Durham University
Research Methods Centre

MATH43515: Multilevel Modelling

Lecture 2: Multiple Linear Regression

Module convenor/Tutor:
Andy Golightly

Outline

- Brief recap (simple linear regression)
- Multiple linear regression
- Assumptions/Diagnostics
- Goodness of fit (F-test)

Simple linear regression (SLR) recap

The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Response/outcome/dependent variable Y_i
- Predictor/explanatory/independent variable/covariate x_i
- β_1 and β_0 estimated by b_1 and b_0 by minimising the error sum of squares (SSE) (corresponds to maximum likelihood estimates for normal errors).
- The error term ε_i represents the discrepancy between the outcome and line for the i^{th} unit. Assumed normal with zero mean and constant variance.
- R^2 = SSR/SST (coefficient of determination) is the proportion of variance explained by the regression model.

SLR with categorical predictors

The values of a categorical variable indicate group membership of observations rather than a quantitative measurement

Examples:

- **Binary** – e.g. smoker; “Yes”, “No”
- **Ordered** – e.g. obesity; “underweight”, “normal”, “overweight”, “obese”
- **Unordered** – e.g. region; “Bristol”, “London” and “Stoke”

SLR with categorical predictors: Indicator variables

- A binary variable is a special type of categorical variable called an indicator variable taking the values of 0 and 1.
- Including indicator variables in a regression will model a difference in means between groups.
- Indicator variables can also be used to represent categorical variables which have more than two categories (see Week 7, video 3 of MATH42715).

Multiple linear regression

- In real life, the effect of one variable rarely takes place in isolation.
- In Multiple regression, we explain the outcome (dependent) variable using a particular predictor after accounting for the effect of other predictors (independent variable) included in the model.
- It provides an adjusted estimate for the effect of a specific predictor, which can be different from the simple regression results.
- Each additional variable may explain more variability in the outcome variable and consequently reduces the Error Sum of Squares (SSE).
- Typically

$$\uparrow \text{Predictors} = \uparrow \text{SSR} = \downarrow \text{SSE}$$

Do dogs improve health?

… dog ownership reduces risk of Cardiovascular Disease (CVD) in single-person households and lower mortality in the general population.

Mubanga et al, (2017)

- Is the effect simply due to *owning* a dog?
- Is the effect the same for everyone (e.g., age, employment status, all dog breeds)?
- Could this relationship be *confounded* by something else?

Multiple linear regression can help disentangle this.

Durham University

Multiple linear regression

- Multiple regression allows us to understand the effects of multiple predictors on an outcome variable.

- However, we really want the simplest model that best explains the data (parsimony).

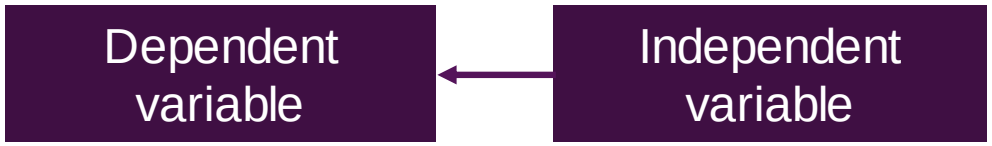
- We need to carefully think about the associations we are testing, what variables should be included in the model, and how to interpret these associations correctly.

Multiple linear regression

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

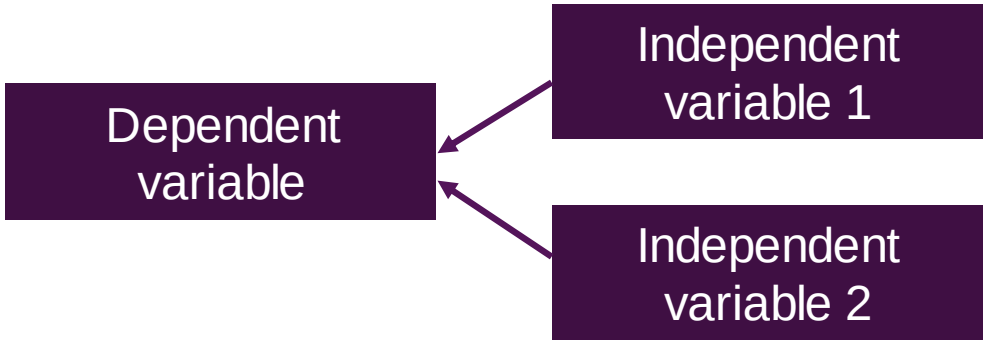
$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Simple linear regression



$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

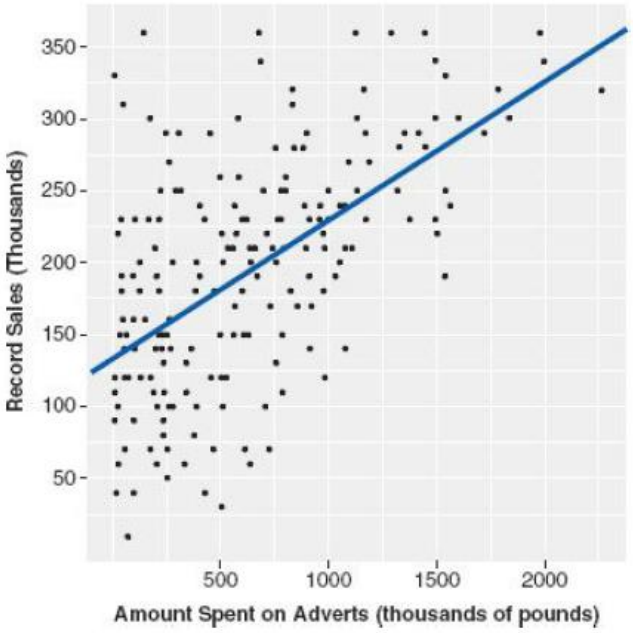
Multiple linear regression



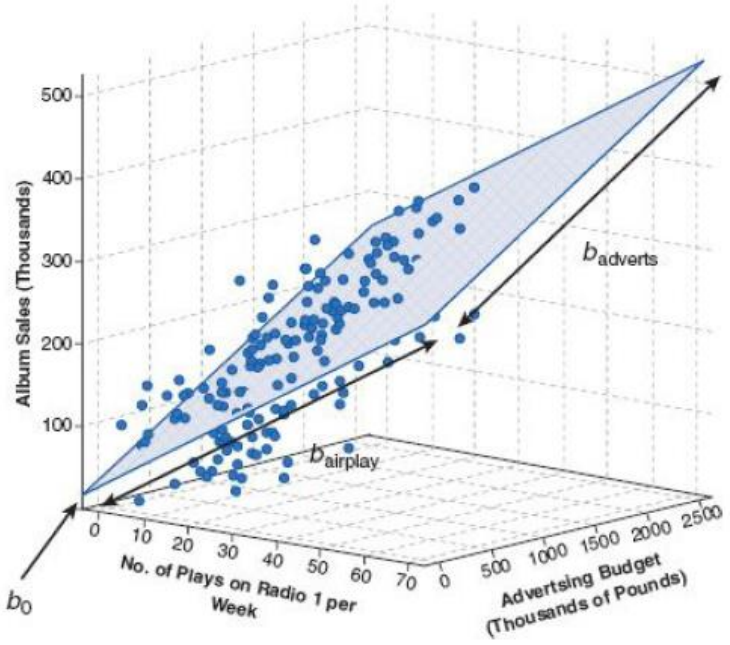
$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

What ‘line’ are we fitting?

