Lesson 1

The Linear Regression Model

Simple linear regression

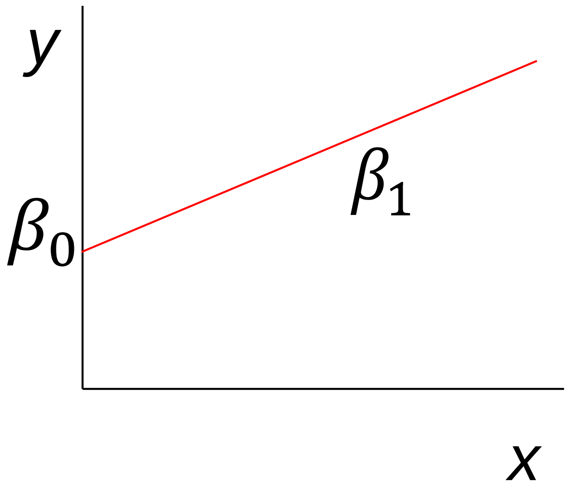
- **The linear model:**
- Can use a linear model to describe the relationship between x and y:

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

y_i is the response variable
x_i is the predictor variable.

β_0 is the intercept (the value of *y* when $x = 0$)

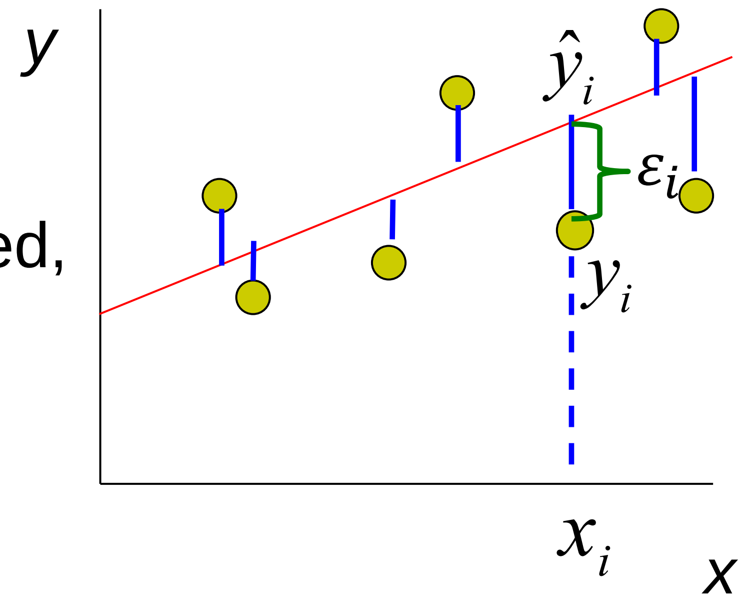
β_1 is the slope of the regression (the change in *y* for every unit of *x*)



Simple linear regression

- **The linear regression model:**
- Can use a linear regression to fit a best linear relationship between x and y for some data:

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



Here:
i are pairs of (x,y) values measured,

y_i is the response variable
x_i is the predictor variable

e_i are the distances between observed y_i and predicted $\hat{y}_i$

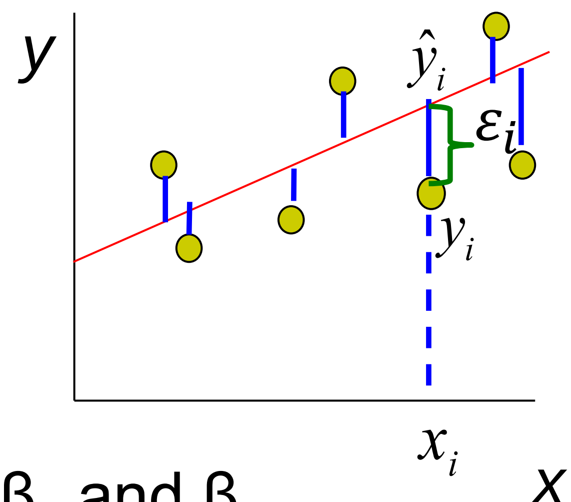
Regression calculation

- Simple regression uses the method of least squares estimation (OLS):

regression fits the best line through the data by minimising the sum of squared errors:

$$Min \sum_{1}^{n} (y_i - \hat{y}_i)^2 = Min \sum_{1}^{n} \varepsilon^2$$

$$Min \sum_{1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$



- This requires differentiation w.r.t. β_0 and β_1.

Finding the coefficients, β_i

- Take partial derivatives wrt to β_0 and β_1

$$\frac{\partial}{\partial \hat{\beta}_0}:\ 2\sum_1^n (y - \hat{\beta}_0 - \hat{\beta}_1 x) = 0 \quad \lll\ggg \quad \hat{\beta}_0 n = \sum_1^n y - \hat{\beta}_1 \sum_1^n x$$

$$\frac{\partial}{\partial \hat{\beta}_1}:\ 2\sum_1^n x\,(y - \hat{\beta}_0 - \hat{\beta}_1 x) = 0 \quad \lll\ggg \quad \hat{\beta}_0 \sum_1^n x = \sum_1^n xy - \hat{\beta}_1 \sum_1^n x^2$$

- Solve for β_1 by equating β_0

$$\beta_1 = \frac{\sum_1^n xy - \frac{\sum_1^n y \sum_1^n x}{n}}{\sum_1^n x^2 - \frac{\sum_1^n x \sum_1^n x}{n}} \qquad n\bar{x} = \sum_1^n x \ \ and \ \ n\bar{y} = \sum_1^n y$$

$$\beta_1 = \frac{\sum_1^n xy - n\bar{x}\bar{y}}{\sum_1^n x^2 - n\bar{x}\bar{x}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

Finding the coefficients, β_i

- To calculate the slope (β_1) we need to calculate:

$$\beta_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

- This can also be written as:

$$\beta_1 = (S_{xx})^{-1} S_{xy}$$

- To calculate intercept (β_0):

