

Linear models

Samraat Pawar

Department of Life Sciences (Silwood Park)

Imperial College
London

LECTURE OUTLINE

- What is a linear model?
- How do we deal with variation?
- Is a linear model appropriate for the data?
- How well does a linear model explain the data?

Concepts:

- Types of variable: continuous versus categorical
- Terms and coefficients of a model
- Model residuals
- Significance testing

WHAT PREDICTS THE WEIGHTS (w) OF LECTURERS?

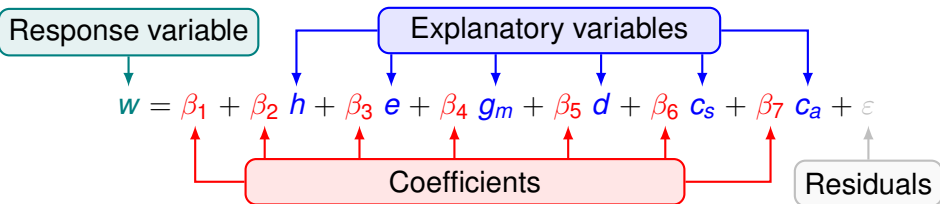
Use our *hypotheses* to identify the *variables* we collect...

- Height (h) in metres
- Exercise per week (e) in hours
- Gender (g)
- Distance from home to nearest Greggs bakery (d) in metres
- Ownership of a games console (c)

... and build a mathematical model:

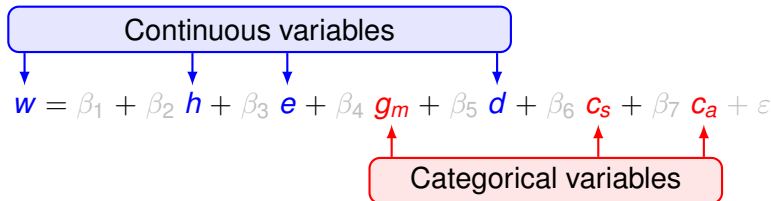
$$w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$$

A COMBINATION OF FOUR COMPONENTS



- A response variable (w)
- A set of explanatory variables (h, e, g, d, c)
- A set of coefficients ($\beta_1 - \beta_7$)
- A set of residuals (ε)

DIFFERENT TYPES OF VARIABLES



- The response variable is always **continuous**.
- The explanatory variables can be a mix of:
 - **Continuous** variables: height, exercise and distance.
 - **Categorical** variables: gender and console ownership.
- **Categorical** variables or *factors* have a number of *levels*:
 - Gender has two levels (Male / Female)
 - Console has three levels (None / Sofa-based / Active)

TERMS AND COEFFICIENTS

The diagram illustrates the relationship between explanatory variables and terms in a linear model. The equation is $w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$. Above the equation, boxes labeled 'Gender' and 'Console' have red arrows pointing to the terms g_m and c_s and c_a respectively. Below the equation, boxes labeled 'Height', 'Exercise', and 'Distance' have blue arrows pointing to the terms h , e , and d respectively.

$$w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$$

- Each explanatory variable is a *term* in the model
- Each term has at least one coefficient
- **Continuous** terms always have one coefficient
- Categorical **Factors** have $N - 1$ coefficients, where N is the number of levels (*where are the missing coefficients??*)

WAIT! WHY $N - 1$? WHAT IS β_1 ?

$$w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$$

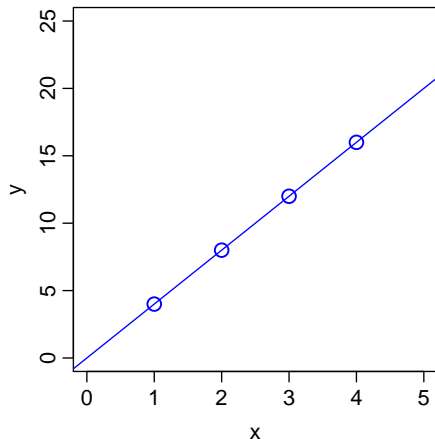
- Two ways of thinking about β_1 :
 - Continuous variables: the *y intercept*
 - Factors: the baseline or *reference* value
- This baseline is the value for the *first levels* of each factor
- All response values start at this baseline
- All the other coefficients measure *differences* from β_1 :
 - along a continuous slope
 - as an offset to a different level

LINEAR MODELS ARE JUST A SUM

$$w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$$

- Find the baseline value for women with no games console (β_1)
- The model tells us how much to add to this. . .
 - for a height of 1.82 metres?
 - for doing 150 minutes of exercise a week?
 - for being male?
 - for living 2416 metres from a Greggs?
 - for owning an Xbox?

