

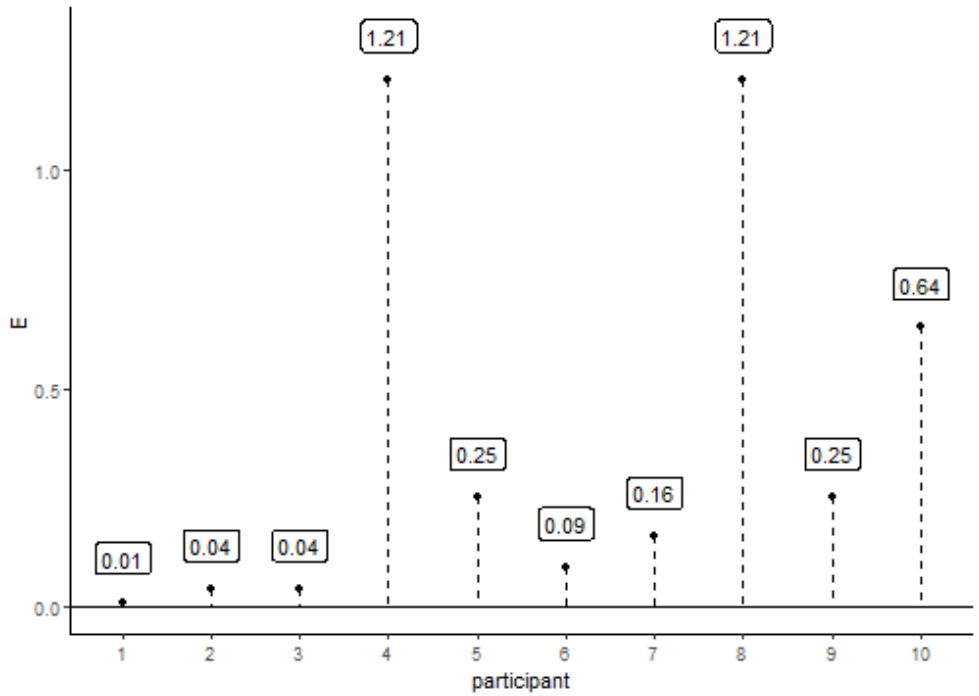
```
sum((crime$E[1:10] - mean(crime$E[1:10]))^2)
```

```
## [1] 0.4333333
```

...or, for short:

```
var(crime$E[1:10])
```

```
## [1] 0.4333333
```



Correlation

Correlation

How related are the variables in the Fear of Crime dataset?

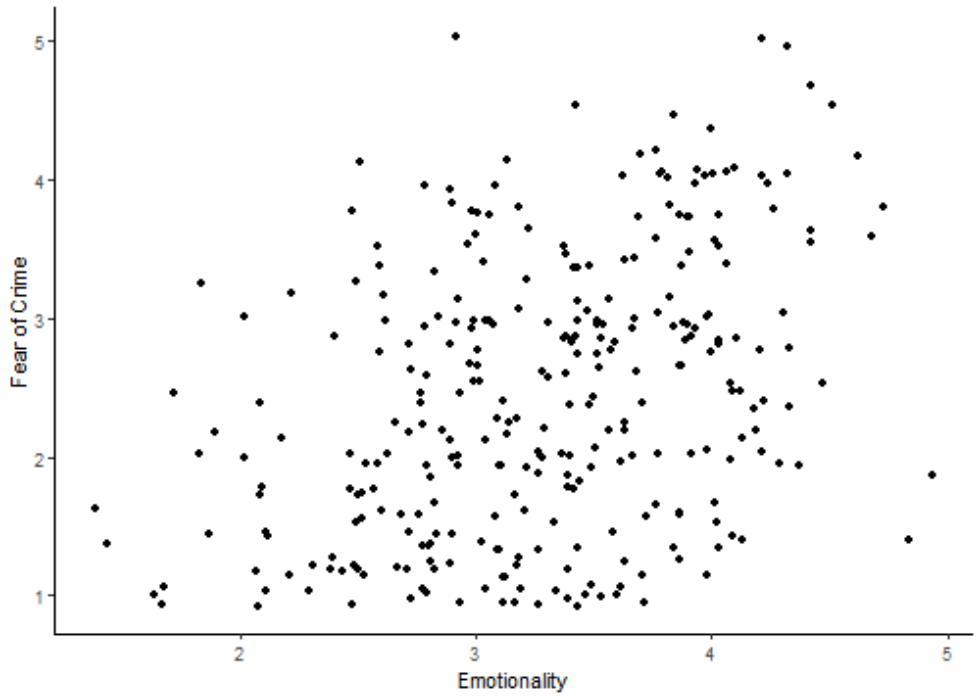
```
head(crime)
```

```
## # A tibble: 6 x 15
##   Participant sex     age victim_crime     H     E     X     A     C     O
##   <chr>       <chr> <dbl> <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 R_01TjXgC1~ male     55 yes            3.7   3     3.4   3.9   3.2   3.6
## 2 R_0dN5YeUL~ fema~    20 no             2.5   3.1   2.5   2.4   2.2   3.1
## 3 R_0DPiPYWh~ male     57 yes            2.6   3.1   3.3   3.1   4.3   2.8
## 4 R_0f7bSsH6~ male     19 no             3.5   1.8   3.3   3.4   2.1   2.7
## 5 R_0rov2RoS~ fema~    20 no             3.3   3.4   3.9   3.2   2.8   3.9
## 6 R_0wioqGER~ fema~    20 no             2.6   2.6   3     2.6   2.9   3.4
## # ... with 5 more variables: SA <dbl>, TA <dbl>, OHQ <dbl>, FoC <dbl>,
## #   Foc2 <dbl>
```

Correlation

What happens to Fear of Crime as Emotionality increases?

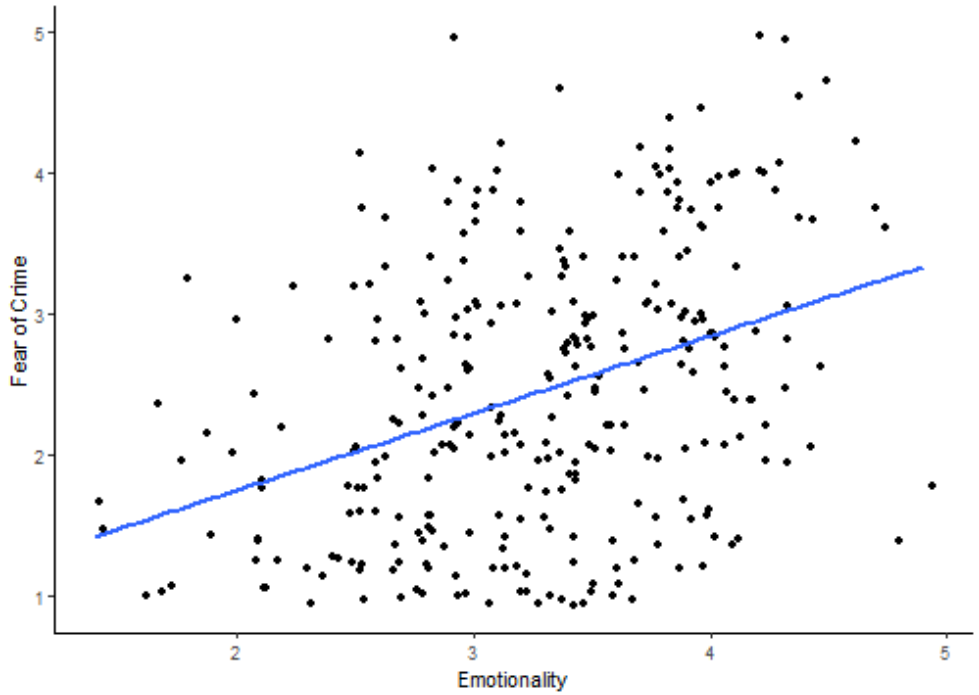
```
ggplot(crime,
       aes(x = E,
           y = FoC)) +
  geom_jitter() +
  theme_classic() +
  labs(x = "Emotionality",
       y = "Fear of Crime")
```



