STAT 110: Week 7

University of Otago

Outline

- Previous
  - Model for linear regression
  - $y = \beta_0 + \beta_1 x + \varepsilon$
- Today:
  - Fitting the model
    - Estimating β_0 and β_1
    - Fitted model
    - Residuals

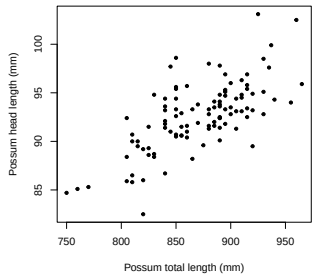
Recall: motivating data

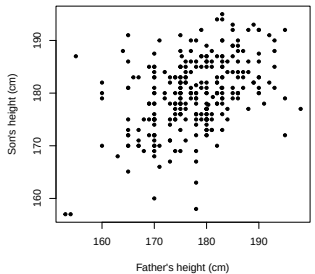
- The size of brushtail possums
  - Exploring relationship between total length (mm) and head length (mm)
- Height of STAT 110 students
  - Compare father's height (cm) and son's height (cm)
- Squat weight of international power lifters
  - Look at the relationship between body weight (kg) and max squat weight (kg)

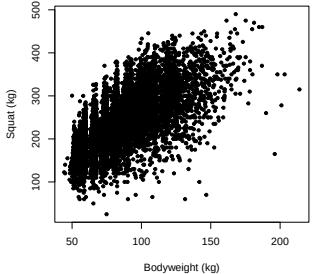
Recall: importing data into R

- Import the data into R

```
possum = read.csv('possum.csv')
height = read.csv('height.csv')
powerlift = read.csv('powerlift.csv')
```







Fitting a regression model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- β_0 and β_1 are parameters
  - Estimate parameters (population) with statistics (sample)
  - What statistics could we use to estimate β_0 and β_1?

Fitting a regression model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- β_0 and β_1 are parameters
  - Estimate parameters (population) with statistics (sample)
  - What statistics could we use to estimate β_0 and β_1?
    - We could guess by eye: use paper, pencil and ruler (or electronic equivalents)
    - Later in the lecture: find general approach for estimating β_0 and β_1
- For now: assume we have some way to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$
- Work through using the possum data to illustrate concepts

Fitted model

- The (simple) linear regression model is

$$y = \underbrace{\beta_0 + \beta_1 x}_{\text{mean response}} + \varepsilon$$

- Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ we can write the fitted model

$$\hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The fitted model is commonly written as

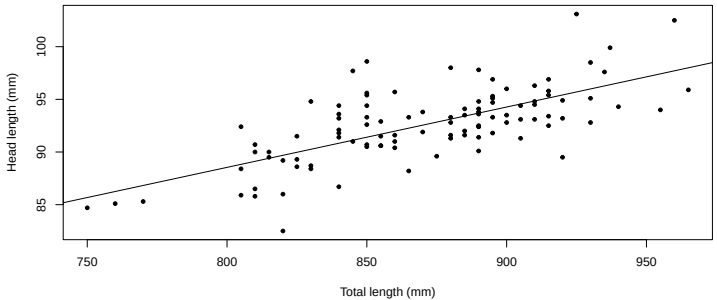
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The fitted model gives the estimate of the mean at a given x value

Fitted model: possum data

- Use estimates $\hat{\beta}_0 = 42.7$ and $\hat{\beta}_1 = 0.057$
- Fitted model is

$$\hat{y} = 42.7 + 0.057x$$



Residuals

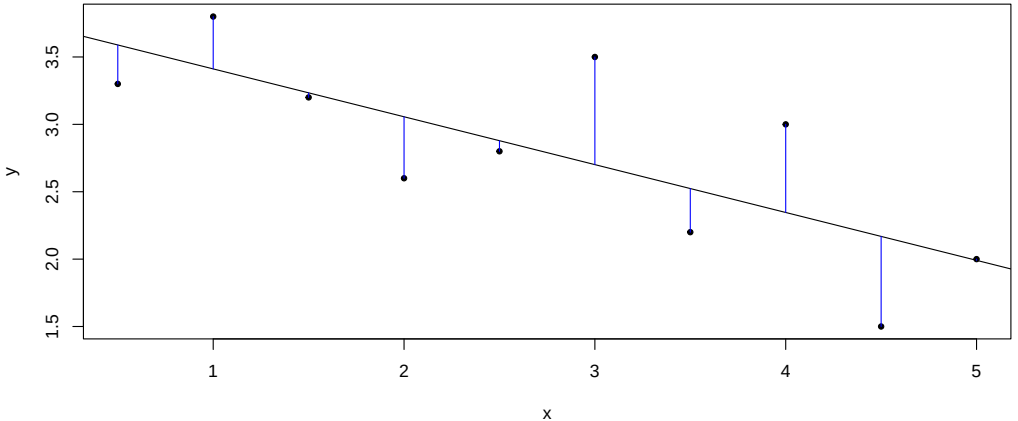
- The statistical model can be expressed as
  - observation = mean response + error
- After fitting the model, we have
  - observation = fitted model + residual
- The residual $\hat{\varepsilon}$ is our best guess (estimate) of the error ε
  - It is the difference between the observation (y) and the mean response ($\hat{y}$)

$$\hat{\varepsilon} = y - \hat{y}$$

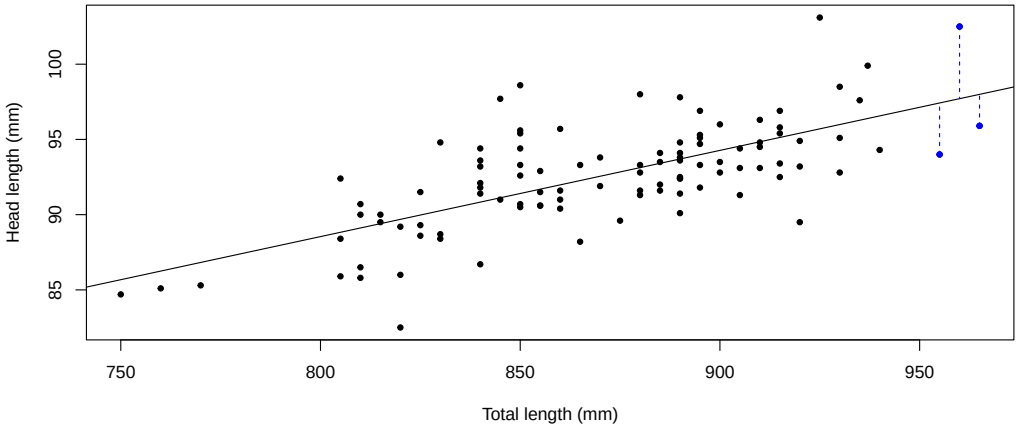
- We often index by i: for the ith observation (x_i, y_i) the residual is

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Residuals: blue lines



Residuals: possum data (three points in blue)



How do we fit the model?