EXAMPLES - ONE CONTINUOUS VARIABLE



$$y = \beta_1 x$$

$$4 = 4 \times 1$$

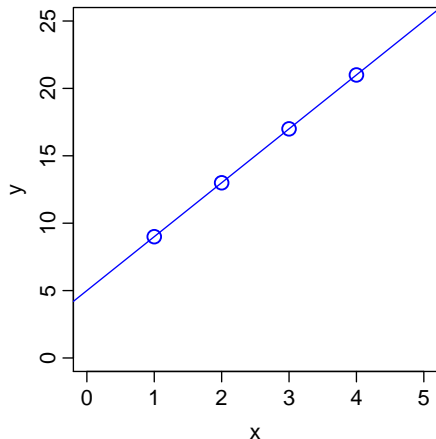
$$8 = 4 \times 2$$

$$12 = 4 \times 3$$

$$16 = 4 \times 4$$

$$\beta_1 = 4$$

EXAMPLES - ONE CONTINUOUS VARIABLE



$$y = \beta_1 + \beta_2 x$$

$$9 = 5 + 4 \times 1$$

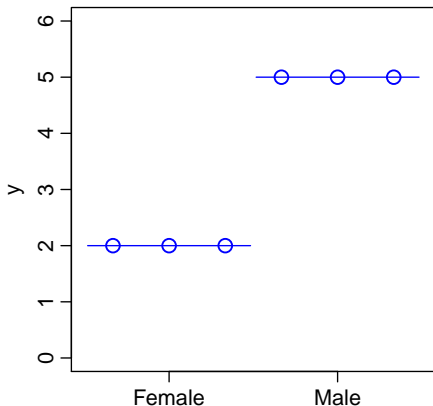
$$13 = 5 + 4 \times 2$$

$$17 = 5 + 4 \times 3$$

$$21 = 5 + 4 \times 4$$

$$\beta_1 = 5; \beta_2 = 4$$

EXAMPLES - ONE FACTOR



$$y = \beta_1 + \beta_2 g_m$$

$$2 = 2 + 3 \times 0$$

$$2 = 2 + 3 \times 0$$

$$2 = 2 + 3 \times 0$$

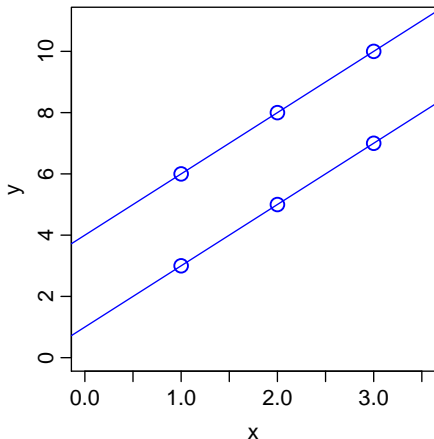
$$5 = 2 + 3 \times 1$$

$$5 = 2 + 3 \times 1$$

$$5 = 2 + 3 \times 1$$

$$\beta_1 = 2; \beta_2 = 3$$

EXAMPLES - ONE CONTINUOUS VARIABLE AND ONE FACTOR



$$y = \beta_1 + \beta_2 x + \beta_3 g_m$$

$$3 = 1 + 2 \times 1 + 3 \times 0$$

$$5 = 1 + 2 \times 2 + 3 \times 0$$

$$7 = 1 + 2 \times 3 + 3 \times 0$$

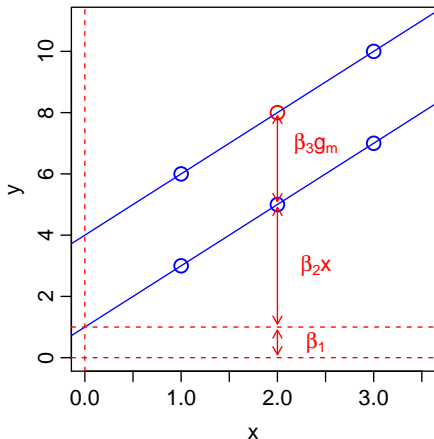
$$6 = 1 + 2 \times 1 + 3 \times 1$$

$$8 = 1 + 2 \times 2 + 3 \times 1$$

$$10 = 1 + 2 \times 3 + 3 \times 1$$