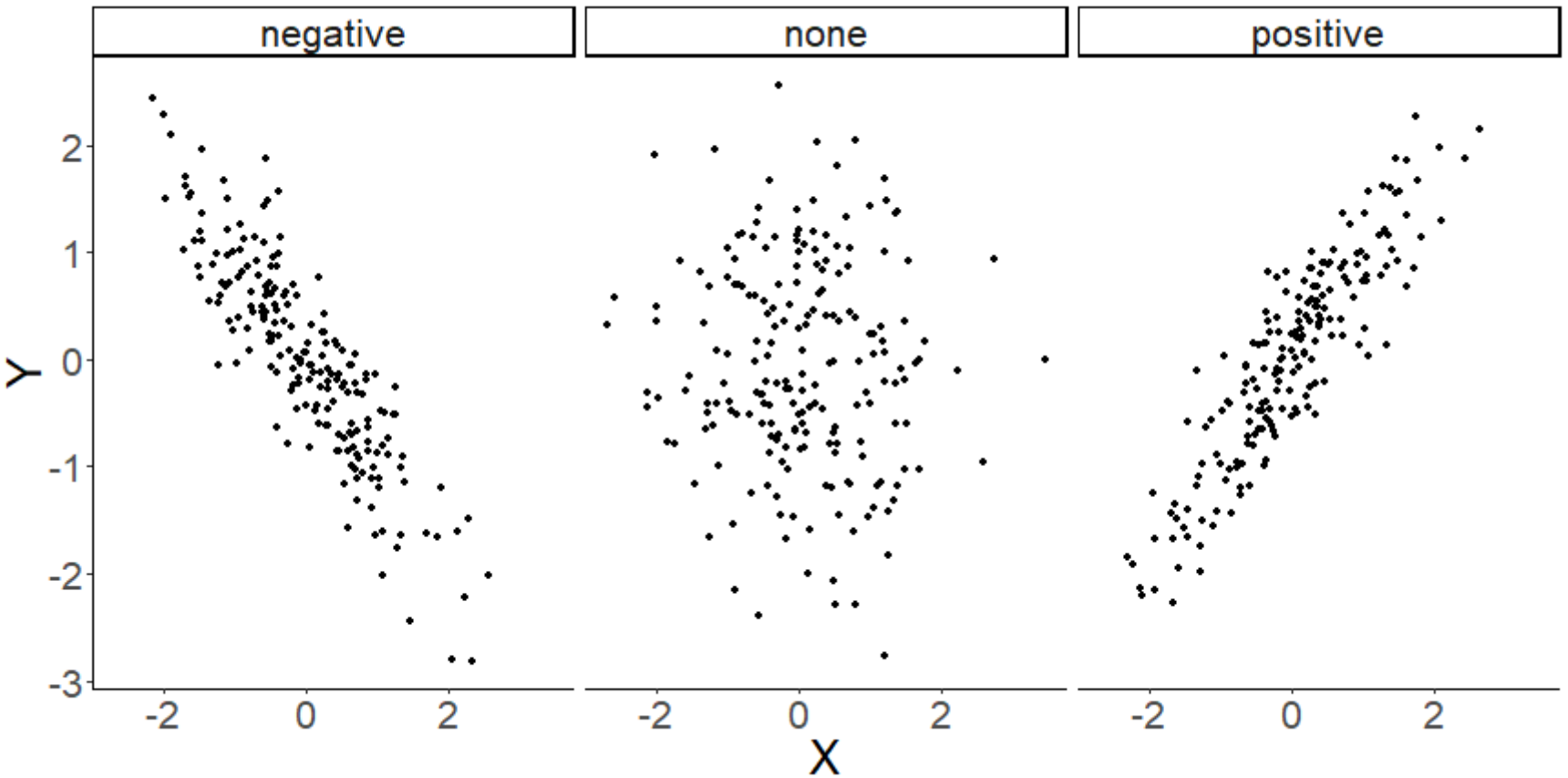
Correlation

There is a positive relationship between Emotionality and Fear of Crime.

```
ggplot(crime,
       aes(x = E,
           y = FoC)) +
  geom_jitter() +
  theme_classic() +
  labs(x = "Emotionality",
       y = "Fear of Crime") +
  stat_smooth(method = "lm", se = FALSE)
```



Different relationships



Correlation and covariance

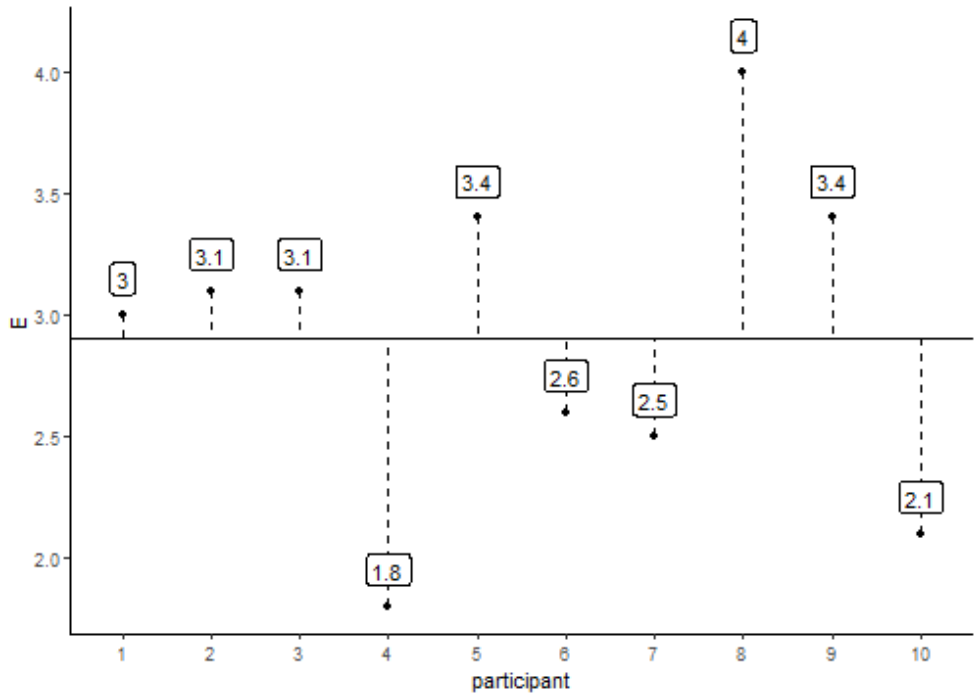
Correlation measures the strength and direction of an association between two continuous variables.

But how do we quantify it? We need to look at how the variables *vary* together.

Let's look at some of the statistics of the Emotionality variable.

The plot shows the Emotionality values of the first ten participants.

The line across the middle is the mean of those values - 2.9.