Simple linear regression



Multiple linear regression



Durham University

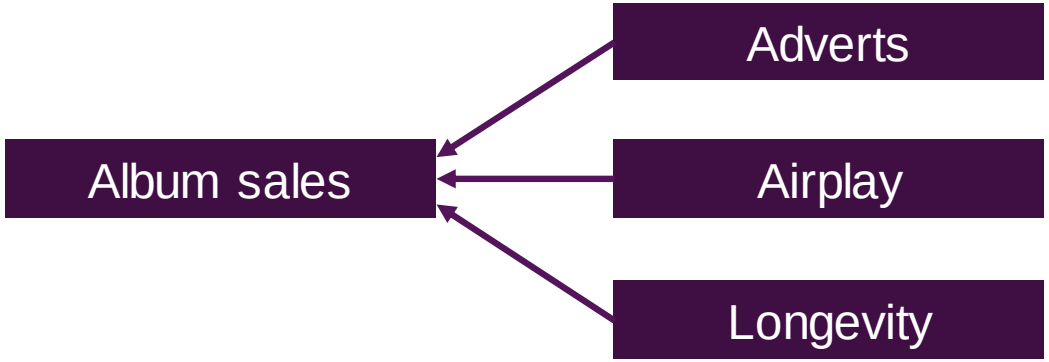
Multiple linear regression

- Allows us to test an association between an outcome and predictor after adjustment for other relevant predictor variables.
- Allows us to identify the relative importance of multiple predictors.
- Allows us to test whether the association between e.g. a continuous predictor and the response differs between groups.

Multiple continuous predictor variables

A record executive wants to know whether airplay and attractiveness are also important predictors of album sales:

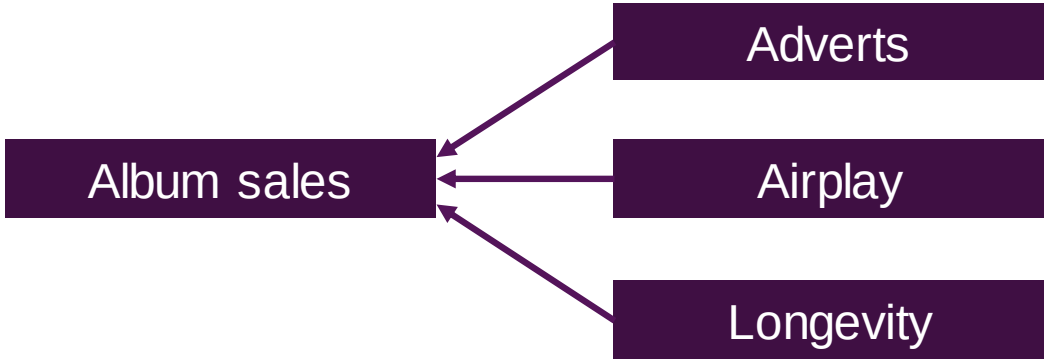
- **Adverts**: money spent on adverts (in thousands of pounds).
- **Airplay**: number of times the song was played on the radio in the week before the album was released.
- **Longevity** of the band: years since the band formed.



Multiple continuous predictor variables

The questions we can address with this model are:

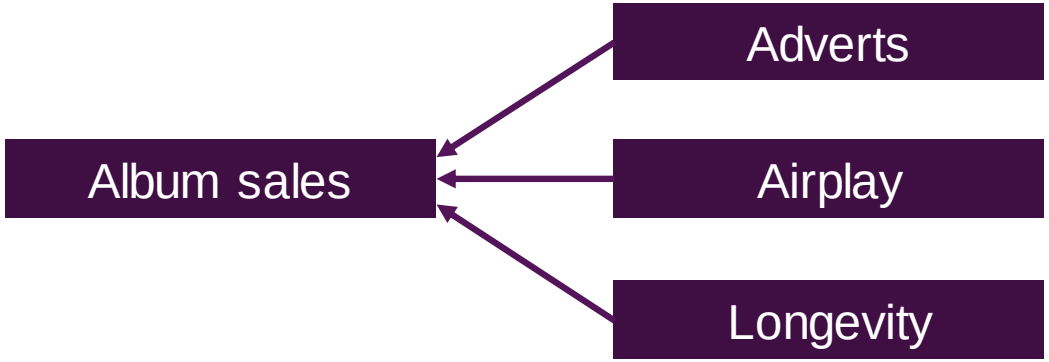
1. Which is the most important predictor of album sales: advert spending, airplay, or longevity?
2. Is the incorporation of airplay and longevity necessary (over and above advert spending?



Multiple continuous predictor variables

$$\text{Album sales} = \beta_0 + \beta_1 \text{adverts}_i + \beta_2 \text{airplay}_i + \beta_3 \text{longevity}_i + \varepsilon_i$$

- β_0 - constant value when adverts, airplay, and longevity are all zero,
- β_1 – slope between album sales and adverts after accounting for airplay and longevity,
- β_2 – slope between album sales and airplay after accounting for adverts and longevity.



Interpretation: multiple continuous predictors

	Estimate	Std error	t-value	p-value
(Intercept)	-26.61	17.35	-1.53	0.13
Adverts	0.08	0.01	12.26	<0.001
Airplay	3.37	0.28	12.12	<0.001
Longevity	11.09	2.44	4.55	<0.001
R^2=0.66				

As spending on adverts increases by one unit, album sales increase by 0.08 units when airplay and longevity are held constant.

Interpretation: continuous predictors

	Estimate	Std error	t-value	p-value
(Intercept)	-26.61	17.35	-1.53	0.13
Adverts	0.08	0.01	12.26	<0.001
Airplay	3.37	0.28	12.12	<0.001
Longevity	11.09	2.44	4.55	<0.001
R^2=0.66				

When only adverts was considered last week, the R^2 was 0.33. How does this model compare?

Adjusted R^2

- Adjusted R^2 compares the descriptive power of regression models.
- Every predictor added to a model increases the R^2 value.
- So, a model that includes several predictors will return higher R^2 values and may seem to provide a better fit. However, this result may be due to including more terms.
- The adjusted R^2 compensates for the addition of predictor variables.

Which predictor(s) is/are most important?

	Estimate	Std error	t-value	p-value
(Intercept)	-26.61	17.35	-1.53	0.13
Adverts	0.08	0.01	12.26	<0.001
Airplay	3.37	0.28	12.12	<0.001
Longevity	11.09	2.44	4.55	<0.001
R^2=0.66				

Which predictor(s) is/are most important?