$$\beta_1 = 1; \beta_2 = 2; \beta_3 = 3$$

EXAMPLES - ONE CONTINUOUS VARIABLE AND ONE FACTOR



$$y = \beta_1 + \beta_2 x + \beta_3 g_m$$

$$3 = 1 + 2 \times 1 + 3 \times 0$$

$$5 = 1 + 2 \times 2 + 3 \times 0$$

$$7 = 1 + 2 \times 3 + 3 \times 0$$

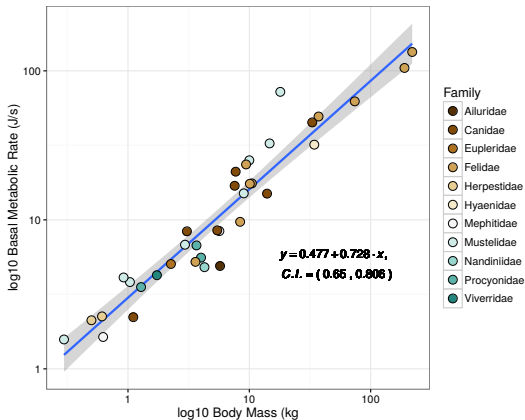
$$6 = 1 + 2 \times 1 + 3 \times 1$$

$$8 = 1 + 2 \times 2 + 3 \times 1$$

$$10 = 1 + 2 \times 3 + 3 \times 1$$

$$\beta_1 = 1; \beta_2 = 2; \beta_3 = 3$$

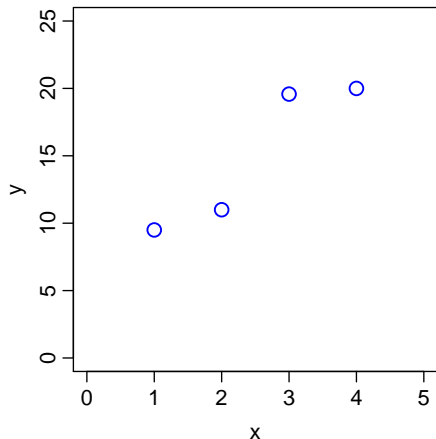
RESIDUALS: VARIATION IS EVERYWHERE



Rizzuto et al. 2017, Nat Ecol Evol

- Data always shows variation from a perfect model (deviations)
 - Missing variables (age, lab vs. field biology, time of day)
 - Measurement error
 - Stochastic variation

RESIDUALS - VARIATION IS EVERYWHERE



$$y = \beta_1 + \beta_2 x$$

$$9.50 = ? + ? \times 1$$

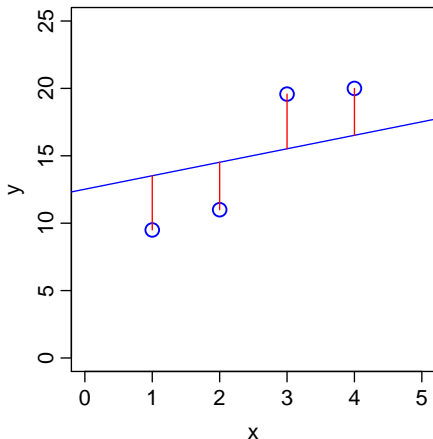
$$11.00 = ? + ? \times 2$$

$$19.58 = ? + ? \times 3$$

$$20.00 = ? + ? \times 4$$

*No unique line through the points
unless we impose some other
constraint or condition*

