MANCHESTER 1824
The University of Manchester

SeM-in-Rs

Regression in R and Mplus

February 25th 2020

Dr Patricio Troncoso

patricio.troncoso@manchester.ac.uk

@ptroncosoruiz

With thanks to Dr Margarita Panayiotou (for the Mplus part)

MIE
S E M

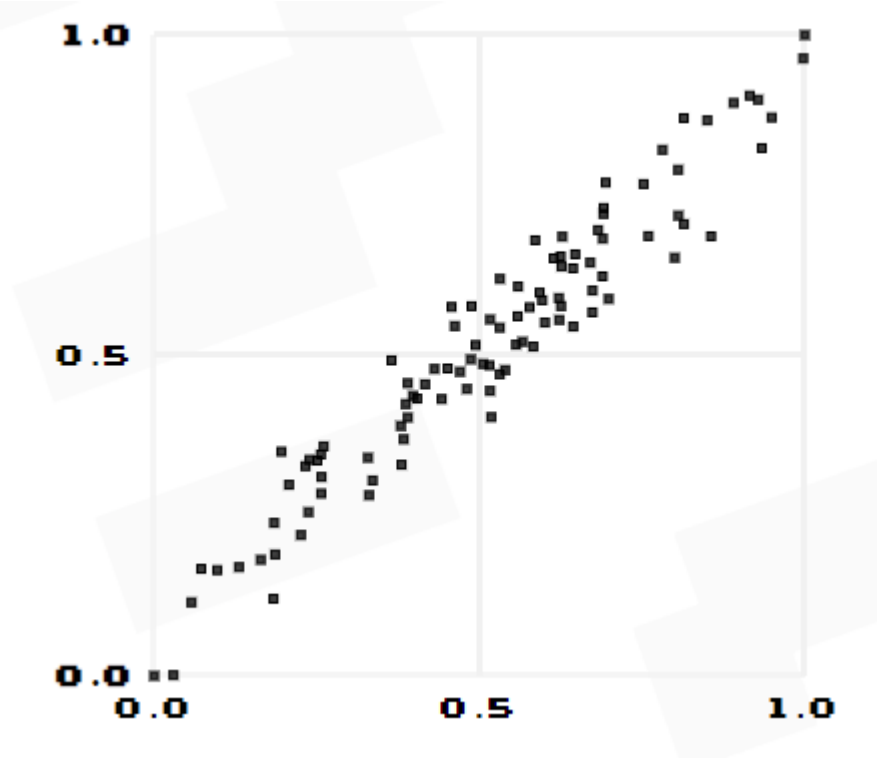
Plan for today

10:00-11:00	Correlation and linear regression
11:00-11:15	Break
11:15-12:00	Practical 4: Linear regression
12:00-12:30	Binary logistic regression
12:30-13:00	Practical 5: Logistic regression

Part One

CORRELATION

Relationship between two variables



The easiest way to see the relationship between two **continuous** variables is to plot the values of one of them against the other.

This is presupposing that you have a **theoretical** reason to think they are indeed related

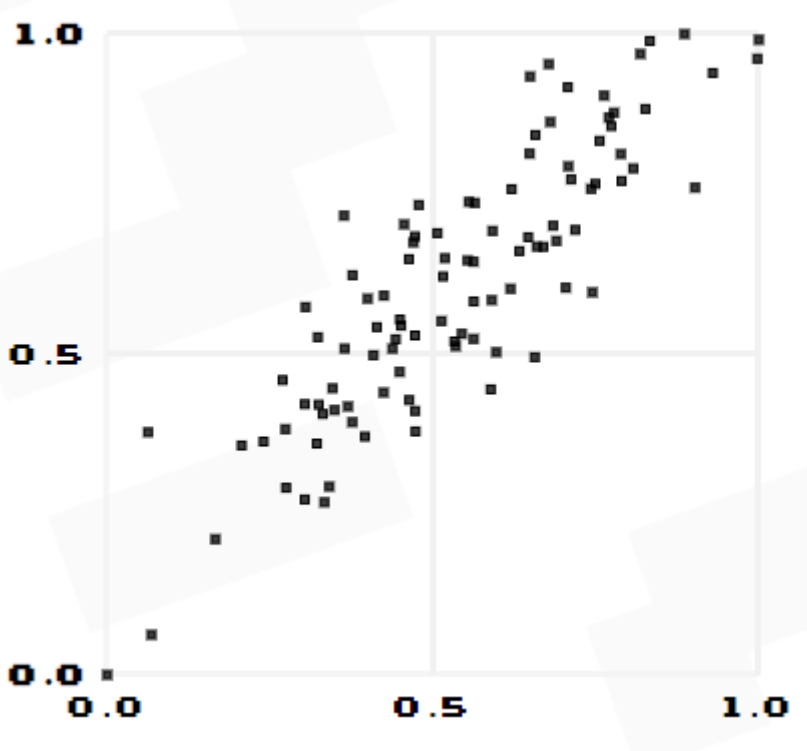
Hold this thought!

Image source: http://guessthecorrelation.com/

Pearson correlation (R)

- The typical measure of association between two continuous variables if the Pearson correlation.
  - It goes from -1 to 1
    - 1 means **perfect positive association**
    - 0 means **no association at all**
    - -1 means **perfect negative association**
  - The **strength** of the association is determined by how **far** the values is from **zero**

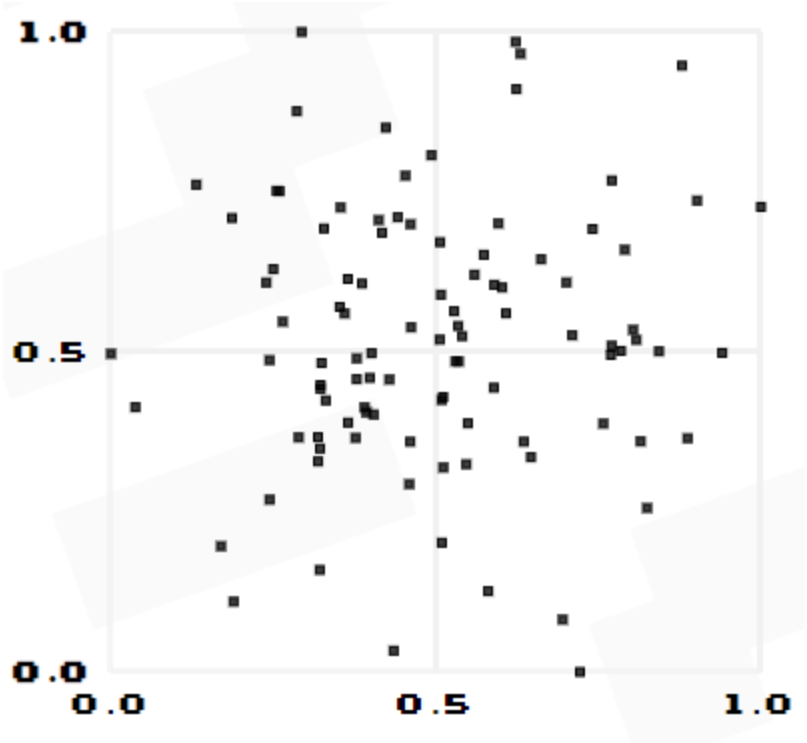
Types of relationships (1)



