Linear models

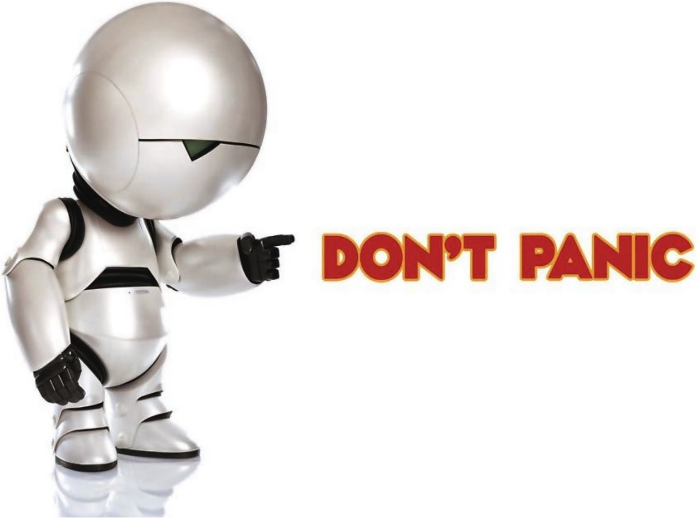
Course aims

- Learn a core set of statistical and visualization skills in R
- Develop ability to build, criticise and interpret linear models

The aim of this lecture:

- Explain the concepts underlying linear models
- Introduce concepts and techniques to be developed during practicals

Aim for today



Lecture structure

- 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

Response variable

Explanatory variables

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

Coefficients

Residuals

- 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

Continuous variables

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

Categorical 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

$$w = \beta_1 + \beta_2 \underset{\text{Height}}{\underset{\uparrow}{h}} + \beta_3 \underset{\text{Exercise}}{\underset{\uparrow}{e}} + \beta_4 \overset{\text{Gender}}{\overset{\downarrow}{g_m}} + \beta_5 \underset{\text{Distance}}{\underset{\uparrow}{d}} + \beta_6 \overbrace{c_s + \beta_7 c_a}^{\text{Console}} + \varepsilon$$

- 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?

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

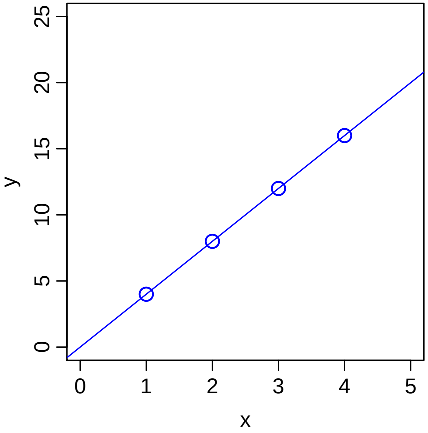
- Two ways of thinking about β_1:
    - Continuous variables: the y *intercept*
    - Factors: the baseline or *reference* value
- This baseline is the value for the *first levels* of each factor
- All response values start at this baseline
- All the other coefficients measure *differences* from β_1:
    - 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

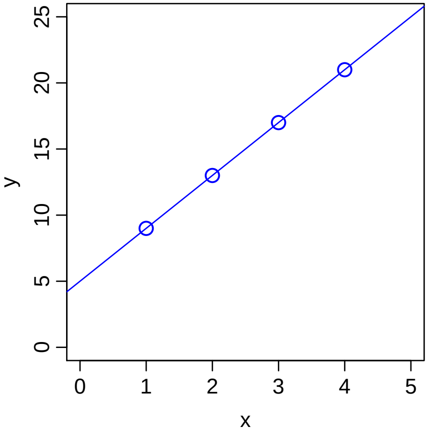


$$y = \beta_1 x$$

$$4 = 4 \times 1$$
$$8 = 4 \times 2$$
$$12 = 4 \times 3$$
$$16 = 4 \times 4$$

$$\beta_1 = 4$$

Examples - one continuous variable

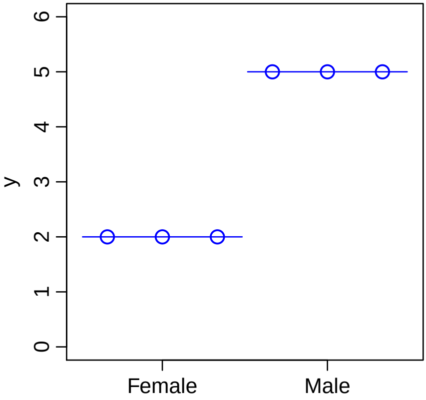


$$y = \beta_1 + \beta_2 x$$

$$9 = 5 + 4 \times 1$$
$$13 = 5 + 4 \times 2$$
$$21 = 5 + 4 \times 3$$
$$29 = 5 + 4 \times 4$$

