STAT 110: Week 8

University of Otago

Outline

- R^2: the proportion of variance explained
- Another look at estimating the mean response
- Predicting a new observation
- Extrapolation

Recall: possum data

- The size of brushtail possums
  - Exploring relationship between total length (mm) and head length (mm)
- If we have a total length measurement
  - Can we predict the head length?
- Import the data into R

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

- Fit a simple linear regression

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

Output

```
summary(m_possum)

##
## Call:
## lm(formula = head_l ~ total_l, data = possum)
##
## 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 ***
## 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
```

R^2: Coefficient of determination

- R^2 is a commonly used measure of how well a regression model describes the data
  - In R summary: `Multiple R-squared` $=$ `0.478`
- Look at two descriptions of R^2
  - Give us different perspectives on what it represents

R^2: squared correlation

- R^2 is the squared correlation between y and $\hat{y}$

```
y = possum$head_l # y values
yhat = fitted(m_possum) # y-hat values
R = cor(y, yhat)
R^2 # correlation^2

## [1] 0.478
```

- Since $-1 \leq r \leq 1$ we have $0 < R^2 < 1$
  - The larger the value of R^2, the better the regression model describes the data
    - The fitted values are ‘close’ to the observations

R^2: percentage of variance explained

- The total sum of squares is $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - Measures the variability of the outcome variable
- (Recall) the residual sum of squares $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Measures the variability of the outcome variable after fitting regression model
- The explained sum of squares $ESS = TSS - RSS$
  - Amount of variation in the outcome variable that is explained by the regression model
- R^2 can be expressed as

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

- The proportion of variance explained by the model
  - R^2 is often reported as a percentage: $R^2 = 47.8\%$

Interpreting R^2

- R^2 is often reported when fitting a linear regression
- No absolute rule for what a good (or bad) R^2 value is
  - In one particular area of application: an R^2 of 0.3 might be good
  - In another area of application: an R^2 of 0.8 might be poor

Mean response

- Recall: linear regression model

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

- Mean response at a given x value: $\mu_y = \beta_0 + \beta_1 x$
- The fitted model is an estimate of the mean response

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

- How precise is this estimate?
- Can we find a confidence interval for μ_y?
  - e.g. what is the confidence interval for mean head length of the subpopulation of possums with total length 850 mm

Confidence interval for mean response

- Goal: find a confidence interval for μ_{y0}, the mean response when $x = x_0$
- Confidence interval will have the form

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

- Estimate: $\hat{\mu}_{y0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- The (estimated) standard error for $\hat{\mu}_{y0}$ is

$$s_{\hat{\mu}_{y0}} = s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

- Multiplier: t-distribution with $\nu = n - 2$ degrees of freedom

Confidence interval for mean response

- A $100(1-\alpha)\%$ confidence interval for μ_{y_0} is given by

$$\hat{\mu}_{y_0} \pm t_{(1-\frac{\alpha}{2},n-2)} \times s_{\hat{\mu}_{y_0}}$$

- This is an interval estimate for the mean response μ_{y_0}
- Finding this confidence interval by hand is tedious
  - Use R to help us
  - `predict` function
- The `predict` function requires a data frame
  - Contains x_0: the predictor variable values where we want to find the mean response

Excursion: data frames in R

- You have been using data frames all semester
- When we import data into R: it is in a data frame
  - Rows: Each row is an observation or data record
  - Columns: Each column is a variable (typically with a name)
- We can construct a data frame using function `data.frame`

```
first_df = data.frame(name = c("Bob","Mary","Lucy"), age = c(19,17,23),
                      height = c(173, 168, 176))
first_df
```

```
##   name age height
## 1  Bob  19    173
## 2 Mary  17    168
## 3 Lucy  23    176
```

Data from for `predict`: possum data

- We need to construct a data frame in R
  - Contain the x (predictor variable) values where we want to find the mean response
  - Same variable name as was used to fit the model in `lm`
- Recall:

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

- Predictor variable name: `total_l`
- Let's say we want to estimate the mean response at 850 mm

```
predictor1 = data.frame(total_l = 850)
```

- If we wanted to find the mean response at 850 mm and 900 mm

```
predictor2 = data.frame(total_l = c(850,900))
```

Mean response in R

- Use the `predict` function, with option `interval = "confidence"`

```
mean_resp = predict(m_possum, newdata = predictor1, interval = "confidence")
mean_resp
##    fit  lwr upr
## 1 91.4 90.8  92
```

- First argument: model we are using (`m_possum`)
- Second argument (`newdata`): data frame of predictor values
- Third argument (`interval`): the kind of interval
  - Confidence interval for mean response: `interval = "confidence"`

Mean response: possum

- The estimated mean response is

$$\hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 42.71 + 0.057 \times 850 = 91.4$$

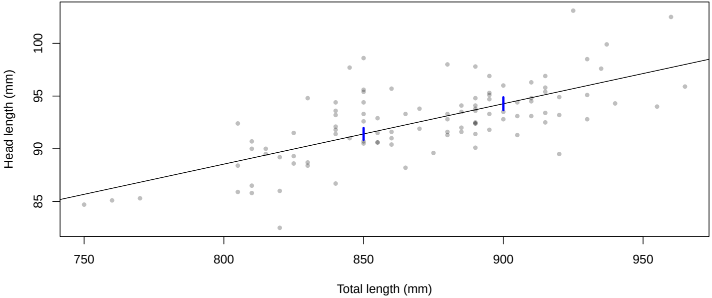
- Estimated mean head length for possums with total length 850 mm is 91.4 mm
  - Given by `fit` from `predict` output

```
mean_resp

##    fit  lwr upr
## 1 91.4 90.8  92
```

- We are 95% confident that the mean head length for possums with total length 850 mm is between 90.8 mm and 92 mm
  - Given by `lwr` and `upr` in `predict` output

Mean response: visual

```
mean_resp2 = predict(m_possum, newdata = predictor2, interval = "confidence")
mean_resp2

##    fit  lwr  upr
## 1 91.4 90.8 92.0
## 2 94.3 93.7 94.9
```



Prediction

- We can also use the model to predict a new observation y_0
- At a given value of $x = x_0$ (say $x_0 = 850$ mm)
  - ▸ The prediction ($\hat{y}_0$) is the same as the estimated mean response ($\hat{\mu}_{y_0}$)
    - Recall: fitted line was $\hat{y} = \hat{\mu}_y = \hat{\beta}_0 + \hat{\beta}_1 x$
- That means that at $x_0 = 850$ mm we have

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 = 42.71 + 0.057 \times 850 = 91.4$$

- We predict that a (new) possum of 850 mm would have a head length of 91.4 mm
  - ▸ What about the possible error in the prediction?
  - ▸ We want to find a prediction interval?

Prediction error

- The prediction uncertainty is larger than the uncertainty about mean response
  - It needs to combine uncertainty about the mean response and individual variability
- Eg. if we are predicting the head length of a possum with total length 850 mm
  - The mean head length among the subpopulation of possums with total length 850 mm is uncertain
    - Standard error for mean response
  - There is possum to possum variability in head length among the subpopulation of possums with total length 850 mm
    - Not all possums with total length 850 mm will have the same head length
    - Given by the error ε in the linear regression model

Prediction error

- The prediction error takes account of both sources of uncertainty
- For prediction at $x = x_0$, the prediction error is

$$PE(\hat{y}_0) = s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

  - Looks like standard error for mean response
    - Has an extra term in the square root: $1+$
    - Accounts for individual variation about the mean
- A $100(1-\alpha)\%$ prediction interval for y_0 is $\hat{y}_0 \pm t_{(1-\frac{\alpha}{2}, n-2)} \times PE(\hat{y}_0)$
- The prediction interval is a probability interval
  - There is a probability of $(1-\alpha)$ that y_0 will lie in this interval