This is a **strong positive** association. The correlation here is 0.87.
The higher the values of X (horizontal axis); the higher the values of Y (vertical axis)

Image source: http://guessthecorrelation.com/

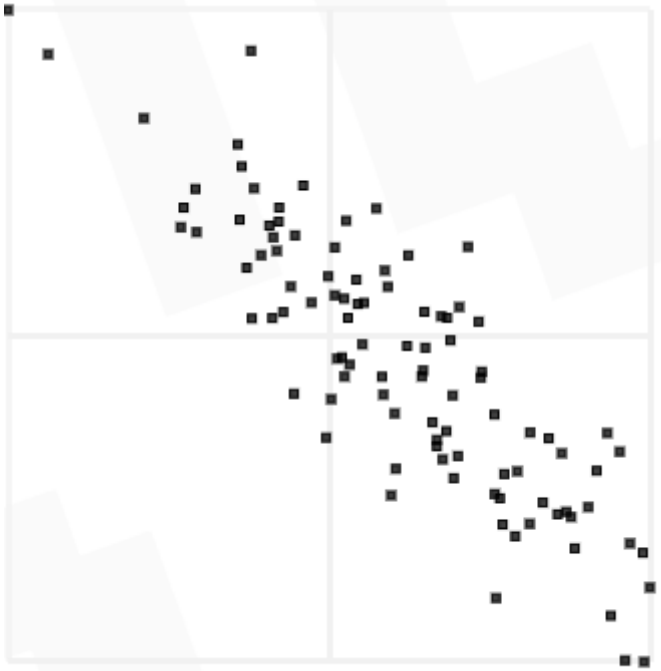
Types of relationships (2)



This is a **very weak** association. The correlation here is 0.06.
So close to zero that is probably meaningless

Image source: http://guessthecorrelation.com/

Types of relationships (3)



This is a **strong negative** association.
The higher the values of X (horizontal axis); the lower the values of Y (vertical axis)

Image source: http://guessthecorrelation.com/ (This is just image in slide 6 flipped)

Why is there a correlation?

- There might be a causal link:
  - X causes Y or vice versa
- Both variables are affected by another variable
  - We call this “confounding”
- Both variables measure the same thing
  - More on this in session 4 of the SEM<-in->Rs
- Last but not least…

Correlation doesn't mean…

- Causation
  - Trends over time can look as if there is correlation, but they can be a natural process.
    - Reading comprehension and age
- Low correlation (close to zero) does not mean there is no association
  - A "non-linear" relationship might exist!

Part Two

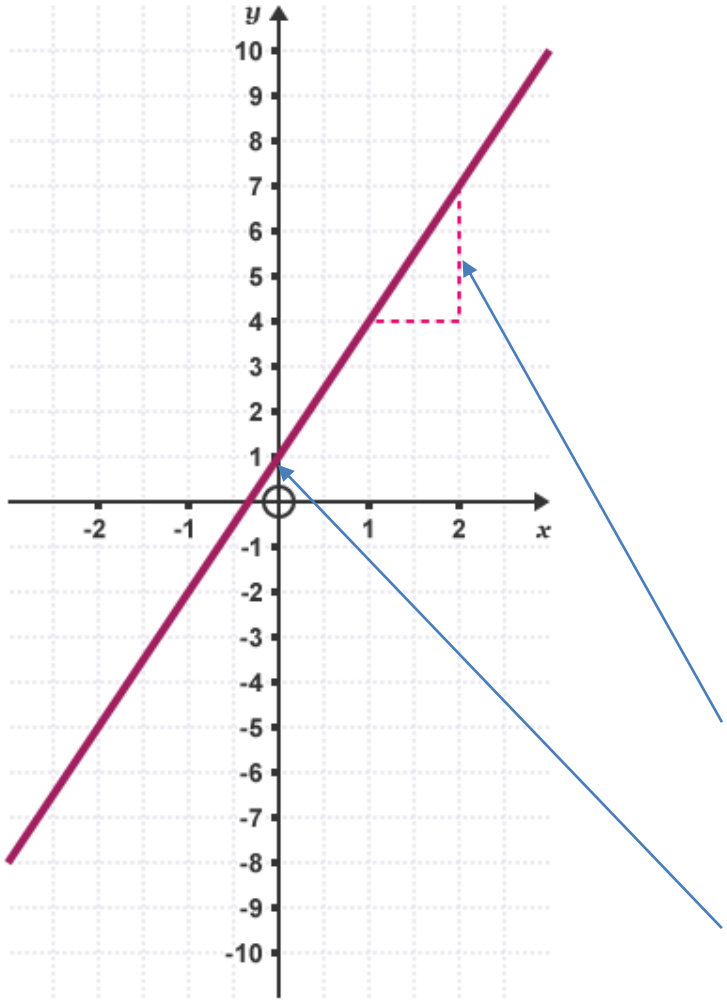
LINEAR REGRESSION

What is regression?

- It's the process by which we systematise the relationship between two (or more) variables.
- We look for the "best fit" to the data we have.
- We obtain an equation that describes how much our dependent variables changes as our independent variable changes.
  - Remember this from high school?

$$y = c + mx$$

The equation of the line



Linear regression owes its name to the fact that we're looking for the equation of the line that describes the best fit of the data

The equation of the line here is:

$$y = 1 + 3x$$

In words: to determine the value of y given x, we multiply x by 3 and add 1.

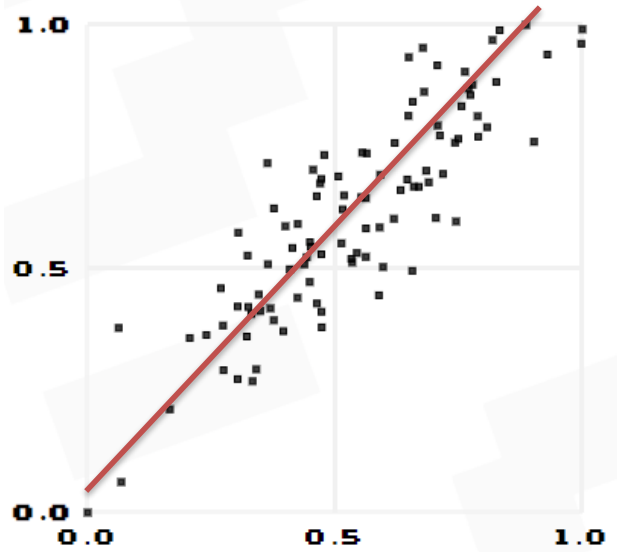
The gradient or slope is the increase in y for each one-unit increase in x. In this case, it's 3.