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

Estimating the residual variance

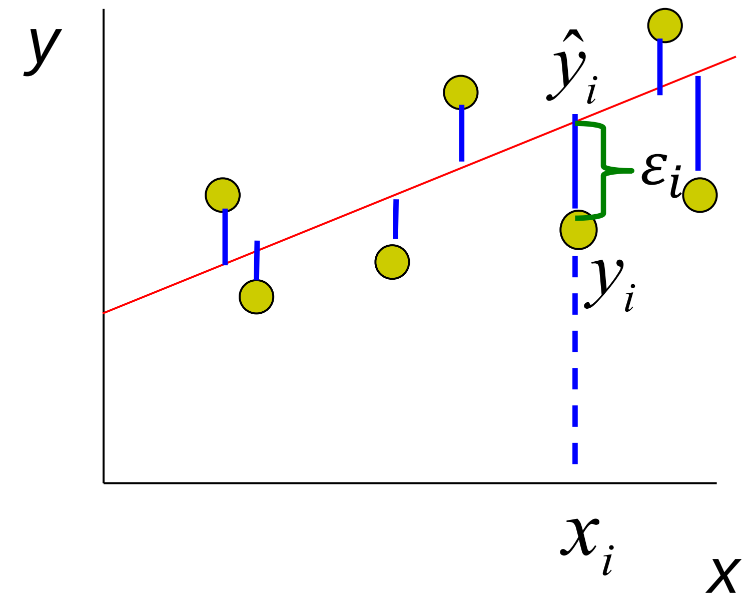
- Statistical tests require the residual variance of the model to quantify the uncertainty (s.e.) of β_i

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}$$

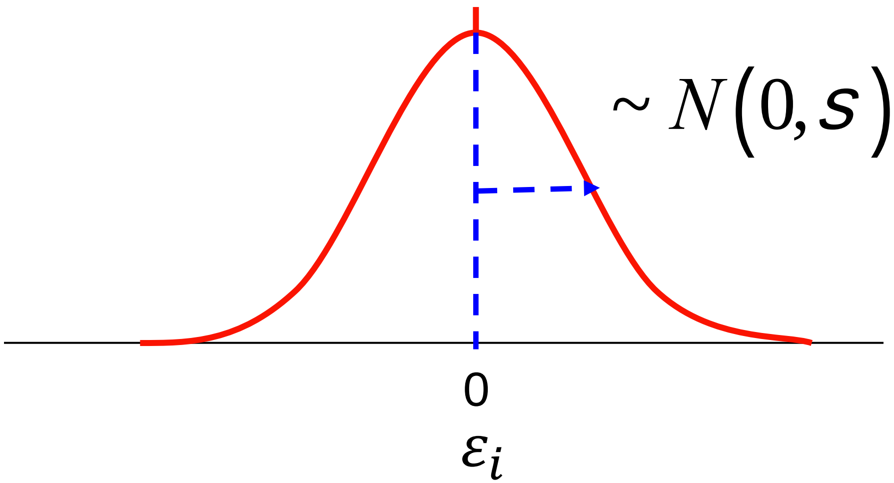
$$= \frac{\sum\left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{n-2}$$

The normal distribution

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



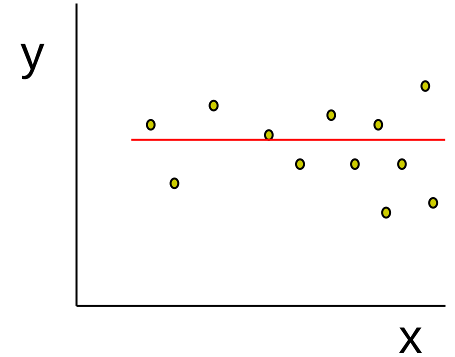
- The residuals, ε_i, of this model are assumed to follow the normal distribution with mean μ, and variance σ^2.



$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left(-\frac{1}{2\sigma^2}\varepsilon_i^2\right)}$$

Hypothesis testing

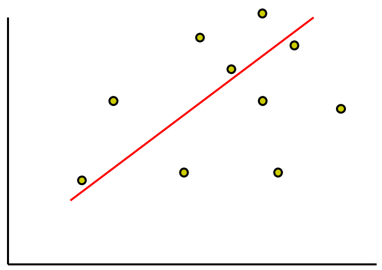
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



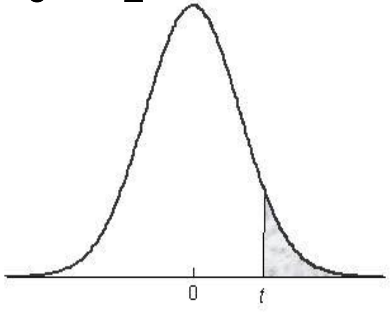
- Main hypothesis: the slope, $\beta_1 \neq 0$

Coefficients:

	Estimate	s.e.	t value	Pr(>|t|)
(Intercept)	22.440	3.4841	6.441	1.55e-05 ***
Temperature	1.015	0.2379	4.268	0.000781 ***



- H_0: $\beta_1 = 0$



$$t_{n-2\,df} = \frac{\widehat{\beta_1}}{se(\widehat{\beta_1})} = \frac{\widehat{\beta_1} - 0}{\sqrt{\frac{\sigma^2}{S_{xx}}}}$$

Interpreting linear models

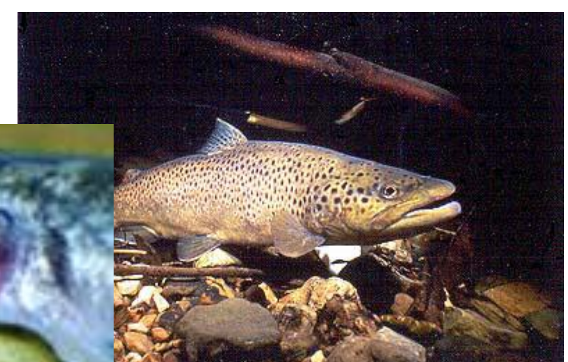
- Linear models can deal with both continuous and categorical predictor variables simultaneously