Prediction in R

- Use the `predict` function, with option `interval = "prediction"`

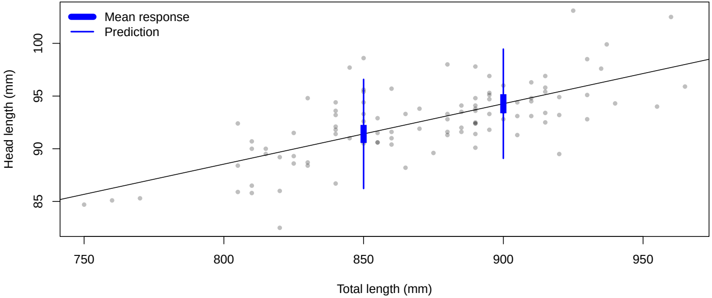
```
pred = predict(m_possum, newdata = predictor1, interval = "prediction")
pred
##    fit  lwr  upr
## 1 91.4 86.2 96.6
```

- There is a probability of 0.95 that a possum with total length 850 mm will have head length between 86.2 mm and 96.6 mm
- Note: we can find a 90% or 99% interval by including the argument `level`
  - Also applies when finding confidence interval for mean response

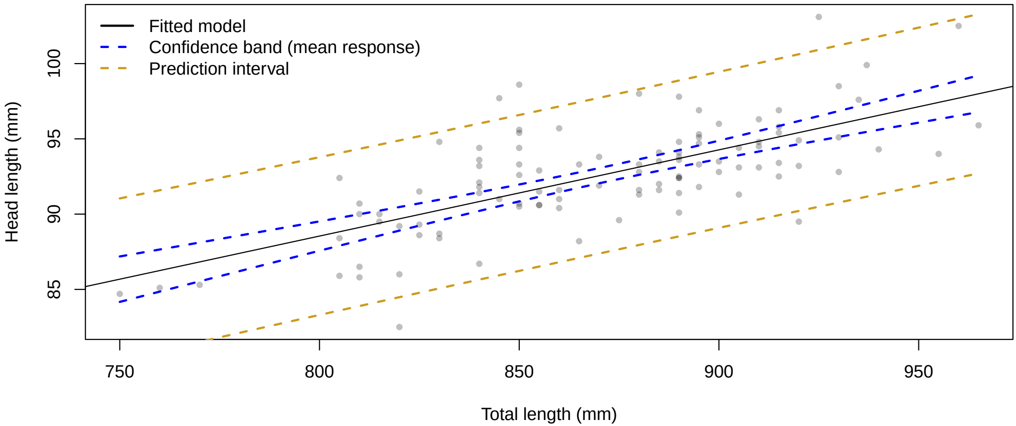
```
predict(m_possum, newdata = predictor1, interval = "prediction", level = 0.99)
##    fit  lwr  upr
## 1 91.4 84.6 98.3
```

Prediction: visual

```r
pred2 = predict(m_possum, newdata = predictor2, interval = "prediction")
pred2

##    fit  lwr  upr
## 1 91.4 86.2 96.6
## 2 94.3 89.1 99.5
```



Mean response and prediction: visual



Mean response and prediction

- The mean response is most precise in middle of plot
  - Confidence interval is narrower
- Same is true of prediction interval (harder to see on plot)
- The standard error and prediction error both include the term

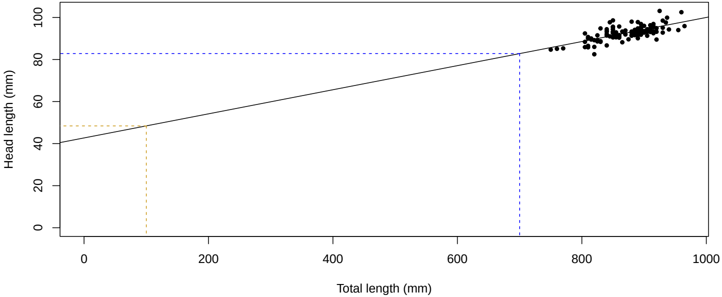
$$\frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- This is smallest when $x_0 = \bar{x}$
  - Estimation of mean response and prediction is most precise at $x_0 = \bar{x}$
  - Errors increase the further x_0 is from sample mean $\bar{x}$

