

# Linear models

Samraat Pawar

*Department of Life Sciences (Silwood Park)*

**Imperial College**  
London

# LECTURE OUTLINE

## Topics:

- What is a linear model?
  - Regression
  - ANOVA
  - Multiple explanatory variables (ANCOVA)
- Fitting linear models to your data
- Is the fitted linear model appropriate for the data?
- How well does a fitted linear model explain the data?

## Concepts:

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

# WHAT PREDICTS THE WEIGHTS ( $w$ ) OF LECTURERS?

Use *intuition* and *prior knowledge* to identify the *variables* to 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 ( $c$ )

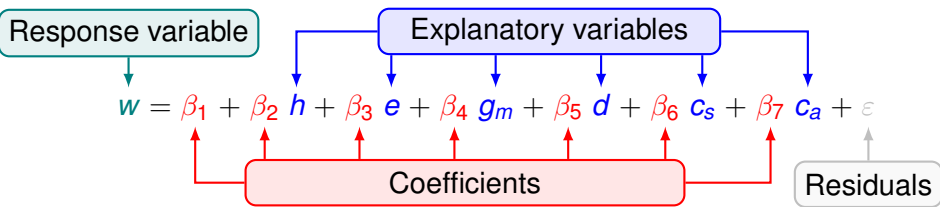
... and build a mathematical model:

Lecturer weight ( $w$ ) = *Combination of Independent Variables* (that determine  $w$ )

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

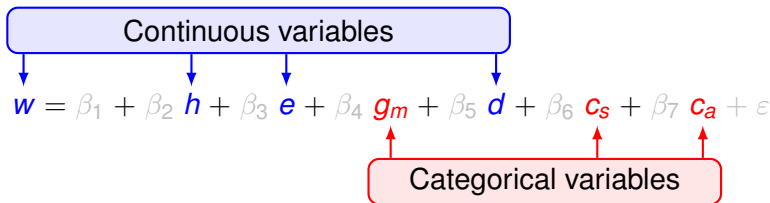
# THE LINEAR MODEL

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** ( $\epsilon$ )

# THE 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)

# THE TERMS AND COEFFICIENTS

The diagram illustrates a linear model equation:  $w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$ . Above the equation, explanatory variables are shown in boxes: 'Height' (blue), 'Exercise' (blue), 'Distance' (blue), 'Gender' (red), and 'Console' (red). Arrows point from these boxes to the corresponding terms in the equation: 'Height' to  $h$ , 'Exercise' to  $e$ , 'Distance' to  $d$ , 'Gender' to  $g_m$ , and 'Console' to  $c_s$  and  $c_a$ . The terms  $h$ ,  $e$ , and  $d$  are blue, while  $g_m$ ,  $c_s$ , and  $c_a$  are red.

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

- 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 $N - 1$ COEFFICIENTS? WHAT IS $\beta_1$ ?

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

- Two ways of thinking about  $\beta_1$ :
  - 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  $\beta_1$ :
  - along a continuous slope
  - as an offset to a different level

## SO, TO PUT IT SIMPLY,

Linear models are just a sum of *terms* that are *linear* in the *coefficients*:

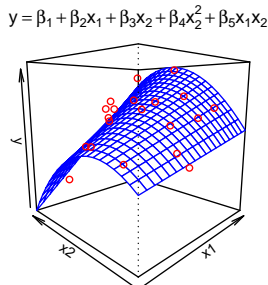
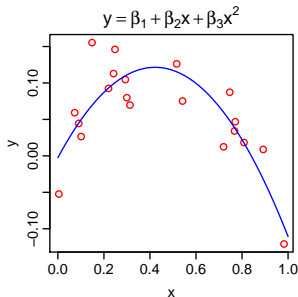
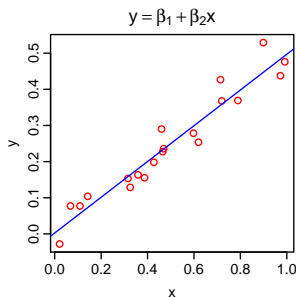
$$w = \beta_1 + \beta_2 h + \beta_3 e + \beta_4 g_m + \beta_5 d + \beta_6 c_s + \beta_7 c_a + \varepsilon$$

What our example linear model means (literally):

- The coefficient  $\beta_1$  is the baseline value of *weight* for *women* with *no games console*
- The model tells us how much (determined by the other coefficients) to add to this baseline weight ...
  - 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?

