Errors

Every measurement carries an error

- It is important to have an idea of how large the error associated with your result is
- It is important to design your experiments to reduce error where possible

What types of error are there?

We often use the **mean** (or **average**) of a set of measurements as the ***best estimate*** of a quantity.

Examples where we use the mean of data to find the best estimate?

Two examples

Here are two examples of data:

- Timing a ball falling a fixed distance (to estimate acceleration due to gravity).

Attempt	1	2	3	4	5	6
Time (s)	1.11	1.18	1.02	1.09	1.10	1.13

Best estimate = …

- Weekly business profit (to estimate future cashflow).

Week	1	2	3	4	5	6
Profit ($)	310	425	545	600	520	300

Best estimate = …

For n data points $x_1, x_2, \dots, x_n$, the average is given by

$$\mu = \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x}{n}$$

Estimating errors from the spread of data

How do we measure the (different) spread in these two sets of data?

- $x_i - \bar{x}$ tells us how far each point is from the mean, but we can't just average these values (why?)
- Instead, we "average"[1] $(x_i - \bar{x})^2$, which gives us the *variance* of our sampled values

For n data points $x_1, x_2, \dots, x_n$, the variance

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\Sigma(x - \bar{x})^2}{n-1}$$

Also, we can consider the result above and rewrite as

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{n}{n-1}\frac{\Sigma(x-\bar{x})^2}{n} = \frac{n}{n-1}\left[\overline{x^2} - \bar{x}^2\right]$$

The variance is in units of $[x^2]$ rather than $[x]$ – if the x values are measured in time (sec), then s^2 is a sum of time-squared (sec^2) values.

[1] It's not quite an average: we divide by $n - 1$, not n, because only $n - 1$ of them are independent values

To measure the spread in the correct units, we use the **standard deviation** s rather than s^2:

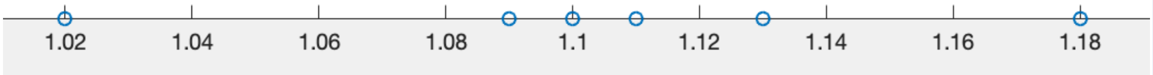
For n data points $x_1, x_2, \ldots, x_n$, the standard deviation

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}} = \sqrt{\frac{n}{n-1}\left[\overline{x^2} - \bar{x}^2\right]}$$

Consider two groups timing an experiment of a ball falling a fixed distance:

- Group 1

Attempt	1	2	3	4	5	6
Time (s)	1.11	1.18	1.02	1.09	1.10	1.13

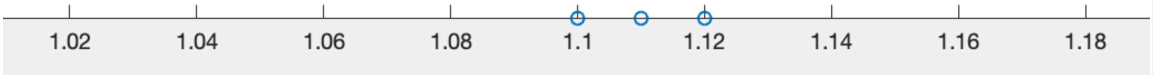


$$\overline{(x^2)} = \frac{1.11^2 + 1.18^2 + 1.02^2 + 1.09^2 + 1.10^2 + 1.13^2}{6} = 1.2233$$

$$s = \sqrt{1.2 \times [0.1.2233 - (1.105)^2]} = \sqrt{1.2 \times [1.2233 - 1.2210]} = \sqrt{0.0028} = 0.053$$

- Group 2

Attempt	1	2	3	4	5	6
Time (s)	1.12	1.10	1.11	1.10	1.11	1.10



$$\overline{(x^2)} = \frac{1.12^2 + 1.10^2 + 1.11^2 + 1.10^2 + 1.11^2 + 1.10^2}{6} = 1.2248$$

$$s = \sqrt{1.2 \times [1.2248 - 1.1067^2]} = \sqrt{1.2 \times [1.2248 - 1.2247]} = \sqrt{0.000066} = 0.0082$$

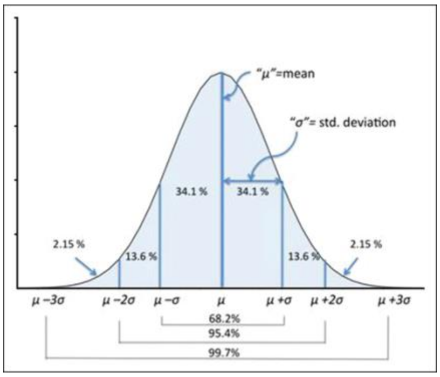
Error in the mean (best estimate)