RESIDUALS - GUESS 1



$$y = \beta_1 + \beta_2 x + \varepsilon$$

$$9.50 = 12.52 + 1 \times 1 - 4.02$$

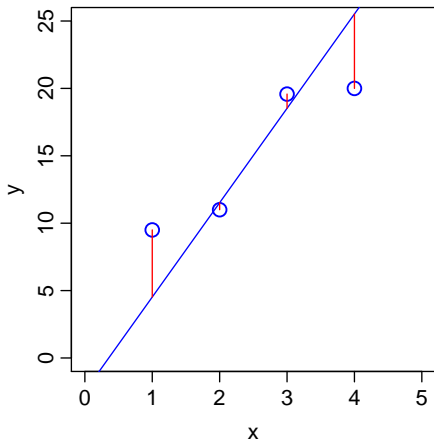
$$11.00 = 12.52 + 1 \times 2 - 3.52$$

$$19.58 = 12.52 + 1 \times 3 + 4.06$$

$$20.00 = 12.52 + 1 \times 4 + 3.48$$

$$\beta_1 = 12.52; \beta_2 = 1$$

RESIDUALS - GUESS 2



$$y = \beta_1 + \beta_2 x + \varepsilon$$

$$9.50 = -2.48 + 7 \times 1 + 4.98$$

$$11.00 = -2.48 + 7 \times 2 - 0.52$$

$$19.58 = -2.48 + 7 \times 3 + 1.06$$

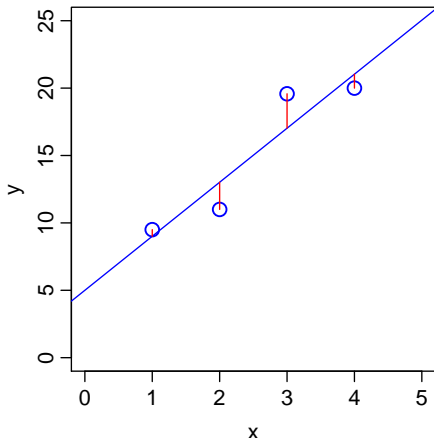
$$20.00 = -2.48 + 7 \times 4 - 5.52$$

$$\beta_1 = -2.48; \beta_2 = 7$$

RESIDUALS: (ORDINARY) LEAST SQUARES SOLUTION

Minimize the *sum* of the *squared* residuals

WHY GUESS?: THE (ORDINARY) LEAST SQUARES SOLUTION



$$y = \beta_1 + \beta_2 x + \varepsilon$$

$$9.50 = 5 + 4 \times 1 + 0.50$$

$$11.00 = 5 + 4 \times 2 - 2.00$$

$$19.58 = 5 + 4 \times 3 + 2.58$$

$$20.00 = 5 + 4 \times 4 - 1.00$$

$$\beta_1 = 5; \beta_2 = 4$$

MODEL AS A MATRIX - TERMINOLOGY

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

Observed values



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

Coefficients



$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

+

$$\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \end{bmatrix}$$

Model matrix



$$= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

Residuals



MODEL AS A MATRIX - TERMINOLOGY

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

Observed values

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}$$

Coefficients

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Residuals

$$\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

MODEL AS A MATRIX - TERMINOLOGY

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

Observed values

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}$$

Coefficients

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Residuals

$$\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

MODEL AS A MATRIX - TERMINOLOGY

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

Observed values

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}$$

Coefficients

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Residuals

$$\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

MODEL AS A MATRIX - TERMINOLOGY

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

Observed values

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}$$

Coefficients

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Residuals

$$\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

MODEL AS A MATRIX - TERMINOLOGY

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

Observed values

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}$$

Coefficients

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Residuals

$$\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

MODEL AS A MATRIX - TERMINOLOGY

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

Given these ...

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}$$

=

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

... find the set of these...

$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

+

$$\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

... that minimize the sum of the squares of these.

MODEL AS A MATRIX - PREDICTIONS

$$\hat{\mathbf{Y}} = \mathbf{X}\beta$$

Predicted or fitted values



$$\begin{bmatrix} 9 \\ 13 \\ 17 \\ 21 \end{bmatrix}$$

=

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

Coefficients

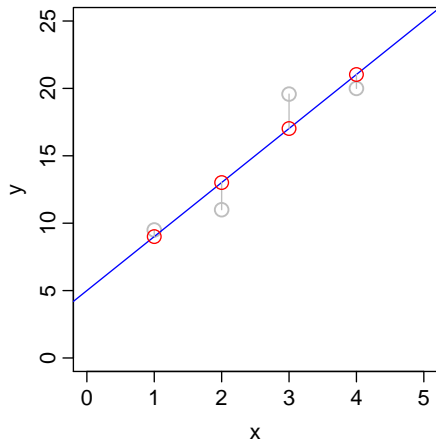


$$\begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix



PREDICTED VALUES



$$\hat{y} = \beta_1 + \beta_2 x$$

$$9 = 5 + 4 \times 1$$

$$13 = 5 + 4 \times 2$$

$$17 = 5 + 4 \times 3$$

$$21 = 5 + 4 \times 4$$

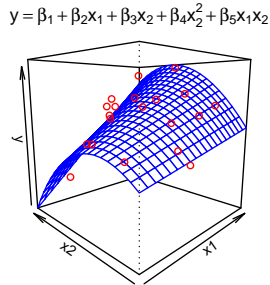
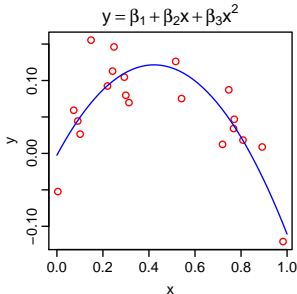
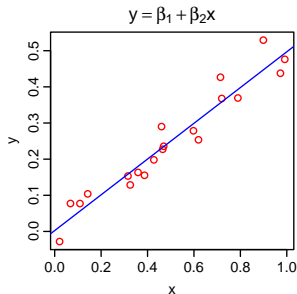
ASSUMPTIONS

- Linear models have the following assumptions:
 - No measurement error in explanatory variables
 - The explanatory variables are not very highly correlated
 - The model is linear
 - The model has constant normal variance
- **If these assumptions are not met, the model can be very wrong**

ASSUMPTIONS

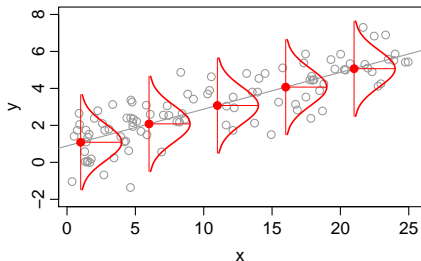
- Linear models have the following assumptions:
 - No measurement error in explanatory variables
 - The explanatory variables are not very highly correlated
 - The model is linear
 - The model has constant normal variance
- If these assumptions are not met, the model can be very wrong
- The last two need some further explanation