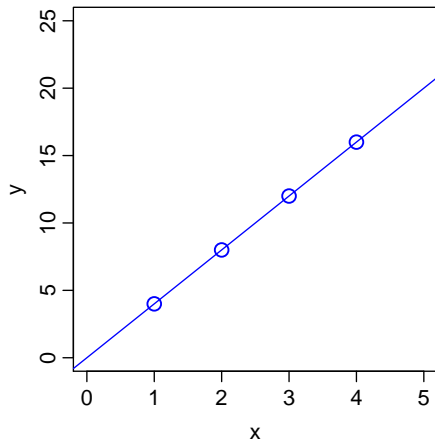


# EXAMPLES OF LINEAR MODELS



- These are *all* linear models (fitted to data)
- Each model a *sum of terms that are linear in coefficients*
- *Linear models can include curved relationships (e.g. polynomials) — this is a common point of confusion!*

# LINEAR MODEL WITH ONE CONTINUOUS VARIABLE



$$y = \beta_1 x$$

$$4 = 4 \times 1$$

$$8 = 4 \times 2$$

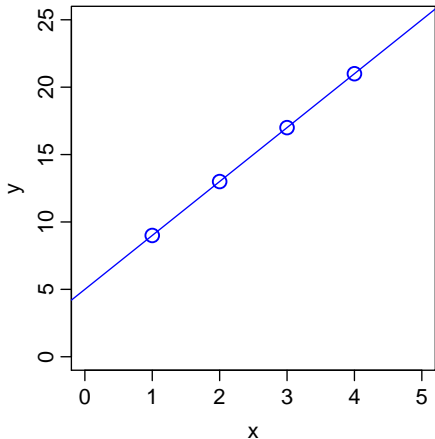
$$12 = 4 \times 3$$

$$16 = 4 \times 4$$

$$\beta_1 = 4$$

**Regression** with *known baseline value (intercept)*

# LINEAR MODEL WITH ONE CONTINUOUS VARIABLE



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

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

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

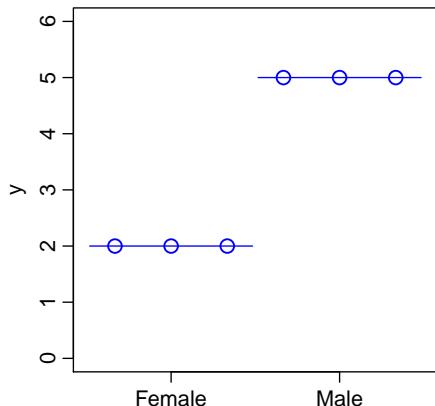
$$17 = 5 + 4 \times 3$$

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

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

**Regression** with *unknown baseline value (intercept)*

# LINEAR MODEL WITH ONE FACTOR (CATEGORICAL VARIABLE)



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

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

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

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

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

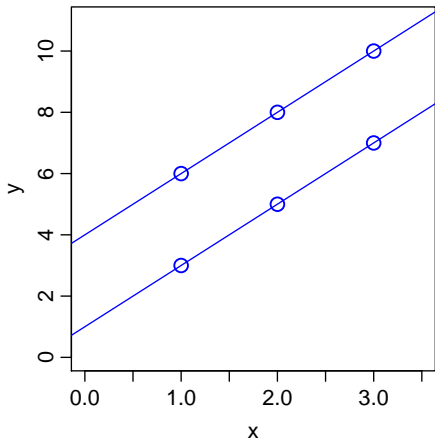
$$5 = 2 + 3 \times 1$$

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

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

*Analysis of Variance (ANOVA)*

# LINEAR MODEL WITH ONE CONTINUOUS VARIABLE AND ONE FACTOR



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

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

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

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

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

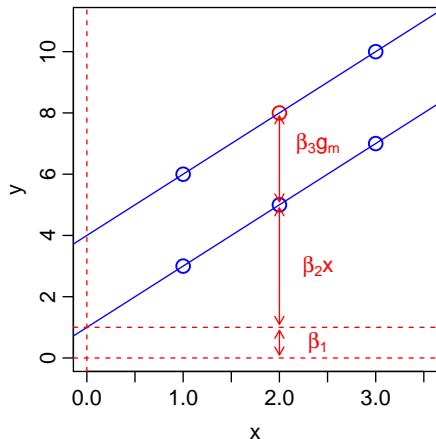
$$8 = 1 + 2 \times 2 + 3 \times 1$$

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

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