The intercept is the point at which the line cross the y axis. In this case, it's 1.

Source: https://www.bbc.co.uk/bitesize/guides/zy2xsg8/revision/4

In real life… (1)

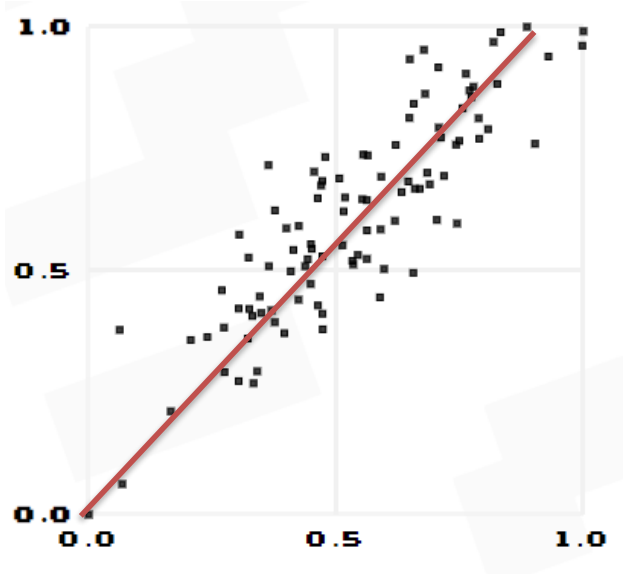
- The equation of line is a useful tool, but when we measure variables in reality, there's this is pesky little thing called "error"
- You plot two variables, as such:



- And you realise there's no way to fit a **straight** line that touches all the data points

Simple linear regression

- Simple linear regression is when you have one dependent variable “y” and only one independent variable “x”

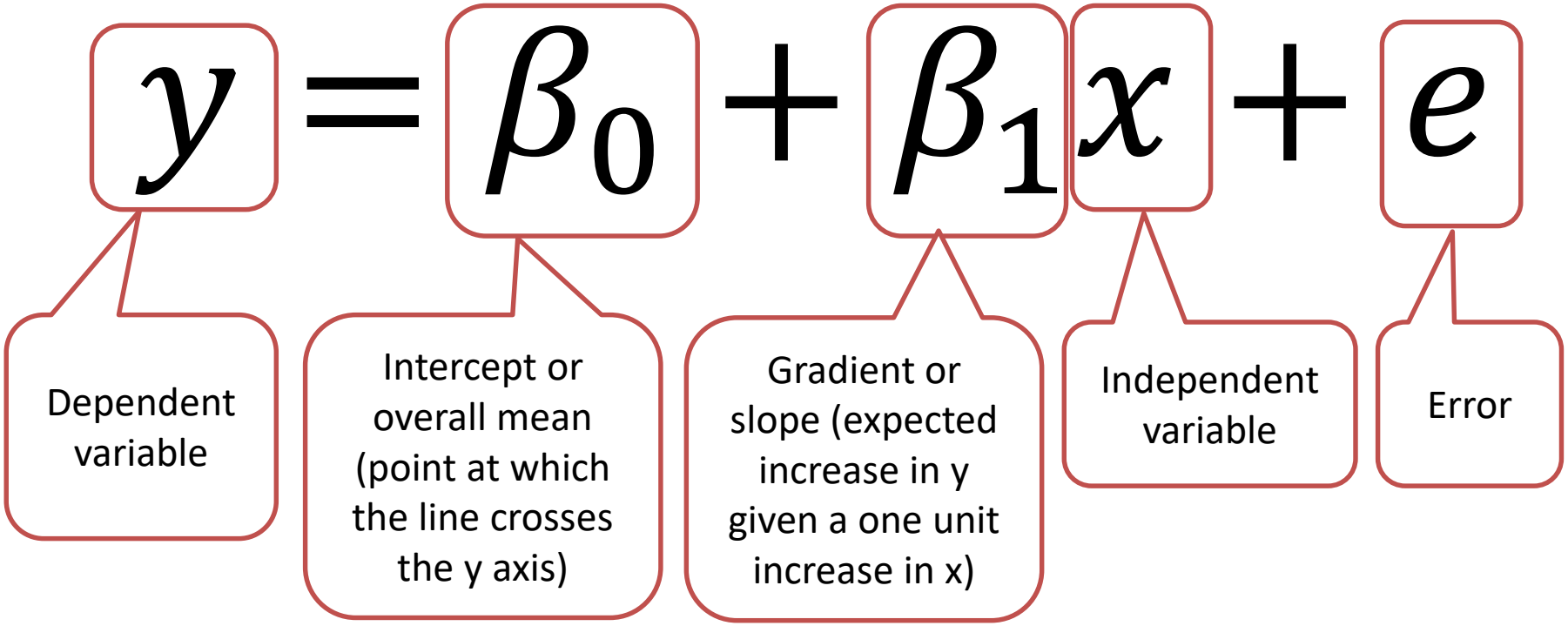


The equation of the line here takes this form:

$$y = \beta_0 + \beta_1 x + e$$

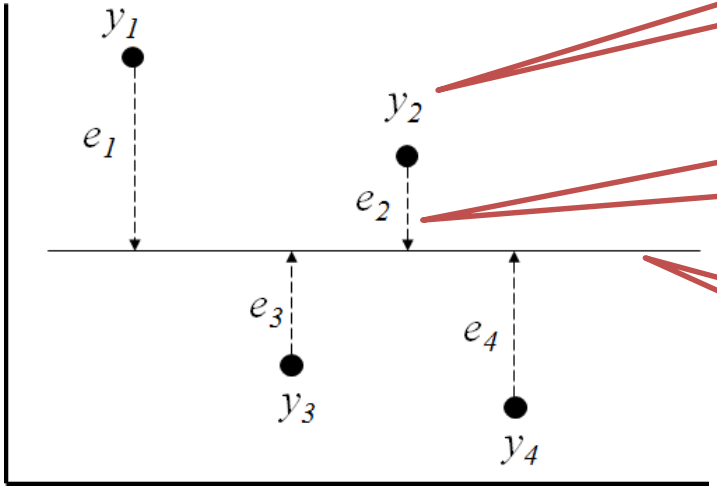
But what does all this mean?