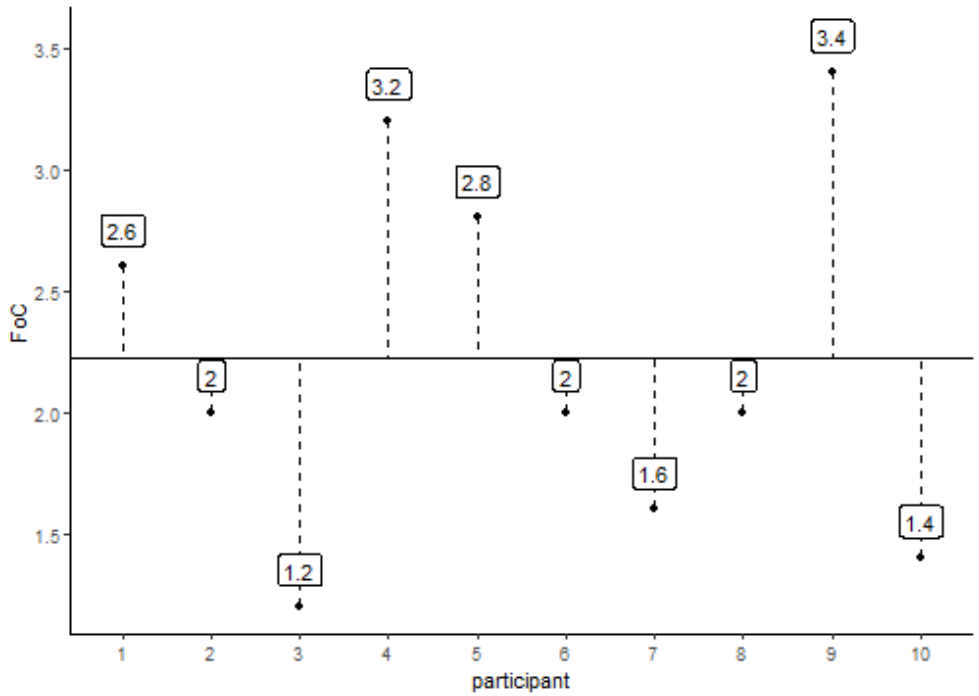


Correlation and covariance

Now let's look at the same plot for Fear of Crime (FoC).

Again, the these points and labels are individual ratings of Fear of Crime.

The line across the middle shows the mean, which is 2.22.



Correlation and covariance

Now let's look at these previous two plots as differences from their respective means.

What we want to now is to what extent the values *vary together*. I.e. as one goes up, does the other?

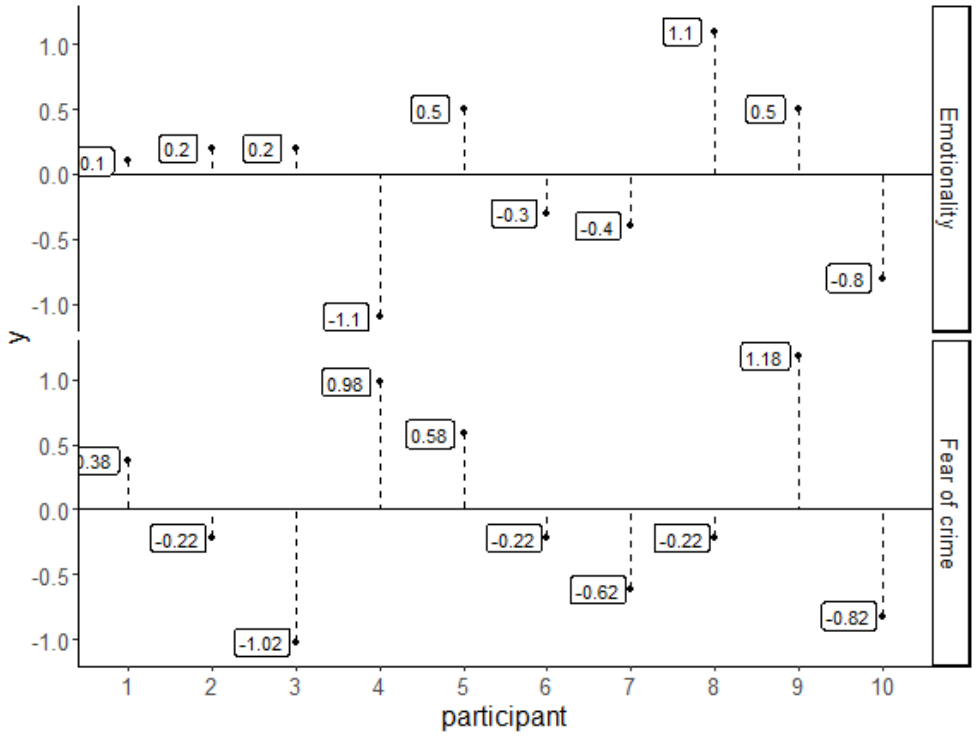
This is *covariance*.

Here's the scary formula:

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

Here's the not-so-scary R function:

```
cov(x, y)
```



Correlation and covariance

Covariance gives us a measure of how much two variables vary together.

But the numbers it gives us can be hard to interpret when the variables are on very different scales.

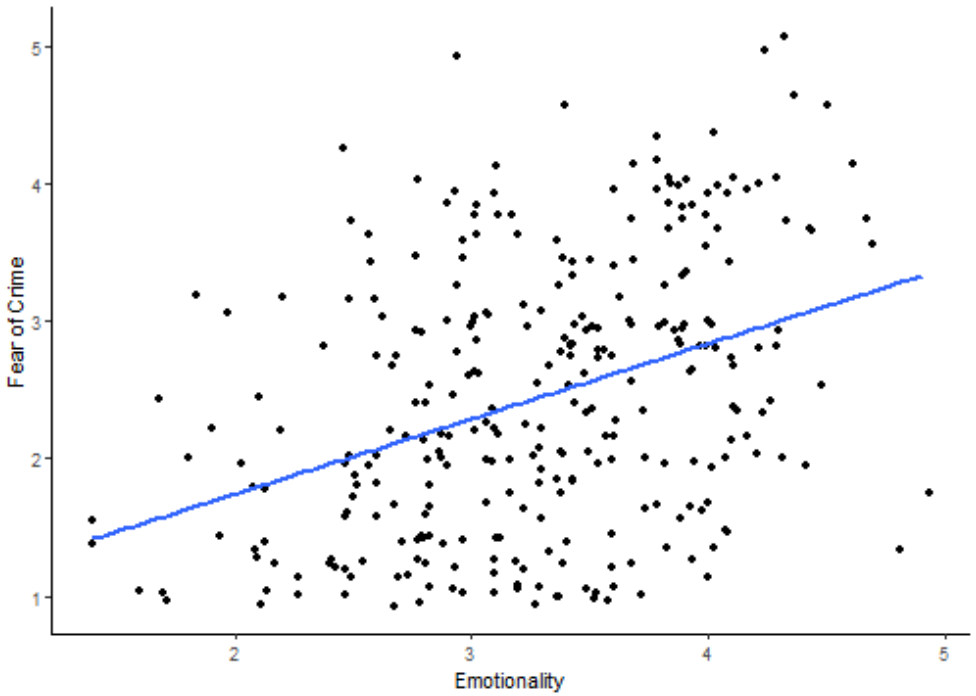
So we rescale the covariance using the standard deviations of each variable.

$$corr(x, y) = r = \frac{cov(x, y)}{\sigma^x \sigma^y}$$

This gives us the *correlation coefficient*, or r.

```
cor(crime$E, crime$FoC)
```

```
## [1] 0.369891
```