Extrapolation

- When using linear regression models
  - Care is needed if extrapolating!
- Extrapolation: predicting values outside the range of the observed data
- Why is this a problem?
  - The linear regression model has limitations
    - It approximates the relationship between x and y across the range of data we observe
    - We don't necessarily believe it describes the true relationship between x and y
    - We don't know how data will behave outside the range we have observed
- If we decide to extrapolate
  - Important to know the risks and limitations

Extrapolation: possum



- The linear regression model provides a description of the relationship between total length and head length across the range of observed data
  - Total length between 750 mm and 950 mm
- We don't believe it describes the true relationship
  - We wouldn't use it to predict head length when total length is 100 mm
  - What about predicting head length when total length is 700 mm?

Summary

- Model summary: R^2
  - Squared correlation between fitted values and observations
  - Gives the percentage of variance explained by regression
- Looked again at mean response
  - Found confidence interval for mean response at $x = x_0$
- Looked at predicting a new observation
  - $\hat{y} = \hat{\mu}_y$
  - Prediction interval wider that confidence interval for mean response
- Looked at dangers of extrapolating

Outline

- Explore multiple linear regression
  - Where there is more than one predictor variable
- How to fit in R
- How to interpret the estimates
- How to find confidence intervals and conduct hypothesis tests
- Estimating mean response and predicting new observation
- Assessing model fit

Neurocognitive scores

- Neurocognitive function evaluated with MATRICS Consensus Cognitive Battery[1]
  - Measures cognitive performance in seven domains
- To start, we will focus on one domain: speed of processing
  - Explore how does it relate to age?
- We will use data from 145 'healthy' participants
  - Screen for medical and psychiatric illness
  - No history of substance abuse
- Subset of a larger study that had different aims[2]
  - Assess how cognitive scores varied between individuals with schizophrenia, individuals with schizoaffective disorder, and healthy controls

[1] *American Journal of Psychiatry*, **165**, 203–213, 2008.
[2] *Schizophrenia Research: Cognition*, **2**, 227–232, 2015.

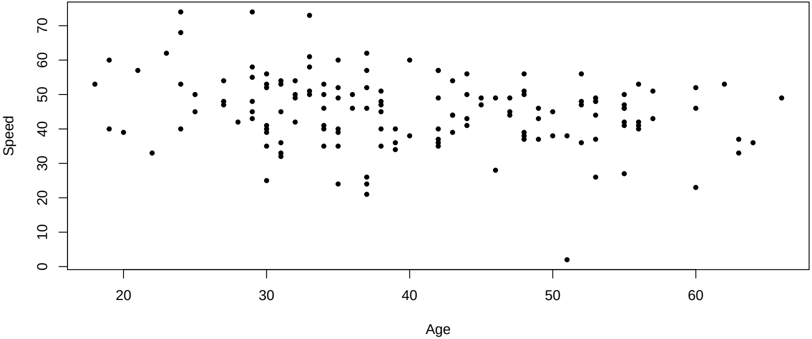
Neurocognitive scores: data

- Import the data

```r
neuro = read.csv('neuro.csv')
```

- Look at scatterplot of speed score and age

```r
plot(neuro$age, neuro$speed, xlab = "Age", ylab = "Speed", pch = 20)
```



Neurocognitive scores: regression model

- Consider the model: $\text{speed} = \beta_0 + \beta_1\,\text{age} + \varepsilon$
  - Score in the speed of processing test: outcome variable y
  - Age of participant: predictor variable x
- If we take $y = \text{speed}$ and $x = \text{age}$ we have the usual model: $y = \beta_0 + \beta_1 x + \varepsilon$
- The parameters:
  - β_0 is the expected outcome when the predictor variable is 0
    - How useful (or meaningful) the parameter is, depends on application
    - Neurocognitive example: expected speed score when age is 0 (not meaningful to interpret)
  - β_1 is the change in the expected outcome for a one unit increase in the predictor
    - Change in the expected speed score for a one year increase in age
    - Comparing two subpopulations that are one year apart in age