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

Examples - one factor

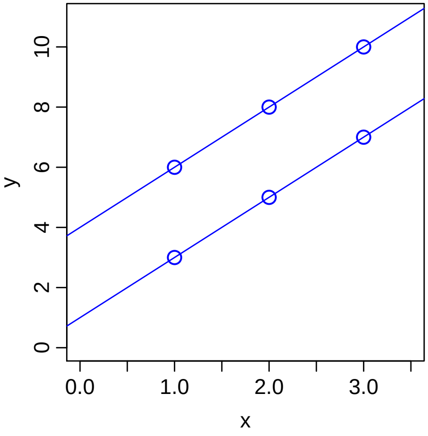


$$y = \beta_1 + \beta_2 g_m$$

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

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

Examples - one continuous variable and one factor

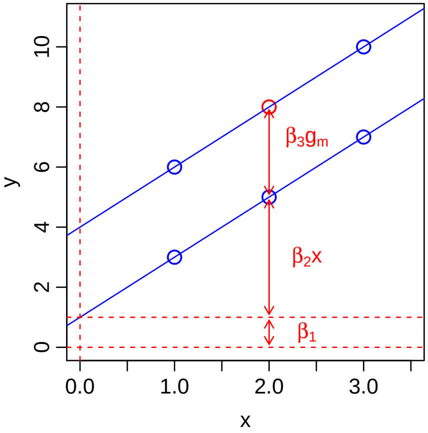


$$y = \beta_1 + \beta_2 x + \beta_3 g_m$$

$$3 = 1 + 2 \times 1 + 3 \times 0$$
$$5 = 1 + 2 \times 2 + 3 \times 0$$
$$7 = 1 + 2 \times 3 + 3 \times 0$$
$$6 = 1 + 2 \times 1 + 3 \times 1$$
$$8 = 1 + 2 \times 2 + 3 \times 1$$
$$10 = 1 + 2 \times 3 + 3 \times 1$$

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

Examples - one continuous variable and one factor

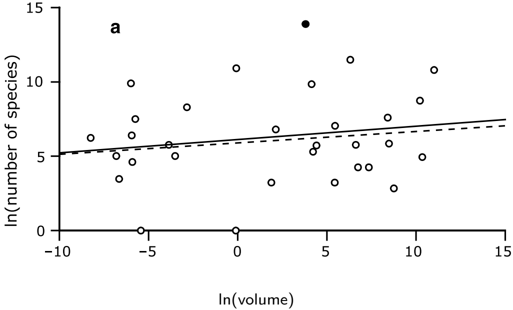


$$y = \beta_1 + \beta_2 x + \beta_3 g_m$$

$$3 = 1 + 2 \times 1 + 3 \times 0$$
$$5 = 1 + 2 \times 2 + 3 \times 0$$
$$7 = 1 + 2 \times 3 + 3 \times 0$$
$$6 = 1 + 2 \times 1 + 3 \times 1$$
$$8 = 1 + 2 \times 2 + 3 \times 1$$
$$10 = 1 + 2 \times 3 + 3 \times 1$$

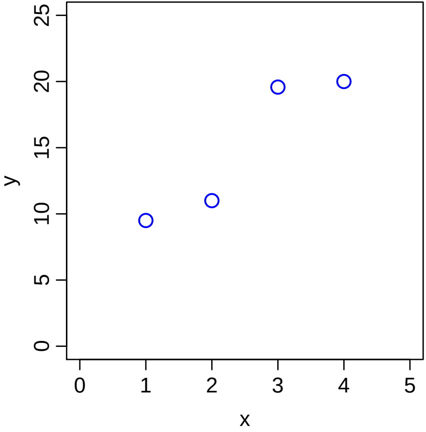
$$\beta_1 = 1; \beta_2 = 2; \beta_3 = 3$$