Simple linear regression



The error (1)

- We saw before that linear regression **minimises the error**.
- But what is the error?



y2 is the observed value of y for individual 2

e2 is the distance between y2 and the regression line. These are also called “residuals”

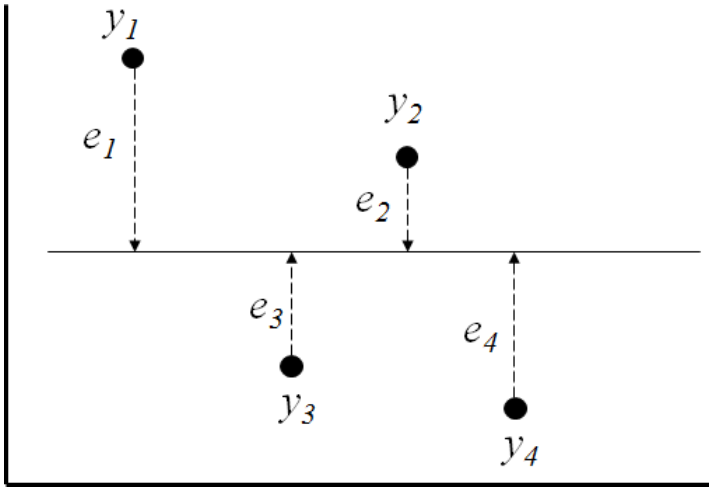
This is the regression line

The procedure by which linear regression minimises these distances is called “ordinary least squares”.

NB: You may see the term “OLS regression” thrown around

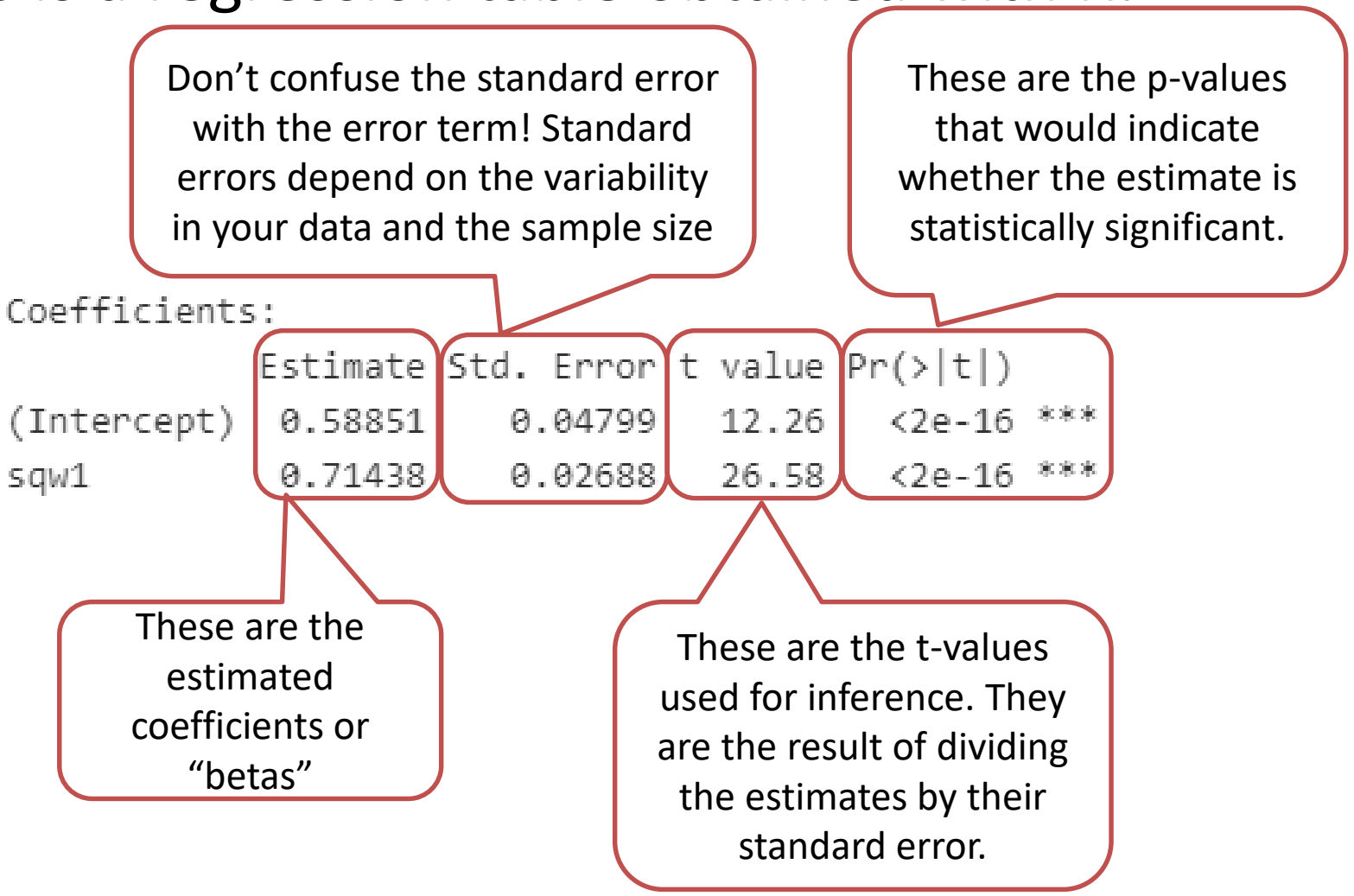
The error (2)

- It's also called "residual term"
- It's a continuous random variable
- Its mean is 0
  - We impose this by applying linear regression
- It's got a variance estimated from the data



Example output (1)

- This is a regression table obtained with R:



Example output (2)

These are the intercept and slope estimated coefficients. The regression line crosses the y axis at 0.58851. For each unit increase in “sqw1”, there is a 0.71438 increase in the dependent variable

Both coefficients are statistically significant, since p<0.001

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.58851    0.04799   12.26   <2e-16 ***
sqw1         0.71438    0.02688   26.58   <2e-16 ***
```

What does “statistically significant coefficients” mean?

It means that there is enough evidence to say that they are different from zero.

H0: Coefficient is zero
H1: Coefficient is different from zero
p<0.001
Conclusion: H0 is rejected

Example output (3)

Remember this equation?:

$$y = \beta_0 + \beta_1 x + e$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.58851    0.04799   12.26   <2e-16 ***
sqw1         0.71438    0.02688   26.58   <2e-16 ***
```

We can use the estimates table and write our regression equation:

$$y = 0.58851 + 0.71438 sqw1 + e$$

How good is my model?

- The goodness of fit of a linear regression model can be evaluated by using the "R-squared"
  - It measures the amount of variance in y that is accounted for by x
    - The closer to 1, the better.

```
Residual standard error: 0.3632 on 501 degrees of freedom
Multiple R-squared:  0.585,   Adjusted R-squared:  0.5842
F-statistic: 706.3 on 1 and 501 DF,  p-value: < 2.2e-16
```