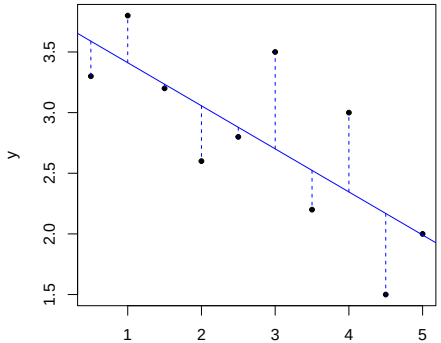
- The (simple) linear regression model is

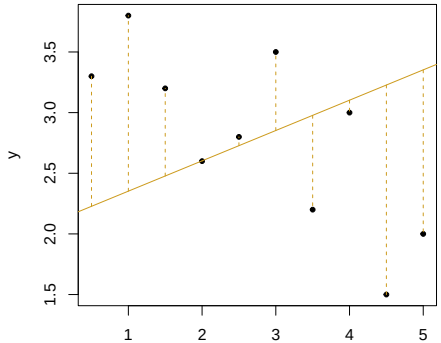
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- Estimate parameters β_0 and β_1
  - Find β_0 and β_1 that give the 'best' description of relationship between x and y
- Suppose we had a choice between two possible fitted models
  1. One of them has many large residuals (large positive and large negative residuals)
  2. The other one has mostly small residuals (small positive and small negative residuals)
- Which is better?
  - Look graphically

Graphical representation

- Same data, two possible fitted models
  - One with larger residuals (magnitude): gold
  - One with smaller residuals (magnitude): blue
- Which describes the relationship between x and y better?



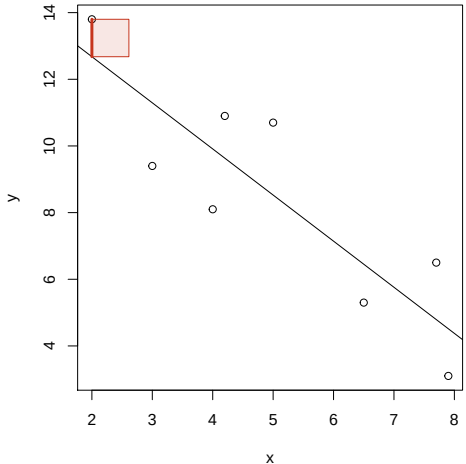


Least squares

- We want the (magnitude of the) residuals to be as small as possible
- We will find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ using the method of least squares
  - Find the values $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared residuals
- Explain the process graphically

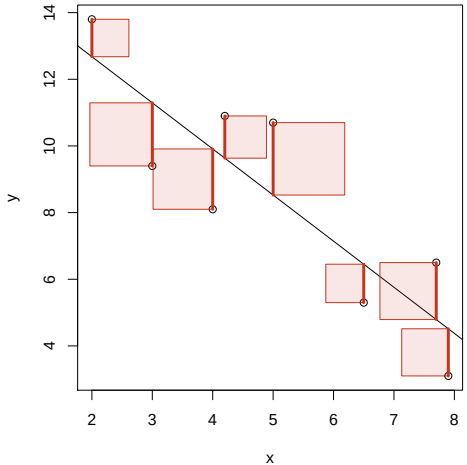
Least squares

- We can visualise the squared residual by drawing a square!
  - Squared residual is the area of red square



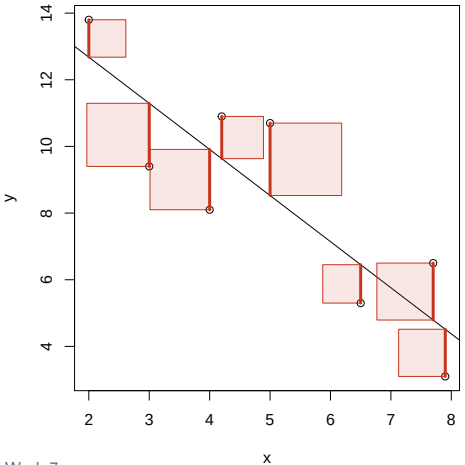
Least squares

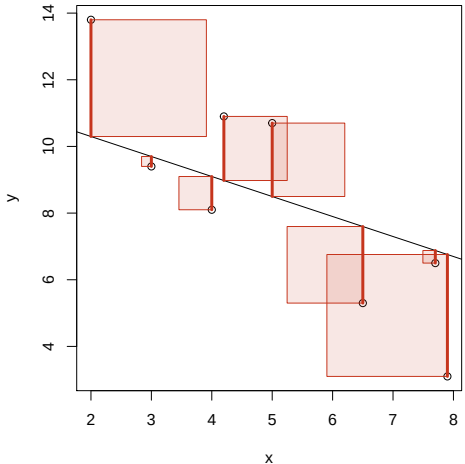
- The sum of squared residuals
  - Combined area of the red squares



Least squares

- Minimise the sum of squared residuals (minimise combined area)
  - Left plot: better fit (to the same data)





Least squares

- The sum of squared residuals:

$$\begin{aligned}\sum_{i=1}^{n} \hat{\varepsilon}_i^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2\end{aligned}$$

  - Note: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that make $\sum \hat{\varepsilon}_i^2$ as small as possible

Parameter estimates

- We can use calculus to find estimates
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimise sum of square residuals

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_y}{s_x} r$$

  - s_y: sample standard deviation of outcome y
  - s_x: sample standard deviation of predictor x
  - r: sample correlation between x and y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Details of how to find these: outside the scope of the course

In R

- We can find the least squares estimates using R
- The R code is

```
lm(y ~ x)
```

- Look at each piece in turn:
  - `lm`: function for fitting a **l**inear **m**odel
  - `y`: outcome variable
  - `x`: predictor variable
  - `~`: thought of as 'is modelled by'
  - `lm(y ~ x)`: is saying that we are fitting a linear model where the outcome variable y is modelled in terms of the predictor variable x

Fitting the possum data

```
m_possum = lm(possum$head_l ~ possum$total_l) # assigned the output to object m_possum
summary(m_possum) # shows a summary of the results

##
## Call:
## lm(formula = possum$head_l ~ possum$total_l)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.188 -1.534 -0.334  1.279  7.397
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    42.70979    5.17281    8.26  5.7e-13 ***
## possum$total_l  0.05729    0.00593    9.66  4.7e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.6 on 102 degrees of freedom
## Multiple R-squared:  0.478, Adjusted R-squared:  0.472
## F-statistic: 93.3 on 1 and 102 DF,  p-value: 4.68e-16
```