Pearson's product-moment correlation

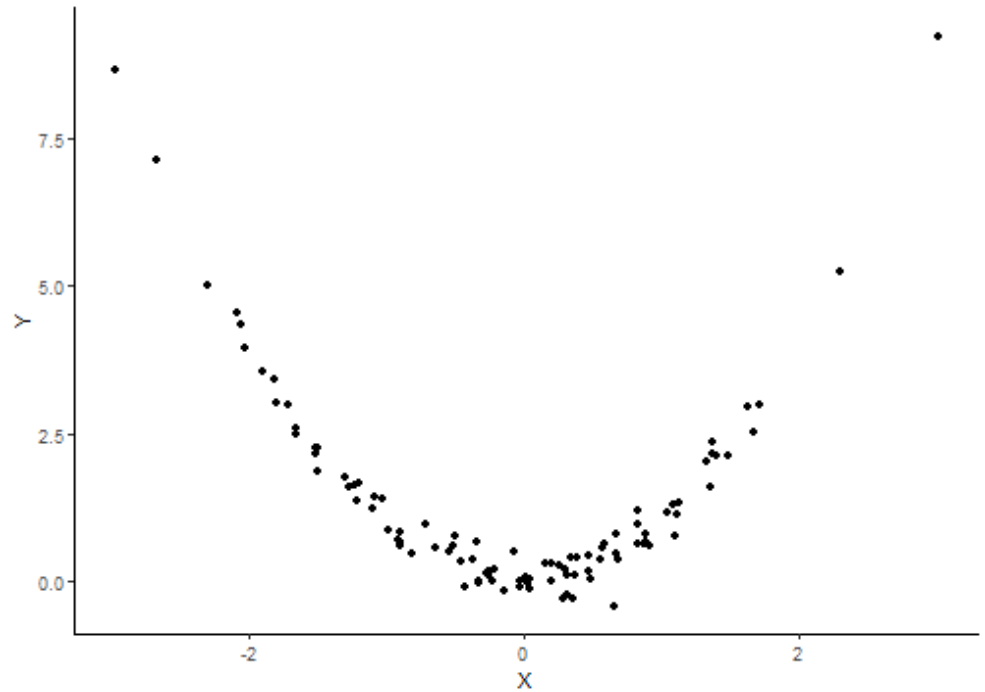
The **cor.test()** function can be used to test the *significance* of a correlation.

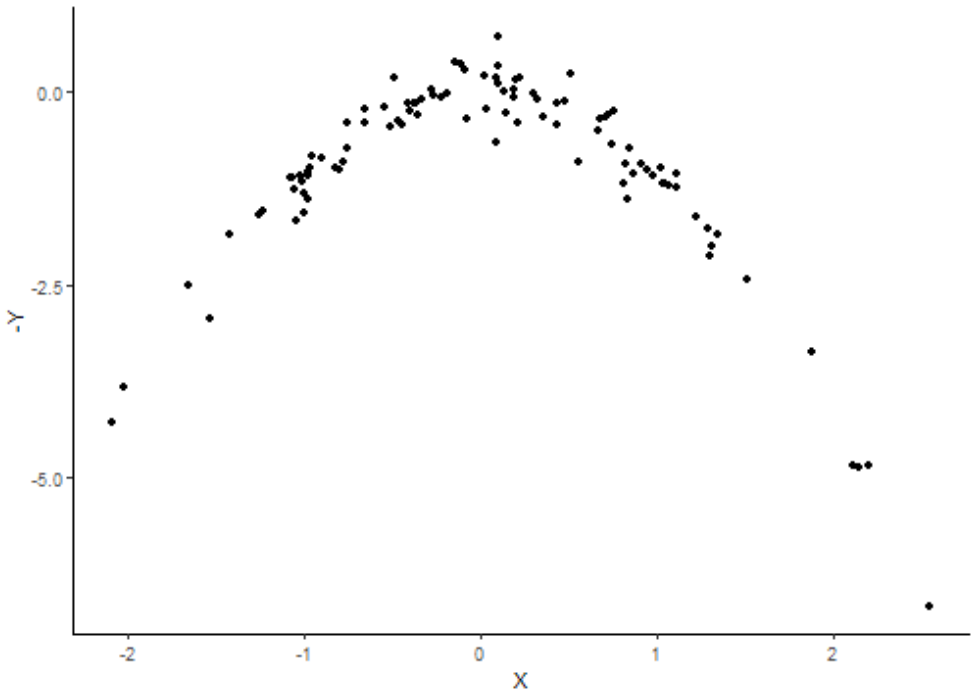
```
cor.test(crime$E, crime$FoC,
         method = "pearson")
```

```
##
##     Pearson's product-moment correlation
##
## data:  crime$E and crime$FoC
## t = 6.8843, df = 299, p-value = 3.421e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2680476 0.4635586
## sample estimates:
##      cor
## 0.369891
```

Curved or non-linear relationships

If your data look like this:

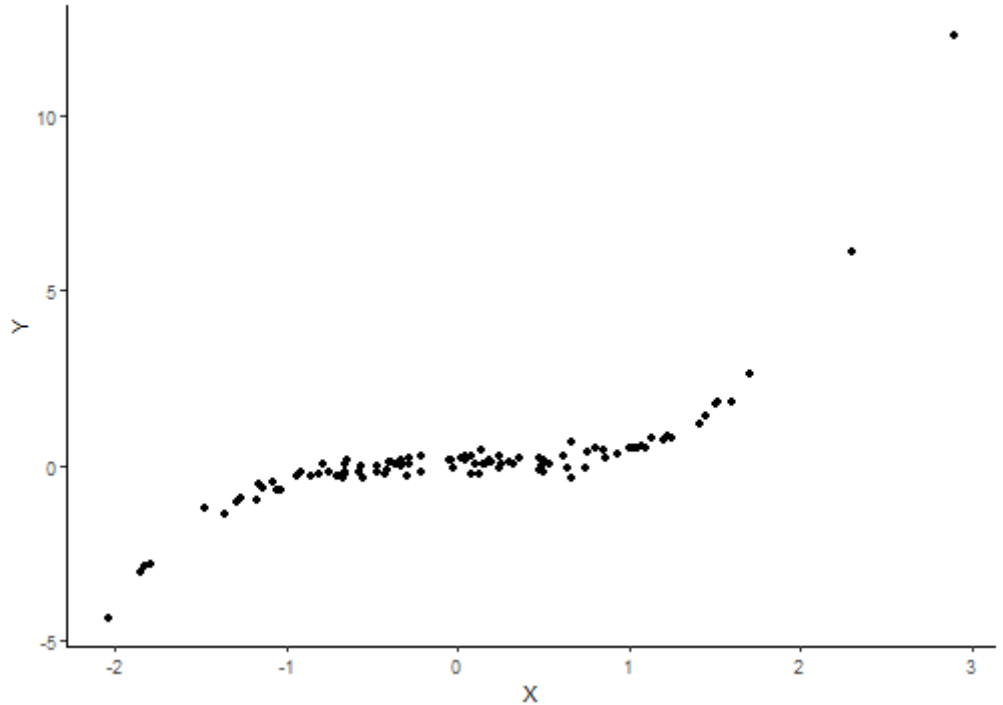




...forget about correlation.

Curved or non-linear relationships

...but if your data look like this:



...there is hope!

Spearman's rank correlation

Spearman's correlation is used to measure *monotonicity*, and is the non-parametric equivalent to Pearson's correlation.

If the data is not already ranks then it is converted to ranks; then the ranks are correlated across variables.