*Multiple Explanatory variables, Analysis of Covariance (ANCOVA)*

# CLOSER LOOK AT THE ANCOVA EXAMPLE



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

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

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

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

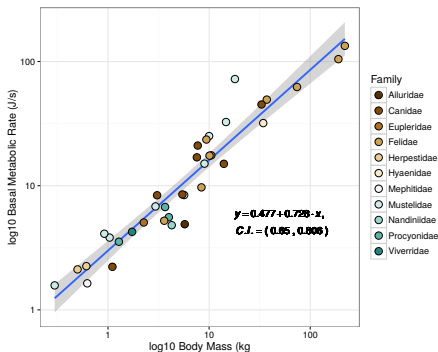
$$6 = 1 + 2 \times 1 + 3 \times 1$$

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

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

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

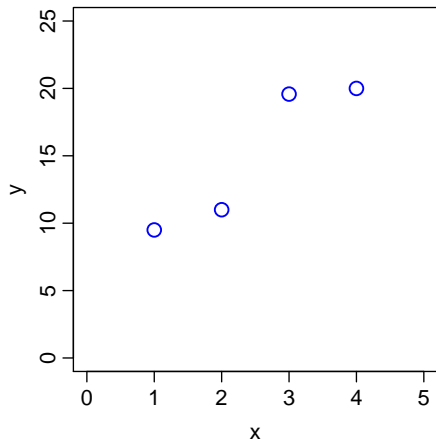
# “FITTING” A LINEAR MODEL TO DATA



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

# FITTING A LINEAR MODEL TO DATA



*What line best passes through  
(describes) these data?*

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

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

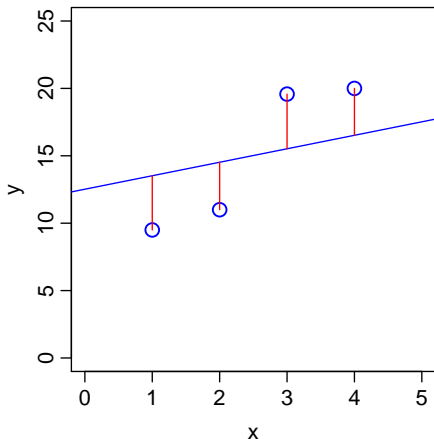
$$11.00 = ? + ? \times 2$$

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

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



# FITTING A LINEAR MODEL TO DATA: GUESS



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

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

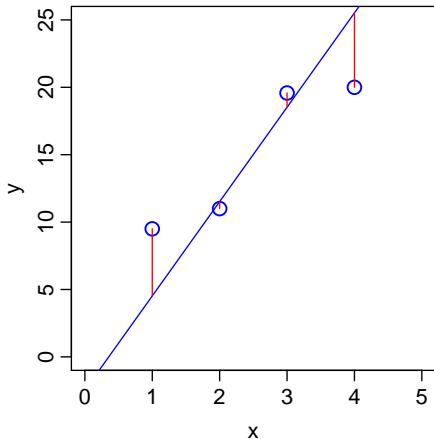
$$11.00 = 12.52 + 1 \times 2 - 3.52$$

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

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

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

# FITTING A LINEAR MODEL TO DATA: GUESS AGAIN!



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

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

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

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

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

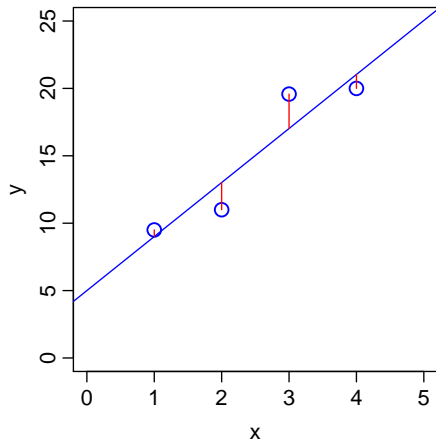
$$\beta_1 = -2.48; \beta_2 = 7$$

*There must be a better way to do this!*

# FITTING A LINEAR MODEL: LEAST SQUARES SOLUTION

Minimize the *sum* of the *squared residuals*:

# THE (ORDINARY) LEAST SQUARES FITTING SOLUTION



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

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

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

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

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

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

# THE MATHS MAGIC UNDER THE HOOD

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

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



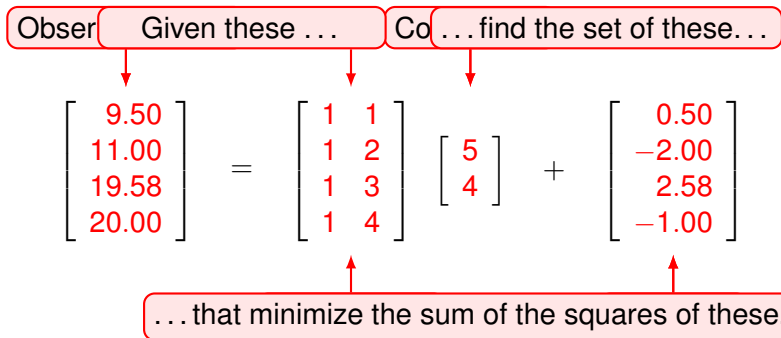
$$= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

Residuals



# THE MATHS MAGIC UNDER THE HOOD

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



# THE MATHS MAGIC UNDER THE HOOD

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

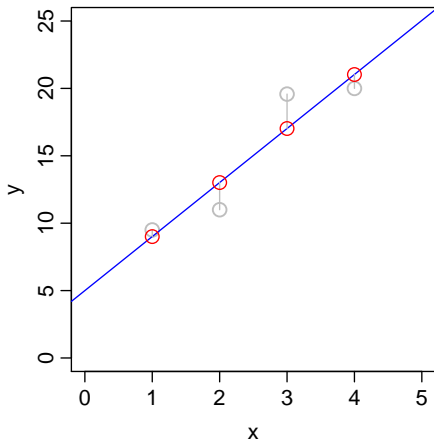
Predicted or fitted values

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 AND RESIDUALS



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

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

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

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

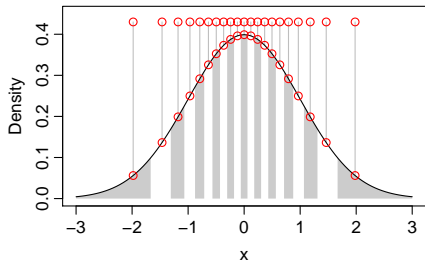
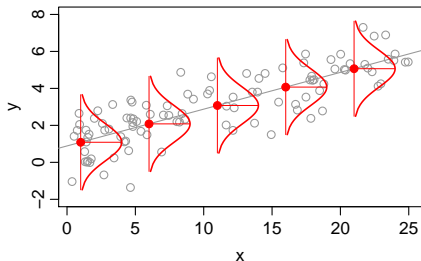
$$21 = 5 + 4 \times 4$$



# FITTING A LINEAR MODEL: ASSUMPTIONS

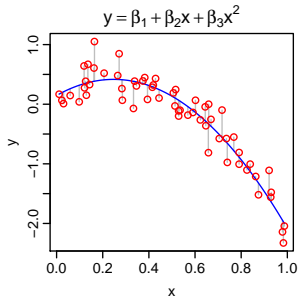
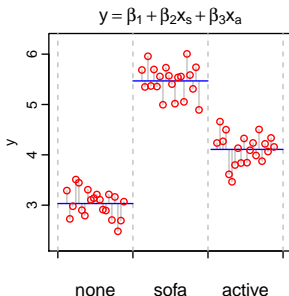
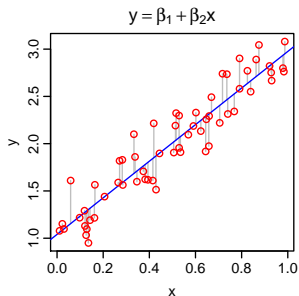
- Linear models are fitted with the following assumptions:
  - No measurement error in explanatory variables
  - The explanatory variables are not very highly (inter-) correlated
  - **The model has constant normal variance**
- **If these assumptions are not met, the model can be very wrong**
- The first two you will should consider *before* even fitting a linear model
- The last one needs can be tested *after* fitting a linear model