Estimates in R

- The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are given in column headed `Estimate`
  - $\hat{\beta}_0 = 42.71$
  - $\hat{\beta}_1 = 0.057$
- R labels the estimates in terms of the variable names
  - `(Intercept)`
  - `possum$total_l`

Detour: `data` option in `lm`

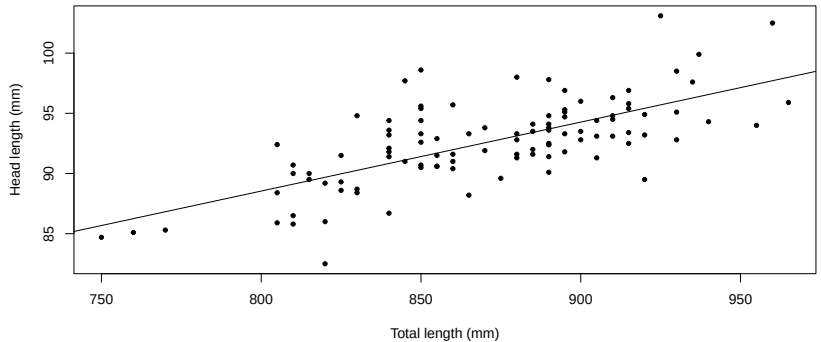
- The `lm` function includes a `data` option that can make specification easier
- Separate the variable (e.g. `head_l`) from the data frame object (`possum`)
- The code is

```
m_possum2 = lm(head_l ~ total_l, data = possum)
```

- This is fitting the same model as in the slide above

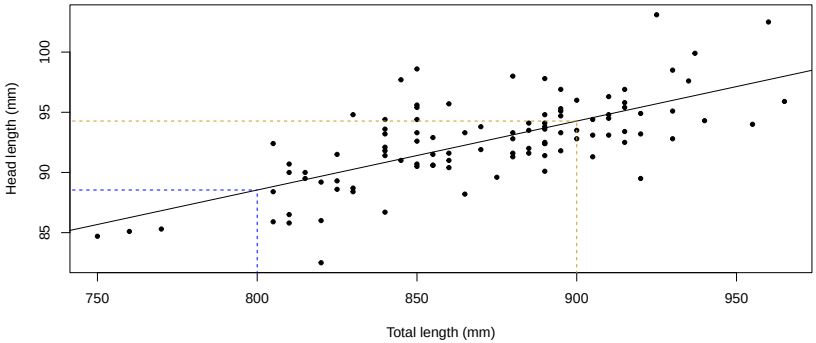
Fitted model: possum data

- The fitted model is $\hat{y} = 42.71 + 0.057x$
  - Recall: y is head length, x is total length
  - We could also write: $\widehat{\text{head}} = 42.71 + 0.057\,\text{total}$



Fitted model: possum data

- Fitted model is $\hat{y} = 42.7 + 0.057x$
  - For $x = 800$ we have $\hat{y} = 42.7 + 0.057 \times 800 = 88.5$
  - For $x = 900$ we have $\hat{y} = 42.7 + 0.057 \times 900 = 94.3$



Interpretation

- Fitted model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - For the possum data: $\hat{y} = 42.7 + 0.057x$
- Our interest is $\hat{\beta}_1$:
  - We estimate that the average head length of a possum will increase by 0.057 mm for a 1 mm increase in total length.
- This is a comparison of two subpopulations
  - If we compare possums whose total length is x mm to possums whose total length is $x + 1$ mm, the estimated increase in their expected (or mean) head length is 0.057 mm.
- $\hat{\beta}_0$ is the estimated mean head length of possums with total length 0 mm
  - Makes no biological sense
  - Do not interpret in this case

Summary

- Fitting a linear regression model
  - Fitted values: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
  - Residuals: $\hat{\epsilon} = y - \hat{y}$
- Method of least squares
  - Minimise the sum of squared residuals
  - Fit the model using `lm` in R: `lm(y ~ x)`

Outline

- Previous:
  - Fitting a statistical model
  - Method of least squares
- Today:
  - Assumptions underlying linear regression
    - What are the assumptions?
    - How do we check the assumptions?

Motivation

- Exploring relationship between total length (mm) and head length (mm) of brushtail possums
- Recall: fitting linear model

```
m_possum = lm(head_l ~ total_l, data = possum) # possum data
```

- Linear regression model allows us to:
  - Estimate the effect of x (total length) on y (head length)
  - Estimate the mean response of y (head length) given x (total length)
    - E.g. estimate mean head length of possums that have total length $x = 820$ mm
- Problem: the model relies on assumptions
  - Interpretations and conclusions may be invalid if assumptions are badly wrong
- We need to test the model assumptions (so far as possible)

Assumptions for Simple Linear Regression

- Recall that the linear regression model is

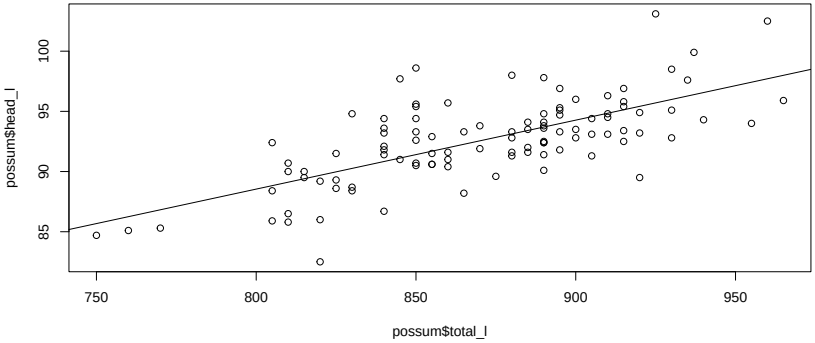
$$y = \underbrace{\beta_0 + \beta_1 x}_{\mu_y} + \varepsilon$$

- The underlying assumptions are:
  - **Linearity:** The mean response μ_y is described by a straight line
  - **Independence:** The errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent
  - **Normality:** The error terms ε are normally distributed
  - **Equal variance:** The errors terms all have the same variance, σ_ε^2 ('homoscedastic')
- These are often remembered using the mnemonic **LINE**.

