UNIVERSITY OF BATH

Data Science and Statistics in Research: unlocking the power of your data

Session 3.2:

Linear regression

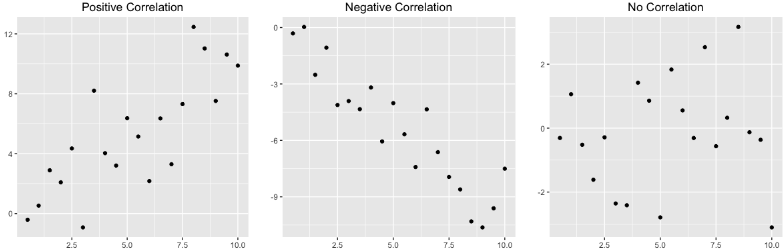
Outline

Correlation

CORRELATION

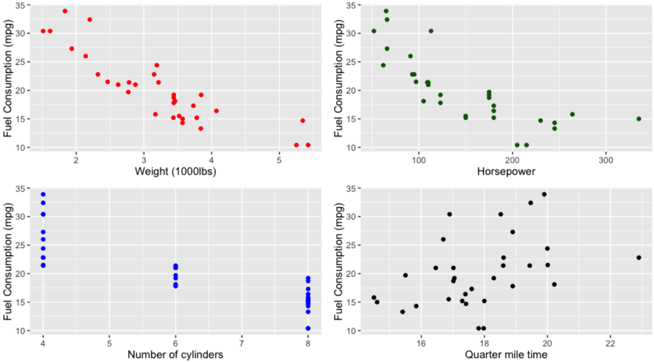
- Correlation is a measure of the association between two variables.
- The correlation coefficient quantifies the strength of any association
  - takes values between -1 and +1
  - values between 0 and +1 indicate a positive association, i.e. as one variable increases the other increases
  - values between -1 and 0 indicate a negative association, i.e. as one variable increases the other decreases
  - values around 0 indicates no relationship.

CORRELATION



Example: Motor Trend Car Road Tests

- The mtcars dataset in R contains fuel consumption and 10 other aspects of 32 cars from 1973-74.
- To check for correlation, we plot fuel consumption against weight, horsepower, number of cylinders and quarter mile time.



Example: Motor Trend Car Road Tests

- We can also calculate correlation coefficients between these variables.

	mpg	wt	cyl	hp	qsec
mpg	1.00	-0.87	-0.85	-0.78	0.42
wt		1.00	0.78	0.66	-0.17
cyl			1.00	0.83	-0.59
hp				1.00	-0.71
qsec					1.00

- We can see that there is
  - strong negative correlation between fuel consumption and weight, horsepower and number of cylinders.
  - low positive correlation between fuel consumption and quarter mile time.

Linear Regression

Response and explanatory variables

- Variables can be classified as **explanatory** or **response**.
- **Response** – we are interested in changes in the response, i.e. our variable of primary interest.
- **Explanatory** – variables that may explain changes the response variable.

Example: Motor Trend Car Road Tests

- Suppose we are interested in how fuel consumption (in miles per gallon) is affected by weight, number of cylinders and transmission.
- Fuel consumption is our response variable.
- Weight, number of cylinders or transmission are explanatory variables

Linear Regression

- Linear regression allows us to analyse the relationship between two variables.
- The most straightforward relationship is a linear, or straight line, relationship.
- This involves fitting a straight line through the data points.
- We could do this by hand, but this would introduce error.

Linear Regression

- Straight lines are described by the formula

$$y = a + bx.$$

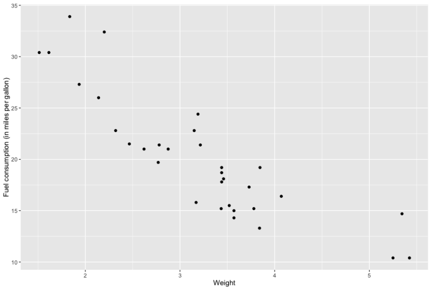
- y - the **response** variable.
- x - the **explanatory** variable.
- a is the intercept; the point at which the line cuts the y axis
  - tells us what response we would expect if the explanatory variable was equal to zero.
- b is the slope of the line
  - is the increase (or decrease) of the response variable per unit increase in the explanatory variable.
- The line is constructed to by choosing a and b to be the best possible fit
  - least squares.

Checking for linearity

- Before attempting to fit a linear model, you should check if the relationship between the response and explanatory variable are linear.
- A scatter plot can be useful to assess this.
- The correlation coefficient indicates how well a straight line fits the data.

Example: Motor Trend Car Road Tests

- We are interested in modelling fuel consumption (response variable), in relation to weight (explanatory variables).
- The correlation coefficient between weight and fuel consumption is -0.87.



Example: Motor Trend Car Road Tests

- We describe the fuel consumption (mpg) of a car using the formula

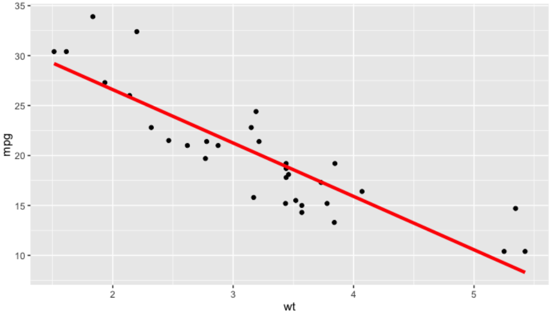
$$\text{mpg} = a + b * \text{weight}.$$

- To fit this model in `R` we use the following

```
formula <- mpg ~ 1 + wt
mod <- lm(formula, data = mtcars)
```

Example: Motor Trend Car Road Tests

- This is the resulting straight line.



