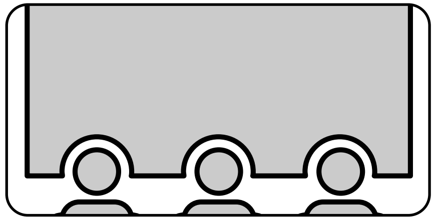
Assessing the Model

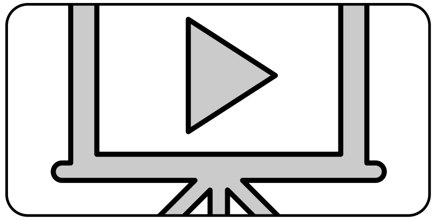
CHRIS SUTTON, JANUARY 2024

Last week



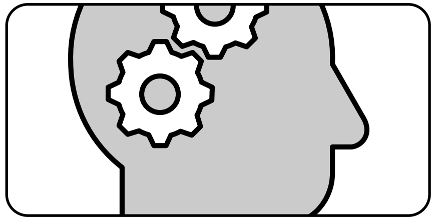
Lectures introducing the module and the model

- Simple linear regression model
- Estimation of betas by Least Squares



6 short video lectures to watch

- 1 & 2 on Principles of Statistical Modelling
- 3 – 6 on the Simple linear regression model



Your ideas

- 3 areas you would be interested in modelling
- 2 variables and a draft modelling question for one of them

This week

Lectures on assessing the model

- Fitted values and residuals
- ANOVA

More short video lectures to watch

- Properties of the estimators [not covered in the campus lectures]
- Details on assessing the model

Your tasks

- Find your own data set
- Upload it to QM Plus before next Monday

Topics in first few weeks of Statistical Modelling

1. Principles of statistical modelling
2. The Simple Linear Regression Model
3. Least Squares estimation
4. Properties of estimators
5. Assessing the model
6. Inference about the model parameters
7. Matrix approaches to simple linear regression
8. Multiple Linear Regression Models

From last week

The Simple (Normal) Linear Regression Model can be written

- $y_i \sim N(\mu_i, \sigma^2)$ where $\mu_i = \beta_0 + \beta_1 x_i$
- $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where the ε_i are iid $\varepsilon_i \sim N(0, \sigma^2)$

From last week (continued)

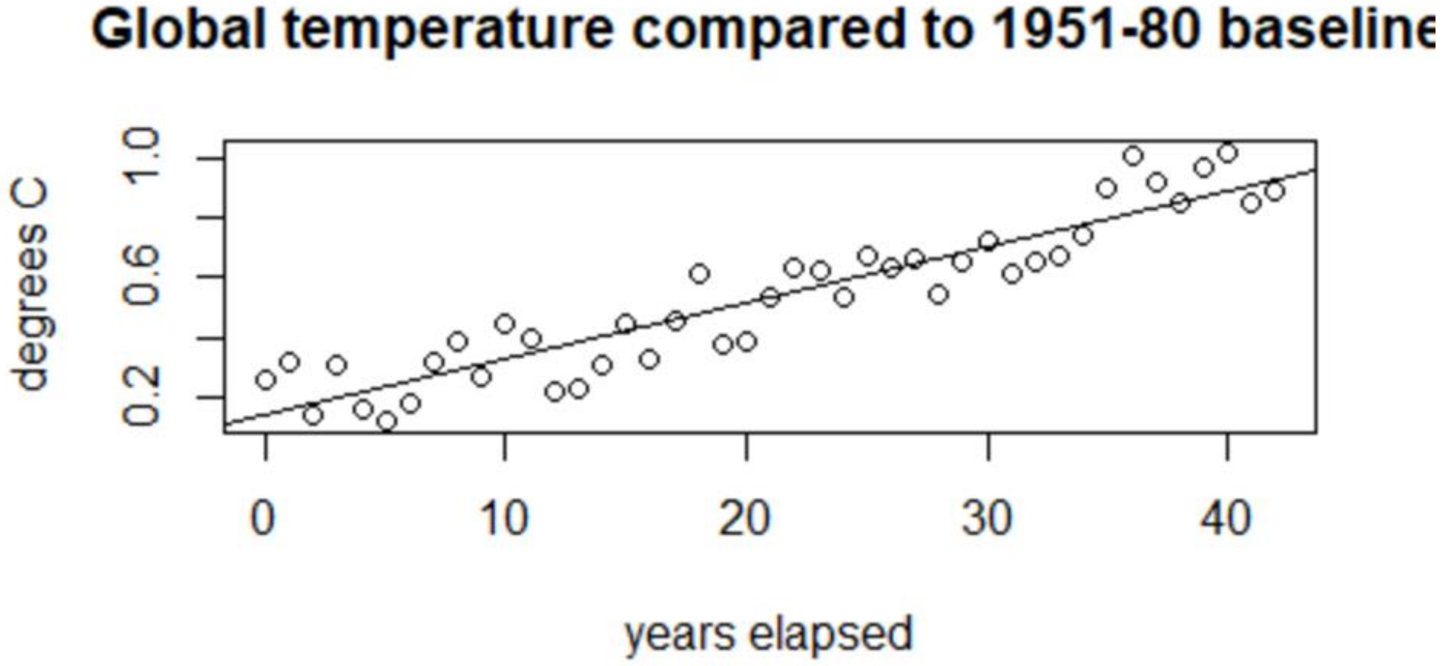
By Least Squares Estimation our simple regression model parameter estimators are