- Ordinal predictors = **variates (e.g. age)**
  - (continuous/discrete)

- Categorical predictors = **factors (e.g. sex)**

Example 1A

- A researcher measures the maximum swimming speed of 10 brown trout and 10 Canterbury galaxias at a range of temperatures.
- Response (Y) = swimming speed
- Predictor Factor = species → dummy variable (D)
- Predictor Variate (X_1) = temperature



Example 1A

		X_1	Y	D
Obs. No.	Species	Temperature (ºC)	Speed (cm/s)	Dummy Variable
1	Trout	3	48.1	0
2	Trout	6	51.2	0
3	Trout	11	73.1	0
4	Trout	12	78.1	0
5	Trout	15	81.1	0
6	Trout	21	94.5	0
7	Trout	24	99.0	0
8	Trout	26	115.3	0
9	Trout	28	113.7	0
10	Trout	30	118.7	0
11	Galaxias	4	26.1	1
12	Galaxias	9	33.7	1
13	Galaxias	12	31.4	1
14	Galaxias	14	38.5	1
15	Galaxias	15	36.9	1
16	Galaxias	17	39.1	1
17	Galaxias	18	42.2	1
18	Galaxias	21	43.3	1
19	Galaxias	25	58.9	1
20	Galaxias	28	54.3	1

Example 1.1

Open the data “fishspeed.csv” in R

(1) Run speed as a function of temperature using lm(). Write out the model.

(2) Run speed as a function of species using lm(). Write out the model.

(3) Compare fishspeed of species using t.test() (please specify: var.equal=TRUE)

(4) Compare (2) and (3)

Example 1A: some output

- Note that the lm() output gives you the β estimates with standard errors plus t-tests evaluating significance

Coefficients:

	Estimate	s.e.	t value	Pr(>|t|)
(Intercept)	17.340	11.967	1.449	0.1667
Temperature	2.979	0.677	4.401	0.0004 ***

- The output also gives you an R^2 value, which is a goodness-of-fit measure (how well model explains data)

- It also gives you the residual standard error = sqrt(σ^2), from which the coefficient s.e.'s are derived.

Matrix calculations

- How regression models actually calculate the parameters for linear models

Linear models in Matrix notation

- Linear models written in expanded form

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

- Or in compound matrix notation

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

- Where y is the vector of response values, capital X is a matrix of predictor values, β is a vector of regression coefficients ($\boldsymbol{\beta}_0$,$\boldsymbol{\beta}_1$), and **ε** is a vector of residuals.
- Matrix algebra allows us to solve linear regressions simply and powerfully

Finding the coefficients, β_i

- Recall for simple linear regression (one predictor):

$$\beta_1 = (S_{xx})^{-1} S_{xy}$$

- This can be solved for p predictors using matrix algebra:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccc} \text{Int.} & \text{X1} & \text{X2} \end{array} \\ \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 450 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{bmatrix} \end{array}, \quad \mathbf{y} = \begin{array}{c} \text{Y} \\ \begin{bmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \\ 8.00 \\ 17.83 \\ 79.24 \\ 21.50 \\ 40.33 \\ 21.00 \\ 13.50 \\ 19.75 \\ 24.00 \\ 29.00 \\ 15.35 \\ 19.00 \\ 9.50 \\ 35.10 \\ 17.90 \\ 52.32 \\ 18.75 \\ 19.83 \\ 10.75 \end{bmatrix} \end{array}$$

The $\mathbf{X'X}$ matrix is

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 150 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10{,}232 \\ 219 & 3{,}055 & 133{,}899 \\ 10{,}232 & 133{,}899 & 6{,}725{,}688 \end{bmatrix}$$

and the $\mathbf{X'y}$ vector is

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 7 & 3 & \cdots & 4 \\ 560 & 220 & \cdots & 150 \end{bmatrix} \begin{bmatrix} 16.68 \\ 11.50 \\ \vdots \\ 10.75 \end{bmatrix} = \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

The least-squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

or

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10{,}232 \\ 219 & 3{,}055 & 133{,}899 \\ 10{,}232 & 133{,}899 & 6{,}725{,}688 \end{bmatrix}^{-1} \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 0.11321518 & -0.00444859 & -0.00008367 \\ -0.00444859 & 0.00274378 & -0.00004786 \\ -0.00008367 & -0.00004786 & 0.00000123 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7{,}375.44 \\ 337{,}072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix}$$

Finding the variance, $\hat{\sigma}^2$

- We can calculate the residual variance for n sampling units and p coefficients $\underline{\hat{\beta}}$:

$$\hat{\sigma}^2 = \frac{SS_{res}}{n-p} = \frac{\boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{n-p}$$

- Why?

$$\begin{aligned} SS_{res} &= \left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}\right)'\left(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X}\right) \\ &= \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}' - \boldsymbol{y}'\widehat{\boldsymbol{\beta}}\boldsymbol{X} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \quad \text{where} \quad \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y} \end{aligned}$$

Thus: $\hat{\sigma}^2 = \frac{SS_{res}}{n-p} = \frac{\boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{n-p}$

Example1.2

Matrix OLS solution of LM

Use the fishspeed data to solve the beta coefficients for:

(1) Speed ~ Temperature

(2) Speed ~ Species

Multi-linear models

- The underlying linear models can be extended from simple linear models

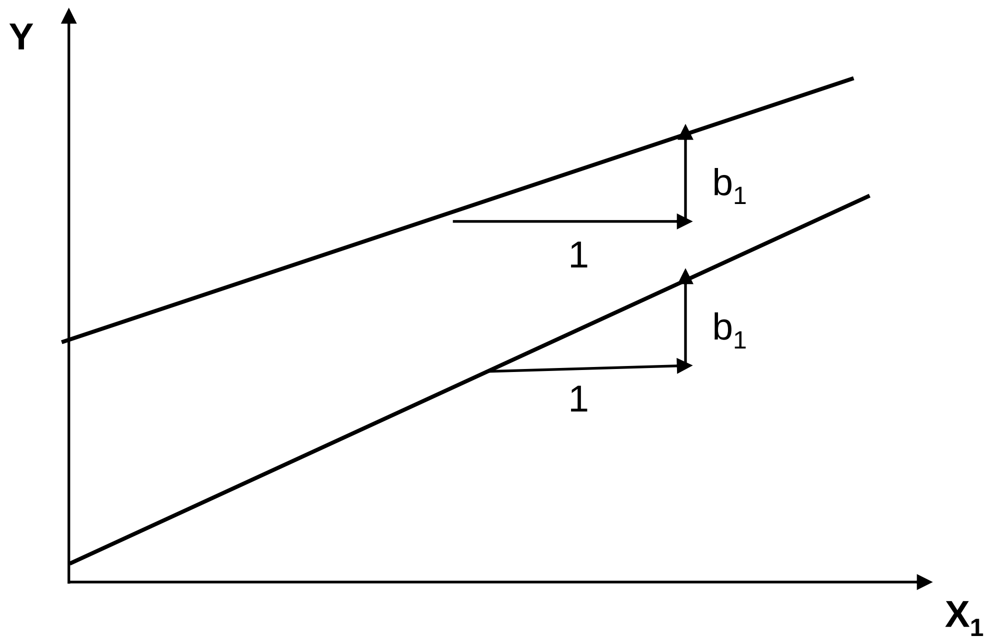
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- to models with multiple variables and interactions.

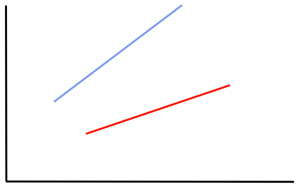
$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{ijk}$$

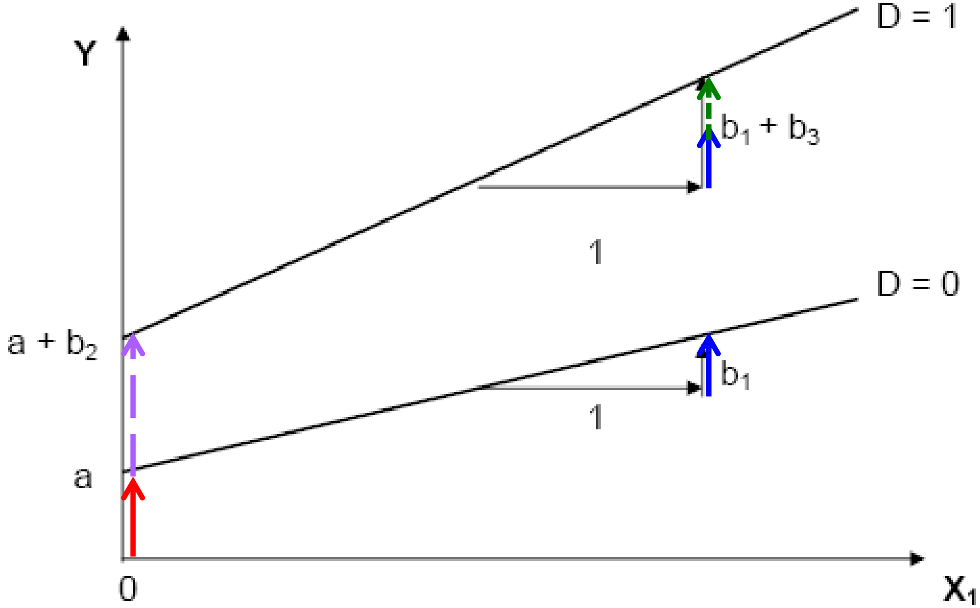
Interpreting multilinear models

- Regression models with factors and variables have separate regression lines for each factor level



Modelling interactions





- a = intercept for level 1 (D=0)
- b_1 = regression slope for level 1
- b_2 = difference in intercepts of the two lines = $Y_1 – Y_2$ when X =0
- b_3 = difference in slopes of the two lines

Fishspeed Example

- A researcher measures the maximum swimming speed of 10 brown trout and 10 Canterbury galaxias at a range of temperatures.
- NOW he wants to know whether the two fish species respond differently to temperature.

=> INTERACTION

Recall the data

		X_1	Y	D
Obs. No.	Species	Temperature (°C)	Speed (cm/s)	Dummy Variable
1	Trout	3	48.1	0
2	Trout	6	51.2	0
3	Trout	11	73.1	0
4	Trout	12	78.1	0
5	Trout	15	81.1	0
6	Trout	21	94.5	0
7	Trout	24	99.0	0
8	Trout	26	115.3	0
9	Trout	28	113.7	0
10	Trout	30	118.7	0
11	Galaxias	4	26.1	1
12	Galaxias	9	33.7	1
13	Galaxias	12	31.4	1
14	Galaxias	14	38.5	1
15	Galaxias	15	36.9	1
16	Galaxias	17	39.1	1
17	Galaxias	18	42.2	1
18	Galaxias	21	43.3	1
19	Galaxias	25	58.9	1
20	Galaxias	28	54.3	1

Example: 1.3
The Fishspeed data

- Formulate the linear model
- Run the regression model in R
- Write out the solution model

Results: coefficients

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.4405	3.4841	6.441	1.55e-05 ***
Temperature	1.0152	0.2379	4.268	0.000781 ***
SpeciesTrout	18.3559	4.2037	4.367	0.000645 ***
SpeciesTrout:Temperature	1.6259	0.2660	6.113	2.68e-05 ***

- Model for galaxias:

Speed = 22.440 + (1.015*Temp)

- Model for trout:

Speed = 22.440 + (1.015*Temp) + (18.356) + (1.626*Temp)

Speed = 40.796 + (2.641*Temp)