# 'THE MODEL HAS CONSTANT NORMAL VARIANCE'



- The data have a similar spread around any predicted point in the model
- Overall, the residuals are *normally distributed*: mostly small but a few larger values
- Points *should* be spaced so as to best capture the normal (gaussian) curve

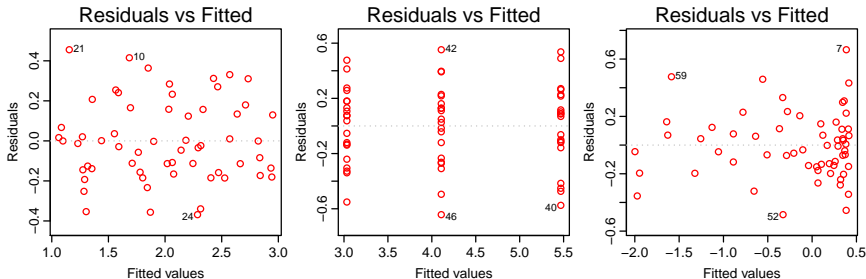
# CHECKING IF THE LINEAR MODEL IS APPROPRIATE



- All these three linear model fits appropriate for the data? Are assumptions of the linear model fit satisfied?
  - The spread of the real data around the fitted line (fitted values) is about the same across the x-axis – good
  - But are the residuals normally distributed?

# DIAGNOSTICS FOR A FITTED LINEAR MODEL

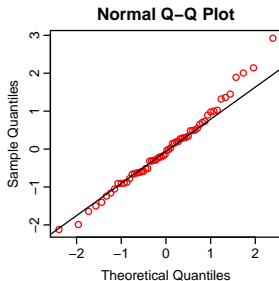
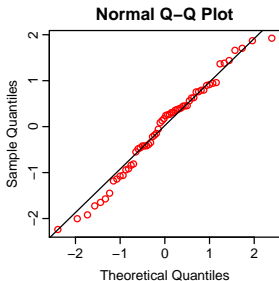
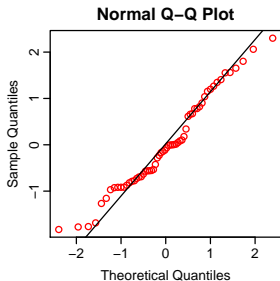
- *The spread of the real data around the fitted line (fitted values) is about the same across the x-axis*



- That is, the residuals have about the same spread irrespective of the fitted values
- The three numbered points in each plot are the three most 'badly behaved' data points.
  - Each number is the datum's row number in the R data frame

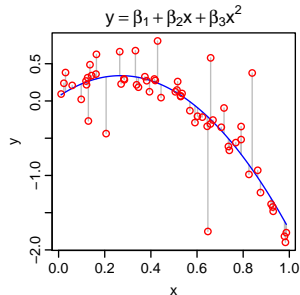
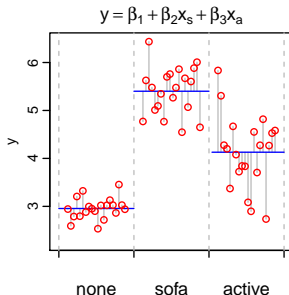
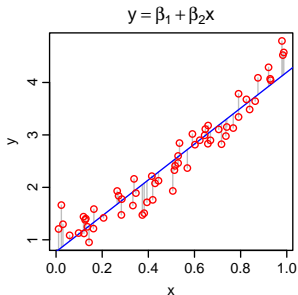
# DIAGNOSTICS FOR A FITTED LINEAR MODEL

- Are the residuals normally distributed?