Residuals - variation is everywhere



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

Residuals - variation is everywhere

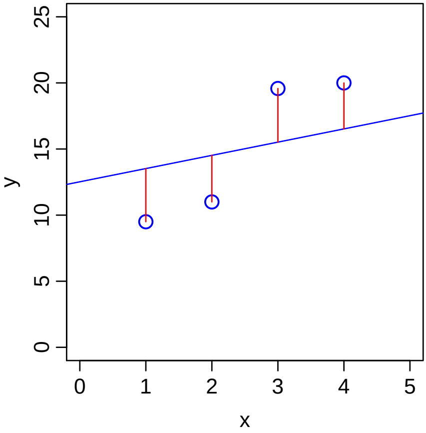


$$y = \beta_1 + \beta_2 x$$

$$9.50 = ? + ? \times 1$$
$$11.00 = ? + ? \times 2$$
$$19.58 = ? + ? \times 3$$
$$20.00 = ? + ? \times 4$$

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

Residuals - Guess 1

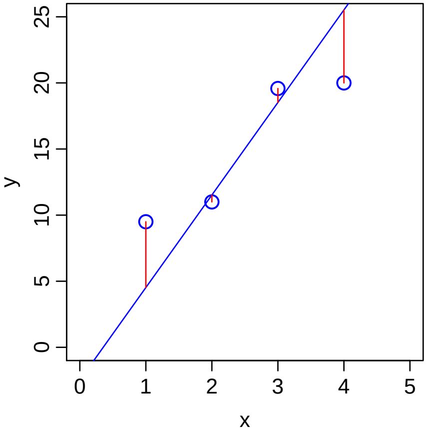


$$y = \beta_1 + \beta_2 x + \varepsilon$$

$$9.50 = 12.52 + 1 \times 1 - 4.02$$
$$11.00 = 12.52 + 1 \times 2 - 3.52$$
$$19.58 = 12.52 + 1 \times 3 + 4.06$$
$$20.00 = 12.52 + 1 \times 4 + 3.48$$

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

Residuals - Guess 2



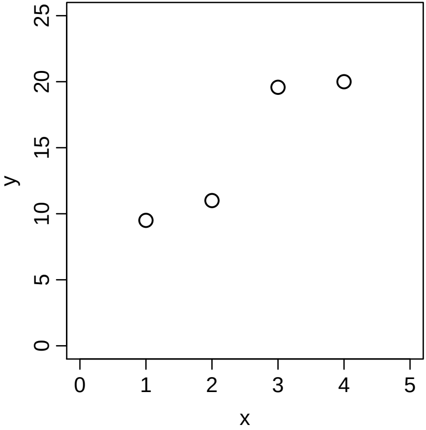
$$y = \beta_1 + \beta_2 x + \varepsilon$$

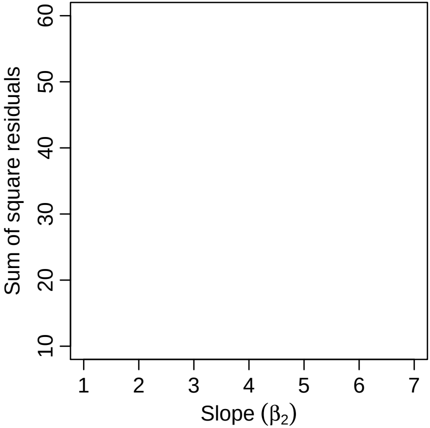
$$9.50 = -2.48 + 7 \times 1 + 4.98$$
$$11.00 = -2.48 + 7 \times 2 - 0.52$$
$$19.58 = -2.48 + 7 \times 3 + 1.06$$
$$20.00 = -2.48 + 7 \times 4 - 5.52$$