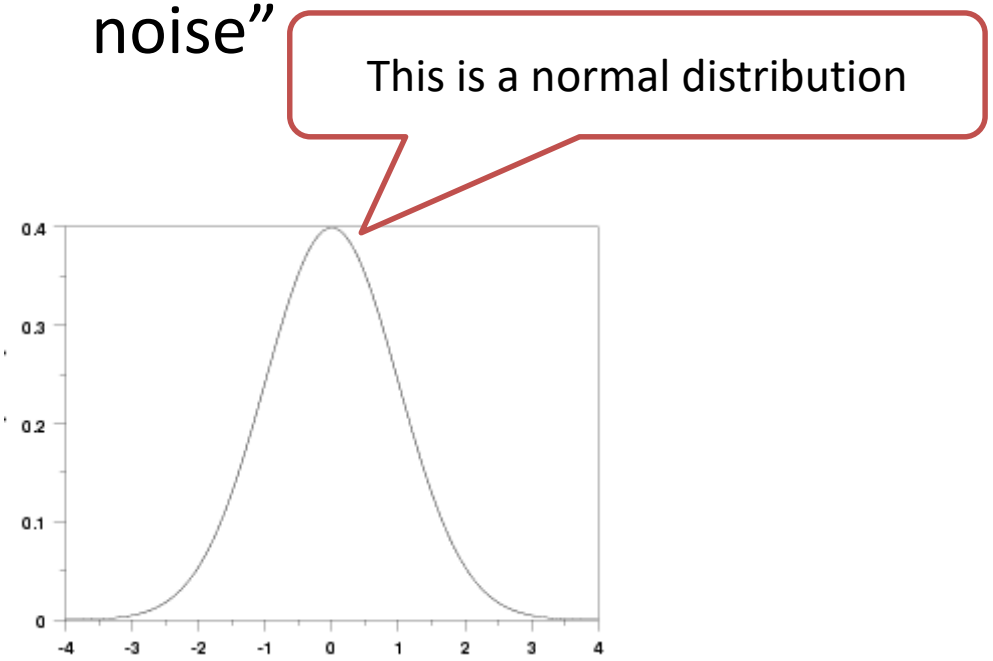
This model explains 58.42% of the variance in y.
It is actually quite a good model

Assumptions

- All statistical models have underlying assumptions.
- Failing to meet them may invalidate the conclusions you reach from the model
- We'll focus on a few statistical assumptions, but…
  - Theory is also very important!

Normality of residuals

- The residual term is a continuous random variable, and as such, it should have a normal distribution
- This roughly means that whatever is left unexplained by our model can be thought of as “random white noise”



Homoscedasticity

- Easier said than done (!)
- The variation of the residuals should remain constant across the predicted values of our model

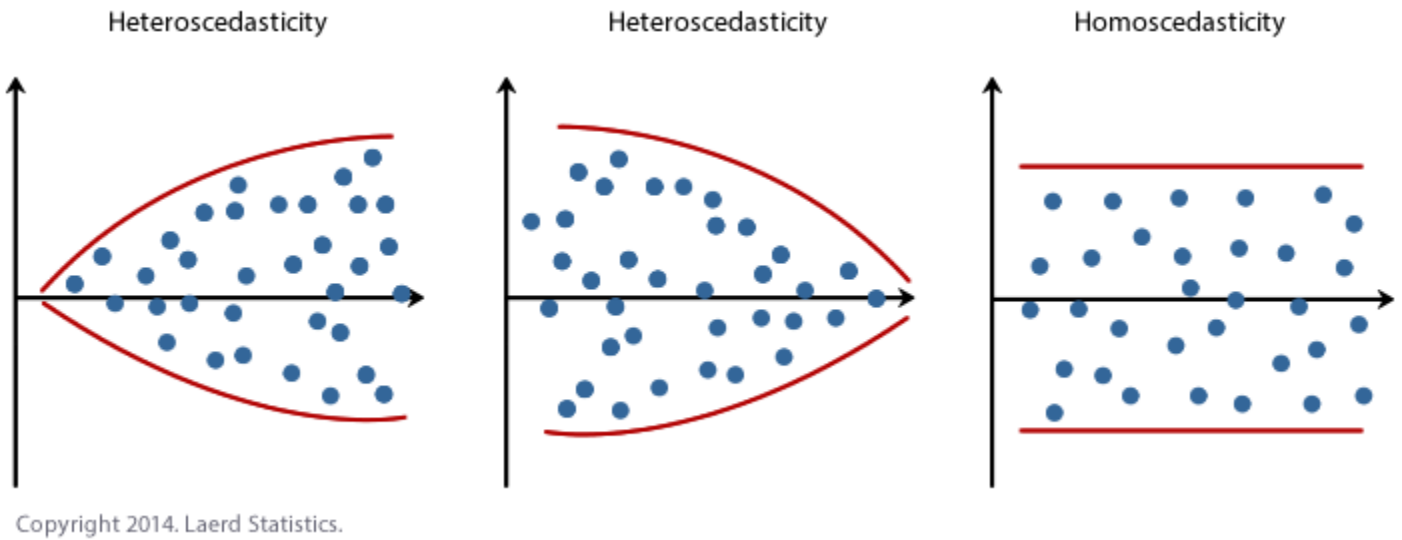


Image source: https://statistics.laerd.com/spss-tutorials/img/lr/heteroscedastic-relationships.png