Tools for checking assumptions

- Fitted line plot: compare the observed data to the fitted model
  - Useful, but not extensively used for checking assumptions

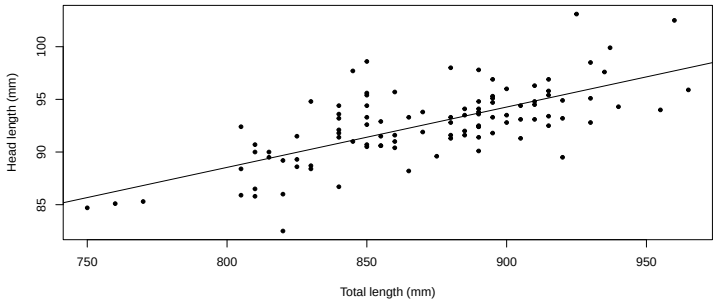
```
plot(possum$total_l, possum$head_l) # plot(x,y): x gives x values, y gives y values
abline(m_possum) # draws the fitted regression line
```



Detour: plotting in R

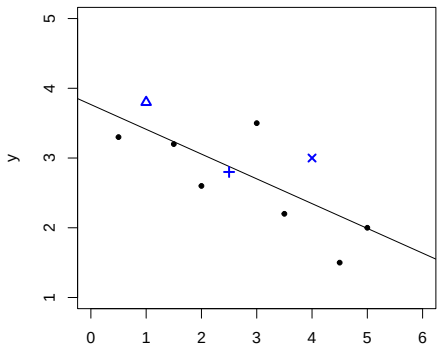
- Show code for 'default' plots: default points, colours, axis labels, etc
  - All that is needed for this course (STAT 260 explores plotting and visualisation of data)
- For interest: present same plot as above with some modifications

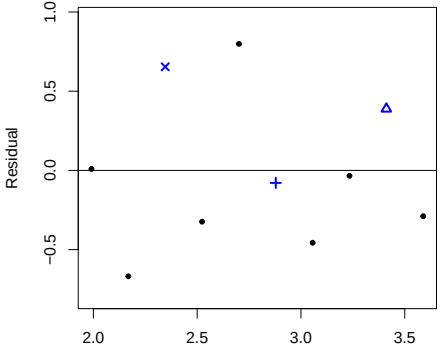
```
plot(possum$total_l, possum$head_l, pch = 20, xlab = "Total length (mm)",
     ylab = "Head length (mm)")
abline(m_possum)
```



Residual plots

- It is more common to use a residual plot
  - Residuals $\hat{\varepsilon}$ are on the y-axis
    - Recall: $\hat{\varepsilon} = y - \hat{y}$
- Look at a small example



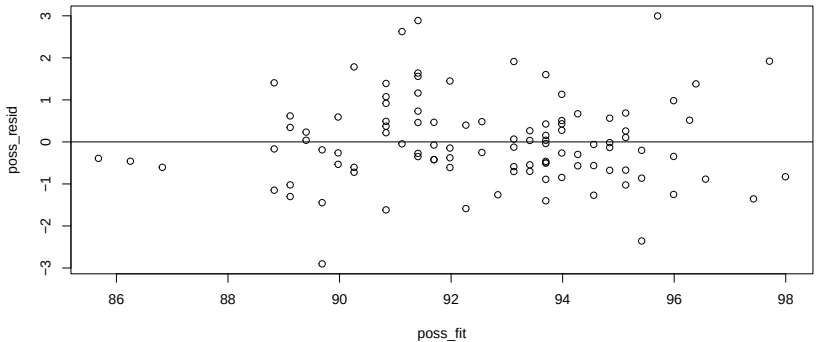


More on residuals: $\hat{\varepsilon} = y - \hat{y}$

- The residual is $\hat{\varepsilon} = y - \hat{\beta}_0 - \hat{\beta}_1 x$
- Residuals are estimates of error terms (ε)
  - Can be used to check assumptions about error terms (ε)
- The residual $\hat{\varepsilon}$ is often called a raw residual
  - Standardised or studentised residuals are often preferred
    - We will use studentised residuals in this course
  - What are studentised (or standardised) residuals?
    - Transformed to have standard deviation ≈ 1
    - (Mathematical) details are beyond the scope of the course
  - Find them in R using function `rstudent`
    - e.g. for model object `m_possum` we find studentised residuals using `rstudent(m_possum)`

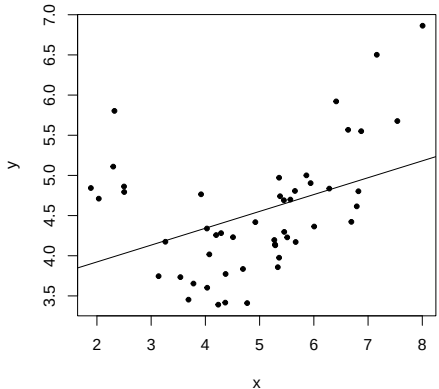
Plotting residuals in R

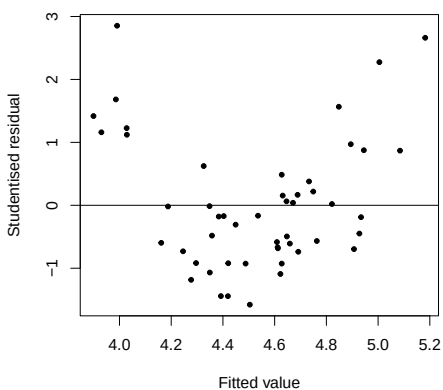
```
poss_fit = fitted(m_possum) # finds the fitted values of the model m_possum
poss_resid = rstudent(m_possum) # finds the studentized residuals of the model m_possum
plot(poss_fit, poss_resid) # plots residuals against fitted values
abline(h=0) # draws a horizontal line at 0
```



Checking the linearity assumption

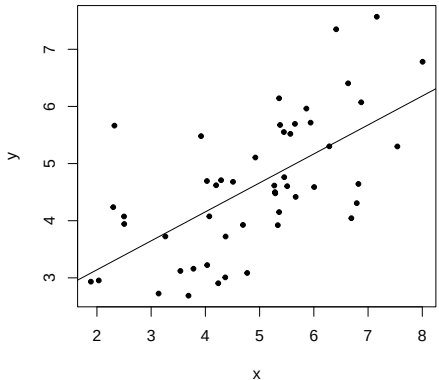
- Looking for clear departure for linearity in trend of data.
  - Look for patterns in plot of residuals against fitted values
- Plots below illustrate failure of linearity assumption (bad)

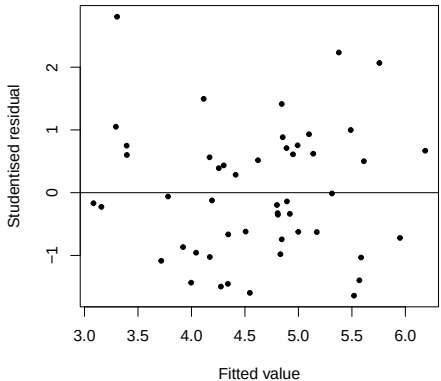




Checking the linearity assumption

- Looking for clear departure for linearity in trend of data.
  - Look for patterns in plot of residuals against fitted values
- Plots below: no evidence of failure of linearity assumption (good)





The independence assumption

- Independence assumption: errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent
- What does it mean that errors ε_1 and ε_2 are independent?
  - Knowing ε_1 tells us nothing about ε_2
    - $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
- For the possum example, independence means
  - Knowing how much above average one possum's head length is, gives no information about how far above average another possum's head length is.