We assume our measurements x have a *normal distribution*

- Symmetric about the mean μ
- Width represented by standard deviation σ
- The bell-shaped curve $f(x) = e^{-x^2/2\sigma^2}$

In this case, the averages of n measurements, $\bar{x}_{(n)}$, also have a normal distribution

- with the same mean μ
- with width represented by standard deviation $\sigma/\sqrt{n}$

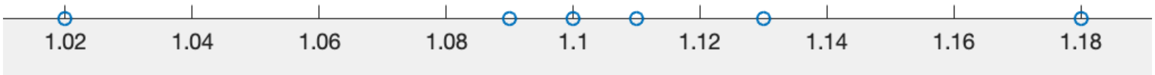


We use this width as the ***standard error in the mean* SE(*μ*)**. We usually don't know what it is, so we estimate it from the standard deviation of our experimental results (sample)

$$\mathrm{SE}(\mu) = s/\sqrt{n}$$

With our two examples:

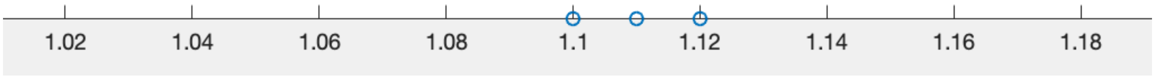
- Group 1



$$n = 6,\ \mu = 1.105,\ s = 0.053$$

Time is 1.105 $\pm \left(\frac{\quad}{\sqrt{\quad}}\right) =$

- Group 2



$$n = 6,\ \mu = 1.1067,\ s = 0.0082$$

Time is 1.1067 $\pm \left(\frac{\quad}{\sqrt{\quad}}\right) =$

We will look next week at how we can use these errors to see if our data is consistent with theoretical predictions or other experiments.

One useful way to think of them is that they provide the width of the bell-shaped curve that describes the probabilities of obtaining different averages from our experiment. This approach provides rules for *propagating* (or carrying through) errors in calculations.

Errors in a quantity need to be carried through in calculations

- When **adding** or **subtracting** quantities, the resulting **absolute error** is the *quadrature sum* (sum of squares) of the **absolute errors**
- When **multiplying** or **dividing** quantities, the resulting **relative error** is the *quadrature sum* (sum of squares) of the **relative errors**

Calculation	Absolute error ΔZ in Z (for quantities $X \pm \Delta X$ and $Y \pm \Delta Y$)
$Z = aX + bY$	$\Delta \boldsymbol{Z} = \sqrt{a^2(\Delta X)^2 + b^2(\Delta Y)^2}$
$Z = aX - bY$	$\Delta \boldsymbol{Z} = \sqrt{a^2(\Delta X)^2 + b^2(\Delta Y)^2}$
$Z = aXY$	$\Delta \boldsymbol{Z} = \lvert aXY \rvert \times \sqrt{\left(\frac{\Delta X}{X}\right)^2 + \left(\frac{\Delta Y}{Y}\right)^2}$
$Z = a\frac{X}{Y}$	$\Delta \boldsymbol{Z} = \left\lvert a\frac{X}{Y} \right\rvert \times \sqrt{\left(\frac{\Delta X}{X}\right)^2 + \left(\frac{\Delta Y}{Y}\right)^2}$
$Z = f(X)$	$\Delta \boldsymbol{Z} = \lvert f'(X) \times \Delta X \rvert$

Examples when adding or subtracting quantities:

I add 20±0.1ml of octanol to 80±1ml of water. How much liquid do I have?

I pipette 25±0.5 ml out of a flash containing 1±0.0002 l of fluid. What do I have left?

What is the concentration of octanol in the octanol/water mixture above?

Significant figures

It is tempting to report values to the same accuracy (number of digits) as we obtain from the calculation, but if we only know a value to a limited accuracy, only the first few digits carry reliably accurate information – the remaining digits *are not significant*. In scientific communication, it is important to recognise the limit of the accuracy of our data, by only giving the significant figures (s.f.) of a result.

The accuracy used to report values should be based on their error

How accurately should we report a value with associated error 1.26842 ± 0.4532?

We take the significant figures (s.f.) from the least s.f. in the given data

How accurately should we report age of the universe, 13.8 billion years, in seconds?