Neurocognitive scores: fitted regression model

```
m_neuro = lm(speed ~ age, data = neuro)
summary(m_neuro)
```

```
##
## Call:
## lm(formula = speed ~ age, data = neuro)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -40.72  -6.17   0.40   5.80  26.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.1468     3.1646   17.11   <2e-16 ***
## age          -0.2240     0.0757   -2.96   0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.2 on 143 degrees of freedom
## Multiple R-squared:  0.0578, Adjusted R-squared:  0.0512
## F-statistic: 8.77 on 1 and 143 DF,  p-value: 0.00359
```

Interpret the effect

- Find confidence intervals

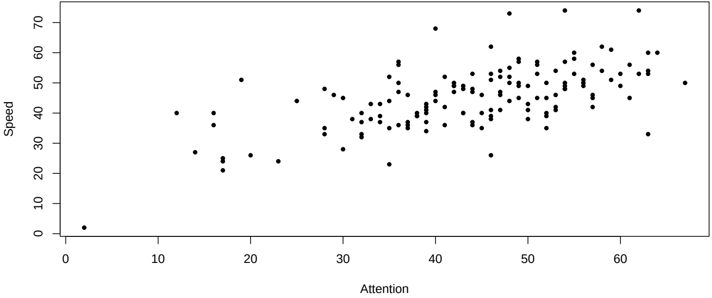
```
confint(m_neuro)
```

```
##                2.5 %  97.5 %
## (Intercept) 47.891 60.4022
## age         -0.374 -0.0745
```

- We are 95% confident that the increase in expected speed score is between -0.374 and -0.074 for a one year increase in age
- As $\hat{\beta}_1$ is negative: represents a decrease in expected score
  - We are 95% confident that the decrease in expected speed score is between 0.074 and 0.374 for a one year increase in age

We have more information...

- The regression is explaining $R^2 = 5.8\%$ of the variation in speed score
- There are other variables that could potentially help explain the speed score
  - e.g. the score on the other domains: we will look at scores from the attention domain



- Can we use attention and age together to describe the speed scores?

Multiple linear regression

- In multiple linear regression we have multiple predictors
  - We call them $x_1, x_2, \ldots, x_k$
  - k denotes the number of predictor variables
- The multiple regression model is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$
  - $\beta_0, \beta_1, \ldots, \beta_k$ are parameters (regression coefficients)
  - ε is an error term following a $N(0, \sigma_\varepsilon^2)$ distribution.
- The mean response is $\mu_y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
  - This is a conditional mean, given the values of the predictor variables $x_1, \ldots, x_k$
- For the neurocognitive scores we have

$$\text{speed} = \beta_0 + \beta_1 \,\text{age} + \beta_2 \,\text{attention} + \varepsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Model fitting

- Once we have parameter estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, the fitted model is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

  - $\hat{y}$ is also an estimate $\hat{\mu}_y$ of the mean response
- We can find the residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
  - Estimate of the error term ε_i
  - Identical to simple linear regression
- We can use least squares to find estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$
  - Minimise the squared residuals $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$
  - Same as with simple linear regression

Multiple regression: in R

- Use the same function to fit multiple linear regression: `lm`
- Add another predictor variable: `+ attention`

```
m_neuro2 = lm(speed ~ age + attention, data = neuro)
```

- We will see that much remains the same with multiple linear regression
  - Highlight differences with simple linear regression
- One difference is that it is much harder to visualise multiple linear regression
  - We now have two predictor variables (and we could potentially have more!)