OUTPUT

- Outputs are estimates of the model coefficients together with standard errors.

```
summary(mod)

Call:
lm(formula = mpg ~ 1 + wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5432 -2.3647 -0.1252  1.4096  6.8727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-squared:  0.7528,Adjusted R-squared:  0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

Statistics for Regression Coefficients

- Hypothesis tests can be used to check whether there is a significant relationship between the response and explanatory variables
  - **null**: $b = 0$
  - **alternative**: $b \neq 0$.
- As default, `R` will test the hypothesis that if the true intercept and slope terms are greater than zero.
- Interest is usually in the effects of the explanatory variables rather than the intercept.

EXAMPLE: MOTOR TREND CAR ROAD TESTS

	Estimate	Std. Error	test Statistics	$P(T > \|t\|)$
Intercept	37.29	1.88	19.86	< 0.0001
Weight	-5.34	0.56	-9.56	< 0.0001

- Testing whether there is a significant association between weight and fuel consumption results in a p-value of <0.0001.
- Therefore, there is a significant association between weight and fuel consumption.
- A 95% confidence interval for the coefficient of weight is $(-6.49, -4.20)$.

R^2 Statistic

- R^2 is a measure of how well the model fits the data.
- Indicates the proportion of variance of the response explained by the model.
- Takes values between 0 and 1, with 0 indicating the model does not explain changes in the response and 1 indicating a 'perfect' model.
- High values of R^2 indicate good model fit.
- R^2 will always increase as more explanatory variables are added to the model
  - Adjusted R^2 penalises models with lots of parameters.
- In the cars example, we have a R^2 of 0.75 and an adjusted R^2 of 0.74.

PREDICTION

- Once we have fitted a model, we can use it predict values of the responses for a set of values for the explanatory variables.
- This is done by plugging the values into the regression equation.
- For example,

$$\text{mpg} = 37.29 - 5.34 * \text{weight}$$

  - for a weight of 3500 lbs we would predict fuel consumption to be 18.58 mpg.

Multiple Regression

Multiple Regression

- Multiple regression is a natural extension of the linear regression model.
- It is used to predict values of a response from **several** explanatory variables.
- Each explanatory variable has its own coefficient.
- The response variable is predicted from a combination of all the variables multiplied by their respective coefficients.

Multiple Regression

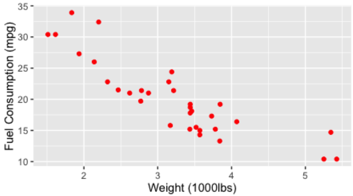
- Multiple regressions are described by the formula

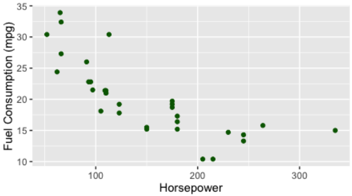
$$y = a + bx_1 + cx_2 + dx_3 + \ldots$$

- y - **response** variable.
- $x_1, \ldots, x_n$ - **explanatory** variables.
- a is the intercept; the point at which the line cuts the y axis
  - tells us what response we would expect if the explanatory variable was equal to zero.
- $b, c, d \ldots$ are the coefficients of the i^{th} explanatory variables
  - is the increase (or decrease) of the response variable per unit increase in the i^{th} explanatory variable.

Example: Motor Trend Car Road Tests

- We saw that both the weight and horsepower are linearly related to fuel consumption.





Example: Motor Trend Car Road Tests

- We can describe the miles per gallon (mpg) of a car using the formula

$$\text{mpg} = a + b * \text{weight} + c * \text{horsepower}.$$

- We extract the intercept and slope estimates and standard errors, as well as the R^2 statistic.

	Estimate	Std. Error	test Statistics	$Pr(T > \|t\|)$
Intercept	37.23	1.60	23.29	< 0.0001
Weight	-3.88	0.63	-6.13	< 0.0001
Horsepower	-0.03	0.01	-3.52	0.0015

Example: Motor Trend Car Road Tests

- Both weight and horsepower are significant predictors of fuel consumption.
- The R^2 (R^2: 0.83, Adjusted R^2: 0.81) indicate that this model is a better fit than one with just weight as the explanatory variable.

PREDICTION

- Once we have fitted a model, we can use this to create predictions of responses for a set of values for explanatory variables.
- This is done by plugging values of the explanatory variables into the equation.
- For example,

$$\text{mpg} = 37.23 - 3.88 * \text{weight} - 0.03 * \text{horsepower}$$

so if we have a weight of 3500 lbs and a horsepower of 150 we would expect fuel consumption to be 19.15 mpg.

Choosing your model

Model Selection

- Often there will be choices of which explanatory variables to include.
- This is known as model selection.
- One of the most common methods is to compare values of the R^2 statistic
  - choose the model which has the largest R^2.
- Others include Akaike Information Criteria (AIC) and Analysis of Variance (ANOVA).

Example: Motor Trend Car Road Tests

- We saw that both the weight and horsepower are linearly related to fuel consumption.
- There are three possible models
  - Weight
  - Horsepower
  - Weight and Horsepower.
- Adjusted R^2 for all models:

Model	R^2
Weight	0.7442
Horsepower	0.5892
Weight + Horsepower	0.8148

- If we use R^2 to select our model, we would choose a model with both weight and horsepower.

Association and Causation

- Correlation and regression allows us to look at the relationship between variables.
- Strong **associations** do not necessarily imply a **causal** relationship.
- The association could be due to another, unmeasured variable (confounder).
- For example, there is a strong relationship between rates lung cancer and owning a washing machine. However, not having a washing machine does not *cause* lung cancer. A possible confounder could be socio-economic status.

Correlation/association does not imply causation!

Any Questions?