[WEEK 4] Regression

CS5701 - Quantitative Data Analysis

Dr Isabel Sassoon

Department of Computer Science
Brunel University London
isabel.sassoon AT brunel.ac.uk

12th October 2022

Week 4 - Housekeeping

- I will start at 10:05 prompt so please be in the Lecture theatre promptly
- The Lecture recordings will be shared in Brightspace but it can take a few hours to make them available
- We will have time for questions but you can also post them on Brightspace Discussion if you prefer ...

rec

Week 4 - Learning Outcomes

This week after the lecture, the lab and the independent practice you should be:

- able to use R to compute the **correlation**, test its significance and interpret the results
- able to use R to build a **linear regression model**
- able to interpret the model output and detect issues with the model

This week's material is also covered in chapter 6 p. 108-110) and Chapter 7 of Crawley Statistics an Introduction using R.

So far in the module...

We have answered questions such as:

- what does the data look like? (using numerical summaries and graphical displays)
- what are the confidence intervals for the mean?
- is the data normally distributed?
- do two samples of data have the same mean and/or variance?
- how do we test a hypothesis for the mean?
- and a few more

All of these were focusing on one variable...

Two variables

- So far we have focused on one variable (Exploratory Data Analysis , measures of location and spread, hypothesis testing and confidence intervals)
- We may also want to compare between two continuous variables (or more)
- This can be achieved graphically and numerically

An example

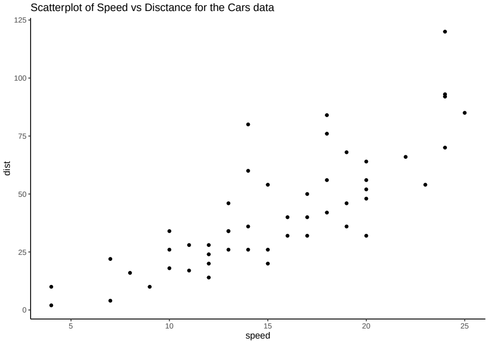
If we have this data:

	speed	dist
1	4.00	2.00
2	4.00	10.00
3	7.00	4.00
4	7.00	22.00
5	8.00	16.00
6	9.00	10.00
7	10.00	18.00
8	10.00	26.00
9	10.00	34.00
10	11.00	17.00

Do x and y behave in a similar way? are they **correlated**?

Graphical summary

One approach is to visualise this relationship using a **scatter plot**



A scatter plot will display the two variables of interest along the *x-axis* and *y-axis*

In this case there seems to be a relation, as lower values of x correspond to lower values of y

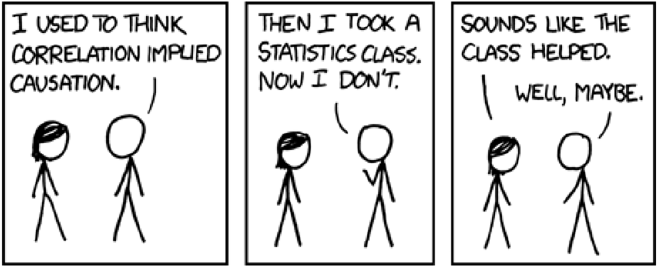
Graphical Summary

What to look out for:

- Is the relation positive, in other words do the values of x go up with higher values of y? or is the opposite occurring?
- Is the relation linear, quadratic or exponential - or other?
- Is the relation clear, or are there plenty of outliers or very different variances?

Correlation does not imply causation!

See this website for some examples:
https://www.tylervigen.com/spurious-correlations