'THE MODEL IS LINEAR'

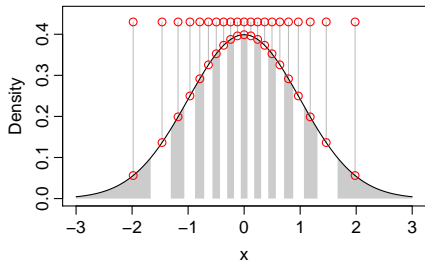


- These are *all* good linear models.
- Linear models can include curved relationships (e.g. polynomials)
- The data can be modelled as a *sum* of components
- A *linear combination* of variables and coefficients

'THE MODEL HAS CONSTANT NORMAL VARIANCE'

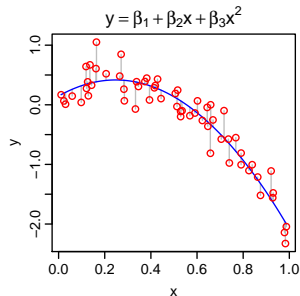
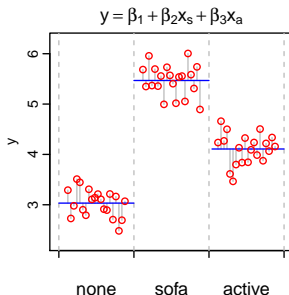
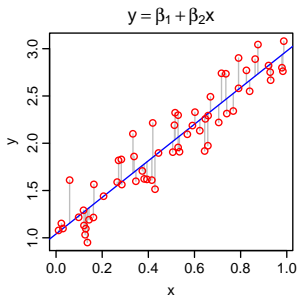


- The data has a similar spread around any predicted point in the model



- The residuals are normal
- Points *should* be spaced equally in the area under the curve
- Expect mostly small but a few larger residuals

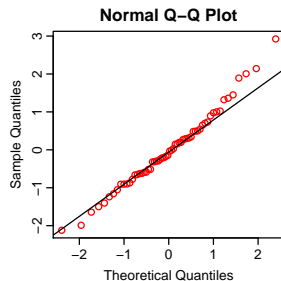
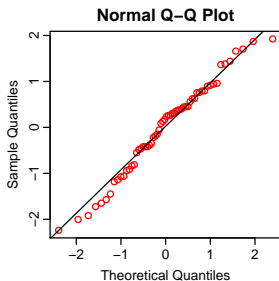
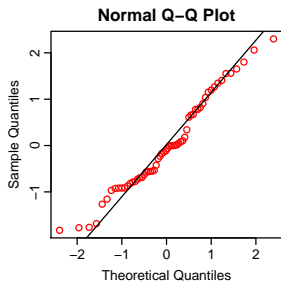
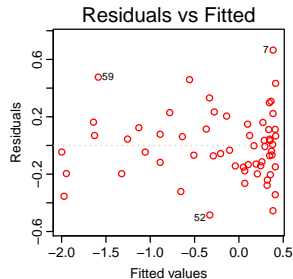
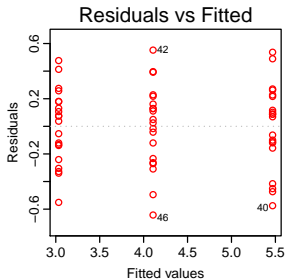
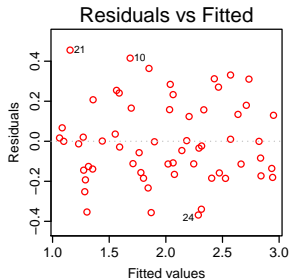
'THE MODEL HAS CONSTANT NORMAL VARIANCE'



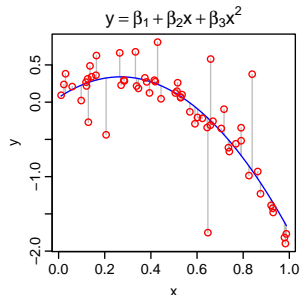
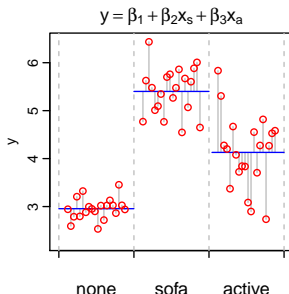
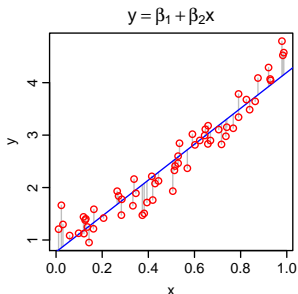
- Three good models

- Is the spread the same for all fitted values?
- Do the residuals match the normal expectation?