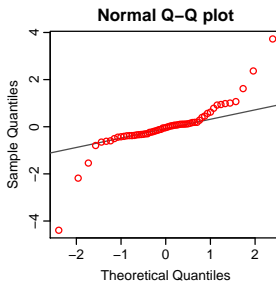
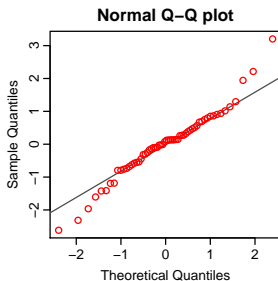
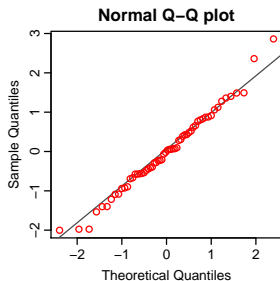
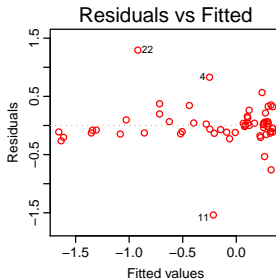
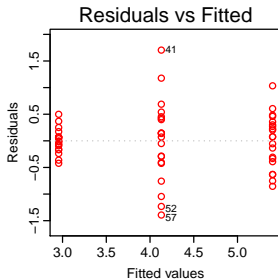
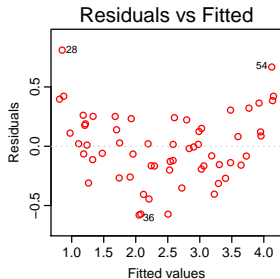
- Residuals from the first (simple regression) and third (polynomial) model's fits show some deviations from normality at the ends (high and low ends of their distributions), but it's acceptable

# THREE BAD LINEAR MODEL FITS



- These are three bad linear model fits
  - The data spread is not the same for all fitted values
  - The first model clearly spread is not the same for all fitted values
  - Are the residuals normally distributed?

# DIAGNOSTICS FOR A (BADLY) FITTED LINEAR MODEL



# IS A LINEAR MODEL APPROPRIATE?

Plot the data!  
Plot the residuals!



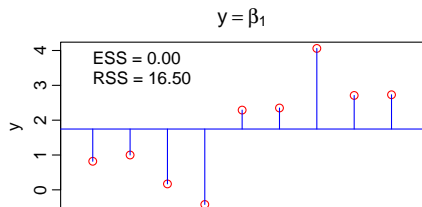
# HOW EXPLANATORY IS THE FITTED LINEAR MODEL?

- The role of  $F$  and  $t$  tests in Linear Model fitting
- Significance of *Terms*:  $F$  test
  - Does the model explain enough variation?
  - Does each term explain enough variation?
- Significance of *Coefficients*:  $t$  tests
  - Are the coefficients different from zero?

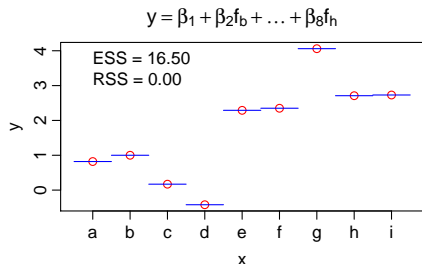
# IS THE FITTED LINEAR MODEL SIGNIFICANT?: $F$ TEST

- **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  $\hat{y}$  (residuals), 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

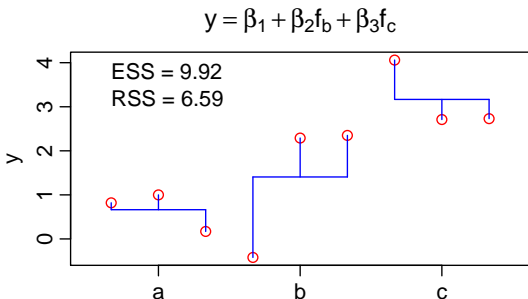


- The null model ( $H_0$ )
- Nothing is going on
- Biggest possible residuals
- Residual sum of squares (RSS) is as big as it can be