(*) image taken from https://imgs.xkcd.com/comics/correlation.png

Correlation

- If we have two continuous normally distributed variables x and y
- **correlation** is defined in terms of $var(x) = s_x^2$, $var(y) = s_y^2$ and the **covariance** of x and y
- The **covariance** is the way x and y vary together and is denoted as $cov(x, y)$
- The **correlation coefficient** r is defined as:

$$r = \frac{cov(x, y)}{\sqrt{s_x^2 s_y^2}}$$

- This is computed in R using `cor()`

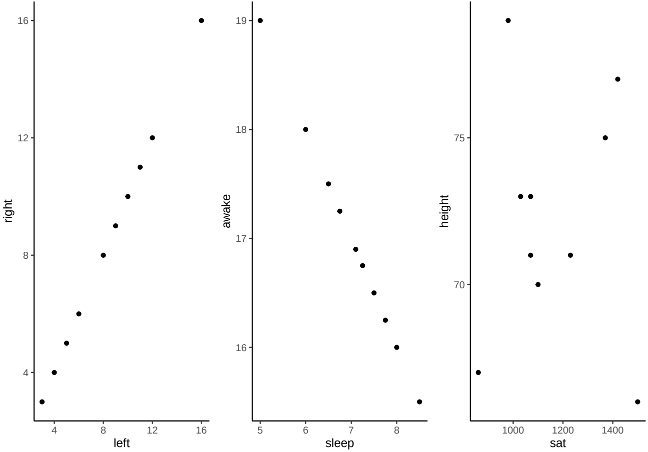
Correlation - interpretation

- The value of r which is the sample estimate for ρ (the correlation) has values between -1 and 1
- a value of $r = 0$ implies that there is no **linear** association or correlation
- a value of $r = 1$ implies that there is a perfect linear relation or correlation where a higher value for x corresponds to a higher value of y
- a value of $r = -1$ implies that there is a perfect linear relation or correlation where a higher value for x corresponds to a lower value of y
- intermediate values imply that there are varying degrees of linear association

When data is not normally distributed then there is the option to use **Spearman's correlation coefficient**

Correlation - examples

Which of these plots shows a correlation of $r = 1$?



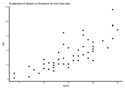
Respond here : pollev.com/isabelsassoon

Computing correlation in R

- `cor(x,y)` computes r - the correlation coefficient
- `cor.test()` returns both the correlation coefficient and the p-value of the correlation.
- The hypotheses being tested are
  - $H_0 : \rho = 0$ (i.e. no correlation)
  - $H_1 : \rho \neq 0$ (i.e. there is a correlation)

Computing correlation - example

```
cor.test(cars$speed, cars$dist)

   Pearson's product-moment correlation

data:  cars$speed and cars$dist
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949
```



The correlation is 0.81 and the p-value is very small. The linear correlation in this case is significant.

Regression

- When we computed the correlation coefficient there was no "difference" in role between the two variables
- Sometimes we need to differentiate between an **explanatory or independent** variable and a **response or dependent** variable
- When both of these are continuous then **Regression Analysis** is a useful approach to model this relationship.

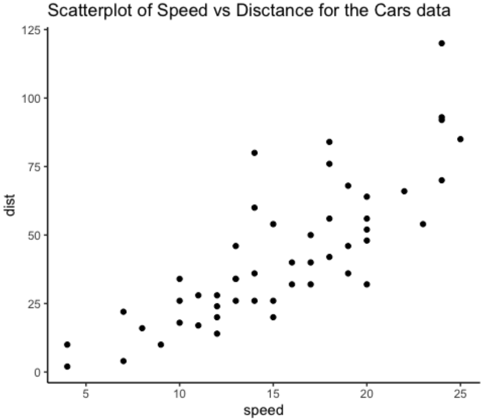
Regression - motivation

- We have some data on the speed of a car and the distance taken to stop
- We may want to answer the question: *"Does a higher speed result in a longer breaking distance?"*

	speed	dist
1	4.00	2.00
2	4.00	10.00
3	7.00	4.00
4	7.00	22.00
5	8.00	16.00
6	9.00	10.00
7	10.00	18.00
8	10.00	26.00
9	10.00	34.00
10	11.00	17.00

Table: First 10 rows of data from the cars data. This data is from 1920

An initial step is to understand what the relation is in the sample between the **response variable** or **dependent** variable and the **explanatory** or **independent** variable is to use a **scatter plot**:



This can give us some visual clues

Regression Analysis

- **Regression Analysis** is the statistical method you use when both the response and the explanatory variable are continuous,
- the essence of regression is to use the sample data to estimate parameter values and their standard errors
- in its simplest form **Linear Regression** the relationship between the response variable (y) and the explanatory variable (x) is:

$$y = a + bx$$

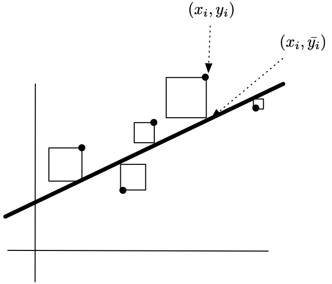
- this has two parameters: a which is the intercept or the value of y when $x = 0$ and b which is the slope

The challenge with Linear Regression is how to get the best fitting line (defined by a and b...)

Linear Regression

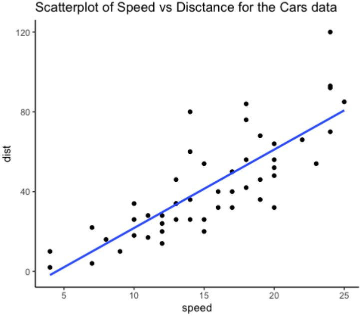
- the idea is to **minimise** the total spread of the y values from this line
- similarly as with the variance, we look at the **squared y distances** from the line, and sum them up to obtain the **Sum of Squared Errors**

$$SSE = \Sigma_{i=1}^{n}(y_i - \bar{y}_i)^2$$



Linear Regression for the example

- Going back the the `cars` data set, and using the **speed as the independent variable** and **distance as the dependent variable**
- Using the `lm` function in R, the coefficients were computed as: $a = -17.63$ and $b = 3.90$



Linear Regression - Interpretation and applications

- using a sample of data to estimate the regression equation can tell us whether there is a positive or negative relation between the dependent and independent variables
- the regression equation can also help us predict values for new data points
- but its important to remember that fitting a line is easy, it does not always makes sense to do so.

Using the regression line to predict

- given a regression line $y = a + b \times x$
- it is also possible to find, for any value y the predicted value the regression line would assign
- this is referred to as $\bar{y}$

For example in the `car` data we may want to find out the estimated stopping distance for speed $=21$

- We know that $y = -17.63 + 3.90 \times \text{speed}$
- so our $\bar{y}_{speed=21} = -17.63 + 3.90 \times 21 = 64.2$

Brunel University London

Residuals

- the difference between each data point and the value predicted by the model for the same value of x is called the **residual**
- a residual d is defined as $d = y - \bar{y}$ and we can substitute the equation line so

$$d = y - (a + b \times x) = y - a - b \times x$$

- residuals can be positive (when the data point is above the line) or negative (below the line)
- the best fit line will is defined by the a and b values that minimise the sum of squares of the ds - the **SSE**(see box 7.2 for the proof)
- residuals help assess how well the regression line fits the data

Diagnostics

- sometimes this is referred to as **goodness of fit**
- this tells us whether the straight line that minimises the **Sum of Squared Errors** SSE gives us an adequate representation of the data
- in order to assess this we can find out the proportion of the **total** variation of the data that is accounted for in the linear trend
- the variation explained by the model is called the **regression sum of squares SSR** and (as we have already seen) the unexplained variation is the **error sum of squares SSE**
- then

$$SSY = SSR + SSE$$

(*) for proof see page 126 crawley

Regression Diagnostics - F ratio

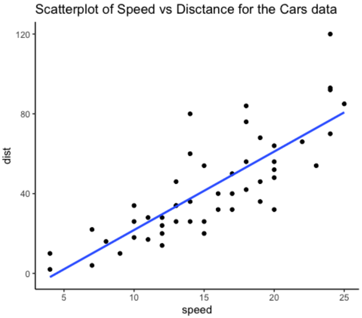
- The **F ratio** is the ratio between two variances
- the treatment variance (variance of SSR) is in the numerator and the variance of SSE is in the denominator
- H_0 under test in linear regression is that the slope of the regression line is 0 (x and y are independent)
- $H_1 : \beta \neq 0$
- F is used to test this ratio between variances.