- Be careful not to underestimate the number of significant figures in whole numbers ending in zeros. Numbers like 10 or 200 which might be correct to 2 or 3 s.f.
- When performing calculations, *don't* use *only* the significant ones – keep the full number. If you only work with the significant figures, you will introduce round-off errors.

Correlation

What does correlation do?

We have *two* quantities we want to compare

- A *response* variable (on the *y*-axis)
- An *explanatory* or *predictor* variable (on the *x*-axis)

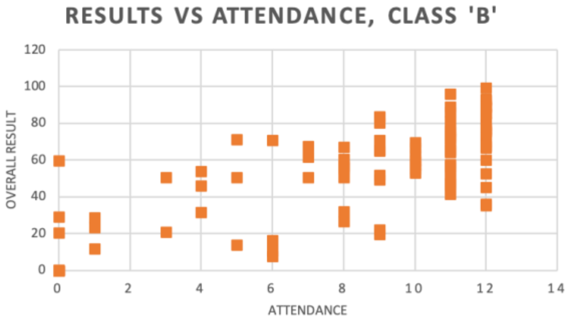
We use correlation to consider: does changing the *predictor* change the *response*?

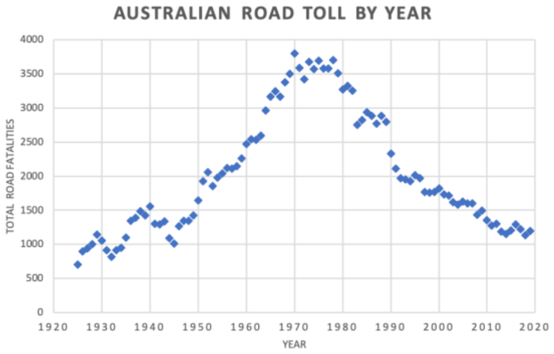
For example

- Do students that attend more tutorials do better in class?
- Do numbers of road deaths continue to decrease year-on-year?

First step: the scatter plot

It is important to plot the data you want to investigate, to make sure our approach to analysing the data is reasonable, e.g





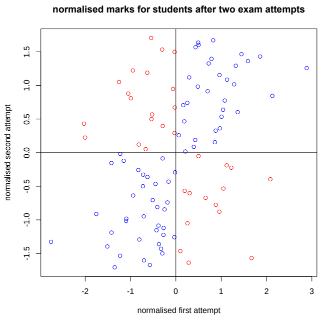
We can produce this data creating a scatterplot chart in Excel, or loading data from a file into R and using the plot command (*more on this in your tutorial next week*)

Where next?

The **correlation coefficient** is a measure a connection in the trends of each variable. It is a rating between 0 and 1, independent of the units of the variables.

We define $z_x = \frac{x-\bar{x}}{s_x}$ and $z_y = \frac{y-\bar{y}}{s_y}$. Each z represents the values about their mean values, scaled by their own spread.

From their definition, we see that z_x and z_y must be dimensionless (they have no units)!



$$\text{Since } z_x z_y \begin{cases} \geq 0 & \text{if } z_x \text{ and } z_y \text{ have the same sign} \\ \leq 0 & \text{if } z_x \text{ and } z_y \text{ have different signs} \end{cases}$$

averaging $z_x z_y$ tells us if there is a positive or a negative trend.

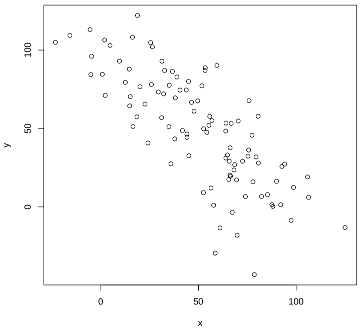
This is the **Pearson correlation coefficient** $\boldsymbol{r = \frac{\Sigma z_x z_y}{n-1}}$

Interpreting the Pearson correlation coefficient *r*

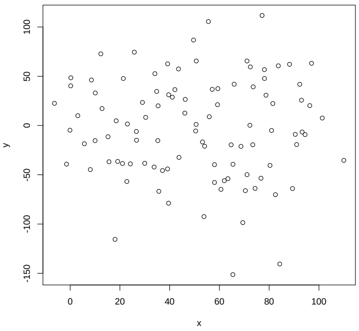
Properties of Pearson correlation coefficient