Checking the independence assumption

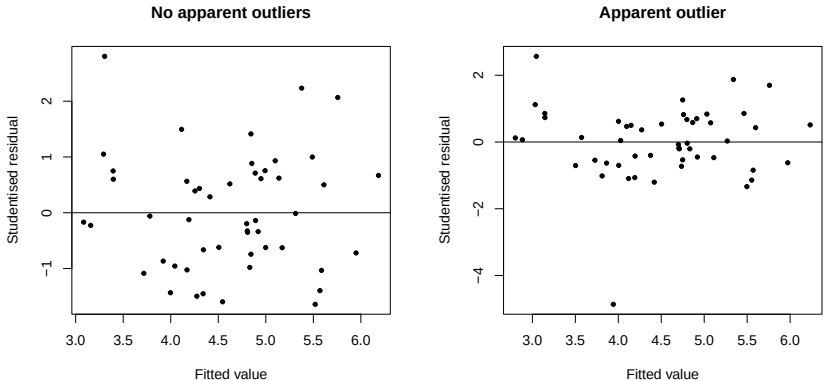
- In general: difficult to assess
  - We are unable to check it by looking at fitted line or residual plots.
- In certain situations, we may be able to check it
  - If the data are collected in time (time series)
    - Expect observations close together in time to be correlated
  - If the data are collected in space (spatial data)
    - Expect observations close together in space to be correlated
  - If there are multiple measurements from each participant (repeated measures)
    - Expect observations from a given participant to be correlated
- We can look at more complex statistical models for each of the cases above
  - Outside the scope of this course

Checking the normality assumption

- Assumption: errors ε are normally distributed
- The importance of the normality assumption depends on sample size
  - Sample size small: important, but hard to check
  - As sample size increases (say $n > 50$) it becomes increasingly less important
    - Looking for large violations of normality
    - Are there one (or more) extreme values: outliers
- We assess outliers using the residual plot

Checking the normality assumption

- Studentized residuals should be approximately normal with standard deviation 1:
  - Most (approx 95%) within ± 2
  - Nearly all ($> 99\%$) within ± 3
  - Values exceeding ± 4 are unusual

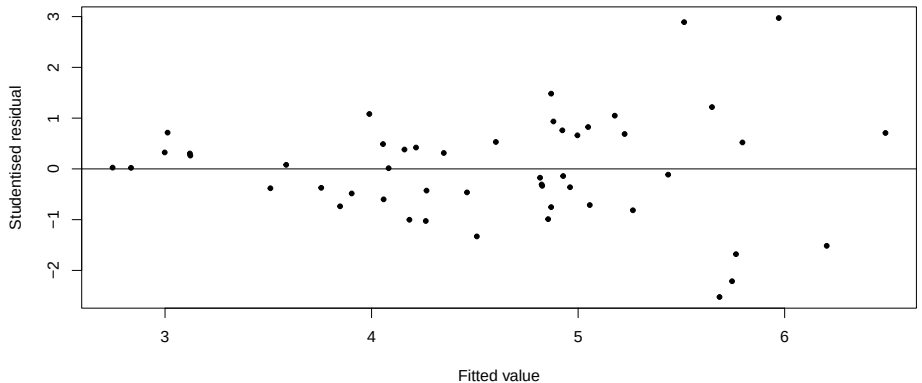


Checking equal variance assumption (homoscedasticity)

- Assumption: error terms $\varepsilon_1, \varepsilon_2, \ldots, \epsilon_n$ have the same variance
  - The magnitude of spread of data about regression line should not change too much with x.
- In contrast, if (say) variance of error terms increases with x
  - We would expect to see data more dispersion as x increases.
- Best seen with residual plot against fitted values.

Checking equal variance

- Example where there is evidence of non-constant variance
  - Variance of residuals increases with fitted value



What to do when assumptions fail: linearity

- Failure of the linearity assumption is critical
  - Conclusions drawn from the model will be invalid
- Paths forward include
  - Consider transforming outcome or predictor variables (where appropriate)
  - Explore more sophisticated models
    - Move beyond a simple linear regression model
- Both of these are outside the scope of the course
  - Considered further in STAT 210, 310

What to do when assumptions fail: independence or equal variance

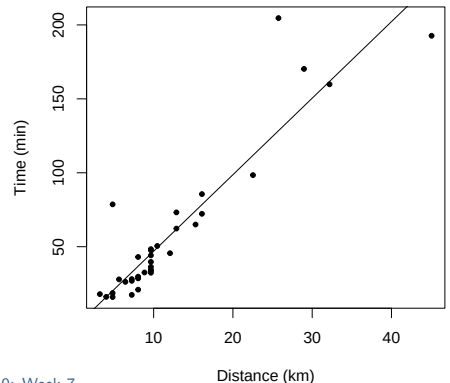
- When independence or equal variance assumptions fail
  - Estimates of parameters remain valid
  - Estimates can be inefficient
    - They can be improved
- Follows that fitted regression line is useable
- Confidence intervals and hypothesis tests will be invalid.
- Failure of assumptions can be rectified by sophisticated modelling techniques.
  - Details beyond this course.