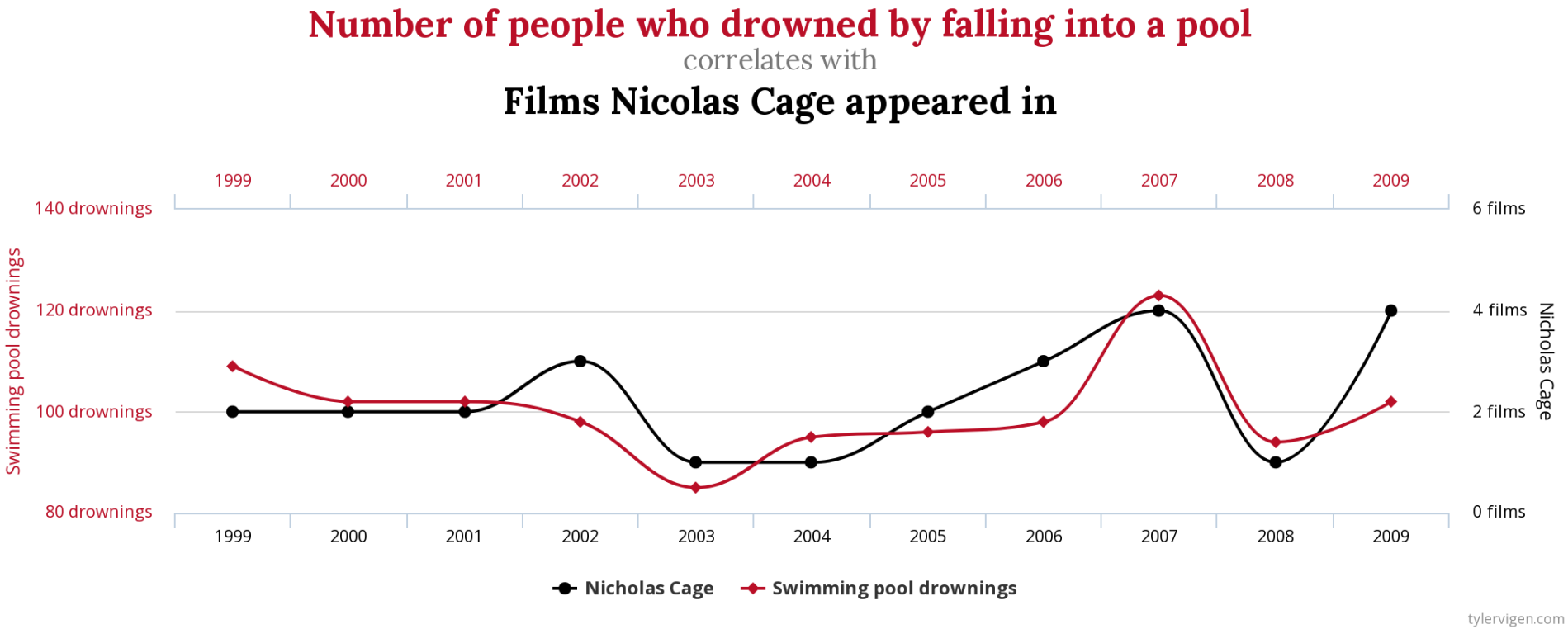
```
cor.test(crime$age,
         crime$E,
         method = "spearman")
```

```
## Warning in cor.test.default(crime$age, crime$E, method = "spearman"):
## Cannot compute exact p-value with ties

##
##     Spearman's rank correlation rho
##
## data:  crime$age and crime$E
## S = 5817900, p-value = 2.402e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.2928709
```

Correlation is not causation

https://www.spuriouscorrelations.com

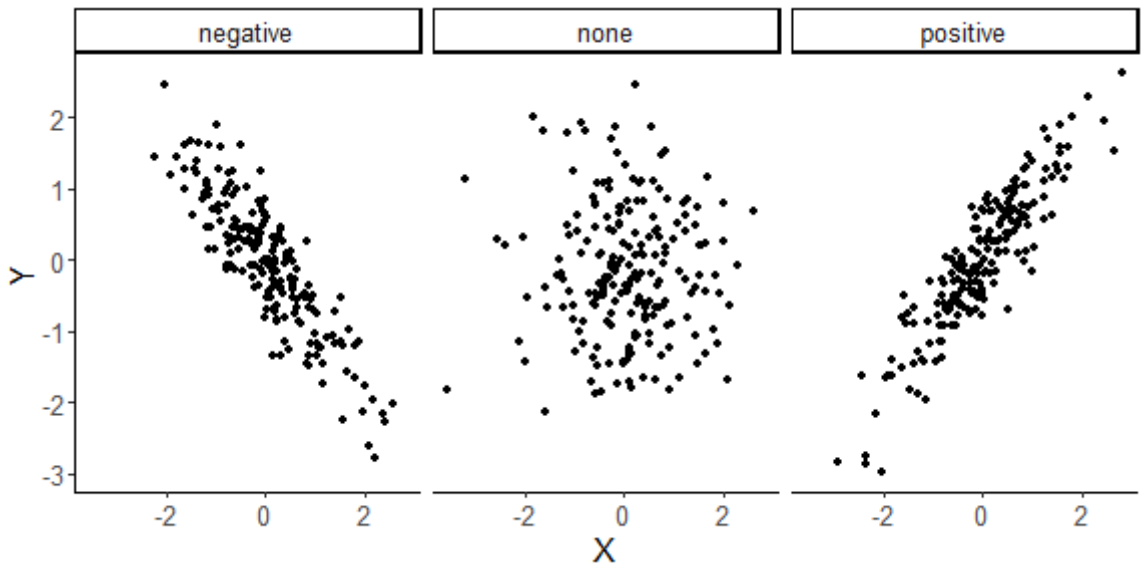


Correlation summary

Correlation coefficients are bound in a range from -1 to 1.

Negative coefficients mean that as one variable increases, the other decreases.

Positive coefficients mean that as one variable increases, the other also increases.



Reporting a correlation

Reporting a correlation is pretty straightforward. Only the correlation coefficient and p-value are typically required. e.g.

"There was a significant positive correlation between emotionality and fear of crime, r = .37, p < .001."

However, it's best to also specify which type of correlation you used (e.g. Pearson's or Spearman's); and a scatterplot showing the relationship should almost always be shown. Typically, the degrees of freedom or number of observations should also be given, e.g. $r(299)$ = .37, p < .001, or r = .37, p < .001, N = 301.

Note that r is considered a measure of *effect size*. An r of .1 is considered a small effect, while an r of .8 is considered a large effect.

Linear regression

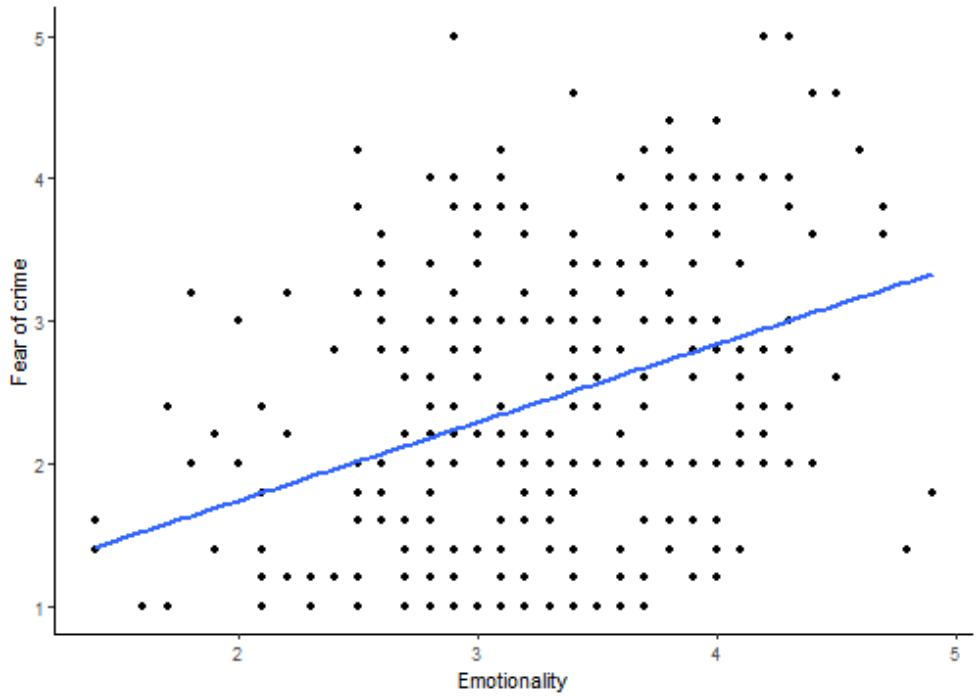
Correlation, regression and prediction

Correlation quantifies the *strength* and *direction* of an association between two continuous variables.