Results: coefficients

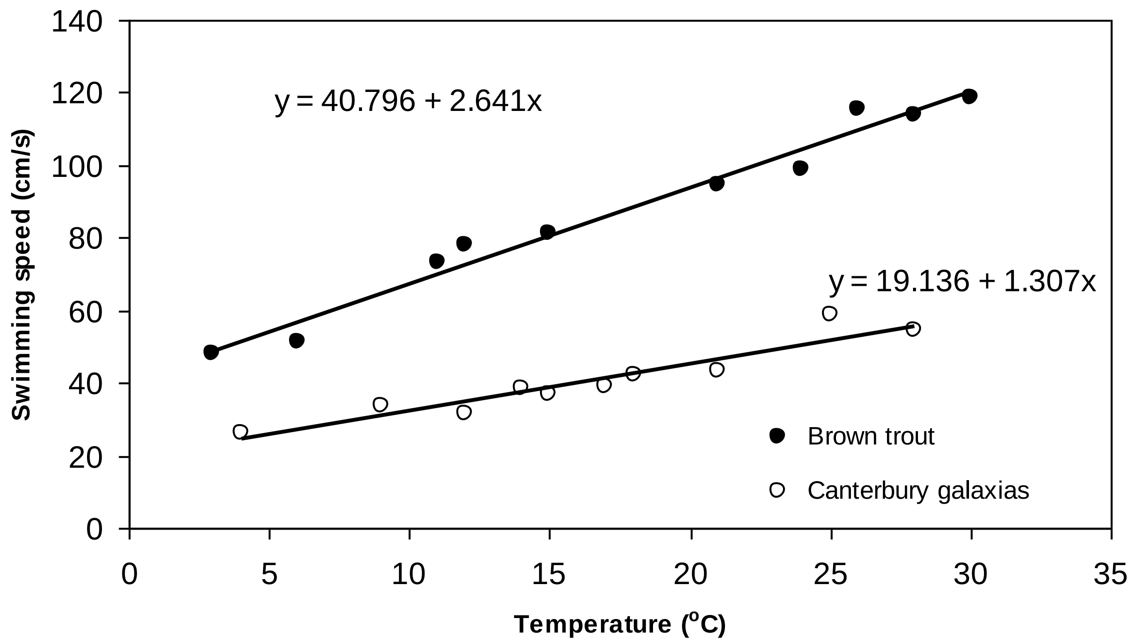
Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	22.4405	3.4841	6.441	1.55e-05 ***
Temperature	1.0152	0.2379	4.268	0.000781 ***
SpeciesTrout	18.3559	4.2037	4.367	0.000645 ***
SpeciesTrout:Temperature	1.6259	0.2660	6.113	2.68e-05 ***

- $b_1 > 0$: swimming speed of galaxias increases with temp.
- $b_2 > 0$: swimming speed of trout is greater than that of galaxias at 0^oC (intercept)
- $b_3 > 0$: temperature has higher effect on swimming speed of trout than on galaxias.

  i.e. Temperature effect <u>depends on</u> species

Results



- $b_1 > 0$: swimming speed of galaxias increases with temp.
- $b_2 > 0$: swimming speed of trout is greater than that of galaxias at 0°C (intercept)
- $b_3 > 0$: temperature has higher effect on swimming speed of trout than on galaxias.

Multilevel Linear Models

- Models with factors are evaluated by considering differences between particular levels and a chosen reference level (the default case)

- When a factor has more than 2 levels, we need to change the default case multiple times to ensure that we cover all pairwise comparisons

Exercise: 1.4

Exercise: 1.5

Principle of marginality

The principle of marginality states that the main effects of a model are marginal to high order terms (such as an interaction). Therefore models should include all lower-order relatives of that higher order term (e.g. the main effects that comprise the interaction).

In other words, the *main effects*, of species and temperature are *marginal* to the species*temperature interaction.

Assumptions of linear models

- IMPORTANT!
- Things we need to check (consider) when using linear regression models

Assumptions of linear models

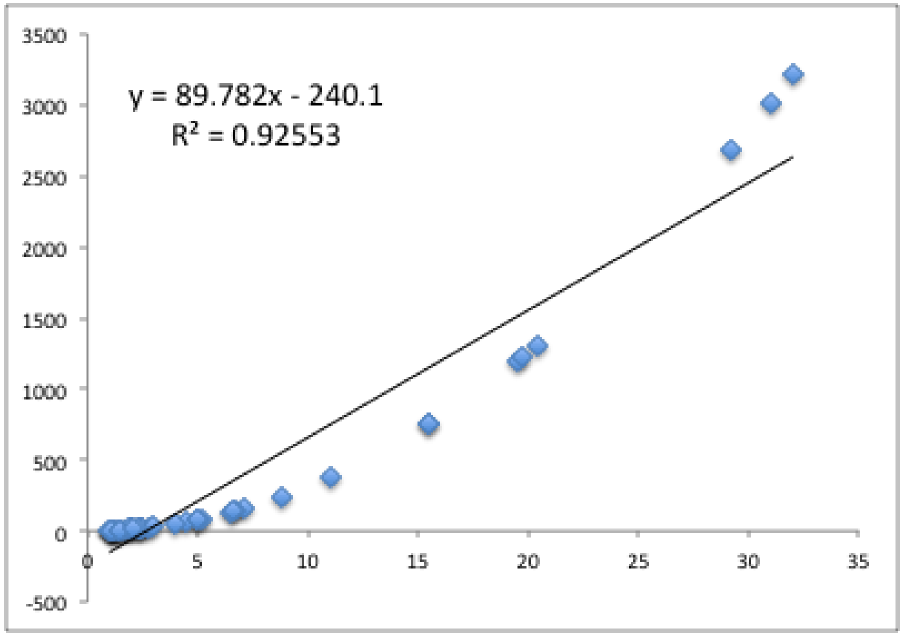
$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon_{ijk}$$

Linear models make many assumptions, including:

1. The model makes biological sense/ physical sense
2. Additivity (terms are added together)
3. Linearity
4. Independence of errors (LATER)
5. Homoscedasticity – equal variance of errors
6. Normality of errors.