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

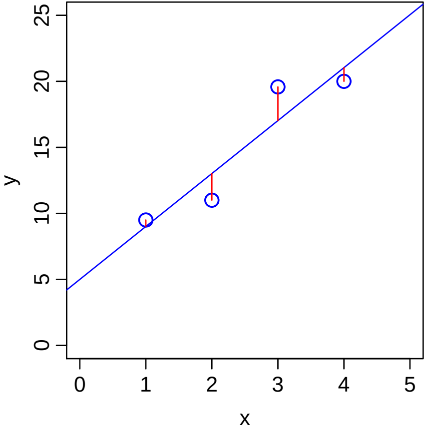
Residuals - least squares solution

Minimize the *sum* of the *squared* residuals





Why guess?: The 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

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

$$\overbrace{\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}}^{\text{Observed values}} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{\text{Model matrix}} \overbrace{\begin{bmatrix} 5 \\ 4 \end{bmatrix}}^{\text{Coefficients}} + \underbrace{\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}}_{\text{Residuals}}$$

Model as a matrix - terminology

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

$$\overbrace{\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}}^{\text{Observed values}} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{\text{Model matrix}} \overbrace{\begin{bmatrix} 5 \\ 4 \end{bmatrix}}^{\text{Coefficients}} + \underbrace{\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}}_{\text{Residuals}}$$

Model as a matrix - terminology

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

$$\overbrace{\begin{bmatrix}9.50\\11.00\\19.58\\20.00\end{bmatrix}}^{\text{Observed values}} = \underbrace{\begin{bmatrix}1&1\\1&2\\1&3\\1&4\end{bmatrix}}_{\text{Model matrix}} \overbrace{\begin{bmatrix}5\\4\end{bmatrix}}^{\text{Coefficients}} + \underbrace{\begin{bmatrix}0.50\\-2.00\\2.58\\-1.00\end{bmatrix}}_{\text{Residuals}}$$

Model as a matrix - terminology

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

$$\underbrace{\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}}_{\text{Observed values}} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{\text{Model matrix}} \underbrace{\begin{bmatrix} 5 \\ 4 \end{bmatrix}}_{\text{Coefficients}} + \underbrace{\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}}_{\text{Residuals}}$$

Model as a matrix - terminology

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

$$\underbrace{\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix}}_{\text{Observed values}} = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{\text{Model matrix}} \underbrace{\begin{bmatrix} 5 \\ 4 \end{bmatrix}}_{\text{Coefficients}} + \underbrace{\begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}}_{\text{Residuals}}$$

Model as a matrix - terminology

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

Given these . . .

. . . find the set of these. . .

$$\begin{bmatrix} 9.50 \\ 11.00 \\ 19.58 \\ 20.00 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 4 \end{bmatrix} + \begin{bmatrix} 0.50 \\ -2.00 \\ 2.58 \\ -1.00 \end{bmatrix}$$

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

Model as a matrix - predictions

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

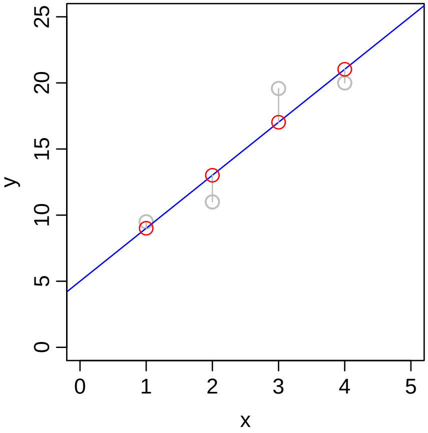
Predicted or fitted values

Coefficients

$$\begin{bmatrix} 9 \\ 13 \\ 17 \\ 21 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 4 \end{bmatrix}$$

Model matrix

Predicted values



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

$$9 = 5 + 4 \times 1$$
$$13 = 5 + 4 \times 2$$
$$17 = 5 + 4 \times 3$$
$$21 = 5 + 4 \times 4$$