- $-1 \leq r \leq 1$
- If points lie on a straight line, $z_x = \pm z_y$, and $r = \pm s_x / s_x = \pm 1$
- There are more efficient ways of calculating r (done in Excel, R, etc.)
- r^2 is the percentage of variability in y that is explained by x

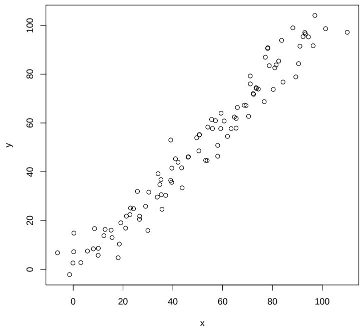
Examples of r values:



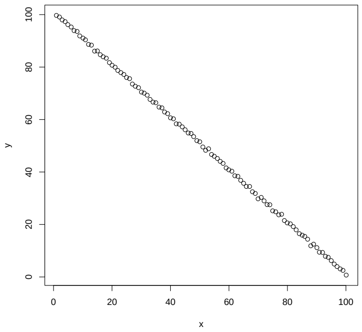
$r < 0$



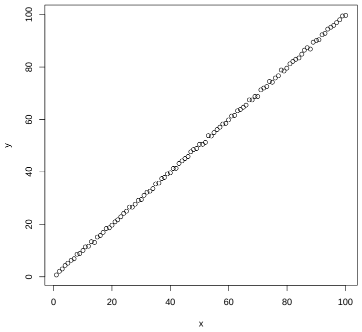
$r = 0$



$r > 0$



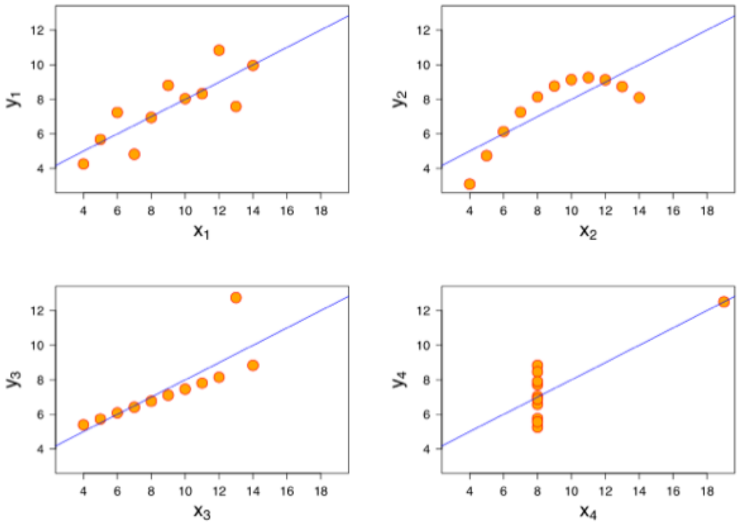
$r \approx -1$



$r \approx 1$

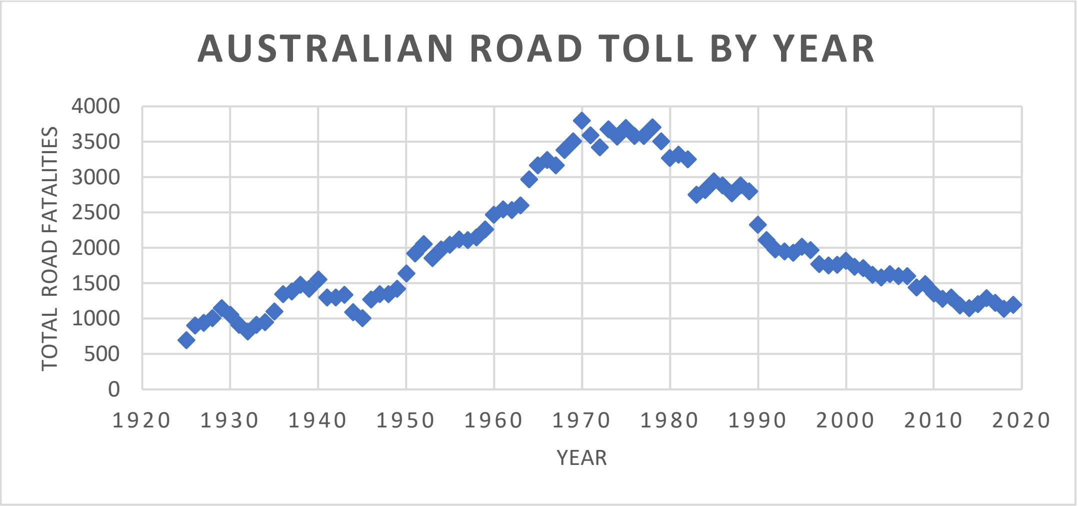
Points of caution when interpreting *r*:

- Outliers can influence *r*
- *r* does not detect non-linear trends
  - and may be high when the trend should be non-linear
- *r* is not a percentage of 'how close the data matches the line'
- Correlation ≠ causation
- Correlation ≠ regression

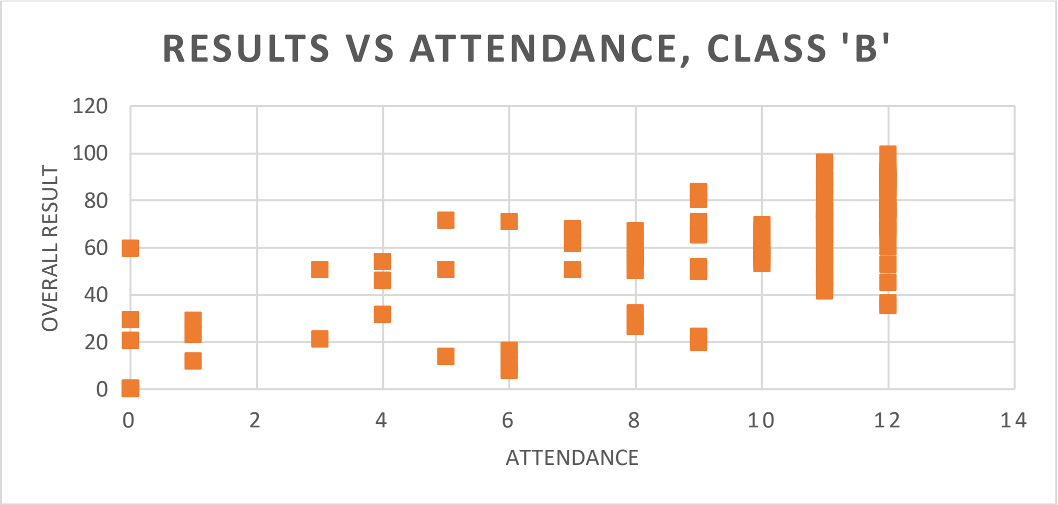


Four data sets with the same correlation coefficient $r \approx 0.8$

Examples



What are the *r* values for the whole data set? Pre 1970? Post 1970?
What events changed road toll behaviour (in what years)?



What is the *r* value for the whole data set?

Conclusions about correlation:

Regression

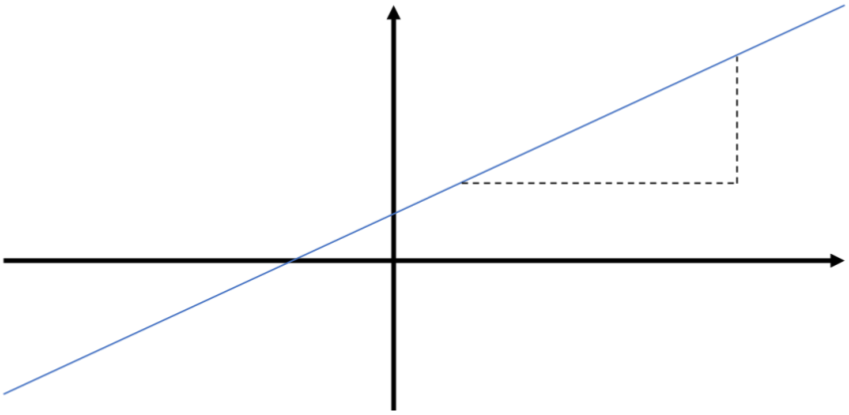
What does regression do?

Correlation tells us if there is a linear association between two quantities: regression tells us what the linear association is. The term "regression" comes from the work of Sir Francis Galton (1822-1911), studying whether sons' heights are related to their fathers': he found that sons' heights tend to be closer to the average than their fathers, which Galton described as "regression to the mean." This terminology, and the use of the word "regression" to describe this style of analysis, comes from Galton's study.