'THE MODEL HAS CONSTANT NORMAL VARIANCE'



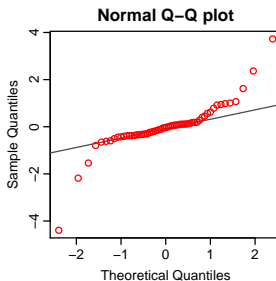
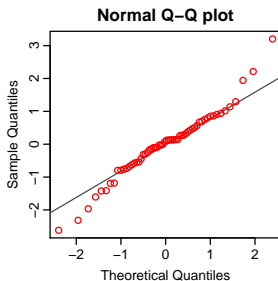
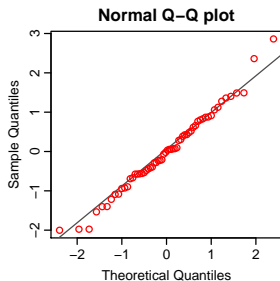
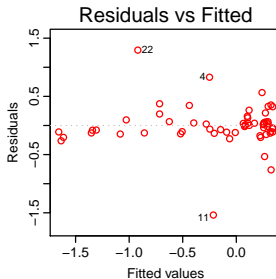
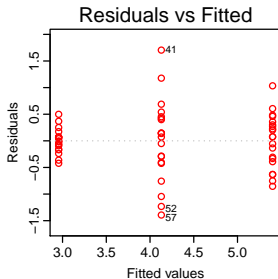
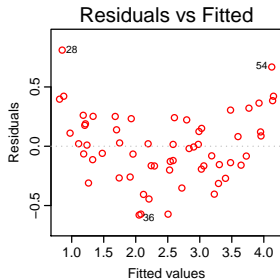
'THE MODEL HAS CONSTANT NORMAL VARIANCE'



• Three bad models

- Is the spread the same for all fitted values?
- Do the residuals match the normal expectation?

'THE MODEL HAS CONSTANT NORMAL VARIANCE'



IS A LINEAR MODEL APPROPRIATE?

Plot the data!
Plot the residuals!

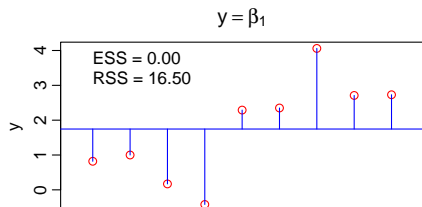
HOW EXPLANATORY IS THE MODEL?

- Back to F and t tests! (Woohoo!)
- *Terms*: analysis of variance
 - Does the model explain enough variation?
 - Does each term explain enough variation?
- *Coefficients*: *t* tests
 - Are the coefficients different from zero?

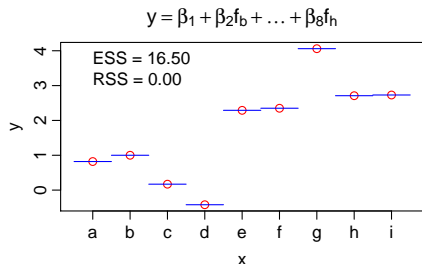
NULL VS. OVER-SPECIFIED MODELS: TWO ENDPOINTS

- **Total sum of squares (TSS):** Sum of the squared difference between the observed dependent variable (y) and the mean of y (\bar{y}), or, $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$
TSS tells us how much variation there is in the dependent variable
- **Explained sum of squares (ESS):** Sum of the squared differences between the predicted y (\hat{y}) and \bar{y} , or, $ESS = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$
ESS tells us how much of the variation in the dependent variable our model was able to explain
- **Residual sum of squares (RSS):** Sum of the squared differences between the observed y and the predicted \hat{y} , or,
 $RSS = \sum_{i=1}^n (\hat{y}_i - y_i)^2$
RSS tells us how much of the variation in the dependent variable our model could not explain
- Of course, $TSS = ESS + RSS$

NULL VS. OVER-SPECIFIED MODELS: TWO ENDPOINTS

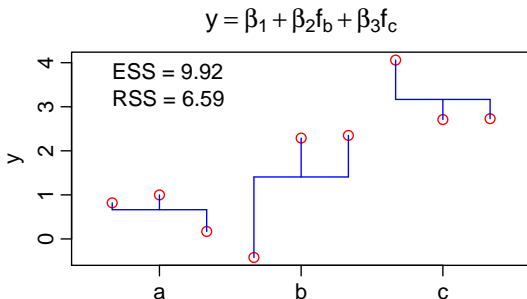


- The null model (H_0)
- Nothing is going on
- Biggest possible residuals
- Residual sum of squares (RSS) is as big as it can be



- The saturated model
- One coefficient per data point
- RSS is zero - all the sums of squares are now explained (ESS)

MORE INTERESTING MODELS



- Added a term with three levels
- Some but not all of the residual sums of squares are explained
- Is this enough to be interesting?