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\,\bar{x}$$

and

$$\widehat{\beta_1} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$ often written in shorthand $$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

Model fitted to global temperature data



R output

```
Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-0.153268 -0.080700 -0.003953  0.080943  0.193511

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.138404   0.028516   4.854 1.79e-05
x           0.018836   0.001169  16.112  < 2e-16

(Intercept) ***
x           ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.09513 on 41 degrees of freedom
Multiple R-squared:  0.8636,  Adjusted R-squared:  0.8603
F-statistic: 259.6 on 1 and 41 DF,  p-value: < 2.2e-16
```

Task for this week
needed for the coursework in week 4

- Find your own data set that could be used for a simple linear regression model
- Link it to one of the three things you said you would like to model
- Observations in (x, y) form with explanatory and response variables
  - Do not make x measure of time in years (2019, 2020, 2021, 2022, 2023 …)
- Don't make it too large: 10 – 50 observations
- Save the data in Excel or csv file and upload that file to the submission point in the week 2 area of the module QM Plus site
- Write down why you chose this data
- There are no prizes for the data but we will use your data in the coming weeks and doing this now will make your first assessed coursework much much easier

Topics in first few weeks of Statistical Modelling

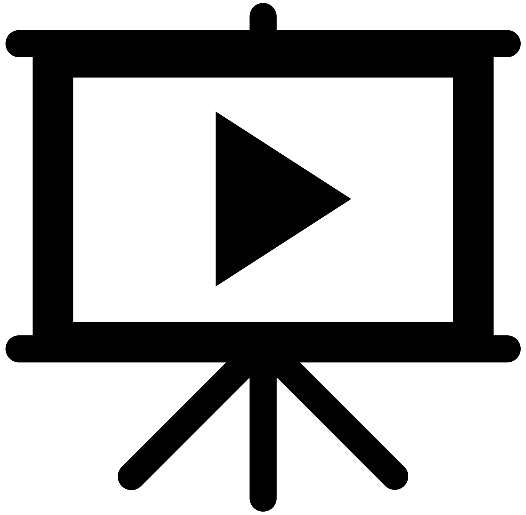
1. Principles of statistical modelling
2. The Simple Linear Regression Model
3. Least Squares estimation
4. Properties of estimators
5. Assessing the model
6. Inference about the model parameters
7. Matrix approaches to simple linear regression
8. Multiple Linear Regression Models

Properties of the estimators

The distribution, mean and variance of

$\widehat{\beta_0}$ and $\widehat{\beta_1}$

Watch the short video lectures on QM Plus before Thursday



How can we evaluate whether our regression model is any good or whether the assumptions we made were reasonable?

Our fitted model

The model we have is

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

By least squares estimation we can find estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$

We can use these to fit the model at *i* = 1, 2, …, *n*

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

These are the *fitted values* and will sit on the *regression line* at the *n* different x_i values

Fitted and Observed values

Now the fitted values

- $\hat{y}_1, \hat{y}_2, \ldots \hat{y}_n$
- on the regression line

are different to the original observed values

- $y_1, y_2, \ldots y_n$
- which do not all sit on the line

The differences between them are the *residuals*

Residuals

The *residuals* (sometimes called the *crude residuals*) are e_i

Residual = Observed Value – Fitted Value

$$e_i = y_i - \hat{y}_i$$

These residuals are the estimates of the random error component ε_i in the original model specification

These residuals are going to be a key tool for assessing how well our linear regression model describes the original observed data

Sum of the residuals

$e_i = y_i - \hat{y}_i$

$= y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$

$= y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})$ from the definition of $\hat{\beta}_0$ under least squares estimation

Therefore

$$\sum_{i=0}^{n} e_i = \sum_{i=0}^{n} (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=0}^{n} (x_i - \bar{x}) = 0 - 0$$

The sum of the residuals e_i is zero

Sum of squares of errors

The sum of the residuals e_i is zero (we'll prove this in a video lecture)