Regression Diagnostics - Degree of Fit r^2

- There is a need to quantify the **degree of fit**
- Such a measure would be 1 if all the data points fit straight on the linear regression line, and 0 when x explains none of the variation in y
- The metric to achieve this is **the fraction of the total variation that is explained by the regression**
- This is $r^2 = \frac{SSR}{SSY}$
- This value is found in the R output using `summary(model.lm)`

Linear Regression for the example

- Going back the the `cars` data set, and using the speed as the independent variable and distance as the dependent variable
- Using the `lm` function in R, the coefficients were computed as: $a = -17.63$ and $b = 3.90$



Diagnostics - what to check (numerically)?

- r^2 - what proportion of the variance in Y is explained by our regression model?
- F-statistic - is the variance that our regression model explains (SSR) significantly different from the one explained by the error (SSE)
- Are the coefficients (a,b) significant?

Summary of the model - Diagnostics

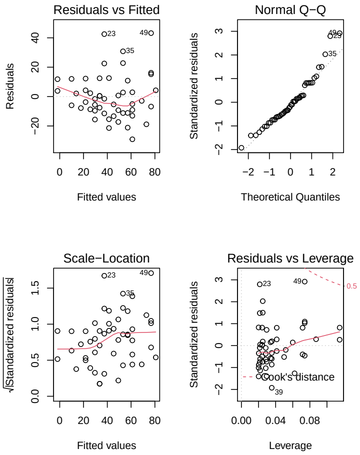
```
> summary(cars.lm)

Call:
lm(formula = cars$dist ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Summary of the model - Diagnostics

- The top of the output has a summary of the residuals
- The next part gives us our **coefficients** (a,b)
- The large **F-value** indicating that we can confidently reject H_0 that the SSR vs SSE are the same.
- Another important number is the r^2 which is best when closest to 1 (as it means most of the variance is explained by the model)

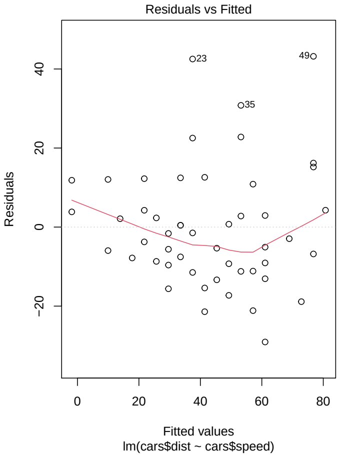
Model Checking

- the final step in appraising a model involves checking the **constancy of variance** and **normality of errors**
- there is an R function that produces the diagnostic plots required `plot(model.lm)`



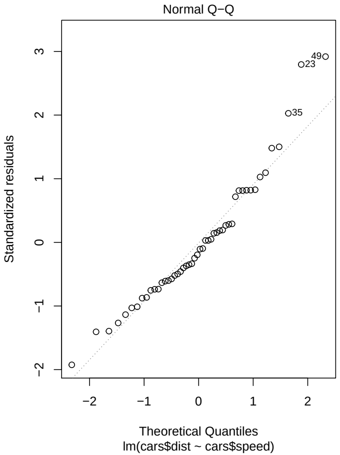
Model Checking - Plot 1

- this shows the residuals on the y-axis vs. the fitted values on the x-axis
- ideally this should look random
- if there are some trends such as larger scatter with larger fitted values then this indicates a problem with the model assumptions



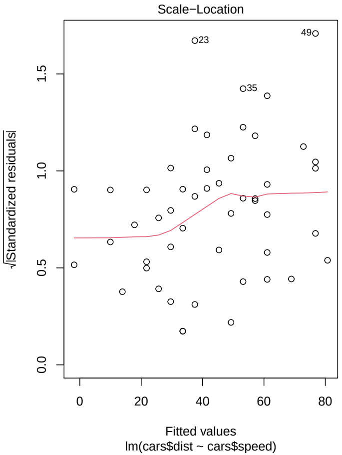
Model Checking - Plot 2

- this shows the **quantile-quantile** or QQ plot
- this should be a straight line if the errors are normally distributed
- if there was an S-shaped or banana shaped pattern a different model would need to be fitted to the data