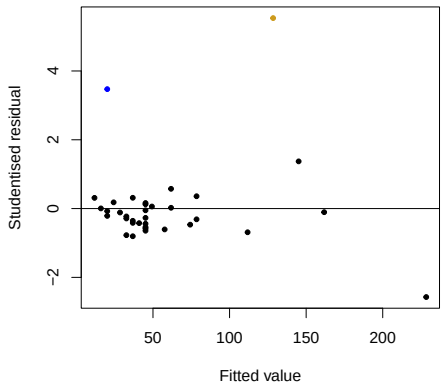
What to do when assumptions fail: normality / outliers

- Outliers can have a dramatic effect on the estimated regression
  - Such values are called influential points
- If outliers are present: check that the data are correctly recorded.
- If outliers remain we may consider removing them, however:
  - Think carefully first
    - Often outliers (or unexpected values in general) are the most interesting
    - They could be revealing something important about what we are studying
  - We should first assess if they are influential
    - If removing them has little effect: leave them in
  - If we do remove observations, we must be transparent
    - It should clear and obvious that values were removed and why
- Look at an example

Scottish hill racing

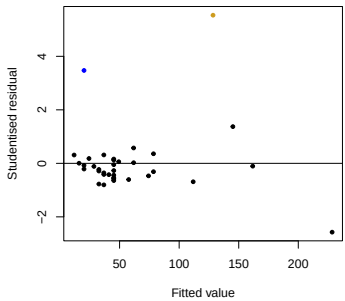
- Data are the record times in 1984 for 35 Scottish hill races (running)
- Interested in the relationship between distance and record time
  - Outcome variable (y): record time (in minutes)
  - Predictor variable (x): distance (in km)





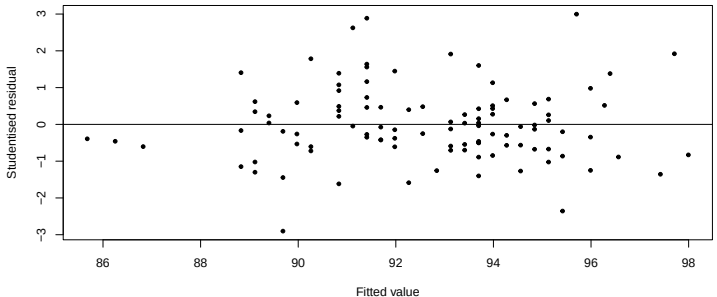
Scottish hill races: Investigate the outliers

- Knock Hill: record incorrectly recorded
  - Recorded as 78 minutes 39 seconds
  - It should have been 18 minutes 39 seconds.
- Bens of Jura: other important information?
  - This race has the largest climb by over 700 m
  - Consider (extended) model that includes climb?
- General: we may want to think about whether it is reasonable to describe the relationship between time and distance as linear for all races between 3 km and 40+ km



Residuals: possum data

```
plot(fitted(m_poss), rstudent(m_poss), pch = 20, xlab = "Fitted value",
     ylab = "Studentised residual")
abline(h = 0)
```



- Linearity: no evidence of a trend
- Outliers: no apparent outliers
- Constant variance: no obvious change in magnitude of spread of residuals

Recall: weightlifting data

- Maximum squat weight of international power lifters
  - Found the maximum squat for each athlete across competitions
- Data from 9045 athletes
- Look at the relationship between body weight (kg) and max squat weight (kg)
  - Outcome variable (y): (best recorded) squat weight
  - Predictor variable (x): body weight
- Import the data

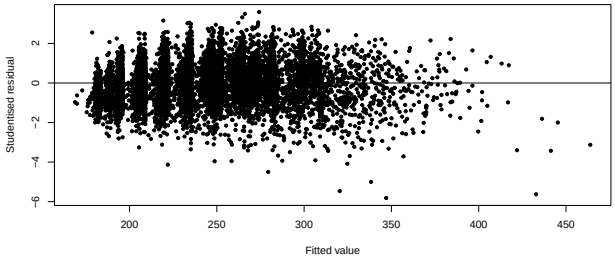
```
powerlift = read.csv('powerlift.csv')
```

- Fit linear regression model

```
m_power = lm(bestsquat ~ bodyweight, data = powerlift)
```

Residuals: powerlift data

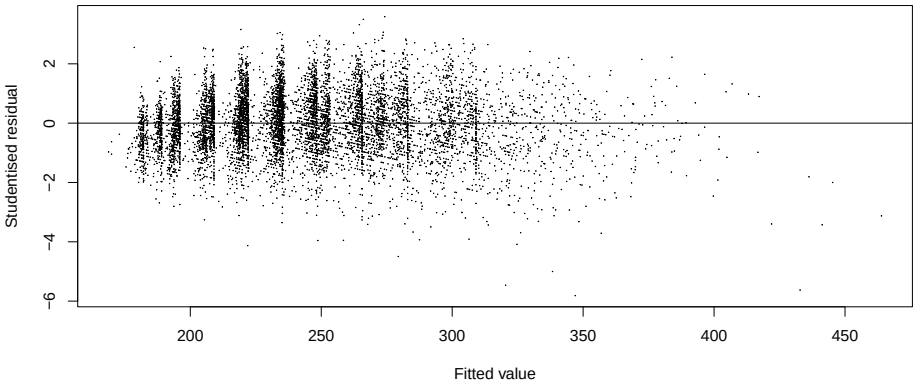
```
plot(fitted(m_power), rstudent(m_power), pch = 20, xlab = "Fitted value",
     ylab = "Studentised residual")
abline(h = 0)
```



- Linearity: very hard to tell
  - Too many points: draw points smaller to distinguish observations
- Outliers: some large negative residuals
- Constant variance: no obvious change in magnitude of spread of residuals

Residuals: powerlift data

- To better assess linearity
  - Draw points smaller (better see the number of points)



Residuals: powerlift data

- There is an apparent trend in the residuals
  - Residuals tend to be negative for low and high fitted values
- A more complex model may be required
  - e.g. there may be an upper 'physiological' limit that a human can squat
    - Consider a model where mean response increases to a maximum value
    - Outside the scope of the course
- Investigate the outliers
  - Data: maximum squat for each athlete across all recorded competitions
  - Outliers may have been from competitors with a single competition
    - Possible option: restrict to competitors with data from at least 5 competitions

Summary

- Assumptions of linear regression
  - LINE
    - Linearity
    - Independence
    - Normality
    - Equal variance
- Introduced residual plots
  - Can be used to check assumptions of linear regression model

Outline

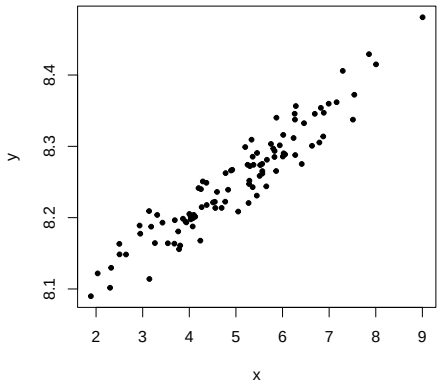
- Previous:
  - Fitting statistical models
  - Checking model assumptions
- Today:
  - Standard error
  - Confidence interval
  - Hypothesis test

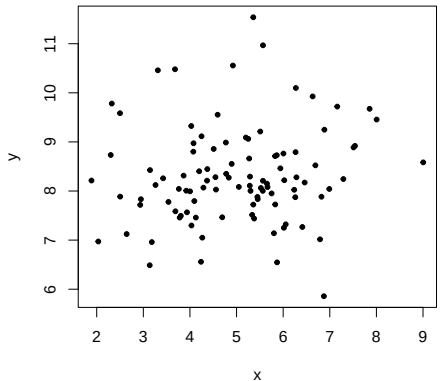
What does a regression model tell us?