Neurocognitive scores: in R

```
summary(m_neuro2)

##
## Call:
## lm(formula = speed ~ age + attention, data = neuro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.176  -5.495  -0.466   4.458  23.770
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  31.6661     3.2885    9.63   <2e-16 ***
## age          -0.2459     0.0579   -4.24    4e-05 ***
## attention     0.5349     0.0529   10.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.79 on 142 degrees of freedom
## Multiple R-squared:  0.452, Adjusted R-squared:  0.444
## F-statistic: 58.6 on 2 and 142 DF,  p-value: <2e-16
```

Interpretation

- There are some (minor) changes in how we interpret the parameters
- β_0: expected outcome when *all* predictor variables are 0
- Other coefficients are specific to the associated explanatory variable
  - e.g. β_2 is the change in the expected outcome when variable x_2 is increased by one unit, *and all other predictor variables remain unchanged*
    - Often say: all else held fixed
- In the neurocognitive scores example: β_2 is the change in the expected speed score when the attention score is increased by one, all else held fixed
  - All else held fixed: age unchanged
- Sometimes expressed as: β_2 is the effect of x_2 *having adjusted for* all other predictor variables

Interpretation: neurocognitive scores

- The fitted model is

$$\widehat{\text{speed}} = 31.67 - 0.25\,\text{age} + 0.53\,\text{attention}$$

- Interpretation of $\hat{\beta}_1$: the decrease in expected speed score is estimated to be 0.25 for a one year increase in age, holding the attention score fixed
- Interpretation of $\hat{\beta}_2$: the increase in average speed score is estimated to be 0.53 for a one year increase in age, having adjusted for age
- It doesn't make sense to interpret $\hat{\beta}_0$, but if we did
  - The average speed score for a participant of age 0, with attention score of 0 is 31.67
  - Why does it not make sense to interpret this?

Confidence interval

- We can find confidence intervals for the parameter β_j
  - Minor changes from simple linear regression
- We still use

$$\text{estimate} \pm \text{multiplier} \times \text{standard error}$$

- The estimate is $\hat{\beta}_j$
- The multiplier comes from a t-distribution with $\nu = n - k - 1$ degrees of freedom
- The (estimated) standard error $s_{\hat{\beta}_j}$ is complicated
  - It can be obtained from R output: column `Std. error`
- We can still find confidence interval directly with `confint`

Confidence interval: neurocognitive scores

- The confidence intervals are

```
confint(m_neuro2, level = 0.9)
```

```
##                  5 %   95 %
## (Intercept) 26.222 37.111
## age         -0.342 -0.150
## attention    0.447  0.623
```

- Interpreting the confidence interval for β_2
  - We are 90% confident that the average speed score will increase by between 0.447 and 0.623 for a one unit increase in the attention score, holding age fixed.

Hypothesis testing

- The multiple linear regression model is

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$$

- The mean response is $\mu_y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$
  - This depends on variable x_j only if β_j is not 0
- Testing $\beta_j = 0$ is equivalent to testing if mean response depends on x_j
  - Having adjusted for all the other variables in the model

Setting up the hypothesis test

- We set up a null hypothesis indicating 'no effect'
  - $\text{H}_0 : \beta_j = 0$
  - $\text{H}_A : \beta_j \neq 0$
- The test statistic is of the usual form:

$$t = \frac{\texttt{estimate} - \texttt{null}}{\text{standard error}} = \frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}$$

- The t statistic, estimate $\hat{\beta}_j$, estimate standard error $s_{\hat{\beta}_j}$ and p-value are all available in the R output
- The p-value quantifies the incompatibility between the data and null hypothesis
  - A small p-value suggests the data are unusual assuming the null hypothesis is true

Prediction and mean estimation in multiple regression

- As with simple linear regression, the fitted model can be interpreted as both
  - An estimate of the mean response $\hat{\mu}_y$, and
  - A prediction of the response for a new data point $\hat{y}$
- If $x_{01}, x_{02}, \ldots, x_{0k}$ give the value of the predictor variables at which we wish to predict/estimate, then

$$\hat{y}_0 = \hat{\mu}_{y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_k x_{0k}$$