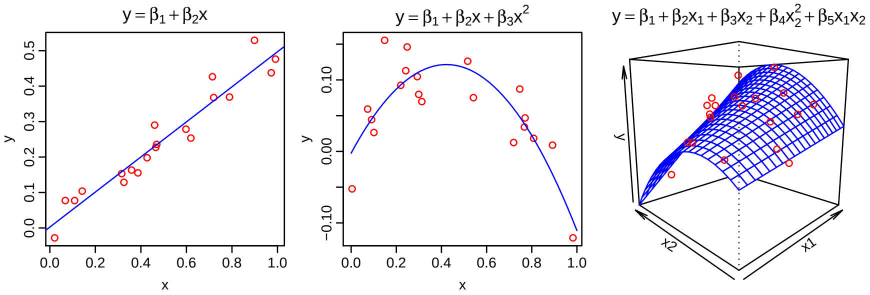
Assumptions

- 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

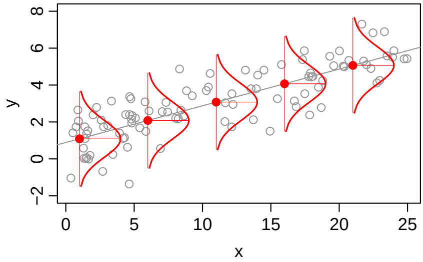
- Linear models have the following assumptions:
  - No measurement error in explanatory variables
  - The explanatory variables are not very highly correlated
  - The model is linear
  - The model has constant normal variance
- **If these assumptions are not met, the model can be very wrong**
- The last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

'The model is linear'

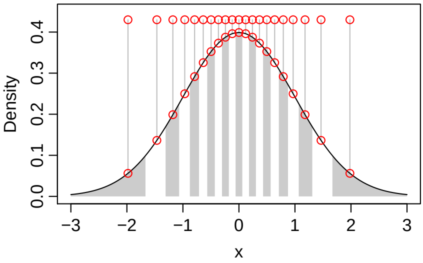


- These are *all* good linear models.
- Linear models can include curved relationships (e.g. polynomials)
- The data can be modelled as a *sum* of components
- A *linear combination* of variables and coefficients

‘The model has constant normal variance’

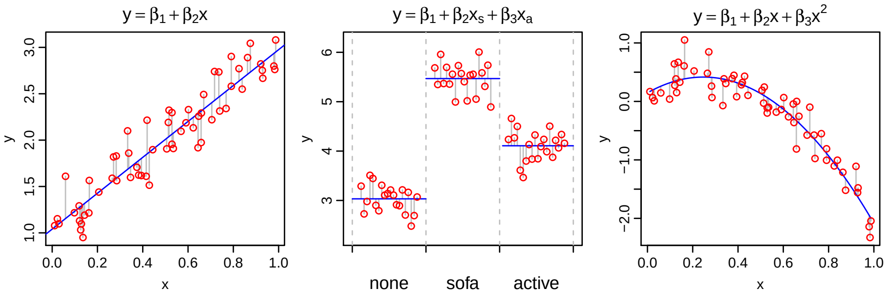


- The data has a similar spread around any predicted point in the model



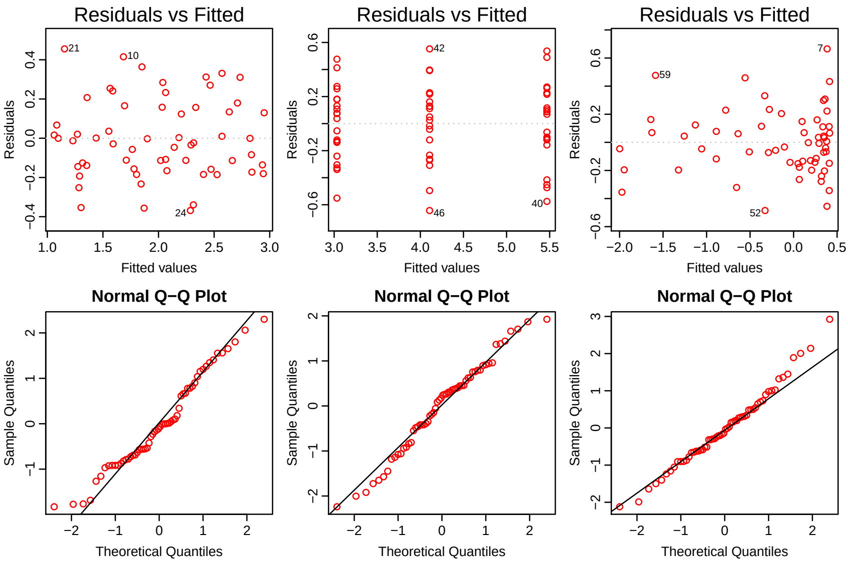
- The residuals are normal
- Points *should* be spaced equally in the area under the curve
- Expect mostly small but a few larger residuals

'The model has constant normal variance'