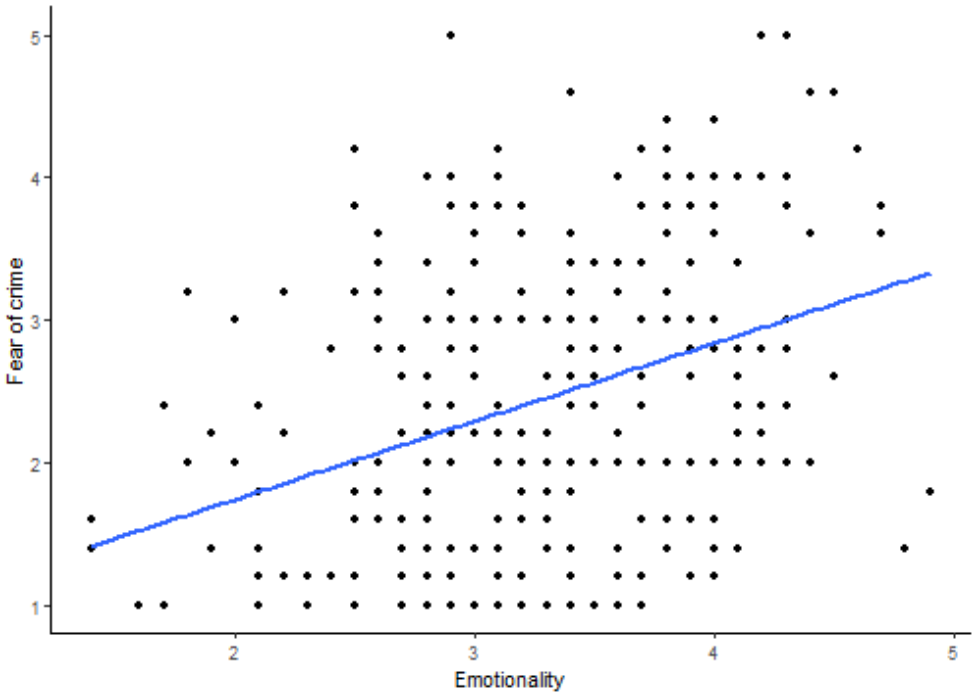
But what if we want to *predict* the values of one variable from those of another?

For example, as Emotionality increases, *how much* does Fear of Crime change?

```
ggplot(crime,
       aes(x = E, y = FoC)) +
  geom_point() +
  stat_smooth(method = "lm", se = FALSE) +
  theme_classic() +
  labs(x = "Emotionality",
       y = "Fear of crime")
```



Linear regression



The line added to this scatterplot is the *line of best fit*.

A line like this can be described by two parameters - the *intercept* and the *slope*.

The *intercept* is where the line crosses the *y-axis*.

The *slope* is the *steepness* of the line.

Given these parameters, we can predict the value of **y** - the dependent variable - at each value of **x** - the independent, predictor variable - using the following formula:

$$y = a + bX$$

Line of best fit demo

The line of best fit can be found by adjusting the *intercept* and *slope* to minimise the *sum of squared residuals*.

Line of best fit demo

The mean as a model

First, let's create a linear model that simply finds the *mean* using the **lm()** function.

```
intercept_only <- lm(FoC ~ 1, data = crime)
intercept_only
```

```
##
## Call:
## lm(formula = FoC ~ 1, data = crime)
##
## Coefficients:
## (Intercept)
##       2.445
```

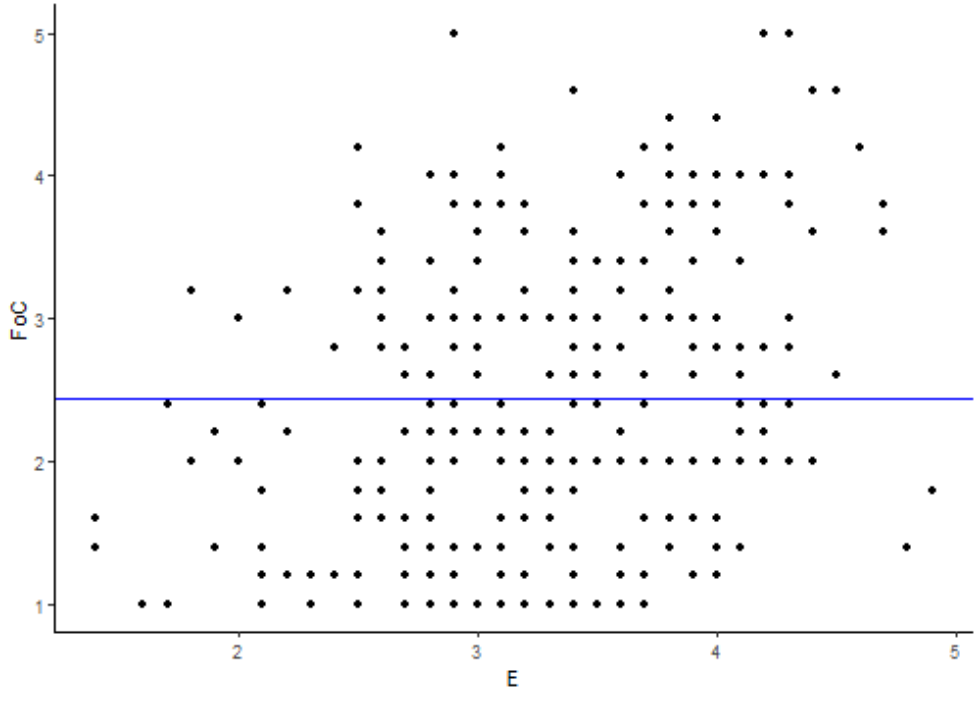
Here the Intercept is equal to the *mean* of FoC.

```
mean(crime$FoC)
```

```
## [1] 2.444518
```

In the formula $y = a + bX$, a is the *Intercept*.

So our prediction for the value of y is $y = 2.44$.