Linearity

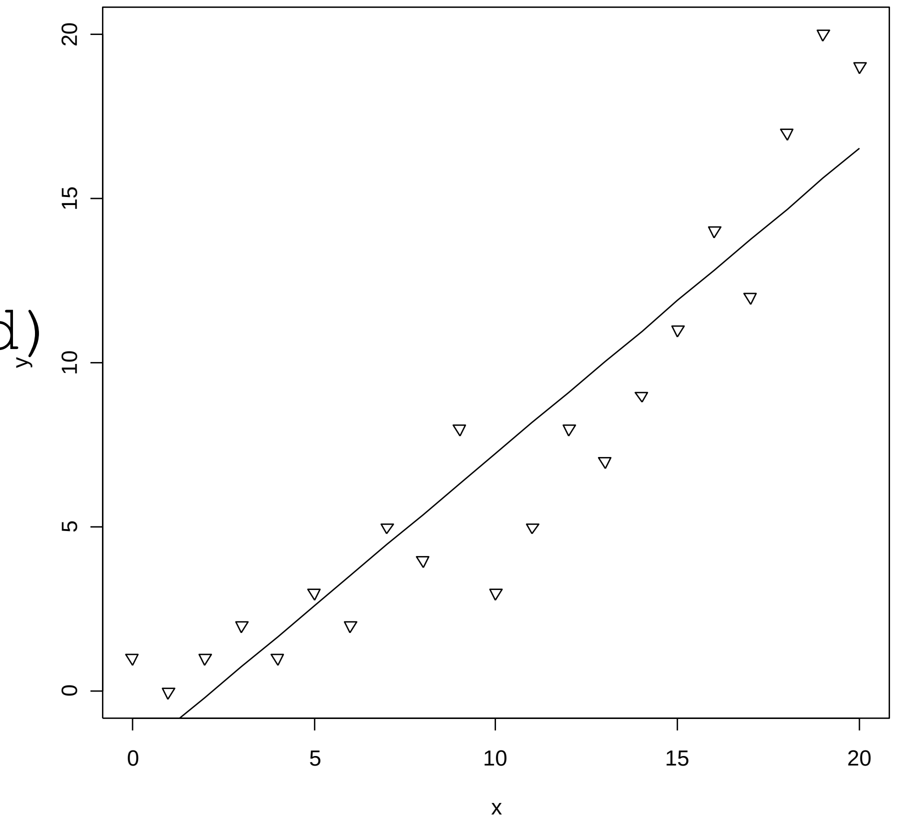
- If the relationship is not linear, then the fitted model will not fit the data properly across the domain of values



Linearity

- Linearity:
- **How do we check it?**
- (1) Plot the raw data:

```
> plot(x,y)
> model<-lm(y~x)
> lines(x,model$fitted)
```



Linearity

- **What's the solution?**
- Transform the response

$$\ln y = a + b_1 x$$

- Transform the predictors

e.g. Polynomial regression

$$y = a + b_1 x + b_2 x^2$$

- THINK ABOUT WHAT RELATIONSHIP IS NATURALLY MEANINGFUL (ASSUMPTION 1)

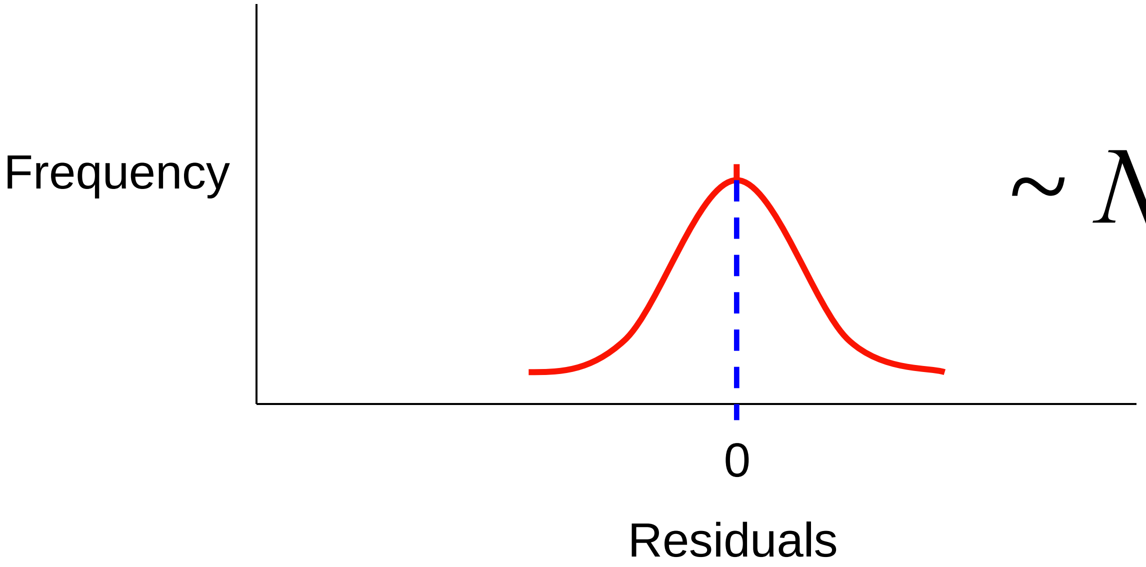
Distribution of residuals

- The parametric tests are mathematically derived, based on assumed distributions (e.g. normal, F)
- Which means the tests can only be trusted when the data being tested follow the assumed distributions
- e.g. t-tests and F-tests assume residuals are approximately normally distributed