- Coefficients (*b*) are interpreted on the original measurement scale.
- This is useful for telling us by how much we can expect a unit change in the outcome given a unit change in the predictor.
- However, this means *b* values cannot be directly compared with each other because they are on different measurement scales.

Compare the t-values…

	Estimate	Std error	t-value	p-value
(Intercept)	-26.61	17.35	-1.53	0.13
Adverts	0.08	0.01	12.26	<0.001
Airplay	3.37	0.28	12.12	<0.001
Longevity	11.09	2.44	4.55	<0.001
R^2=0.66				

Or, standardise regression coefficients (see lecture notes)

DIAGNOSTICS

Residuals

Recall:

$$\mathrm{e_i} = y_i - \mathrm{b}_0 - b_1 x_i$$

residuals = outcome – predicted outcome

The error (noise) accounts for the discrepancy between the data point and the mean response.

Residuals are in fact estimates of the error terms.

Residuals

$$\text{residuals} = \text{outcome} - \text{predicted outcome}$$

In regression, there are some **assumptions** about the error.

All the analysis and interpretations are valid as long as the error meets those conditions.

Thus, the assumptions about the estimates of the errors (i.e. residuals) need to be checked.

Assumptions of Residuals

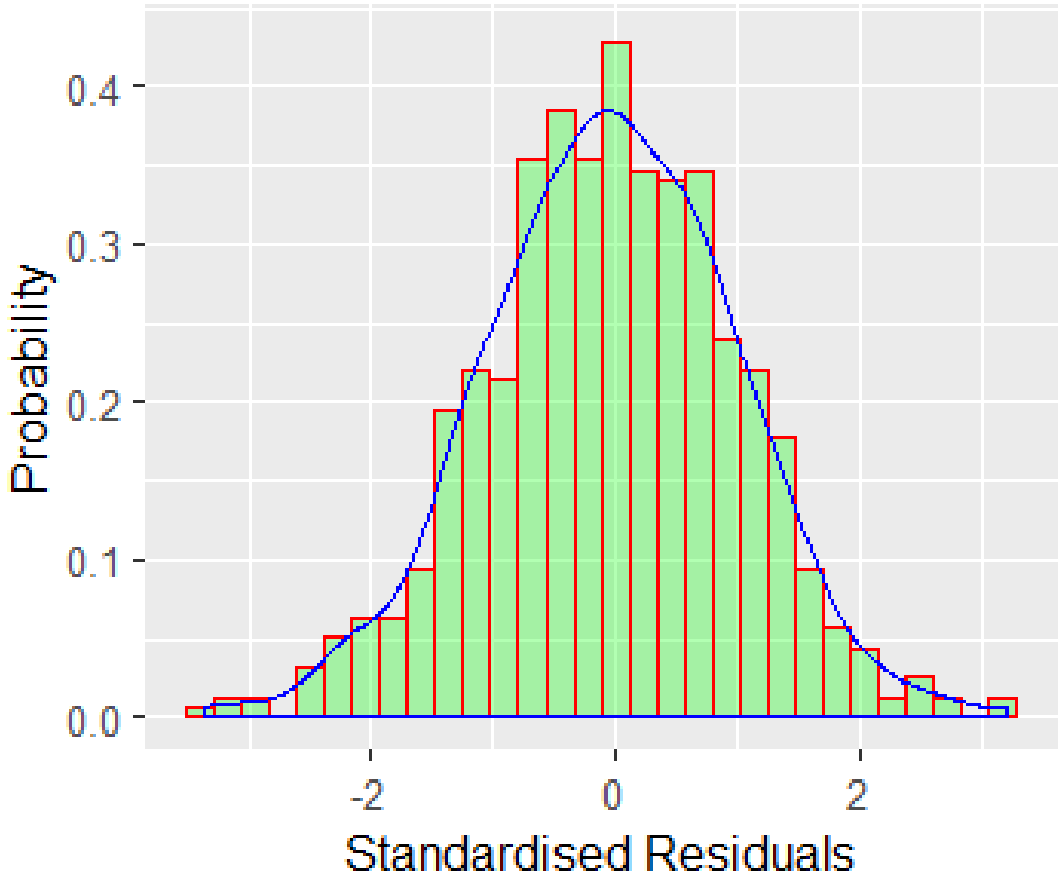
Normality of errors, recall:

Q-Q plot, histogram, normality test.

Should look like a bell

Formal tests:

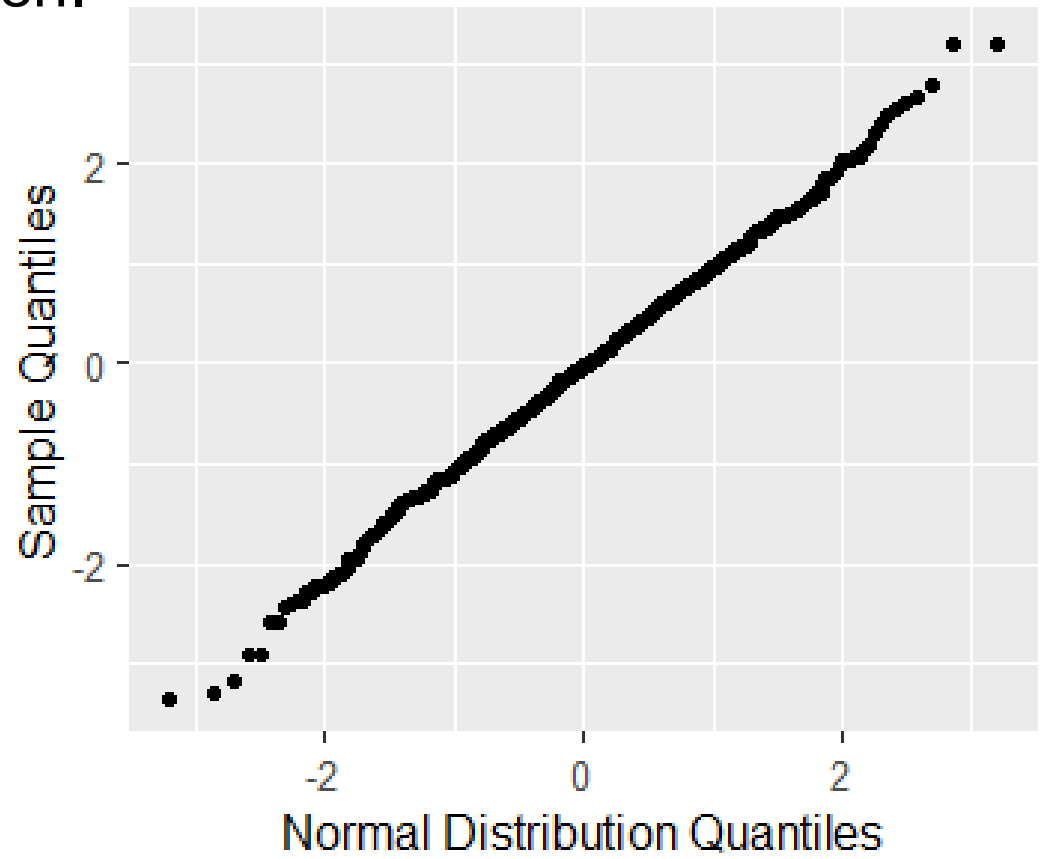
Shapiro-Wilk or Kolmogorov-Smirnov



Assumptions of Residuals

Normality of errors, recall: Q-Q plot, histogram, normality test.