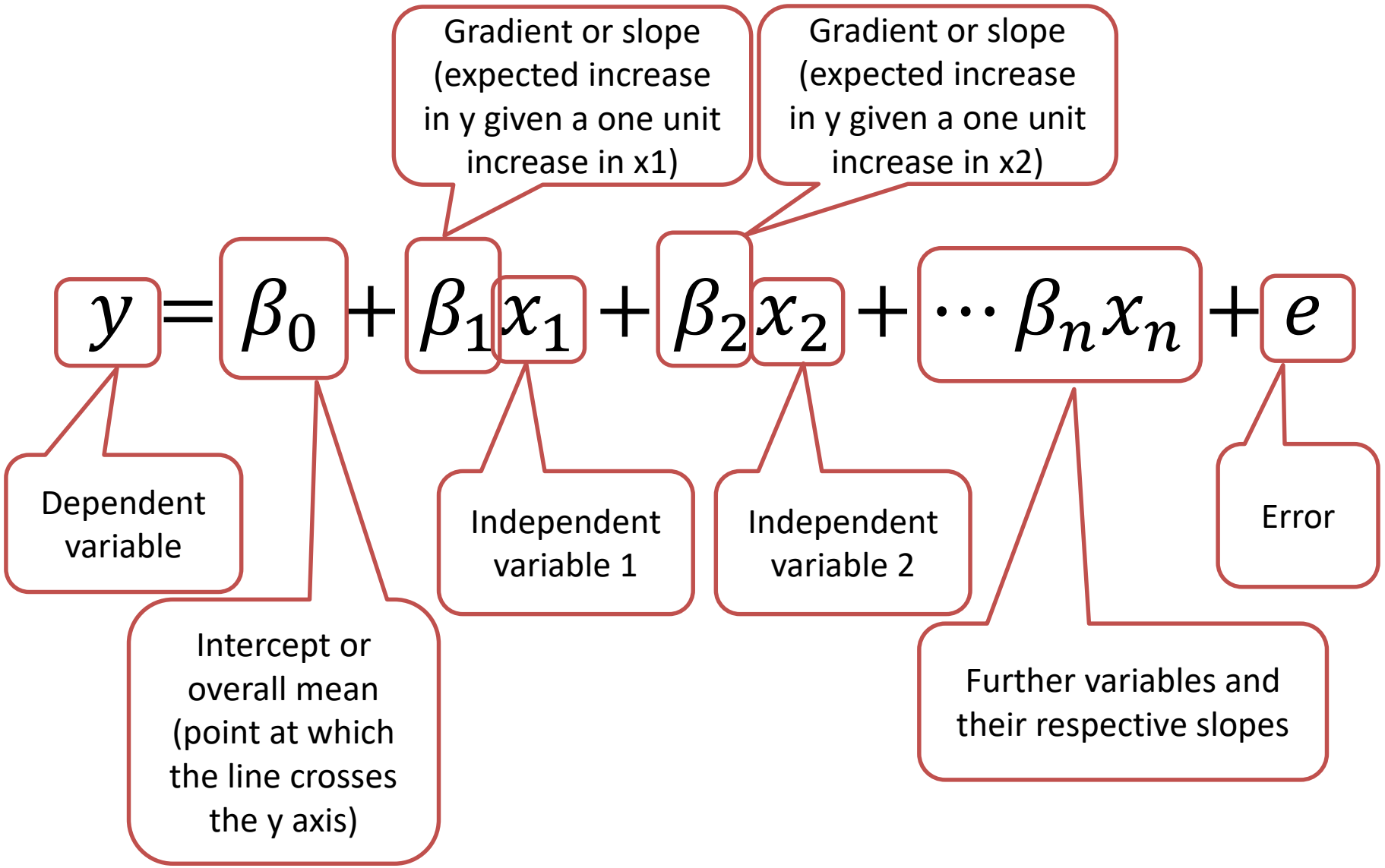
Part Three

MULTIPLE LINEAR REGRESSION

Multiple linear regression

- Why multiple?
  - Phenomena are ever so rarely monocausal
- How?
  - Simply add more x's to the equation

Multiple linear regression



Variable types (1)

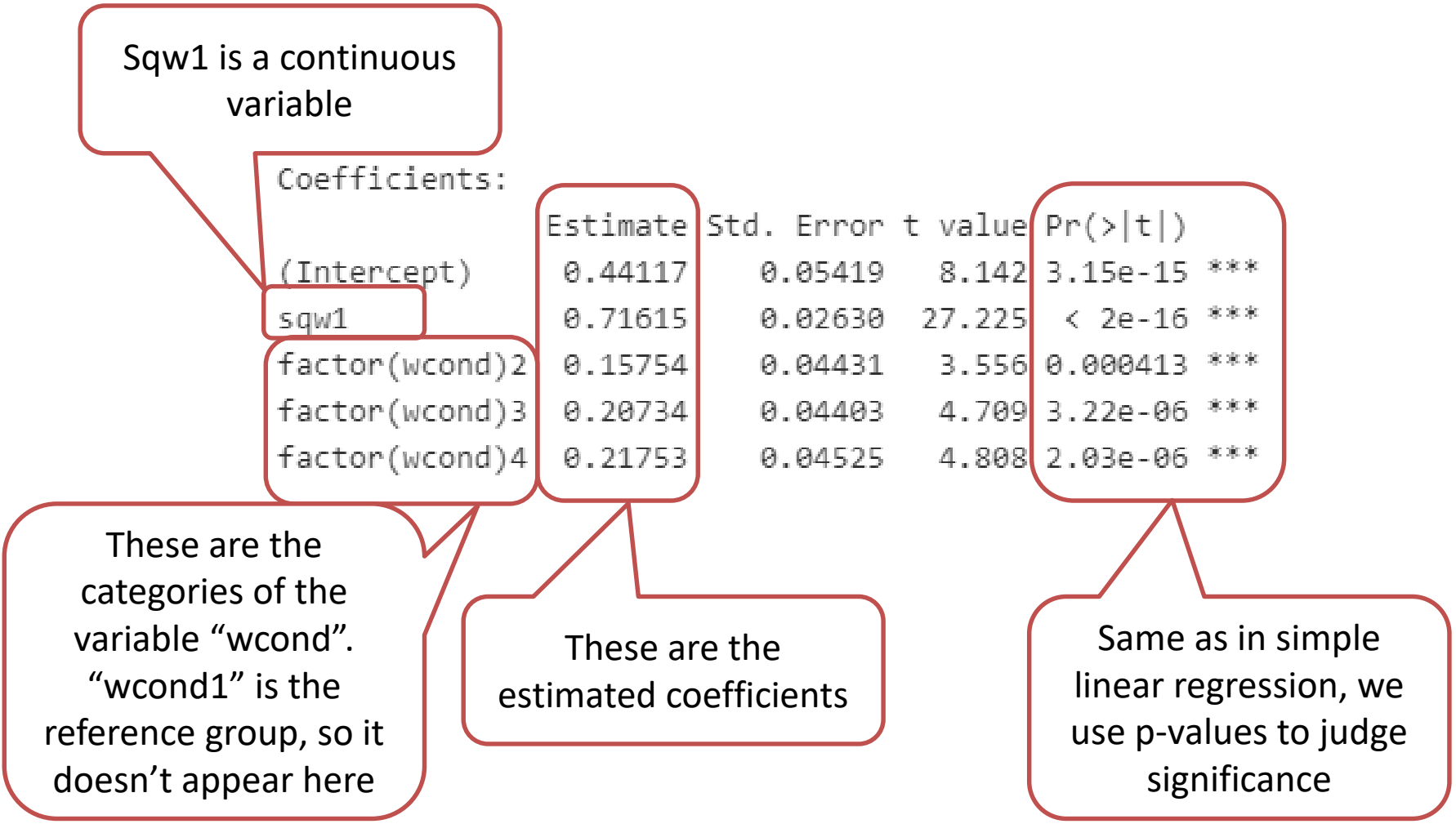
- MLR models can handle two types of independent variables: continuous and binary
  - Examples of continuous: age, height, score at baseline in standardised test
  - Examples of binary: Vote (Yes/No), Disabled (Yes/No), Eligible for free school meals (Yes/No)
- When variables are binary, the estimated coefficient indicates the difference between the “reference group” and the others

Variable types (2)

- MLR models can handle ordinal and unordered categorical variables if they are coded as dummies
  - Examples of ordinal: Health status (Poor, Fair, Good)
  - Examples of unordered categorical: Ethnicity (White, Black, Latin American)
- When “dummy-coded”, multiple coefficients are estimated (number of categories minus 1), except for the reference group.
  - The coefficients indicate the difference with respect to the reference group

Example output (1)

- This is a regression table obtained with R:



Example output (2)

The coefficient for sqw1 is slightly different. This is expected, because now we're controlling for another variable.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.44117    0.05419   8.142 3.15e-15 ***
sqw1             0.71615    0.02630  27.225  < 2e-16 ***
factor(wcond)2   0.15754    0.04431   3.556 0.000413 ***
factor(wcond)3   0.20734    0.04403   4.709 3.22e-06 ***
factor(wcond)4   0.21753    0.04525   4.808 2.03e-06 ***
```

Individuals in group wcond2 have a predicted value for y that is 0.15754 units higher than those in group wcond1 (the reference group). This is statistically significant ($p<0.001$)

Your turn!

- Write the regression equation for this table:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.44117    0.05419   8.142 3.15e-15 ***
sqw1           0.71615    0.02630  27.225  < 2e-16 ***
factor(wcond)2 0.15754    0.04431   3.556 0.000413 ***
factor(wcond)3 0.20734    0.04403   4.709 3.22e-06 ***
factor(wcond)4 0.21753    0.04525   4.808 2.03e-06 ***
```

How good is this model?