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

# THE 'RIGHT' (INTERESTING) MODEL



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

# F STATISTIC OF THE FITTED LINEAR MODEL

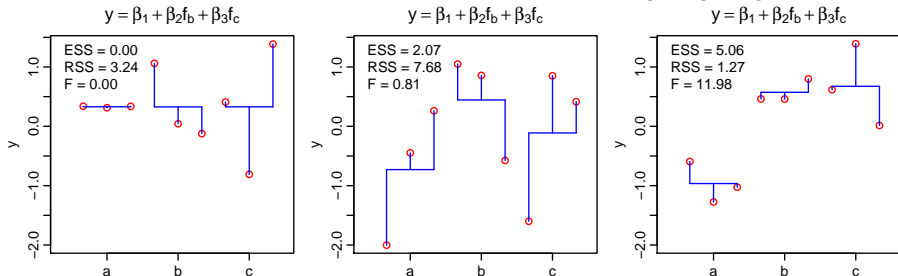
The diagram illustrates the components of the F-statistic formula. Four blue-bordered text boxes provide context for the terms in the formula:

- Large ESS is good**: Points to the ESS term in the numerator.
- Fewer coefficients is better**: Points to the  $N_c$  term in the numerator.
- Small RSS is good**: Points to the RSS term in the denominator.
- Residual degrees of freedom: larger sample size is better**: Points to the  $N_r$  term in the denominator.

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

# WHAT IT REALLY MEANS: $F$ VALUE BY CHANCE?

*What would be 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  $H_1$
- $H_1$  typically has a low  $F$  – but sometimes it is high *by chance*