Straighter line = closer to normal distribution.



It can also be used to identify outliers.

Standardised residuals with absolute value > 3 can be identified as outliers.

Assumptions of Residuals

Outliers and residuals; influential cases:

- Standardised residuals with absolute values of larger than 3 can be identified as outliers.

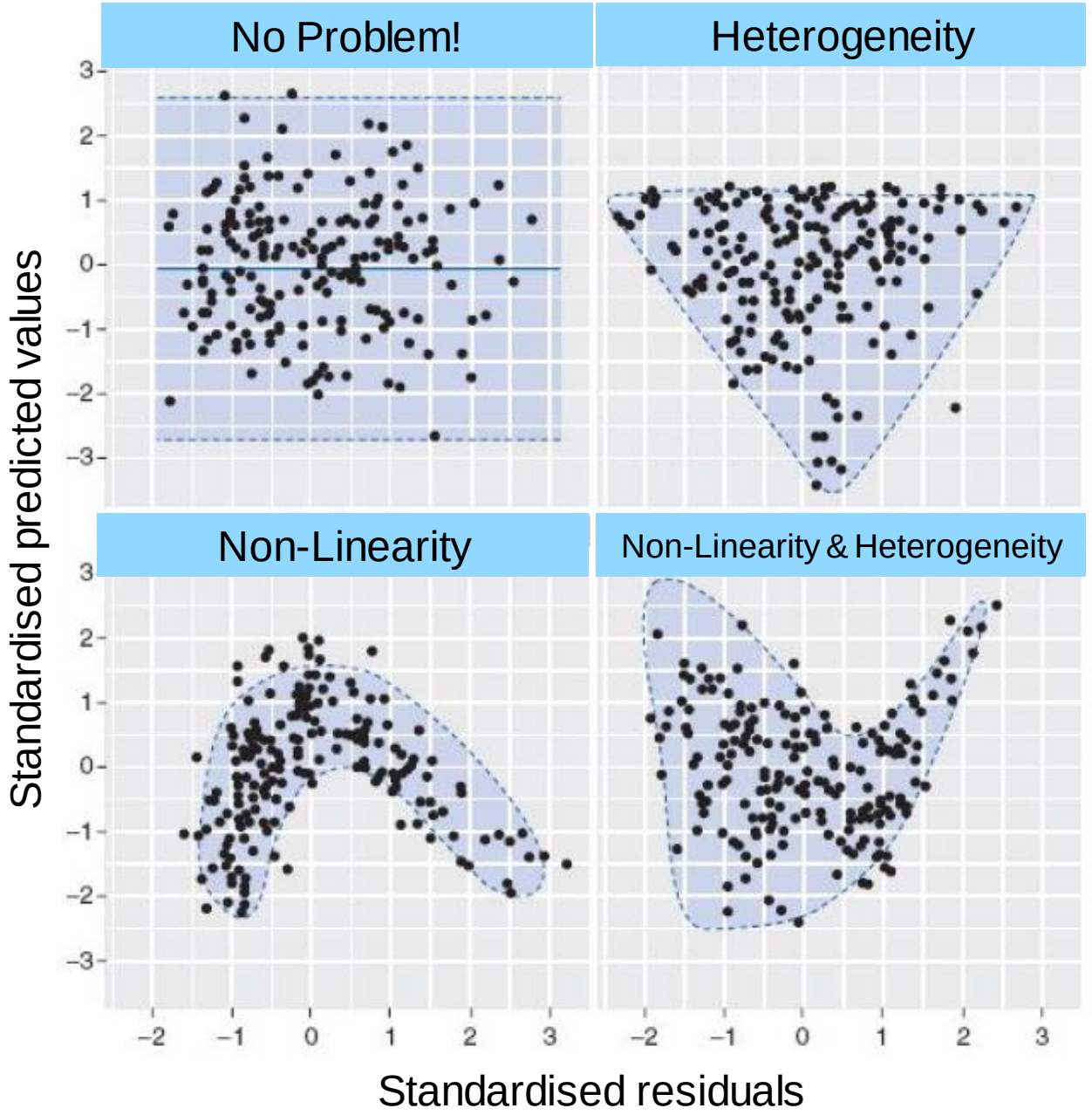
- Cases that are outlying do not mean they are **influential.**

  Cook's distance is a measure of the overall influence of a case on the model; values greater than 1 may be cause for concern.

Assumptions of Residuals

Homogeneity of errors:

scatterplot of fitted values (or outcome) against residuals.