Why straight lines?

Why do we use straight lines $y = mx + c$ to model relationships between variables?

Mark the quantities m and c on the graph

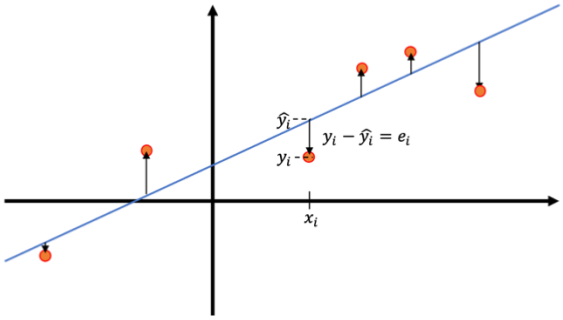


The least-squares method: finding the line of best fit

For the model[2] straight line $\hat{y} = mx + c$, there is a discrepancy between

- the data values y_i; and
- the values predicted by the straight line, $\widehat{y_i} = mx_i + c$

What is the discrepancy?



- For each data point, it is the *residual* $e_i = y_i - \widehat{y_i}$
- But we need to minimize over all points
- We can't just minimize $\Sigma(y - \hat{y})$
    - This is the same problem we encountered with the spread of data
    - Very negative is as bad as very positive
    - Turns out to be constant for fixed slope
- We minimize $\Sigma(y - \hat{y})^2$
    - i.e. we minimize the *spread*
    - This gives our approach its name

Using the method of least squares, we get the following estimates for the line parameters:

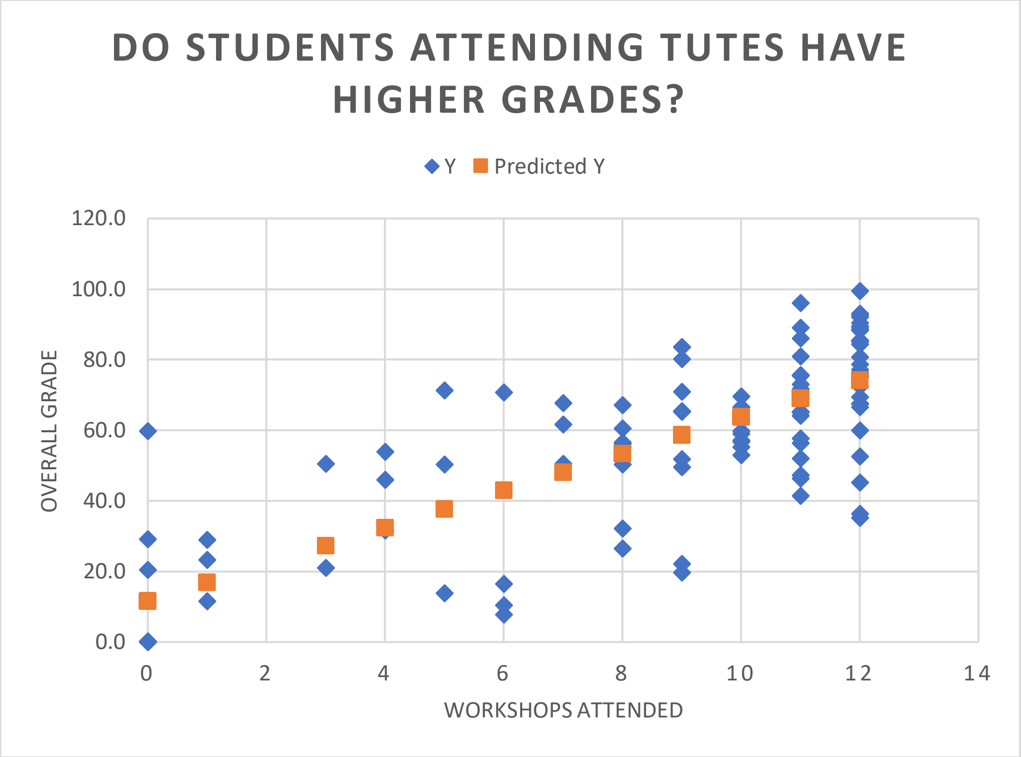
- $m = \frac{rs_x}{s_y}$, where
    - r is the correlation coefficient;
    - s_x is the standard deviation of the predictor variable; and
    - s_y is the standard deviation of the response variable
- $c = \bar{y} - \frac{rs_x}{s_y}\bar{x}$
    - Averaging both sides of our model gives $\bar{y} = m\bar{x} + c$, so
    - $c = \bar{y} - m\bar{x} = \bar{y} - \frac{rs_x}{s_y}\bar{x}$
- s_e, the error in our estimates $\hat{y}$
    - Known as the *standard error of estimate*
    - Also called *residual standard deviation*
    - $s_e = \sqrt{\frac{\Sigma(y-\hat{y})^2}{n-2}}$