- The goodness of fit of a linear regression model can be evaluated by using the "R-squared"
  - In the case of multiple linear regression, it's especially important to use the "Adjusted R-squared", because it penalises the number of variables you put in the model

```
Residual standard error: 0.3536 on 498 degrees of freedom
Multiple R-squared:  0.6091, Adjusted R-squared:  0.606
F-statistic:   194 on 4 and 498 DF,  p-value: < 2.2e-16
```

This model explains 60.6% of the variance in y.
It is actually quite a good model

Model selection

- How do you decide what variables to put in a model?
  - First of all: Theory!
  - Model nesting: start with what is known and add your own hypotheses after that
- If you have multiple models with different variables, how do you decide which one is better?
  - Can you guess?... Theory!
  - Statistical criteria include: R-Squared comparison, comparing goodness of fit criteria, like AIC and BIC values (we're not covering this)

Part Four

SOFTWARE (1)

Linear Regression (1)

- Whether simple or multiple linear regression, R base package has an equation-like syntax:

  y ~ x (y regressed on x)

- In R, that would be:

  lm(mydata$y ~ mydata$x)

- If you want to add more variables, simply add a "+":

  lm(mydata$y ~ mydata$x1 + mydata$x2)

- Alternatively, you can specify the data to avoid typing the name of the data frame for each variable:

  lm(y ~ x1 + x2, data = mydata)

- Also, you should save the model to an object:

  mymodel <- lm(y ~ x1 + x2, data = mydata)

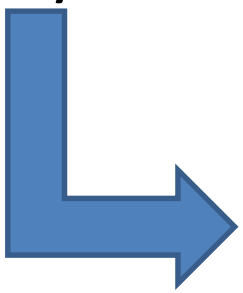
Linear Regression (2)

- The object “mymodel” contains a lot of useful model information, not just the coefficients!

  str(mymodel) # to see all the stored information

- To obtain the table with coefficients and other summary info:

  summary(mymodel)



```
Console //nask.man.ac.uk/home$/
> summary(mymodel)

Call:
lm(formula = weight ~ height, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.12989 -0.42894  0.01745  0.17372  1.43564

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.990095   1.064822   1.869   0.0986 .
height      0.082341   0.006934  11.874 2.32e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7689 on 8 degrees of freedom
Multiple R-squared:  0.9463,    Adjusted R-squared:  0.9396
F-statistic:   141 on 1 and 8 DF,  p-value: 2.323e-06
```

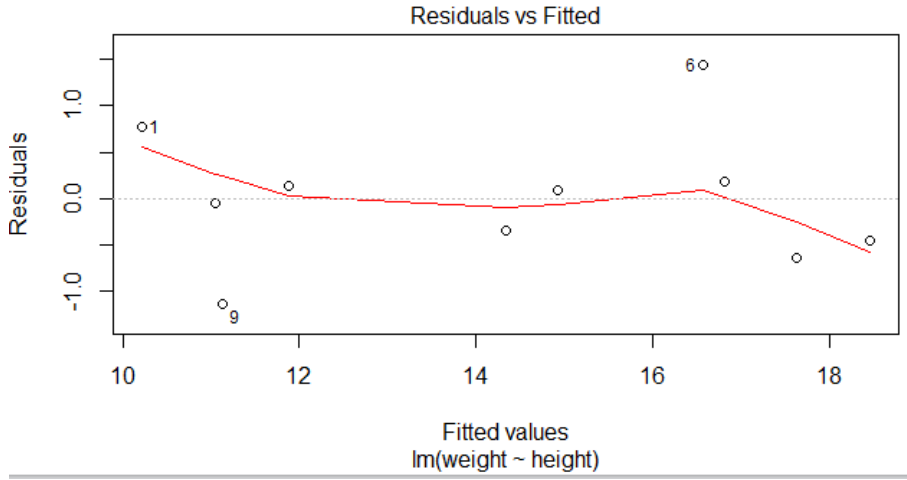
Linear Regression (3)

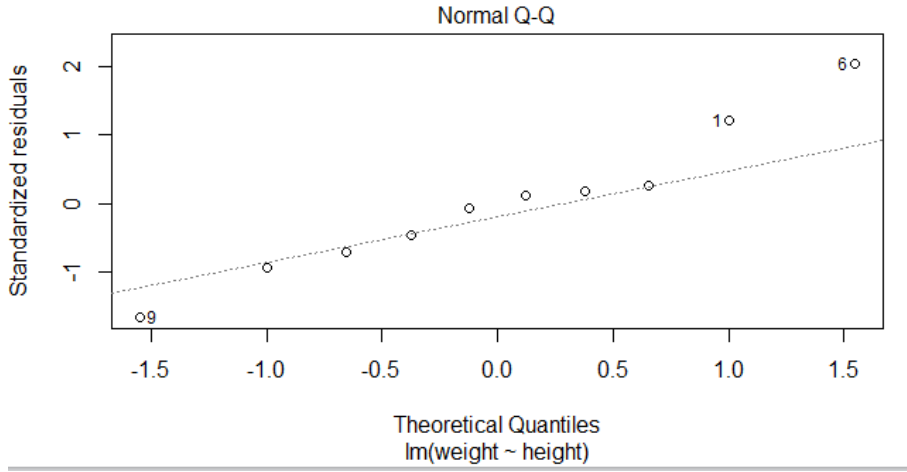
- R produces regression plots to check assumptions.

  plot(mymodel) # you'll be prompted to hit Enter 4 times

```
> plot(mymodel)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
>
```

- These are 2 of the plots you get:





Part Five

BINARY LOGISTIC REGRESSION

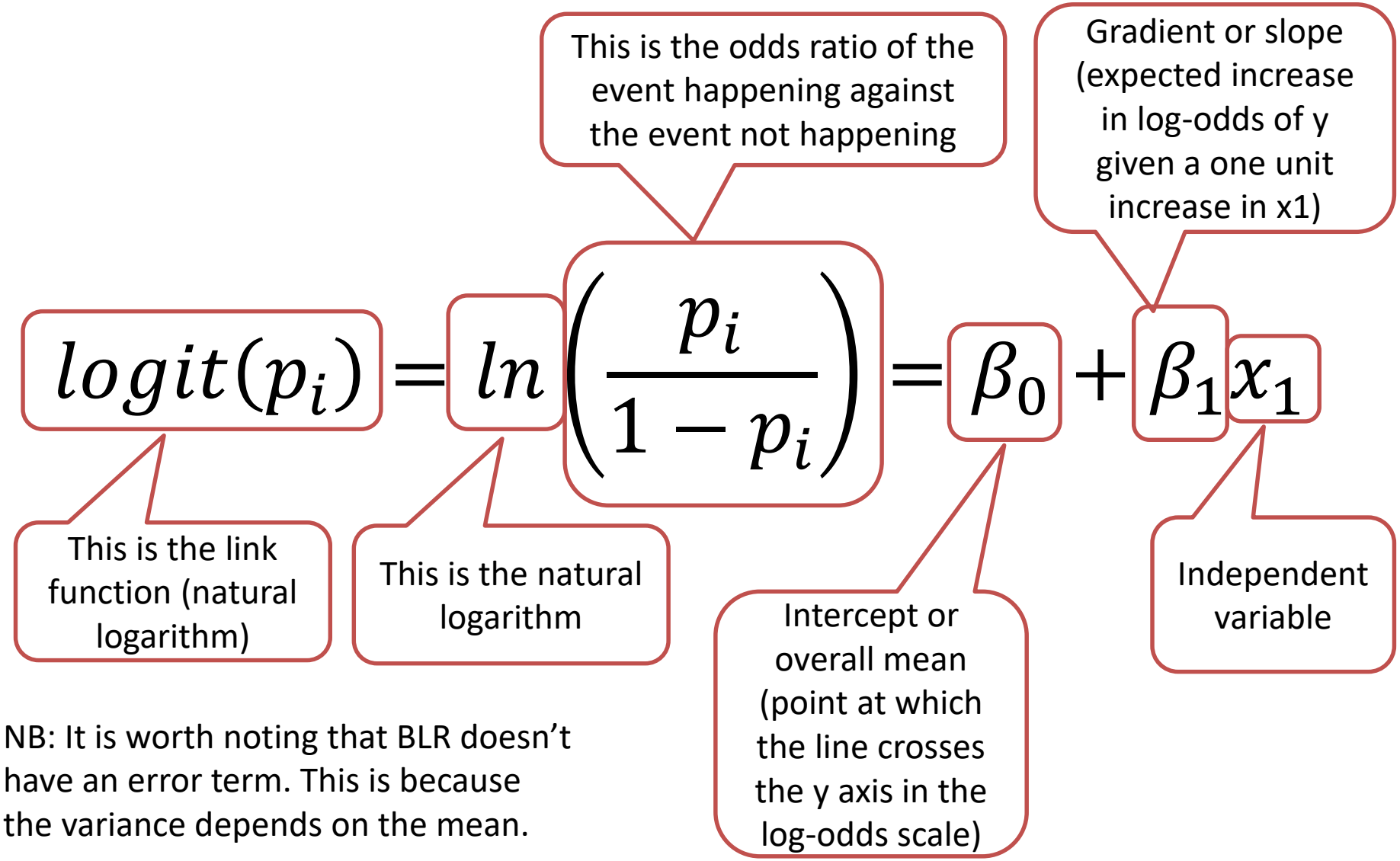
Generalised linear models

- Often in Social Sciences, the outcomes that we're interested in are not continuous, but binary:
  - Voting
  - Passing a test
  - Catching (or recovering from) an illness
- Linear regression cannot be used in those cases, so alternative methods have been developed.
- We'll focus on the case of binary outcomes variables, but bear in mind there are other cases
  - The umbrella term for those models is "generalised linear models"

Binary logistic regression (1)

- Given that our outcome of interest has only zeroes and ones, we need a way to convert that to a continuous measure.
  - For this we use the natural logarithm of the odds ratio.
  - The odds ratio is the probability of the outcome happening against the outcome not happening.
  - How likely is one event to occur compared to the opposite?
- Most of the principles of MLR apply also to BLR
  - Except for the assumptions (no need to check residuals)

Binary logistic regression (2)



NB: It is worth noting that BLR doesn't have an error term. This is because the variance depends on the mean.

Example output (1)

- Most software packages will provide estimates in the log-odds scale by default:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.9697	0.5782	-10.326	< 2e-16	***
sqw1	3.3664	0.3074	10.952	< 2e-16	***
factor(wcond)2	0.7479	0.3304	2.264	0.02359	*
factor(wcond)3	1.1452	0.3196	3.583	0.00034	***
factor(wcond)4	0.8838	0.3417	2.587	0.00969	**

- This is not necessarily easy to interpret, but:
  - when coefficients are positive, the outcome is more likely and vice versa
  - The rest of the table works in the same way as MLR

Example output (2)

- The estimates can be converted from the log-odds scale to odds ratio by using the “exponent” function “exp()”

```
factor(wcond)2 factor(wcond)3 factor(wcond)4
   2.112561059    3.142957336    2.420035894
```

- These odds ratio estimates are in relation to the reference group “wcond1”
  - Those in group wcond2 are 2.11 times as likely to experience the outcomes as those in group wcond1

Example output (3)

- Beware of the way you interpret odds ratios:
  - From the table below:

factor(wcond)2	factor(wcond)3	factor(wcond)4
2.112561059	3.142957336	2.420035894

- This is wrong:
  - Those in group wcond2 are 2.11 times **more likely** to experience the outcomes **than those** in group wcond1
- They are only **1.11 times more likely**.
  - Odds ratio=1 indicates that the probability of the event occurring is 0.5 (as likely to happen as not to happen)

How good is my logistic model?

- There is no exact equivalent for R-squared in binary logistic regression
  - Some software packages provide “pseudo R-squared” measures.
    - For example, McFadden’s pseudo R-squared is based on the log-likelihood
- How do you compare competing models then?
  - You can use… you guessed it… Theory!
  - You can also use **Information Criteria**, such as AIC and BIC (advanced topic, we’re not covering this)

Part Six

SOFTWARE (2)

Binary logistic regression (1)

- This works very similar to linear regression. The model is specified as an equation, although now the function “glm” (generalised linear model) is used.
- Also, we need to specify the link function with the subcommand “family”:

```
logmodel <- glm(y ~ x1 + x2, family = binomial(logit),
data = mydata)
```

Binary logistic regression (2)

- Most of the post-estimation procedures that are used for linear regression are available for binary logistic regression.
- Coefficients are presented by default as log-odds. You may want to convert them to odds ratio:

  exp(logmodel$coefficients)
- Beware that predictions are not in the original scale (zeroes and ones) nor in the log-odds scale, but in the probability scale.

Next SeM-in-Rs session

Multilevel modelling in R and Mplus

March 17th, 10:00 – 13:00 (not a strike day!)

Mansfield Cooper 4.05

- Basics of multilevel modelling
- Model specification for continuous and discrete outcomes
- Interpretation of output
- Diagnostics and assumption checking