Durham University

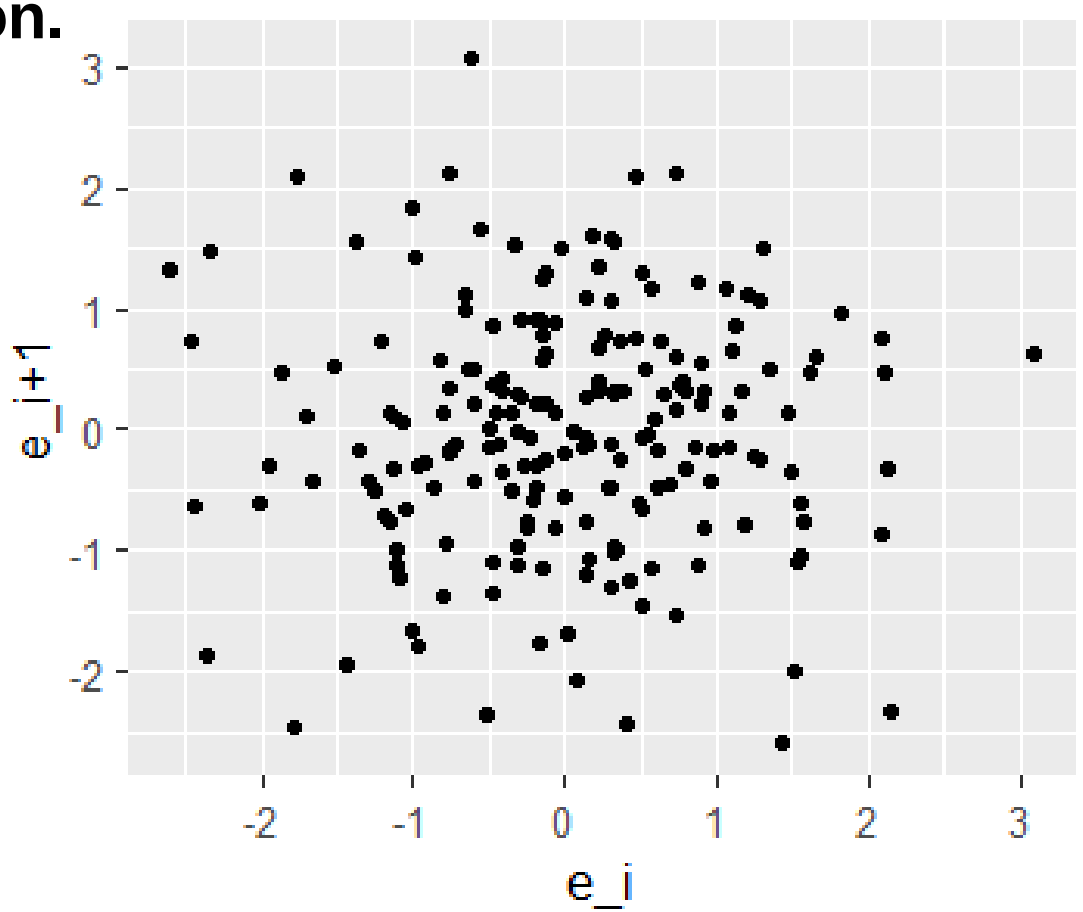
Assumptions of Residuals

Independent errors: for any two outcomes the corresponding residual terms should be uncorrelated (or independent).

To put it another way: lack of **autocorrelation.**

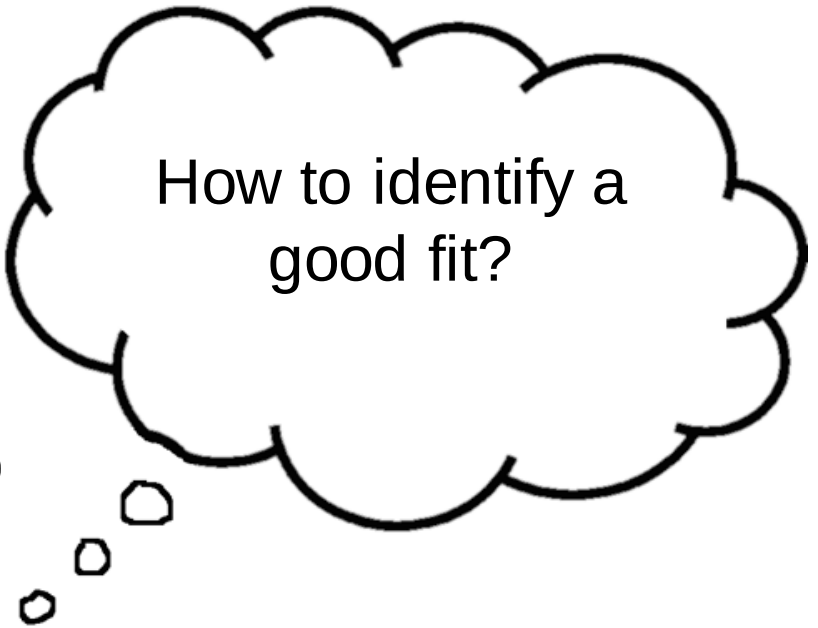
Hope to see no pattern.

Could use **Durbin–Watson (DW) test,** which tests for serial correlation between errors. Specifically, it tests whether adjacent residuals are correlated.



GODNESS OF FIT

Goodness of Fit



Compare two models through the F-test.

Reduced model (p1 predictors) versus full model (p2 predictors)

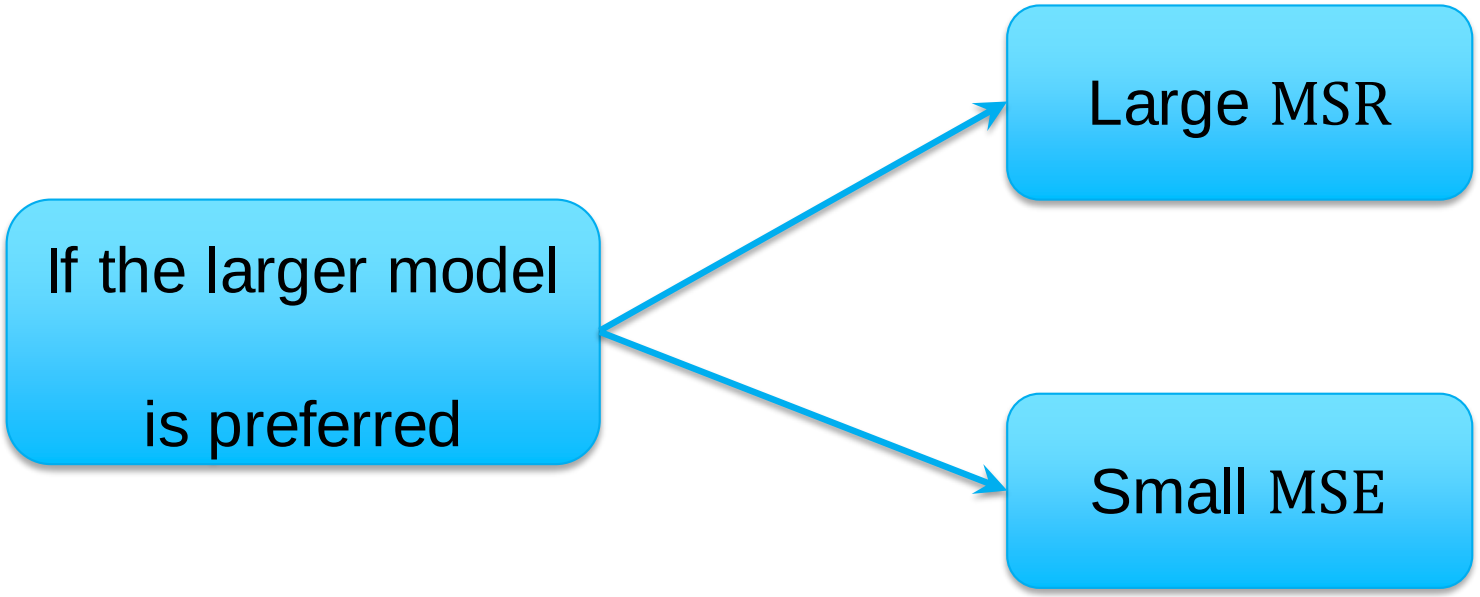
Calculate MSR=SSE(R) - SSE(F) / (p2-p1), MSE=SSE(F)/(n-p2)

The F statistic is F*=MSR / MSE

What it does:

- Tests the **null hypothesis that the additional predictors in the full model have zero coefficients** versus the alternative that at least one coefficient is non-zero
- If SSE(F) is close to SSE(R), it makes sense to use the simpler model.
- If SSE(F) is substantially smaller than SSE(R), it makes sense to use the full model.