- The estimated mean response and predicted value are the same

Prediction and mean estimation: neurocognitive scores

- The fitted model is

$$\widehat{\text{speed}} = 31.67 - 0.25\,\text{age} + 0.53\,\text{attention}$$

- The estimated mean response (and prediction) for participant aged 40, with attention score of 50 is

$$\begin{aligned}\widehat{\text{speed}} &= 31.67 - 0.25 \times 40 + 0.53 \times 50 \\ &= 48.58\end{aligned}$$

Prediction and mean estimation in multiple regression

- The general structure of the intervals is the same as with simple linear regression
  - A $100(1-\alpha)\%$ confidence interval for mean response μ_{y_0} is

$$\hat{\mu}_{y_0} \pm t_{(1-\frac{\alpha}{2}, n-k-1)} \times s_{\hat{\mu}_{y_0}}$$

  - A $100(1-\alpha)\%$ prediction interval for y_0 is

$$\hat{y}_0 \pm t_{(1-\frac{\alpha}{2}, n-k-1)} \times PE(\hat{y}_0)$$

- These are minor changes from simple linear regression:
  - Multiplier degrees of freedom are now $n-k-1$
  - The formulae for standard error $s_{\hat{\mu}_{y_0}}$ and prediction error $PE(\hat{y}_0)$ are more complicated
- The way in which we find these in R remains the same

Mean response and prediction in R

- Mean response and prediction for participant aged 40 with attention score 50
- Set up data frame

```r
to_pred = data.frame(age = 40, attention = 50)
```

- Estimated mean response with confidence interval (`interval = "confidence"`)

```r
predict(m_neuro2, newdata = to_pred, interval = "confidence")

##    fit  lwr upr
## 1 48.6 47.1  50
```

- Prediction with prediction interval (`interval = "predict"`)

```r
predict(m_neuro2, newdata = to_pred, interval = "predict")

##    fit  lwr upr
## 1 48.6 33.1  64
```

Model assumptions

- The multiple linear regression model is

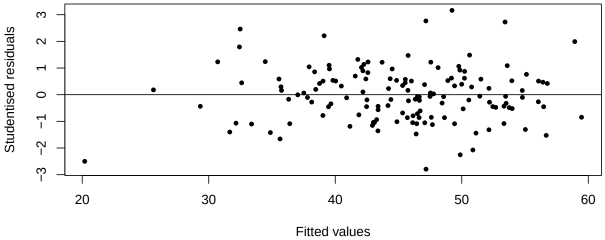
$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\mu_y} + \varepsilon$$

- We are making the following assumptions:
  - **Linearity:** There is a linear line relationship between μ_y and x_j when all other predictor variables are held constant
  - **Independence:** The error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent
  - **Normality:** The error terms $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are normally distributed
  - **Equal variance:** The errors terms all have the same variance, σ_ε^2 ('homoscedastic').

Checking assumptions: same as simple linear regression

- Check assumptions by plotting studentised residuals against fitted values
- Violation of assumptions given by
  - A trend (linearity), changing variance (equal variance), outliers (normality)
- Are there any obvious violations of assumptions?

```
plot(fitted(m_neuro2), rstudent(m_neuro2), xlab = "Fitted values",
     ylab = "Studentised residuals", pch = 20)
abline(h = 0)
```



Coefficient of determination R^2

- Definition of R^2 the same as for simple linear regression
  - The squared correlation between outcome y and fitted values $\hat{y}$
  - The percentage of variance explained by the regression model
- For neurocognitive example:
  - Age (simple linear regression) explains $R^2 = 5.8\%$ of the variation in speed scores
  - Age and the attention score (multiple linear regression) explain $R^2 = 45.2\%$ of the variation in speed scores
- Both of these can be read off the summaries in slides above

Big picture

- Multiple linear regression is an incredibly powerful tool
  - We've only just scratched the surface
- There are a lot of important topics we haven't covered, including
  - Model building
  - Variable selection
  - Collinearity (this is when two predictors explain similar variation)
  - Interactions (when effect of one variable depends on value of another)
  - ...
- There are lots of possible extensions
- There are also lots of ways to get ourselves into trouble
- STAT 210 explores the use of multiple linear regression for scientific problems