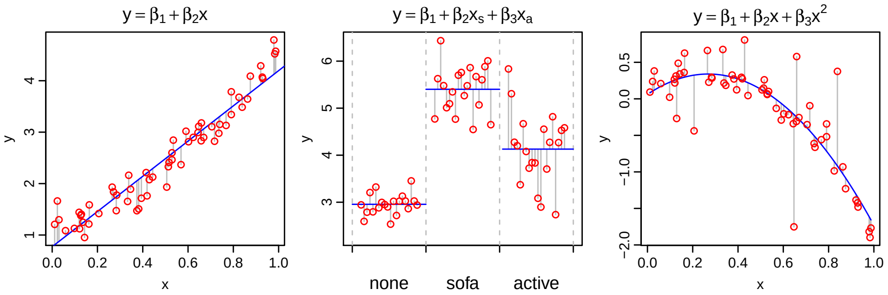


- Three good models
    - Is the spread the same for all fitted values?
    - Do the residuals match the normal expectation?

'The model has constant normal variance'

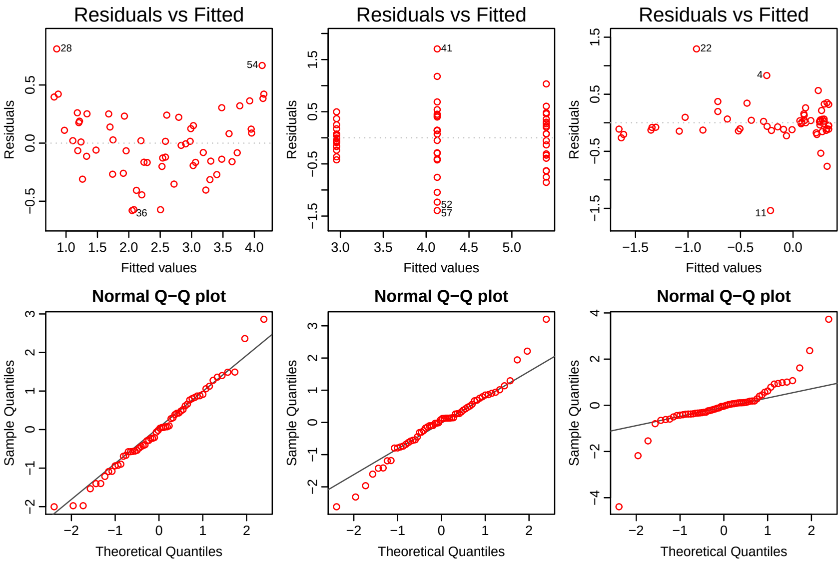


'The model has constant normal variance'



- 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!)

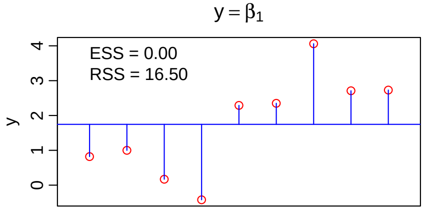
- *Terms*: analysis of variance
    - Does the model explain enough variation?
    - Does each term explain enough variation?

- *Coefficients*: t tests
    - Are the coefficients different from zero?

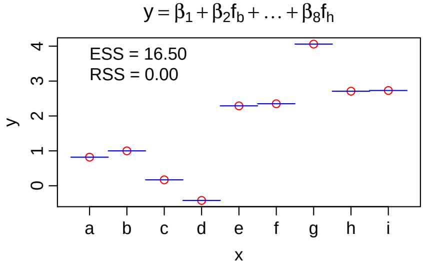
Null and over-specified models - two endpoints

- **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 and over-specified models - two endpoints

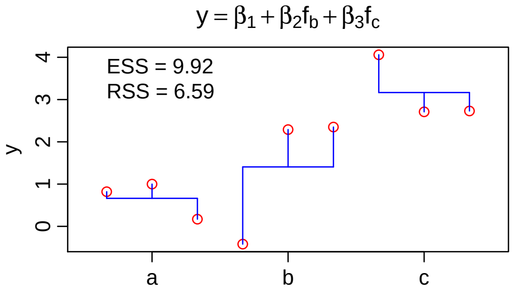


- The null model (H_0)
- Nothing is going on
- Biggest possible residuals
- Residual sum of squares (RSS) is as big as it can be