Goodness of Fit



A good model should have a large F-ratio (greater than 1 at least)

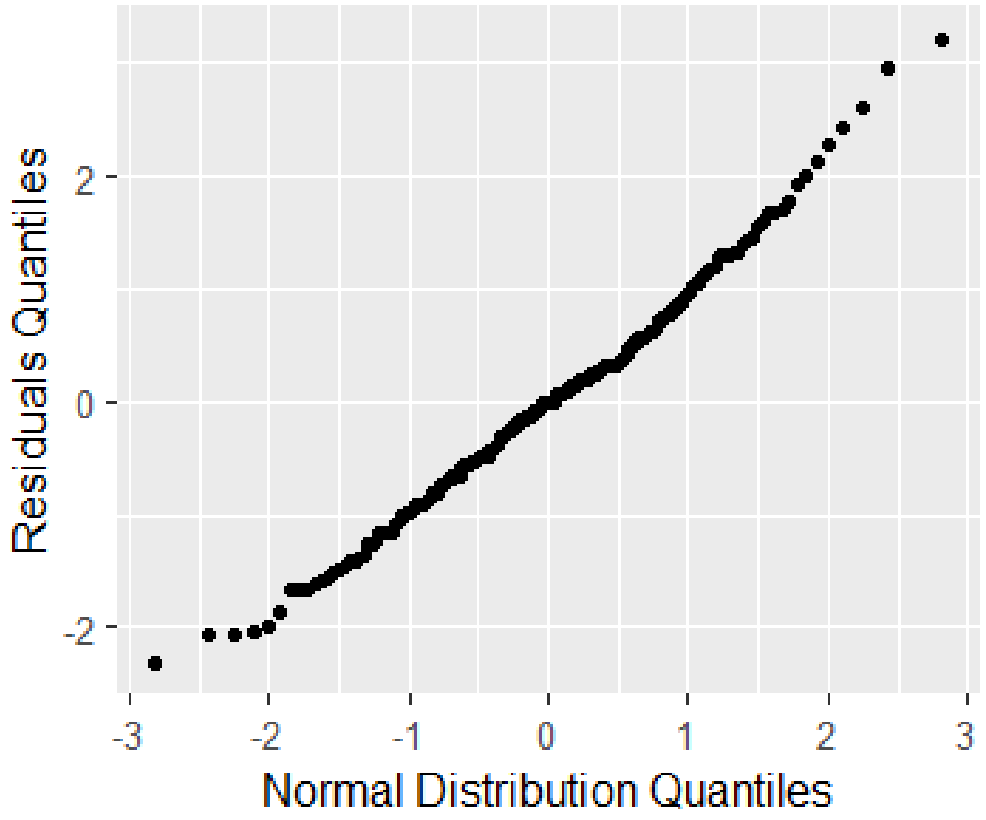
Case Study 1

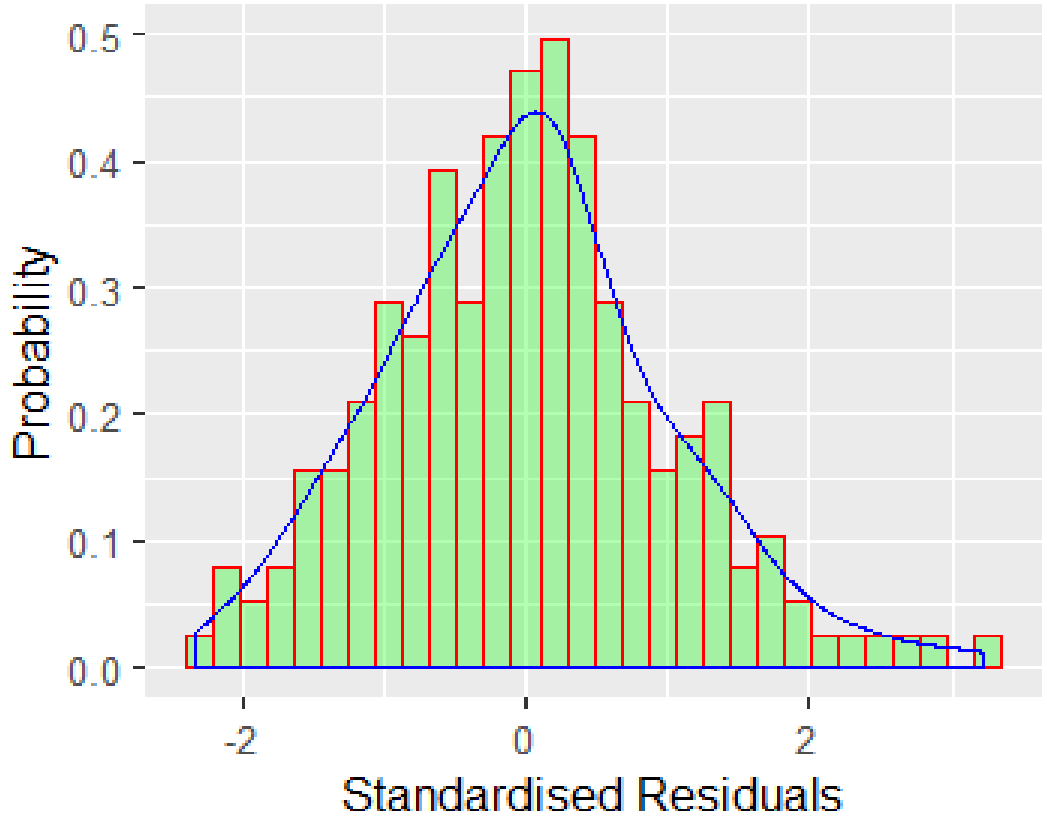
Recall the sales and money spent in adverts example in simple linear regression:
Record sales of albums (per thousands) is the outcome and the amount spent on advertisements (per thousand pounds) is the predictor.

	Coefs	Standard error	t value	Pr(>|t|)	95% CI	
					LB	UB
b_0	134.10	7.54	17.80	0.00***	119.28	149.00
b_1	0.096	0.01	9.98	0.00***	0.08	0.12

What do the residuals to look like?