Fear of crime predicted by emotionality

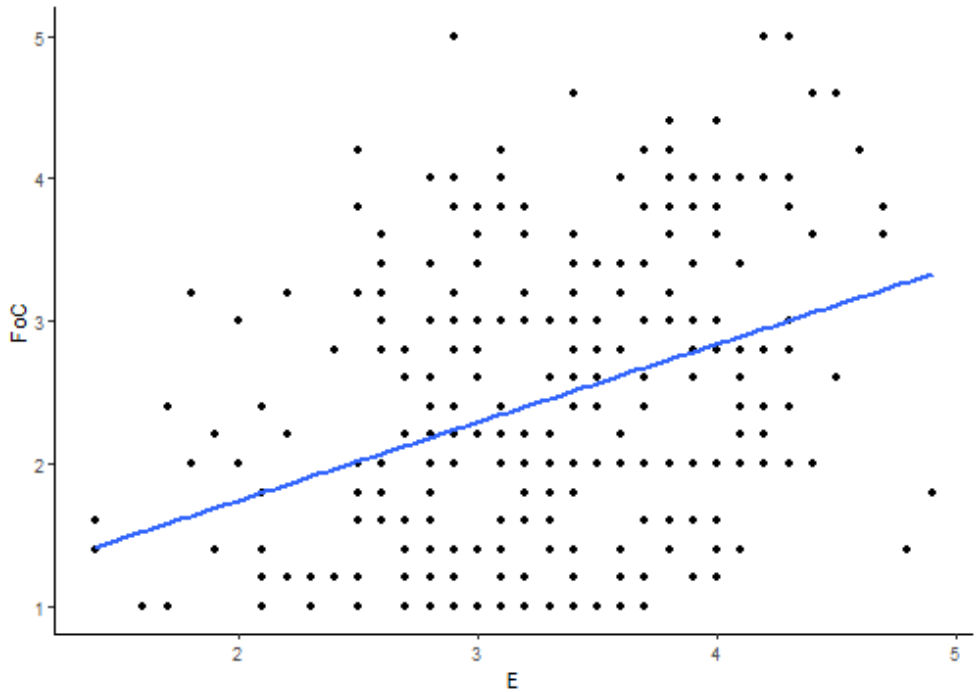
Now let's add Emotionality as a *predictor*.

```
foc_by_E <- lm(FoC ~ E, data = crime)
foc_by_E
```

```
##
## Call:
## lm(formula = FoC ~ E, data = crime)
##
## Coefficients:
## (Intercept)            E
##      0.6492       0.5475
```

We now have two coefficients - one for the intercept, and one for Emotionality.

These are the *intercept* and *slope* of the regression line on the right.



Fear of crime predicted by emotionality

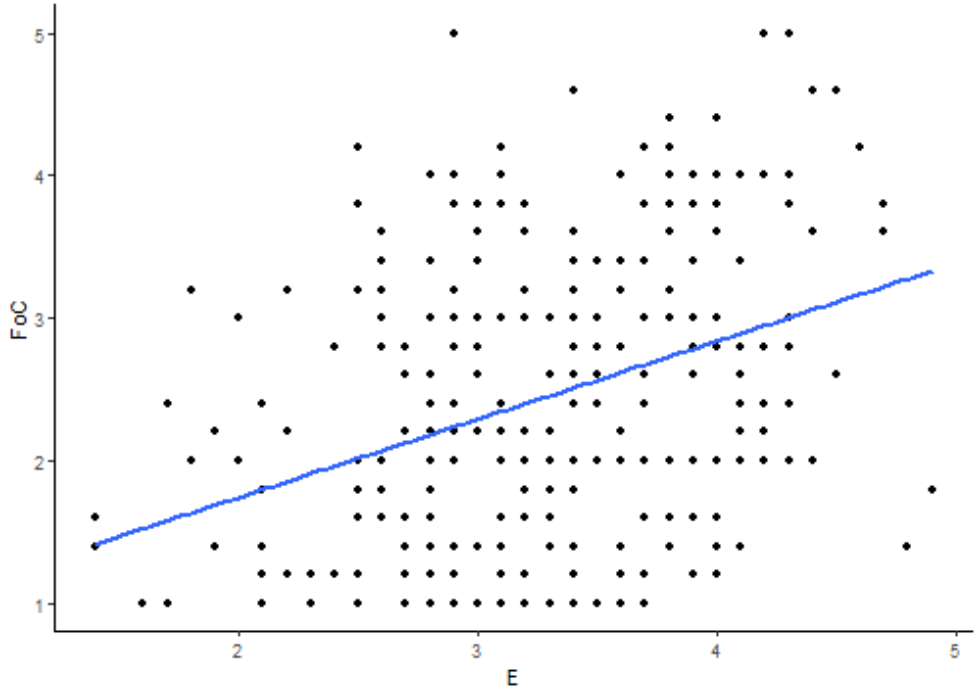
Again, the regression line is described by the formula $y = a + bX$. So we can fill that out with our model coefficients as follows:

Fear of crime = 0.65 + 0.55 * X

X is the value of the *predictor*.

The *intercept* is now the value of y when the value of the predictor is *zero*.

The coefficient for the predictor is the amount that y increases for each 1 unit increase in the predictor.



Is this a good model of Fear of crime?

We can get more information about our model using the **summary()** function.

```
summary(foc_by_E)
```

```
##
## Call:
## lm(formula = FoC ~ E, data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87698 -0.72952 -0.03902  0.70844  2.76319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.64918    0.26621   2.439   0.0153 *
## E            0.54746    0.07952   6.884 3.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9278 on 299 degrees of freedom
## Multiple R-squared:  0.1368,    Adjusted R-squared:  0.1339
## F-statistic: 47.39 on 1 and 299 DF,  p-value: 3.421e-11
```

Nicer formatting using stargazer!

```
stargazer(foc_by_E, single.row = TRUE, type = "html")
```

	Dependent variable:
	FoC
E	0.547*** (0.080)
Constant	0.649** (0.266)
Observations	301
R^2	0.137
Adjusted R^2	0.134
Residual Std. Error	0.928 (df = 299)
F Statistic	47.393*** (df = 1; 299)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fear of crime predicted by emotionality

Let's focus on the coefficients.

```
summary(foc_by_E)$coefficients
```

```
##               Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.6491774 0.26621482 2.438547 1.532835e-02
## E           0.5474598 0.07952319 6.884279 3.421376e-11
```

Estimate is the *coefficient* of each predictor; Std. Error is an estimate of the accuracy of that coefficient.

The significance of each predictor is tested using a t-test; *t value* is thus the t statistic. The *Pr(>|t|)* column is the p-value for that test.

P-values below .05 are considered significant.

Thus, E - Emotionality - is a significant predictor of Fear of Crime.

Since its coefficient is positive, Fear of Crime increases as Emotionality increases.

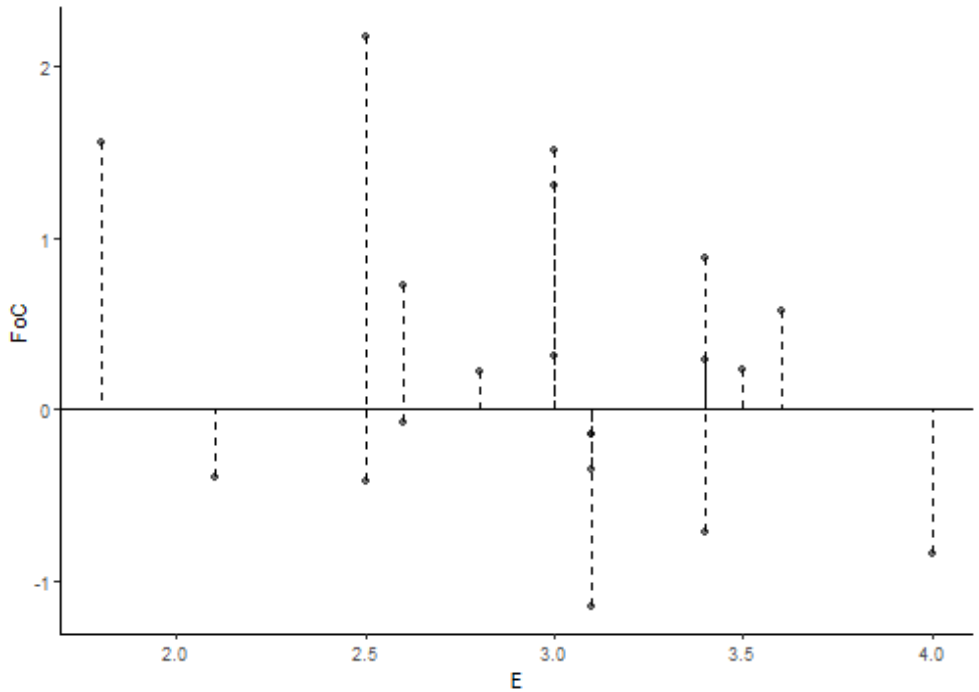
Model-fit

R-squared (R^2) is a measure of model fit. Specifically, it's the proportion of *explained* variance in the data.