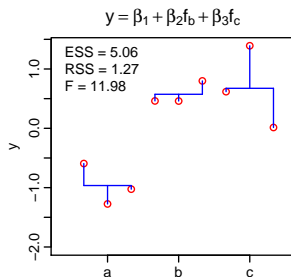
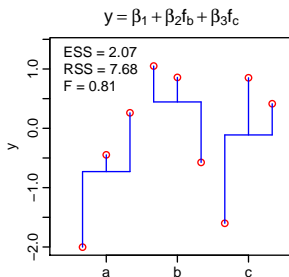
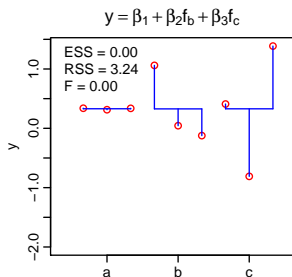
THE F STATISTIC

The diagram illustrates the components of the F-statistic formula. Four blue-bordered boxes with arrows point to the corresponding parts of the formula:

- Large ESS is good** points to the ESS in the numerator.
- Fewer coefficients is better** points to the N_c in the numerator.
- Small RSS is good** points to the RSS in the denominator.
- Residual degrees of freedom** points to the N_r in the denominator.

$$F = \frac{\text{ESS} / N_c}{\text{RSS} / N_r} = \frac{9.92 / 2}{6.59 / 6} = 4.52$$

F VALUES BY CHANCE



- What is the distribution of F if nothing is going on?
- Simulate 10,000 datasets where nothing is going on (H_0 is true)
- Calculate F for each random dataset under H_1
- Mostly H_1 has a low F - but sometimes it is high by chance

DISTRIBUTION OF F

- In our possibly interesting model, $F = 4.52$

DISTRIBUTION OF F

- In our possibly interesting model, $F = 4.52$
- 95% of the random data sets have $F \leq 5.5$
- A model this good is found by chance 1 in 16 times ($p = 0.063$)
- Not quite interesting enough!

ARE COEFFICIENTS DIFFERENT FROM ZERO?

Large is good - bigger changes

↓ ↓

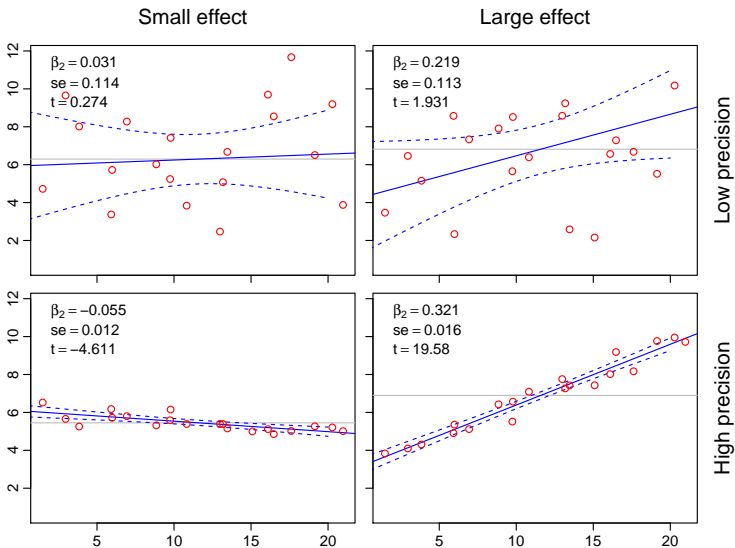
$$t = \frac{\text{Effect size}}{\text{Precision}} = \frac{\text{Coefficient value}}{\text{Standard error}}$$

↑ ↑

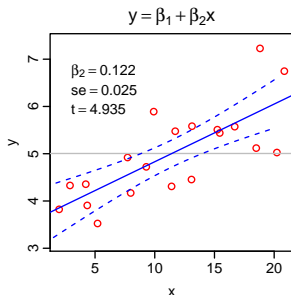
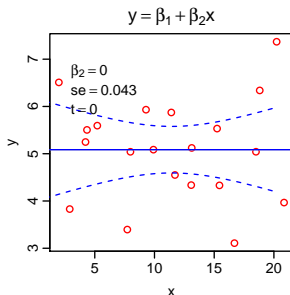
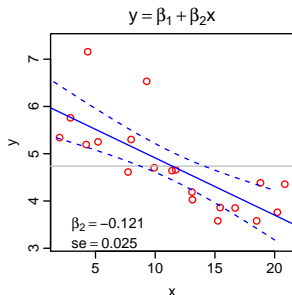
Small is good - known more precisely