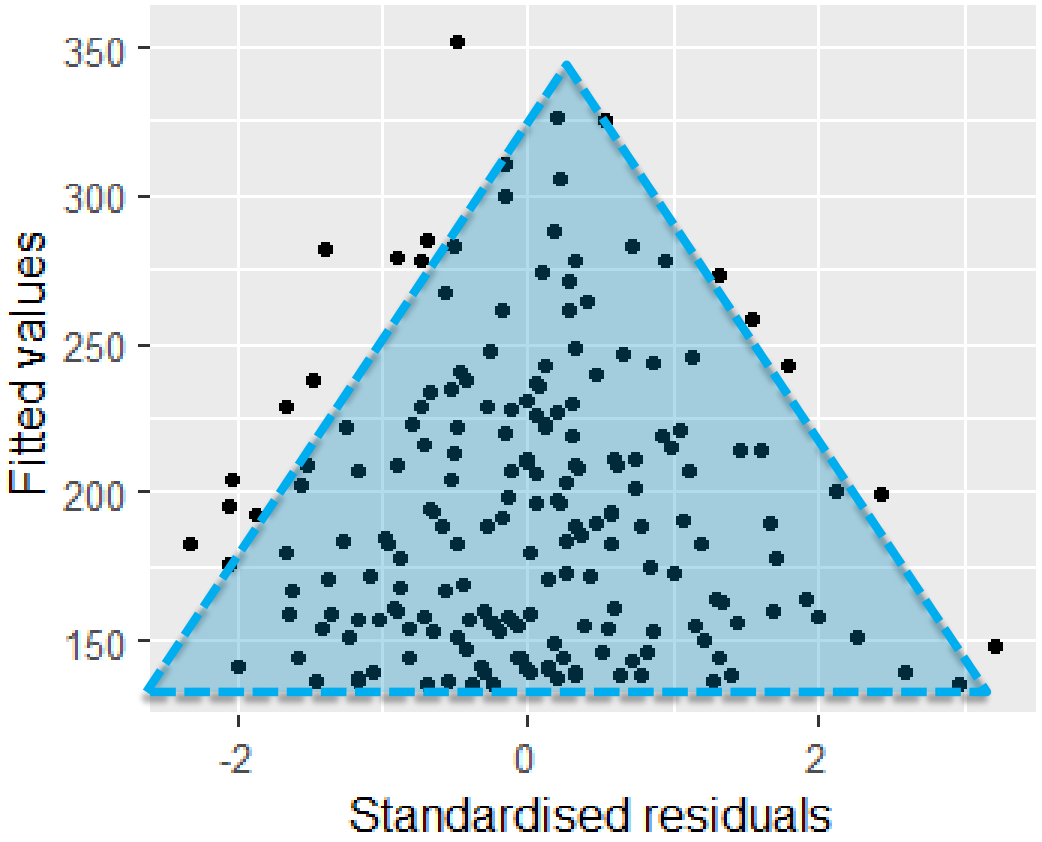
Case Study 1

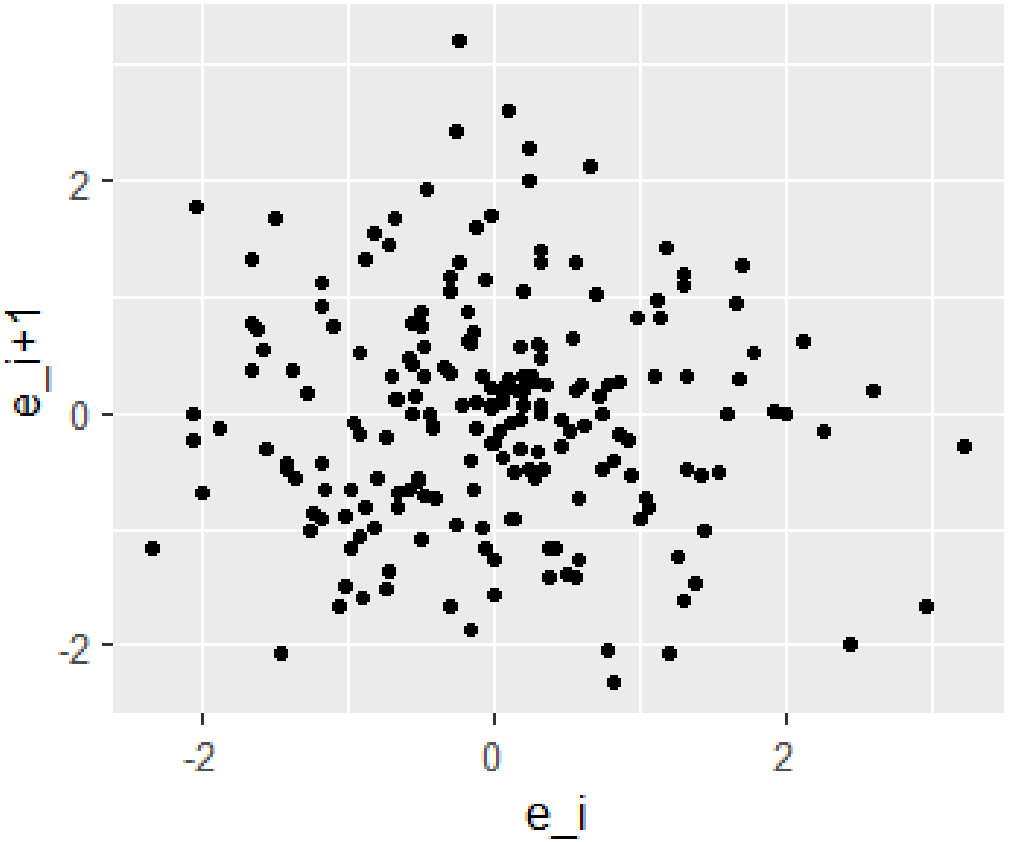




Durham University

Case Study 1


Fitted values
350
300
250
200
150
-2
0
2
Standardised residuals

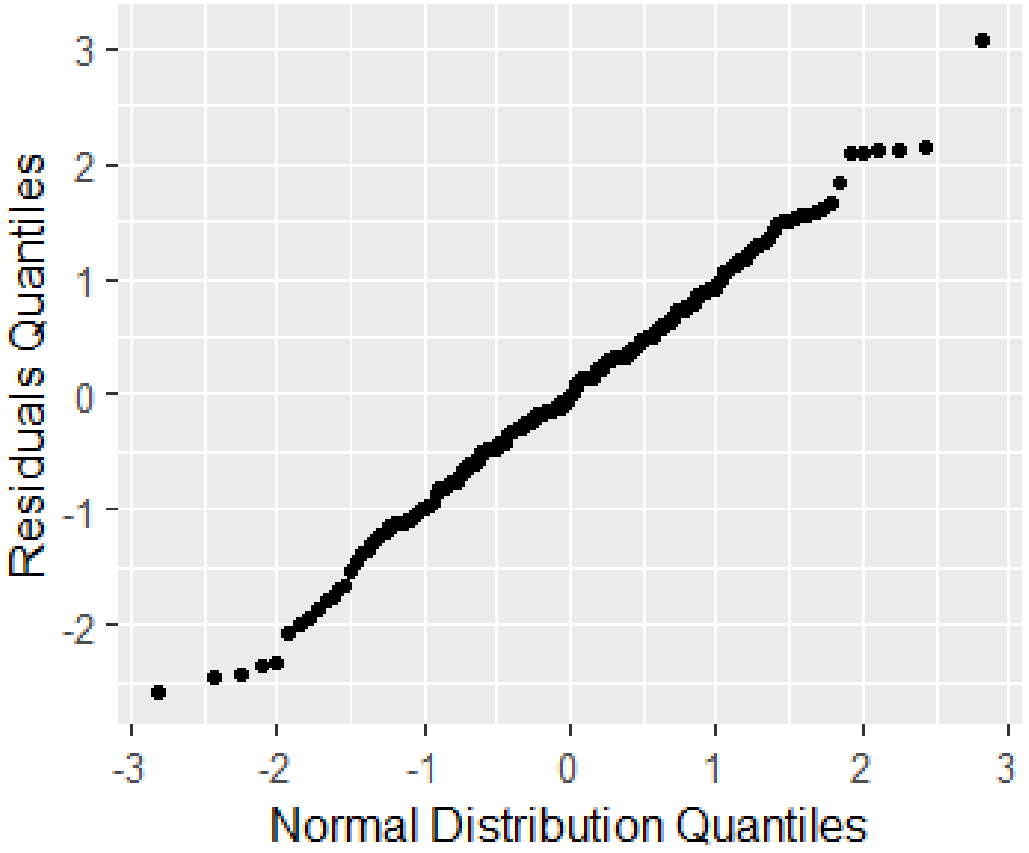

e_i+1
2
0
-2
-2
0
2
e_i

Case Study 2

Recall multiple linear regression with **Airplay** and **Longevity** added
Airplay: number of times the song was played on the radio in the week before the album was released.
Longevity of the band: years since band was formed.