- 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

Large ESS is good

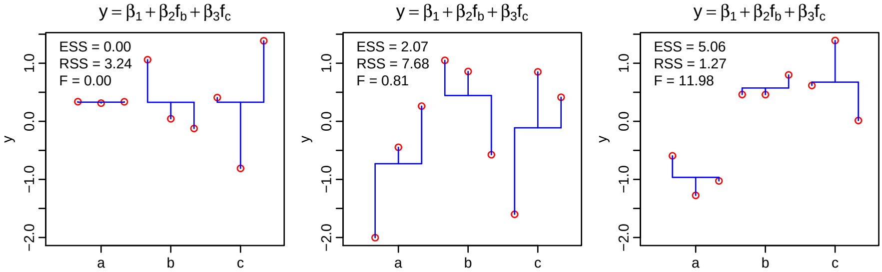
Fewer coefficients is better

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

Small RSS is good

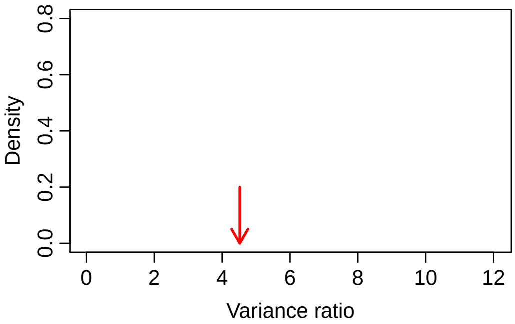
Residual degrees of freedom

F values by chance



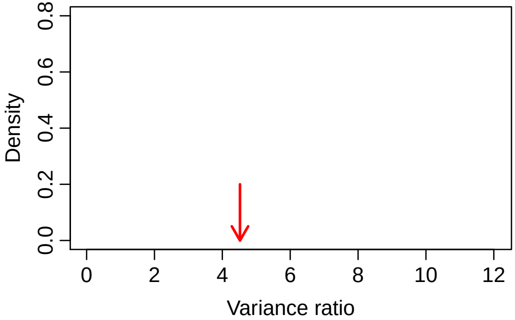
- What is the distribution of F if nothing is going on?
- Simulate 10,000 datasets where nothing is going on (H_0 is true)
- Calculate F for each random dataset under H_1
- Mostly H_1 has a low F - but sometimes it is high by chance

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?