We will be relying on software to calculate these quantities and find the lines of best fit.

You'll get practice at this in your workshops.

[2] In statistics, hats on symbols like $\hat{y}$ indicate the prediction of a model

Do students who attend more tutorials get higher grades?

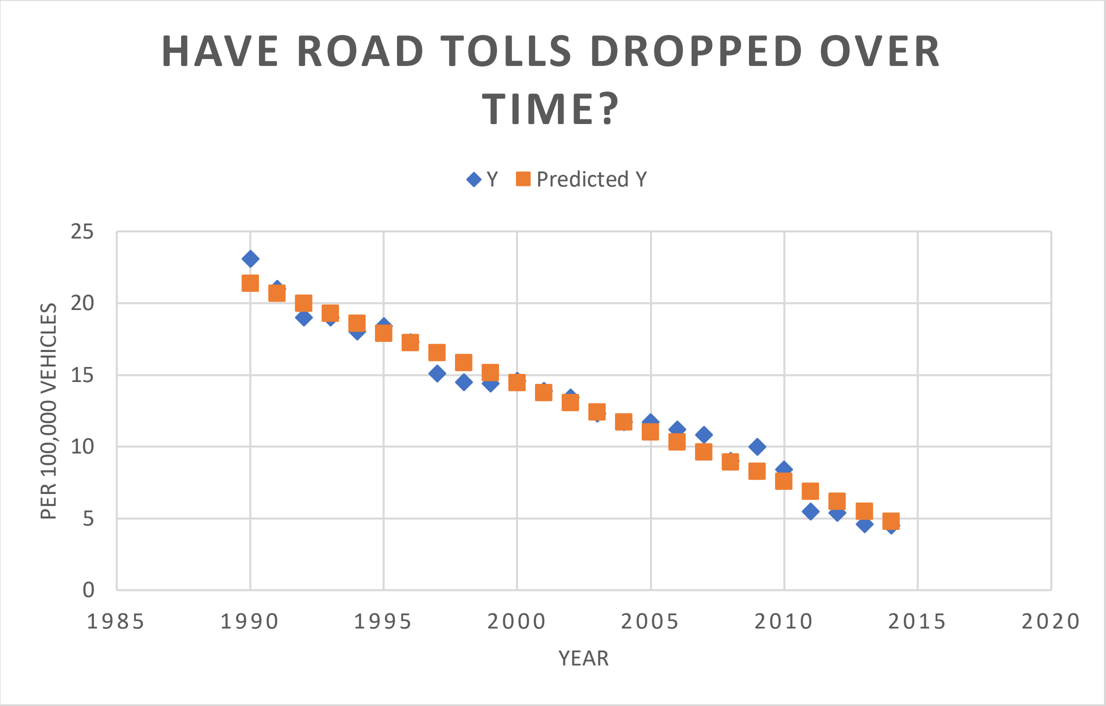


R-squared =

Standard Error in predictions =

Slope (and error) =

Have road tolls dropped over time?



R-squared =

Standard Error in predictions =

Slope (and error) =

The regression fallacy

The response is never more SDs away than the predictor

Beware the *regression fallacy:* interpreting “regression to the mean” as a real effect
Example:

Conclusions

Regression is an important element in applying the scientific method, telling us the linear relationship relating two quantities. But care needs to be taken in applying it:

What are the following assumptions, conditions, etc. to watch out for, when using linear regression?

Linear model assumption

Equal Variance assumption (homoscedasticity)

Quantitative data condition

Outliers

Interpolation requirement

Regression ≠ Causation