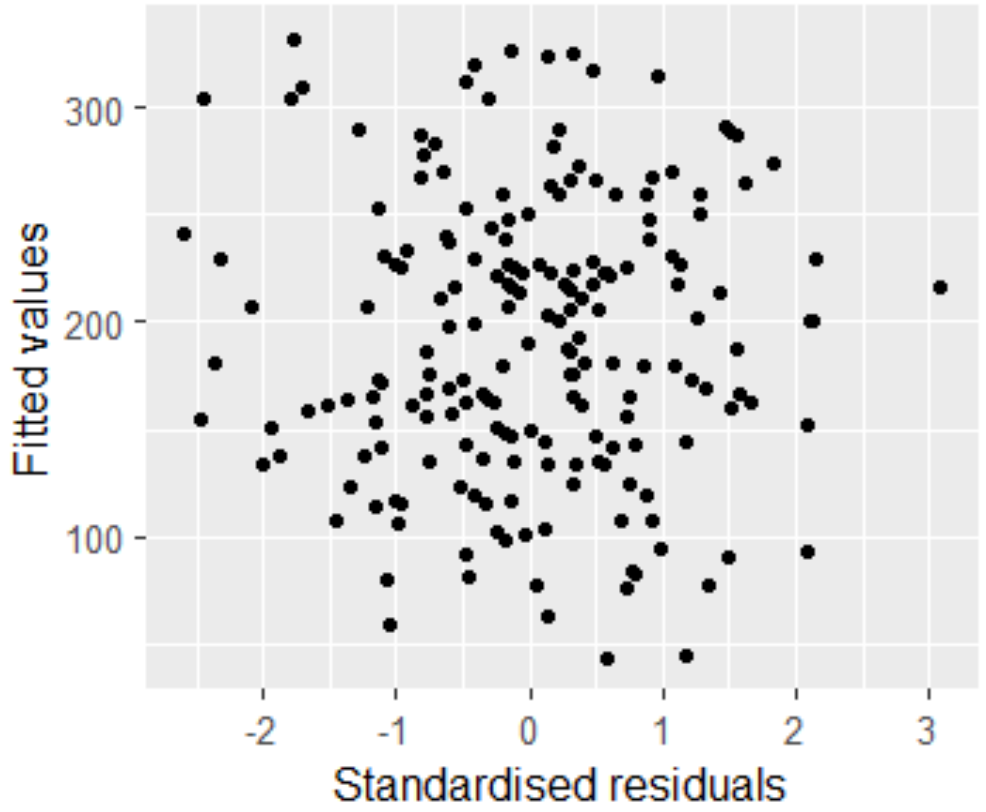
	Estimate	Std error	t-value	p-value
(Intercept)	-26.61	17.35	-1.53	0.13
Adverts	0.08	0.01	12.26	<0.001
Airplay	3.37	0.28	12.12	<0.001
Longevity	11.09	2.44	4.55	<0.001
R^2=0.66				

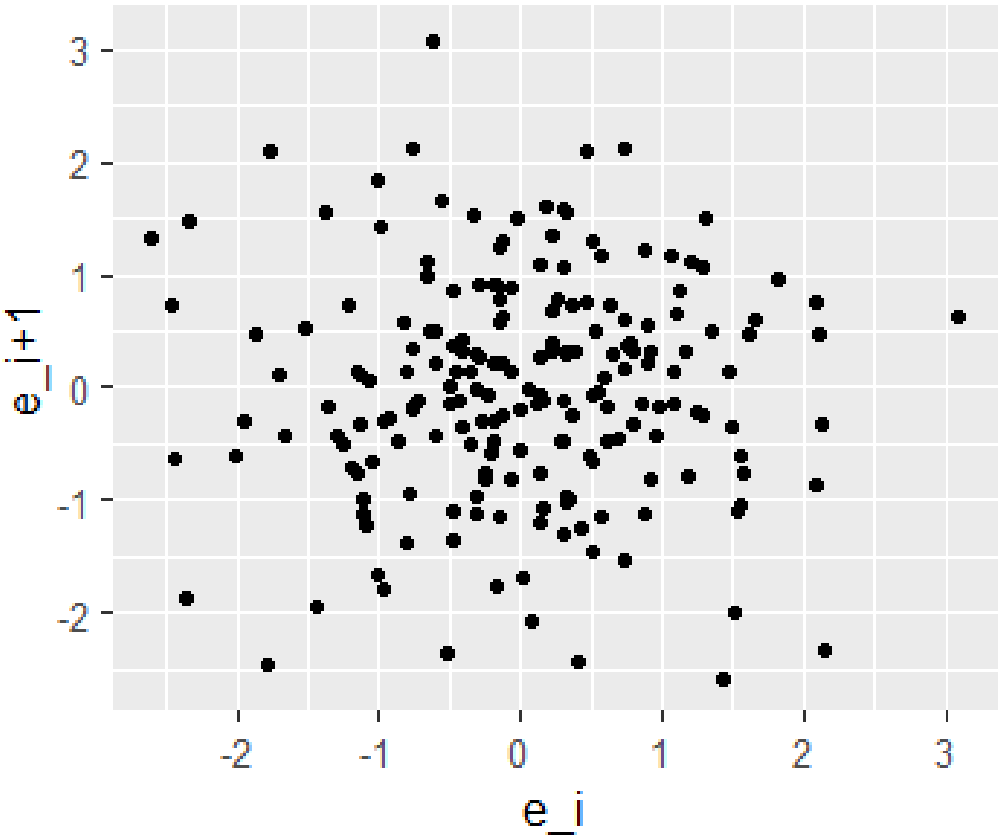
Case Study 2



Durham University

Case Study 2


Fitted values
300
200
100
-2
-1
0
1
2
3
Standardised residuals


e_i+1
3
2
1
0
-1
-2
-2
-1
0
1
2
3
e_i

Case Study 2

VIF:

adverts	airplay	longevity
1.014593	1.042504	1.038455

What do we conclude?

Model	**predictors**	**SSE**	**F***
Simple regression with only adverts	1	866645	
Multiple regression	3	442113	94.6

Which model is better and why?

Assumptions are met, large F* (much larger than the upper 5% critical value of 3.9)

Thanks for your attention!

Durham University
Research Methods Centre