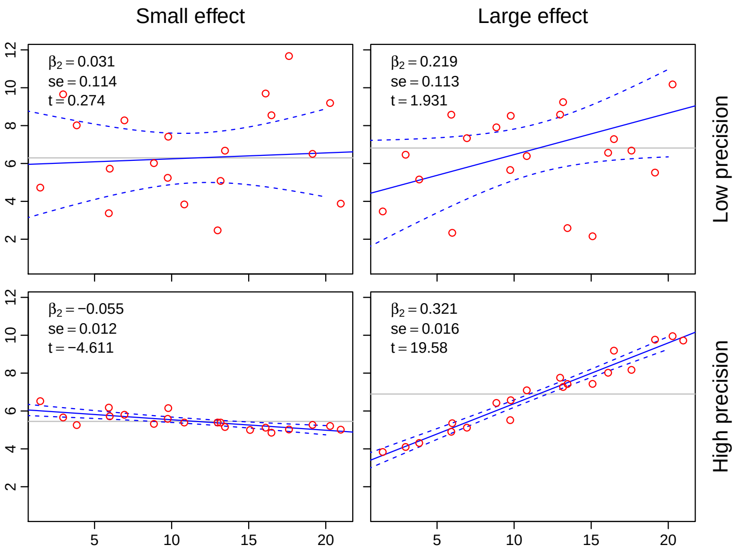
Large is good - bigger changes

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

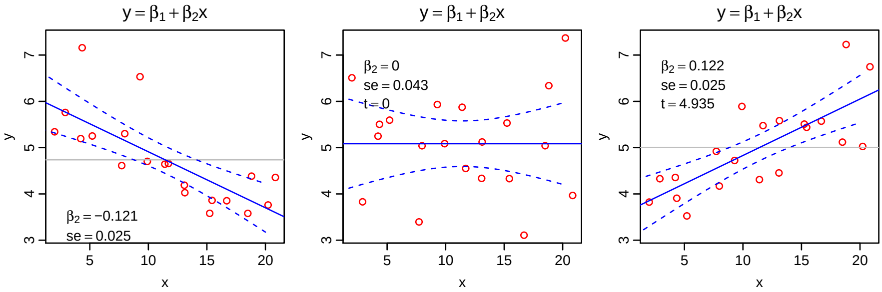
Small is good - known more precisely

- 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

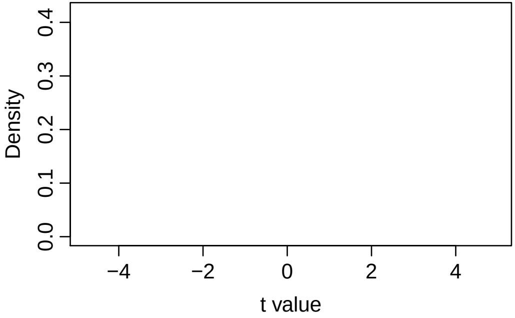


t values by chance



- 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

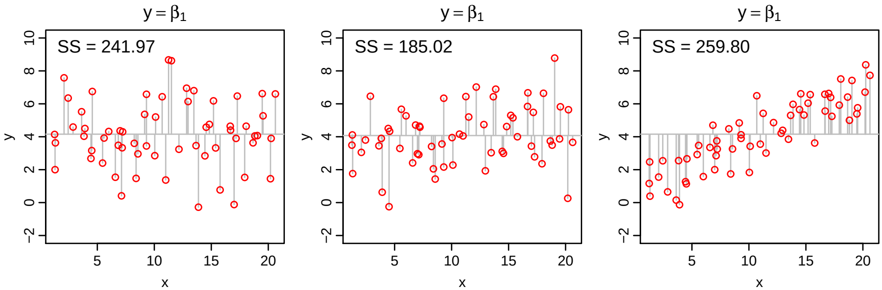


- 95% of the random data sets have $t \leq \pm 2.09$
- Only the two higher precision models are expected to occur less than 1 time in 20 by chance.

Summary

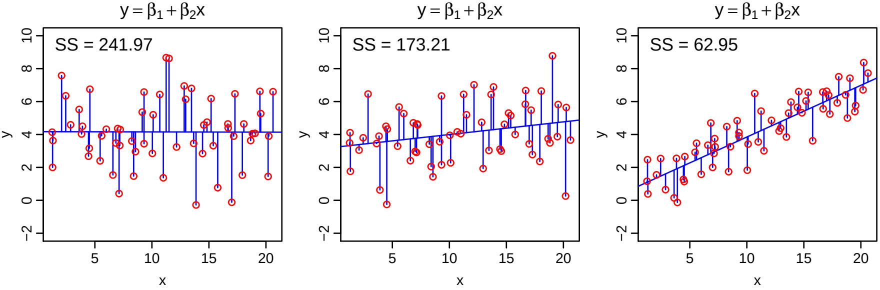
- Linear models predict a continuous response variable
- A sum based on the effect size of explanatory variables
- Estimate the model using least squares residuals
- Need to check if the model is appropriate
- Then check if the model is explanatory

What about Analysis of Variance (ANOVA)?



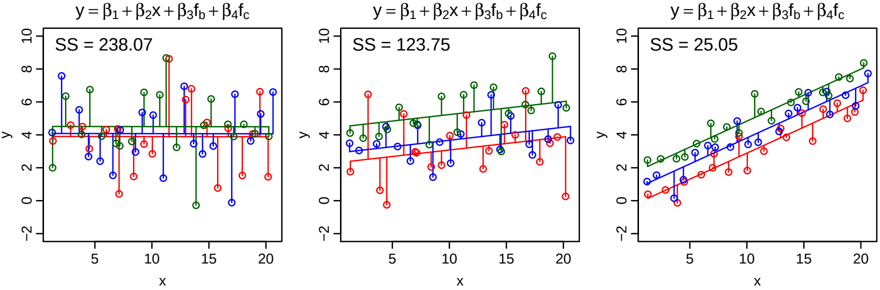
- The null hypothesis (H_0): Nothing is going on
- The residuals have to get smaller as we include terms.
- How much shorter?

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