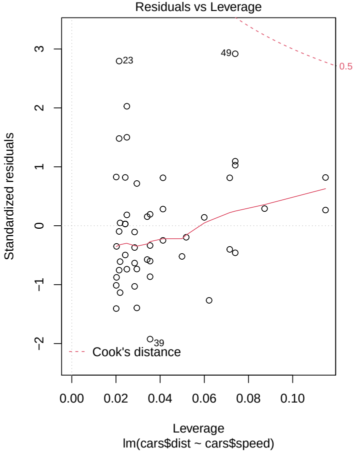
Model Checking - Plot 3

- this is like Plot 1, but on a different scale
- it shows the square root of the standardised residuals against the fitted values
- if there was a problem the points would be distributed inside a triangular shape, with scatter of residuals increasing as the fitted values increase



Model Checking - Plot 4

- this plot highlights the **influential** points, these points having the largest effect on the parameter estimates
- **Cook's distance** (red contours on the plane) shows the standardised residuals vs leverage for each point in the data
- this information is easier to scrutinise using `influence.measures(model.lm)`



Model Checking plots - summary

- The **residuals vs fitted** plot and the **Q-Q** plot are the most important
- It is not enough to look at the r^2 and the F test
- And always plot the data first
- For our example data on cars, the plots did not flag any serious concerns.

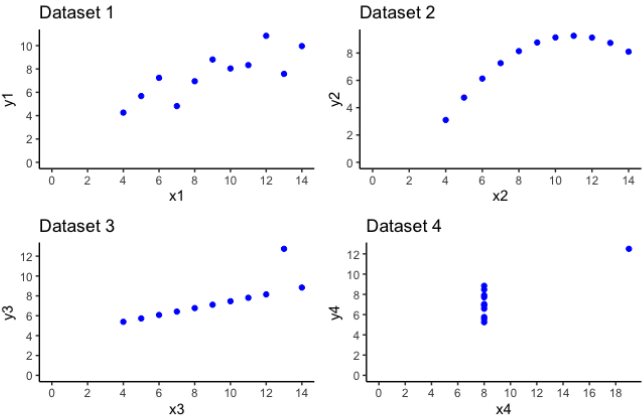
Example - Anscombe Quartet data

Here are four data sets with have similar traditional statistical properties (mean, variance, correlation) that look very different when plotted.

	x1	x2	x3	x4	y1	y2	y3	y4
Min.	4.0	4.0	4.0	8	4.260	3.100	5.39	5.250
Median	9.0	9.0	9.0	8	7.580	8.140	7.11	7.040
Mean	: 9.0	9.0	9.0	9	7.501	7.501	7.50	7.501
Max.	14.0	14.0	14.0	19	10.840	9.260	12.74	12.500

Example - Anscombe Quartet data

When the data is plotted we do get a different view

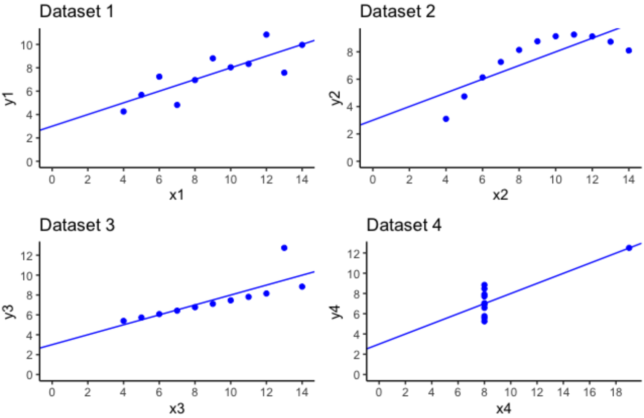


Which ones are suitable for linear regression?

Respond here - pollev.com/isabelsassoon

Example - Anscombe Quartet data

For each of these it is possible not only to build a linear regression model, but the regression line is identical estimates for α and β.



The regression line for each of these is $y_i = 3.001 + 0.5x_i$ where $i = 1, 2, 3, 4$

Example - Anscombe Quartet data

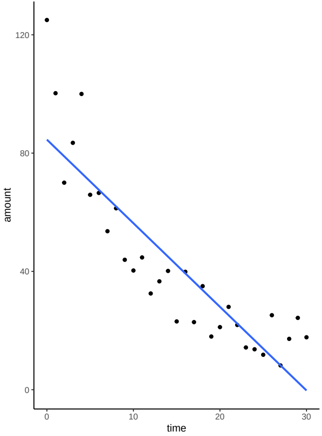
- This example emphasises the importance of **visualising the data**
- Similar range, mean and median do not show the full picture
- This data set is available in R (`anscombe`), take a look at it and see how the regression summaries and plots vary

Transformation

- $y = a + b \times x$ is not the only two parameter model for describing the relationship between a response variable and a single continuous explanatory variable
- other options include
  - log X $y = a + b \times logx$