## WHAT IT REALLY MEANS: $F$ VALUE BY CHANCE?

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

# ARE THE COEFFICIENTS DIFFERENT FROM ZERO?

Large is good: bigger changes

↓ ↓

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

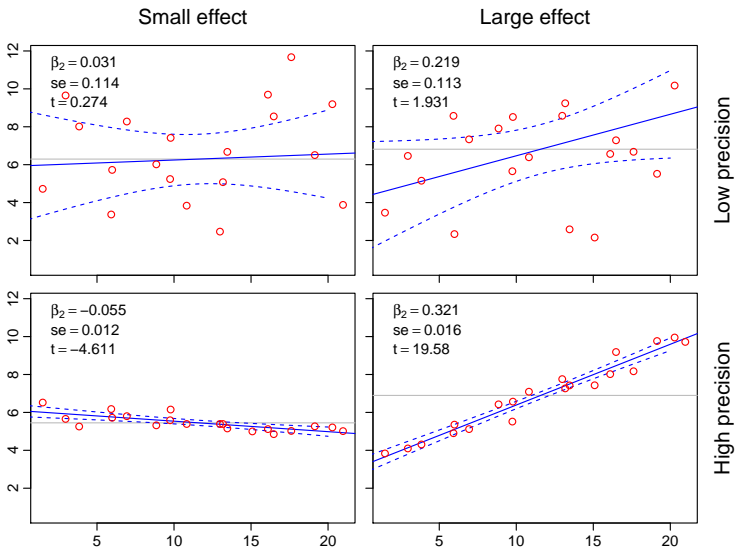
↑ ↑

Small is good: known more precisely

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

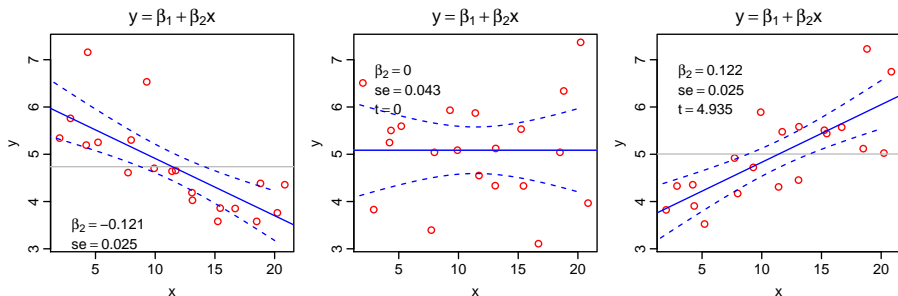


# VARIATION IN EFFECT SIZE AND PRECISION



# WHAT IT REALLY MEANS: $t$ VALUES BY CHANCE

*What is the distribution of  $t$  if nothing is going on?*

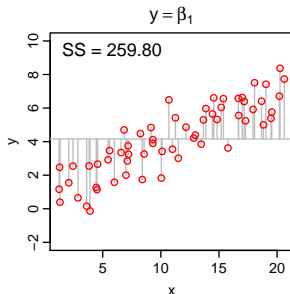
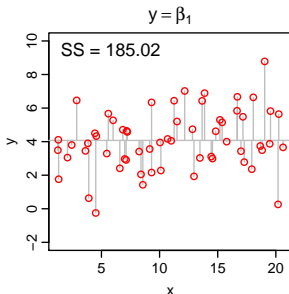
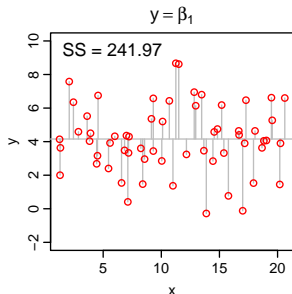


- Simulate 10,000 datasets where nothing is going on ( $H_0$  is true)
- Calculate  $t$  for each random dataset under  $H_1$
- $H_1$  typically has a  $t$  near zero but can be strongly positive or negative *by chance*

# 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.

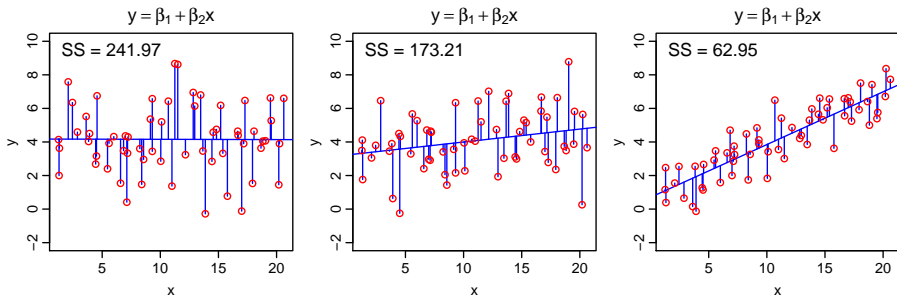
# SOME MORE EXAMPLES OF LINEAR MODEL FITTING



- The null hypothesis ( $H_0$ ): Nothing is going on (model is just  $\beta_1$ !)
- The residuals (and therefore, RSS) will get *smaller* as we include more terms to the model
- *How much smaller is enough?*

# SOME MORE EXAMPLES OF LINEAR MODEL FITTING

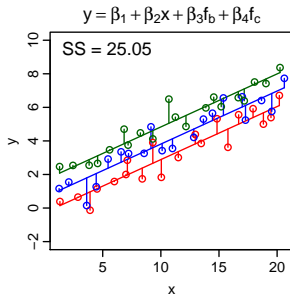
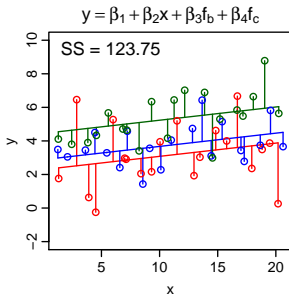
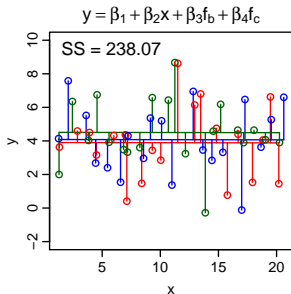
*First try: Add one continuous term*



- Fitted an *alternative* model ( $H_1$ ) using a predictor variable  $x$
- i.e., Added one term ( $x$ ) to the model to give ( $H_1$ )
- Do we reject  $H_0$  and accept this new model?

# SOME MORE EXAMPLES OF LINEAR MODEL FITTING

*Second try: Add one continuous term*



- Fitted another model ( $H_2$ ) with continuous predictor  $x$  and factor  $f$
- The RSS gets still smaller
- Is this *even* better than  $H_1$ ?

# COMPARE THE THREE MODELS

		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

- Which model would you choose between  $H_1$  and  $H_2$ ?
- Every alternative model is an *alternative hypothesis*

# LINEAR MODELS: SUMMARY

- Linear models predict a continuous response variable
- A LM is a sum of terms that are linear in the coefficients capturing the effect sizes of explanatory variables
- LMs are fitted using (ordinary) least squares — minimizes sum of squared residuals
- Need to check if the fitted LM is appropriate
- Then check if the LM is explanatory
- Fitting alternative LMs = Testing alternative hypotheses