- The value of a coefficient in a model is an *effect size*
- How much does changing this variable change the response?
- A *standard error* estimates how precisely we know the value

VARIATION IN EFFECT SIZE AND PRECISION



t VALUES BY CHANCE



- What is the distribution of t if nothing is going on?
- Simulate 10,000 datasets where nothing is going on (H_0 is true)
- Calculate t for each random dataset under H_1
- Mostly H_1 has a t near zero but can be positive or negative

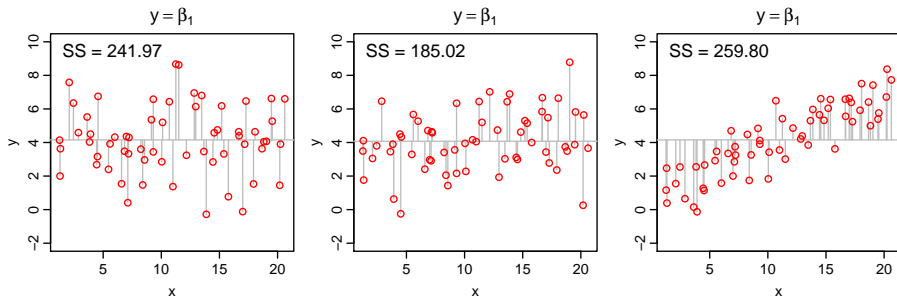
DISTRIBUTION OF t

- 95% of the random data sets have $t \leq \pm 2.09$
- Only the two higher precision models are expected to occur less than 1 time in 20 by chance.

SUMMARY

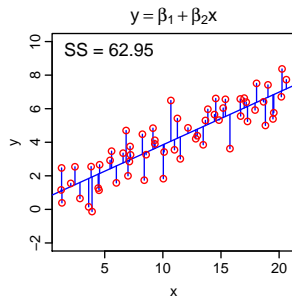
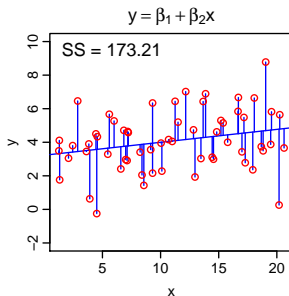
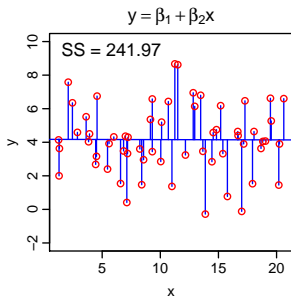
- Linear models predict a continuous response variable
- A sum based on the effect size of explanatory variables
- Estimate the model using least squares residuals
- Need to check if the model is appropriate
- Then check if the model is explanatory

WHAT ABOUT ANALYSIS OF VARIANCE (ANOVA)?



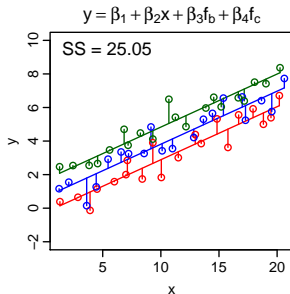
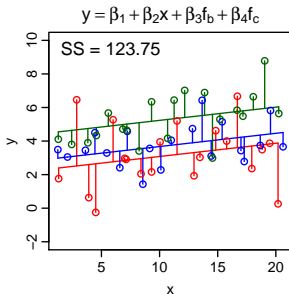
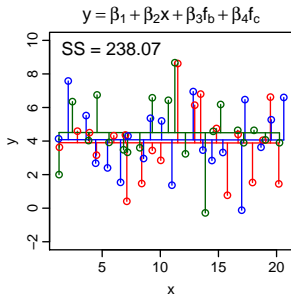
- The null hypothesis (H_0): Nothing is going on
- The residuals have to get smaller as we include terms.
- How much shorter?

EXAMPLES: ONE CONTINUOUS TERM



- An alternative model (H_1) using x
- Added one term (x) to the model to give (H_1)
- Do we reject H_0 and accept this new model?

EXAMPLES: ADDING A FACTOR



- Another model (H_2) using x and a factor f with three levels
- The sum of squares gets smaller again
- We've added one term (f) but two coefficients (f_b and f_c)
- Is this even better than H_1 ?

CHANGE IN VARIANCE

		Model A	Model B	Model C
H_0	Unexplained SS	241.97	185.02	259.80
	Explained SS	0	0	0
H_1	Unexplained SS	241.97	173.21	62.95
	Explained SS	0.00	11.81	196.85
H_2	Unexplained SS	238.07	123.75	25.05
	Explained SS	3.9	61.27	234.75