  - log Y $y = exp(a + b \times x)$
  - asymptotic $y = \frac{ax}{1+bx}$
  - reciprocal $y = a + \frac{b}{x}$
  - power law $y = ax^b$
  - exponential $y = ae^{bx}$

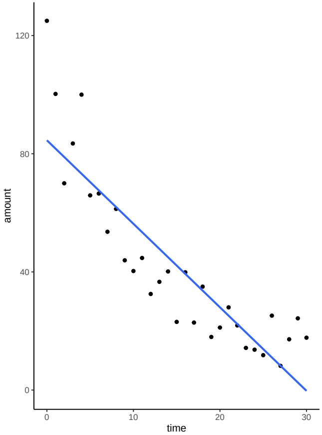
An example

Does this look like a good linear fit?



An example

Does this look like a good linear fit?



There appears to be a curvature in the data, we can see this as most of the residuals at the extremes are positive.

An example - LM summary

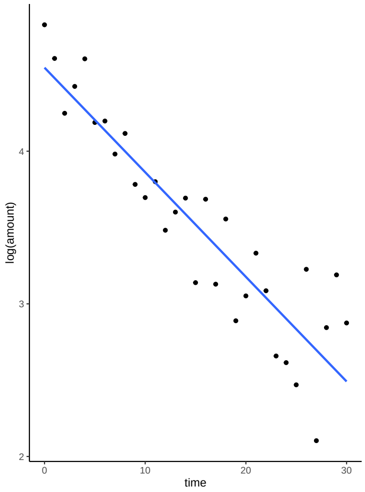
If we just look at the summary of the lm model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	84.5534	5.0277	16.82	0.0000
decay$time	-2.8272	0.2879	-9.82	0.0000

- We may be distracted by the fact that the model does explain 76% of the variance (see r^2). This is not a good measure of model adequacy.
- **Looking at this table is not enough to establish the suitability of a linear model.**

Which transformation?

- as this data is related to a decay process we will try to use the exponential relationship $y = ae^{-bx}$
- if we take logs for both sides: $log(y) = log(a) - b \times x$ we have a linear relationship
- this can be modelled in R using $log(y)$ as the dependent variable instead of y and an even better r^2 is obtained (in this case)



Polynomial, Non Linear regression and GAMs

There are some other approaches to modelling the relationship between a dependent and an explanatory variable some options:

- **polynomial regression**: the idea is that we have just one explanatory variable but can fit higher powers of x such as x^2
- **non linear regression** to the data using `nls()` when non linear models are run in R the exact nature of the equation needs to be specified in R
- When the relationship between x and y is non linear but we don't have a theory to suggest a particular equation then **Generalised Additive Models (GAM)** can be helpful. GAMs work by fitting non parametric smoothers to the data.

For more on Polynomial, Non linear regression and GAM see Chapter 7 of Crawley.

Lecture Summary

In the lecture we looked at:

- Correlation and how to compute and interpret it
- Linear regression, how to compute it, interpret is and detect issues with the model
- Other approaches such as polynomial regression

In the labs you will practice correlation and linear regressions.

This week's material is also covered in chapter 6 p. 108-110) and Chapter 7 of Crawley Statistics an Introduction using R.