Normality

.. the residuals of the model should be normally distributed

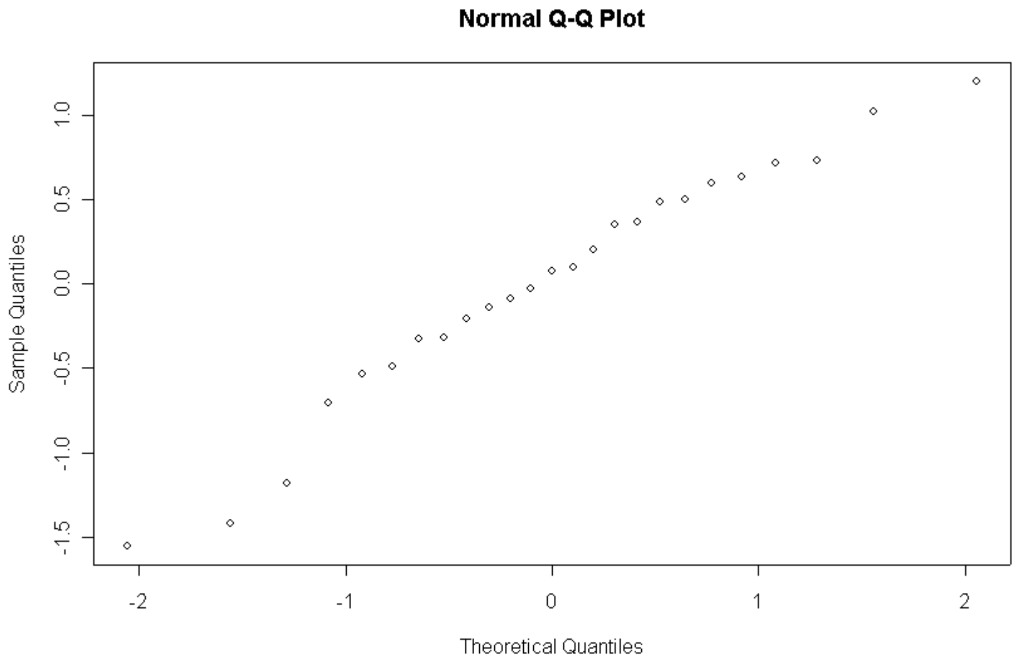
Why? Because t-tests and ANOVA assume normal residuals!



$$\sim N(0, s)$$

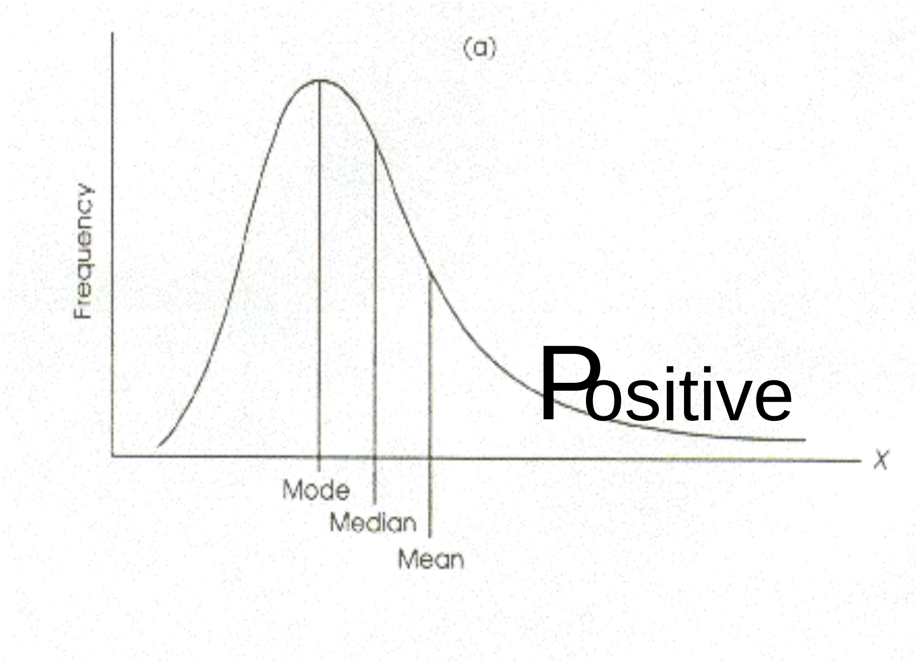
Normality

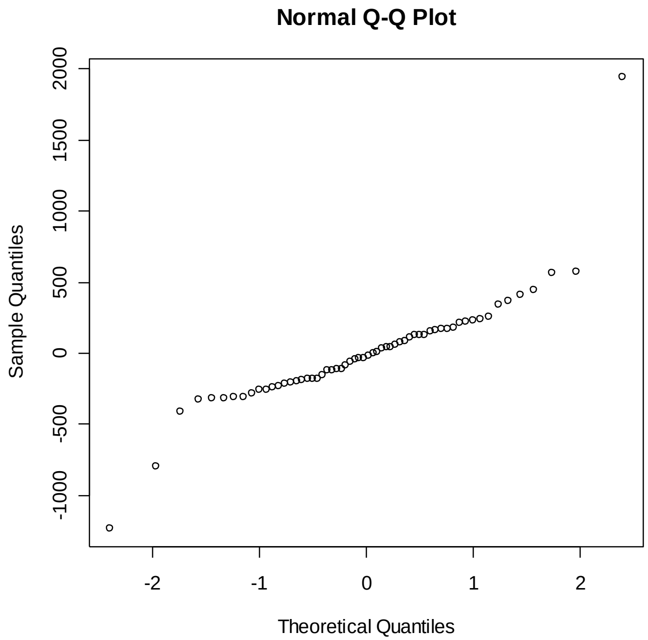
- **How do we check it?**
- Normality of residuals can be checked using a q-q plot



Distribution of residuals

- Most common problem is positive skew:





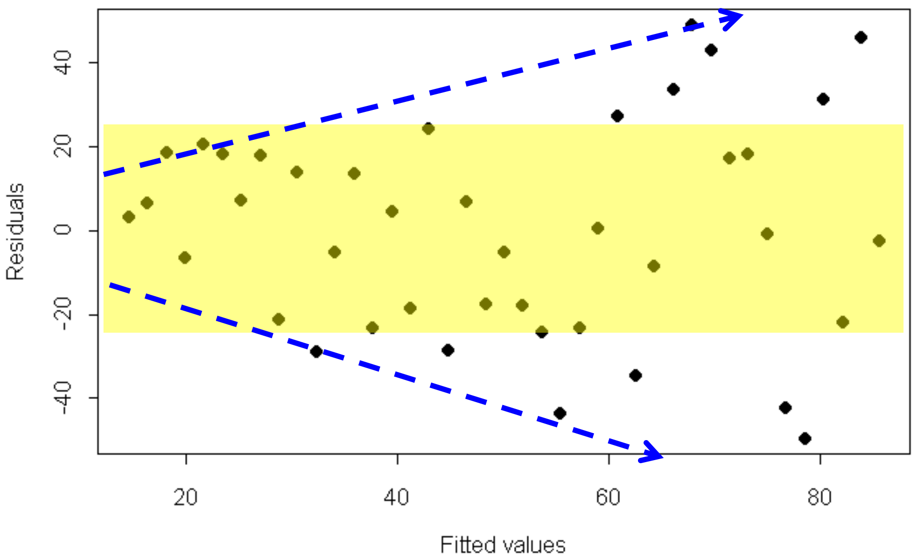
- **Solution?** Transform response variable

Homoscedasticity

- Homogeneity of variances (Homoscedasticity):
- i.e. the residuals (e_i's) should have the same variance across values of the response variable
- If not, then parameter estimates are likely to be unreliable

Homoscedasticity

- **How do we check it?**
- 1) Plot residuals ($y_i - \hat{y}$) against fitted values ($\hat{y}$)
- Residuals should be evenly spread across the range of fitted values



Homoscedasticity

- **What can cause the assumption to be broken?**
- Skew (response or predictors require transformation)
- Outliers

- **What's the solution?**
- Transform the response variable or predictor variables

Checking diagnostics

- This is easily done in R for lm() models

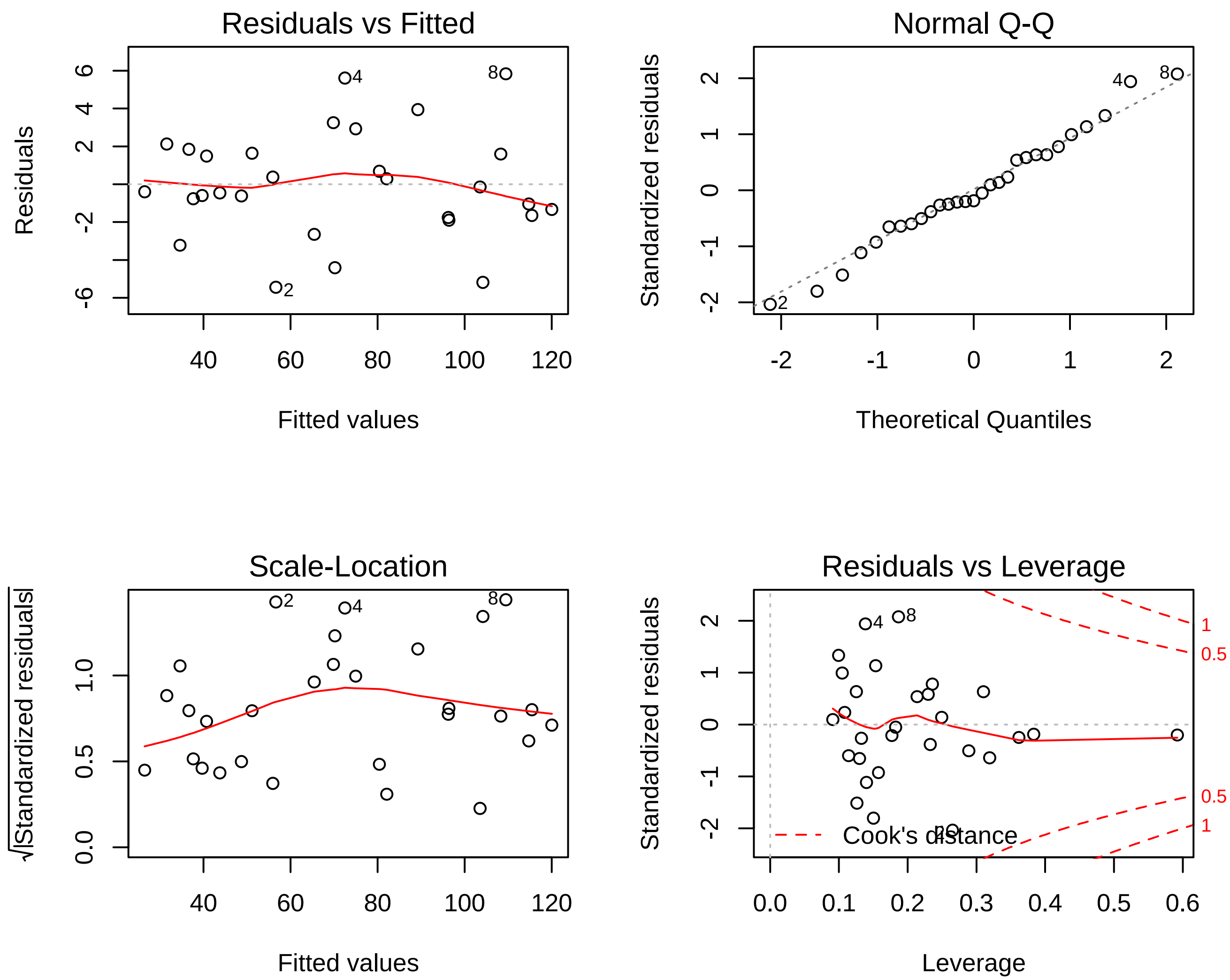
- Use the plot() function on the derived lm() object and it will plot the errors for you.

Example: 1.6

Return to the fishspeed2 problem.

Run diagnostic plots for the model you constructed there

Diagnostics plot for lm() in R



End of Lecture 1