- Consider the height of fathers and sons data
- The fitted model is an estimate of the true regression line in population
  - Population may be all male NZ university students (and their fathers)
- We need to assess the precision of the estimated parameters
  - Standard errors of the regression parameters
- Use standard errors to find confidence intervals and conduct hypothesis tests

The importance of the error variance

- Both sets of data come from populations with identical trend: $\mu_y = 8 + 0.05x$.





The importance of error variance

- The linear regresion model is $y = \beta_0 + \beta_1 x + \varepsilon$
  - The error ε is assumed to be normal with mean 0 and variance σ_ε^2
- The larger the error variance (all else equal)
  - The larger the spread of points around the true regression line
  - The more uncertain we are about the fitted regression line
    - That is, the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are less precise
  - To quantify our uncertainty about a fitted model
    - We need to estimate the error variance σ_ε^2

Estimation of the error variance

- The residuals ($\hat{\varepsilon}$) are estimates of the true errors (ε)
- Good estimate of error variance σ_ε^2: sample variance of the residuals $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$
- We need a few minor technical modifications
- The sample variance of the residuals is $\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$.
  - The sample mean of the residuals is 0: $\bar{\hat{\varepsilon}} = 0$
  - The correct divisor for simple linear regression is $n-2$ (rather than $n-1$)
- So estimate of error variance is

$$s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{RSS}{n-2}$$

  - $RSS = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$ is called the residual sum of squares

In R: father/son height data

- We can get s_ε from the R output (called `Residual standard error`)

```
m_height = lm(son ~ father, data = height)
summary(m_height)

##
## Call:
## lm(formula = son ~ father, data = height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.89  -3.89  -0.41   4.59  15.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.0533     8.4979   13.42  < 2e-16 ***
## father        0.3699     0.0478    7.74  1.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 277 degrees of freedom
## Multiple R-squared:  0.178, Adjusted R-squared:  0.175
## F-statistic: 59.9 on 1 and 277 DF,  p-value: 1.9e-13
```

Standard error of $\hat{\beta}_1$

- In many studies β_1 is the parameter we are most interested in
  - Change in the expected value of y for changing x in the population
- We estimate $\hat{\beta}_1$ from the observed data (sample)
- Measure precision of estimate by standard error $\sigma_{\hat{\beta}_1}$
  - Standard deviation of the sampling distribution of $\hat{\beta}_1$
    - Variation in $\hat{\beta}_1$ if there were many data sets (of the same size) from the population

Standard error of $\hat{\beta}_1$

- The standard error for $\hat{\beta}_1$ is

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  - The standard error is proportional to the error standard deviation σ_ε
    - As σ_ε^2 increases, the standard error of $\hat{\beta}_1$ also increases
- In principle this tells us about the precision of our estimated slope, $\hat{\beta}_1$
- In practice the formula is useless, since we don't know σ_ε
- We can handle that by estimating σ_ε by s_ε
- In practice, we will then use (estimated) standard error

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

In R

- We can get $s_{\hat{\beta}_1}$ from the R output (column called `Std. Error`)

```
summary(m_height)

##
## Call:
## lm(formula = son ~ father, data = height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.89  -3.89  -0.41   4.59  15.92
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.0533     8.4979   13.42  < 2e-16 ***
## father        0.3699     0.0478    7.74  1.9e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.13 on 277 degrees of freedom
## Multiple R-squared:  0.178, Adjusted R-squared:  0.175
## F-statistic: 59.9 on 1 and 277 DF,  p-value: 1.9e-13
```

Confidence intervals and hypothesis tests

- The standard error is needed to find confidence intervals and test statistics
- Earlier in semester we have seen that confidence intervals take the form

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- For testing $H_0 : \beta_1 = \text{null}$ we use the test statistic

$$t = \frac{\text{estimate} - \text{null}}{\text{std. error}}$$

- These continue to apply for a simple linear regression model

Confidence interval for slope

$$\text{estimate} \pm \text{multiplier} \times \text{std. error}$$

- Estimate is $\hat{\beta}_1$
- Multiplier comes from a t-distribution with $\nu = n - 2$ degrees of freedom.
  - Degrees of freedom match denominator in equation $s_\varepsilon^2 = RSS/(n-2)$.
  - So for $100(1-\alpha)\%$ confidence interval, multiplier is $t_{(1-\frac{\alpha}{2},\nu)}$.
- Standard error is

$$s_{\hat{\beta}_1} = \frac{s_\varepsilon}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Confidence interval 'by hand'

$$\hat{\beta}_1 \pm t_{(1-\frac{\alpha}{2},n-2)} s_{\hat{\beta}_1}$$

$$0.37 \pm t_{(0.975,277)} \times 0.048$$

- There are $n = 279$ observations
- From R: `qt(0.975, 277)` $= 1.969$

$$0.37 \pm 1.969 \times 0.048$$

$$0.37 \pm 0.094$$

$$(0.276, 0.464)$$

- We are 95% confident that the true slope is between 0.276 and 0.464
  - We estimate that the expected height of a son will increase by between 0.276 and 0.464 cm for a 1 cm increase in height of father

In R

- We typically find confidence intervals in R using `confint` function
  - It is important to understand how the confidence interval is found

```
confint(m_height)
```

```
##                2.5 %  97.5 %
## (Intercept) 97.325 130.782
## father       0.276   0.464
```

- Confidence interval for `father` is identical to that calculated on previous slide
- For a 99% confidence interval