So we work with the squares of residuals, ${e_i}^2$ instead

When we were estimating $\widehat{\beta_0}$ and $\widehat{\beta_1}$ by least squares last week, we sought to minimise a function *S* of β_0 and β_1

$$S(\beta_0, \beta_1) = \sum_{i=0}^{n} \varepsilon_i^2$$

If we evaluate this function *S* with the *n* data points (x_i, y_i) and the estimates $\widehat{\beta_0}$ and $\widehat{\beta_1}$

This quantity is called the ***Residual Sum of Squares*** denoted SS_E

Residual Sum of Squares

$$SS_E = \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n}(y_i - \hat{y}_i)^2$$

For a particular data set:

- SS_E is the minimum value of $S(\beta_0, \beta_1)$
- it is one measure of how well the model fits the data
- it describes one of the sources of variability of the y_i around their mean $\bar{y}$

Total Sum of Squares

The total variance of y_i around their mean $\bar{y}$ is the ***Total Sum of Squares*** *or* SS_T

$$SS_T = \sum_{i=0}^{n} (y_i - \bar{y})^2$$

In the simple linear regression model

Total Sum of Squares = Regression Sum of Squares + Residual Sum of Squares

$SS_T = SS_R + SS_E$ This equation is called the *Analysis of Variance Identity*

Regression Sum of Squares

SS_R is the ***Regression Sum of Squares***

sometimes called the *Model Fit Sum of Squares*

$$SS_R = \sum_{i=0}^{n} (\hat{y}_i - \bar{y})^2$$

We will prove that $SS_T = SS_R + SS_E$ in one of the short video lectures

Total Sum of Squares

The Total Sum of Squares SS_T is made up of:

- the Regression Sum of Squares SS_R
  - the variability in the y_i around their mean $\bar{y}$
  - which is accounted for by the fitted model
- the Residual Sum of Squares SS_E
  - the variability in the y_i
  - accounted for by the difference between observed and fitted values

This view can be presented in an ***Analysis of Variance Table*** or ***ANOVA Table***

The ANOVA Table

CHRIS SUTTON, FEBRUARY 2023

Source of variation	d.f.	SS	MS	VR
Regression	$v_R = 1$	SS_R	$MS_R = \frac{SS_R}{v_R}$	$F = \frac{MS_R}{MS_E}$
Residual	$v_E = \mathrm{n} - 2$	SS_E	$MS_E = \frac{SS_E}{v_E}$	
Total	$v_T = \mathrm{n} - 1$	SS_T		

ANOVA is a way of presenting the variability found in our model

The ANOVA table

The variability in the y_i is accounted for by 4 quantities

Each is a column of the ANOVA table:

- degrees of freedom (d.f.)
- Sum of Squares (SS)
- Mean Squares (MS)
- Variance Ration (VR)

we have already defined *SS* and will now consider the other three

Degrees of freedom

The *degrees of freedom* (d.f.) are the number of independent observations (out of the *n* in total) that are used in the estimation of a parameter

E.g. if we have *n* observations $y_1, y_2, \ldots, y_n$ and fix their mean or their sum then *n-1* of the observations are free to vary but one of them will need to be a certain value in order to get to that fixed mean (or sum).

Degrees of freedom in ANOVA

SS_T

n observations, one d.f. taken up with calculating $\bar{y}$
so Total Sum of Squares has n – 1 d.f.

SS_E

one d.f. taken up with finding $\hat{\beta}_0$ and one with $\hat{\beta}_1$
so Residual Sum of Squares has n – 2 d.f.

SS_R

as $SS_R = SS_T - SS_E$
Regression Sum of Squares has (n – 1) – (n – 2) = 1 d.f.

Mean Squares

Mean Squares (MS) is a measure of the average variation for Residuals and for Regression

Found by dividing the relevant Sum of Squares (SS) by its degrees of freedom

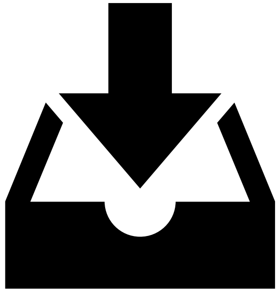
$$MS_R = \frac{SS_R}{v_R} \quad \text{and} \quad MS_E = \frac{SS_E}{v_E}$$

Variance Ratio

The Variance Ratio measures the variance explained by the model fit relative to that explained by the residuals. We usually denote this *F*.

$$F = \frac{MS_R}{MS_E}$$

Task for this week
needed for the coursework in week 4



- Find your own data set that could be used for a simple linear regression model
- Link it to one of the three things you said you would like to model
- Observations in (x, y) form with explanatory and response variables
  - Do not make x measure of time in years (2019, 2020, 2021, 2022, 2023 …)
- Don't make it too large: 10 – 50 observations
- Save the data in Excel or csv file and upload that file to the submission point in the week 2 area of the module QM Plus site
- Write down why you chose this data
- There are no prizes for the data but we will use your data in the coming weeks and doing this now will make your first assessed coursework much much easier