We previously looked at the deviation of values from the mean.

After linear regression, we look at deviation of values from the model predictions.

What's left are the *residuals*.

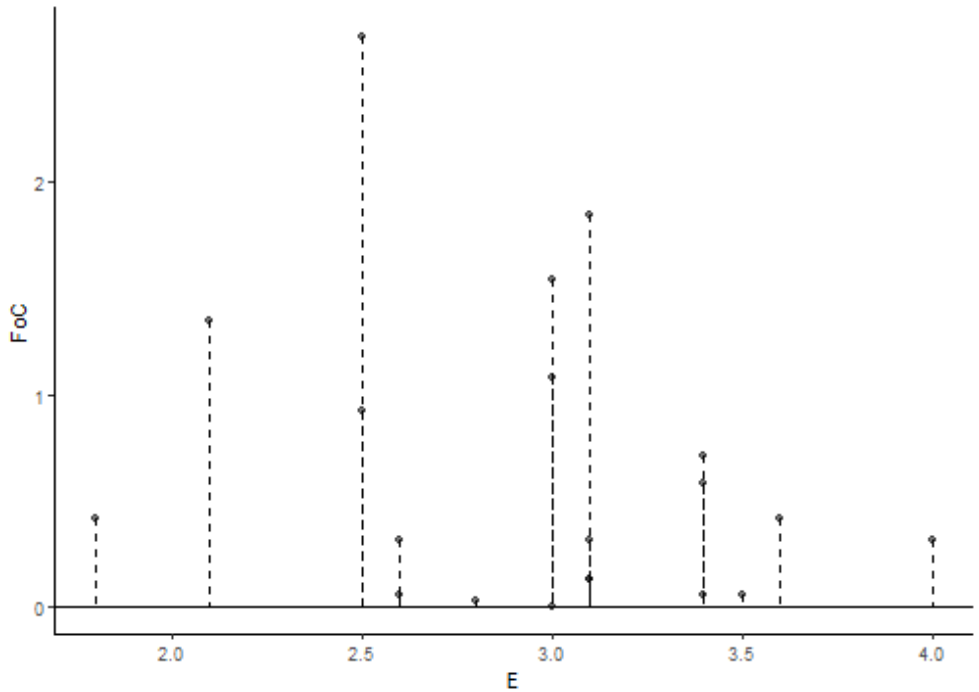


Model-fit

To work out how well our model fits, we first need to know how much *total variation* there is in the data.

For that, we sum the squared differences of the values of the dependent variable from the mean of the dependent variable y - the *total sum of squares*, SS_t:

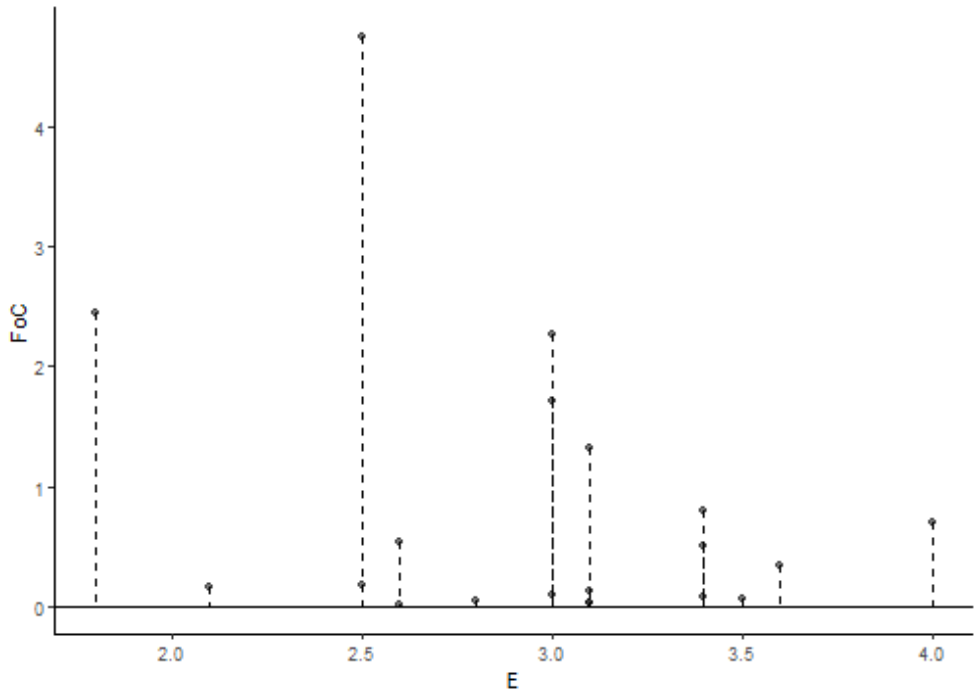
$$SS_t = \sum(y - \bar{y})^2$$



Model-fit

We then calculate the sum of the squared differences of the values of the dependent variable (y) from the model predictions - the sum of the squared residuals, SS_r:

$$SS_r = \sum(y - \hat{y})^2$$



Model-fit

Finally, we calculate *model sum of squares* - SS_m - as the difference between the *total sum of squares* and the *residual sum of squares*. This tells us, roughly, how much better our model is than just using the *mean*:

$$SS_m = SS_t - SS_r$$

R-squared ($R^2 $) can then be calculated by dividing the model sum of squares by the total sum of squares:

$$R^2 = \frac{SS_m}{SS_t}$$

This yields the *percentage of variance explained by the model*.

This is a long-winded way of saying: Higher R^2 means more explained variance, and thus, a better fitting model.

Model-fit

Thankfully, R does all these calculations for us!

```
summary(foc_by_E)$r.squared
```

```
## [1] 0.1368193
```

Our simple regression model of the effect of Emotionality on Fear of Crime explained ~ 14% of the variance.

The F-statistic

The last row here shows if the overall model is significantly better than an "intercept only" model.

```
summary(foc_by_E)
```

```
##
## Call:
## lm(formula = FoC ~ E, data = crime)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87698 -0.72952 -0.03902  0.70844  2.76319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.64918    0.26621   2.439   0.0153 *
## E            0.54746    0.07952   6.884 3.42e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9278 on 299 degrees of freedom
## Multiple R-squared:  0.1368,    Adjusted R-squared:  0.1339
## F-statistic: 47.39 on 1 and 299 DF,  p-value: 3.421e-11
```

Example of reporting a simple regression model

"Simple linear regression was used to investigate the relationship between emotionality and fear of crime. A significant regression equation was found that explained 14% of the variance, R^2 = .14, F(1, 299) = 47.39, p < .001. Fear of crime also increased significantly with increases in Emotionality, β = 0.55, t(6.884), p < .001."

Next week

Next week we'll continue with **regression**, looking at multiple predictors.

We'll also begin with **one-way ANOVA** for comparison of multiple means.

Reading

Chapter 10 - Comparing Several Means - ANOVA (GLM 1)

UNIVERSITY OF
LINCOLN

Additional support

Maths & Stats Help (AKA MASH) are a service offered by the University, based over in the library.

They offer support to both undergraduate and postgraduate students. You'll find their website at

https://guides.library.lincoln.ac.uk/mash

Note that while their website is mostly about other software, they do support R!

And a reminder!

My office hours are 1-2 Mondays and Tuesdays. You'll find me at SSB2226.