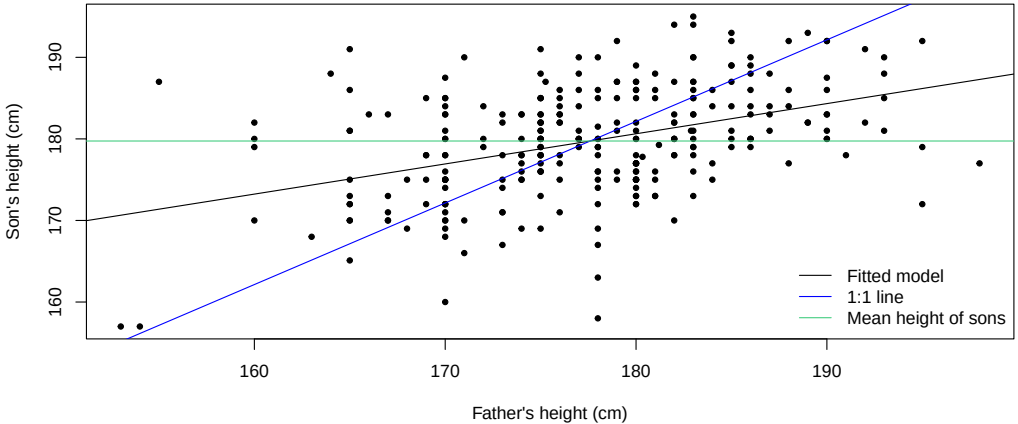
```
confint(m_height, level = 0.99)
```

```
##                0.5 %  99.5 %
## (Intercept) 92.012 136.094
## father       0.246   0.494
```

Regression

- We might have expected the average height of a son to increase by 1 cm for a 1 cm increase in father's height.
- That it does not, is the origin of the label: regression (to the mean)
  - The son of a short father tends to be short, but on average he is taller than his father
  - The son of a tall father tends to be tall, but on average he is shorter than his father
  - Extreme traits tend to regress to the mean
- 'Regression' introduced by Francis Galton when comparing the heights of parents and children
  - Galton, F. (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, **15**, 246–263

Regression to the mean



Hypothesis test for the slope

- Recall: $y = \beta_0 + \beta_1 x + \varepsilon$
  - β_1 describes how the mean response μ_y changes with x at population level
- If $\beta_1 = 0$ then $y = \beta_0 + \varepsilon$
  - $\mu_y = \beta_0$: μ_y does not depend on x
  - Outcome variable is not (linearly) related to the predictor variable
- A hypothesis test about β_1 assesses the hypothesis that two variables are related
  - Null hypothesis: statement of no relationship between x and y
    - $\mathsf{H}_0 : \beta_1 = 0$
  - Alternative hypothesis: relationship exists
    - $\mathsf{H}_A : \beta_1 \neq 0$

The test statistic

- To compute the p-value, we need a test statistic
- The test statistic is

$$t = \frac{\text{estimate} - \text{null}}{\text{std. error}}$$

- The estimate is $\hat{\beta}_1$
- The null value is 0 (previous slide)
- The standard error is $s_{\hat{\beta}_1} = s_\varepsilon / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - See previous lecture
- So for testing hypothesis about β_1, we use the test statistic

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Example: PCB in trout

- Concern that polychlorinated biphenyls (PCBs) polluting waterways and accumulating in food chain
  - PCBs used to be commonly found in transformers, capacitors, paints, etc
  - 28 trout collected[1] from Cayuga Lake, NY in 1970
    - Fish were marked and annually stocked (age was known)
    - Each trout was (mechanically) chopped, ground, and mixed before a 5 gm sample taken
    - Chromatography used to find PCB residue in ppm (parts per million)
- Scientific question: is there evidence that (log) PCB residue increases with age?
  - Null hypothesis: $\mathrm{H}_0 : \beta_1 = 0$
  - Alternative hypothesis: $\mathrm{H}_A : \beta_1 \neq 0$
- Treat it as a confirmatory study (specific hypothesis to assess)

[1] Science (1972), **177**, 1191–1192.

Example: PCB in trout

- Import the data into R

```
pcb = read.csv('pcb.csv')
```

- Look at the data

```
head(pcb)
```

```
##   age logpcb
## 1   1 -0.511
## 2   1  0.470
## 3   1 -0.693
## 4   1  0.182
## 5   2  0.693
## 6   2  0.262
```

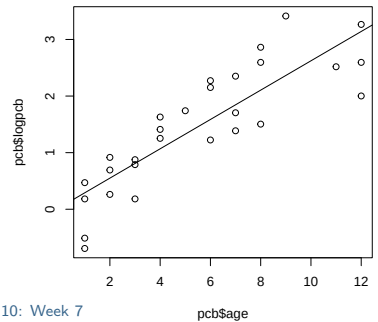
Example: PCB in trout

- Fit simple linear regression

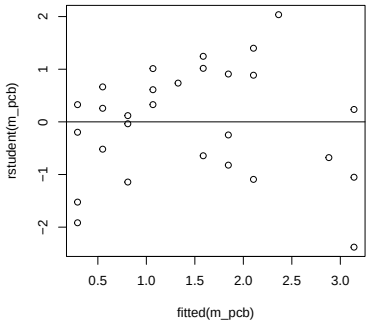
```
m_pcb = lm(logpcb ~ age, data = pcb)
```

- Plot fitted model and residuals: any concerns?

```
plot(pcb$age, pcb$logpcb)
abline(m_pcb)
```




```
plot(fitted(m_pcb), rstudent(m_pcb))
abline(h = 0)
```



R model output

```
summary(m_pcb)

##
## Call:
## lm(formula = logpcb ~ age, data = pcb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1395 -0.3879  0.0957  0.4327  1.0508
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0315     0.2014    0.16     0.88
## age           0.2591     0.0308    8.41  6.8e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.567 on 26 degrees of freedom
## Multiple R-squared:  0.731, Adjusted R-squared:  0.721
## F-statistic: 70.8 on 1 and 26 DF,  p-value: 6.78e-09
```

Interpretation

- For a confirmatory study
  - Formal test
- Compare the p-value to α
  - If p-value $< \alpha$: reject H_0
    - Evidence in favour of H_A
  - If p-value $< \alpha$: fail to reject H_0
- For an exploratory study
  - Interpret the p-value as a degree of incompatibility between data and null hypothesis
    - Use α as a guide
    - Try to avoid making a decision between hypotheses
  - Often prefer to use confidence intervals

Interpretation PCB: $\alpha = 0.05$

- The test statistic t is given in column `t value`: `8.41`
- The p-value is given in the column `Pr(>|t|)`: `6.8e-09`
  - These are found assuming the hypothesis: $H_0 : \beta_i = 0$
- p-value $< \alpha$: evidence of incompatibility between the data and null hypothesis
  - Data are incompatible with assumption of no relationship between PCB and age
  - Data are unusual compared to what we would expect if the null hypothesis were correct
- As this is a confirmatory study, we conclude that
  - There is evidence against H_0
  - There is evidence of a relationship between (log) PCB and age of fish (H_A)

Summary

- We want to quantify how precise our estimate is
  - Estimate of error variance
  - Estimate of standard error for $\hat{\beta}_1$
  - Found confidence interval for β_1
  - Hypothesis test for β_1
- Discussed origin of 'regression'