Summary

- Looked at multiple linear regression
  - Where we have more than one predictor variable
- Only scratched the surface
- We have looked at
  - Fitting the model
  - Interpreting the parameters
  - Finding confidence interval or performing a hypothesis test
  - Estimating the mean response and predicting a new observation
  - Assessing model fit

Outline

- Think again about categorical predictor variables
- Categorical predictors with two levels
  - Include them in a linear regression model
  - Compare to the difference in means of two independent groups
- Categorical predictors with more than two levels
  - Introduce ANOVA (analysis of variance) model

Predictor variables

- We have looked at lots of linear regression examples
- The predictor variables in these examples were
  - Height: father's height
  - Possums: total length of possum
  - Powerlifting: weight of athlete
  - Neurocognitive scores: age and attention score
- All of these are continuous variables
- Linear regression can also be used when the predictor variable is categorical
  - Represent groups or categories, e.g. sex, country of birth, blood type, etc.
  - Start with categorical variables with two levels (or groups)
    - e.g. sex: male and female

Mario Kart

- Ebay auctions for video game: Mario Kart for Nintendo Wii
  - Ebay is similar to trademe
  - Online auction website
- Two variables:
  - Total auction price: continuous outcome variable y
  - Game condition: categorical predictor variable x taking values `used` and `new`
- Another example is comparing EEG frequencies (brain waves) according to sensory deprivation (`control` or `solitary confinement`)
  - Example we considered in an earlier lecture

Hang on a minute...

- We already know how to model these data!
  - Two independent groups
    - Group 1: normally distributed with mean μ_1 and variance σ_1^2
    - Group 2: normally distributed with mean μ_2 and variance σ_2^2
  - Find confidence interval for $\mu_2 - \mu_1$ using `t.test` in R
- Why are we looking at this in the context of linear regression?
  1. Understanding: see how two independent groups is 'special case' of linear regression
  2. Useful: use categorical variables in multiple regression
     - e.g. for Mario Kart auction data: we could explore how auction length, and the number of bids, as well as game condition relate to auction price
- We will look at only one outcome variable and one categorical predictor
  - See STAT 210 for more elaborate models

Data: Mario Kart

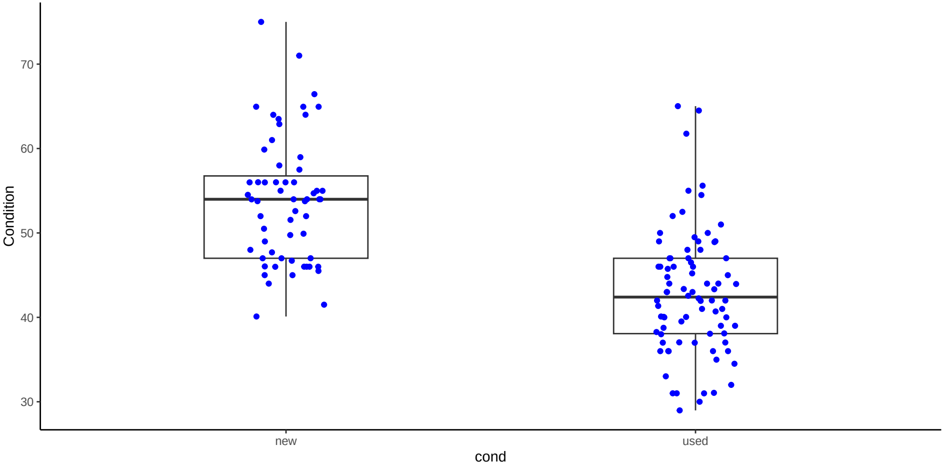
- Import the data into R

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

- The data have had two observations / outliers removed
  - The data are from a full week of auctions in October 2009
  - Removed observations: auctions where multiple games (incl. Mario Kart) were sold
- Look at the data

```
head(mario)
##   cond price
## 1  new  51.5
## 2 used  37.0
## 3  new  45.5
## 4  new  44.0
## 5  new  71.0
## 6  new  45.0
```

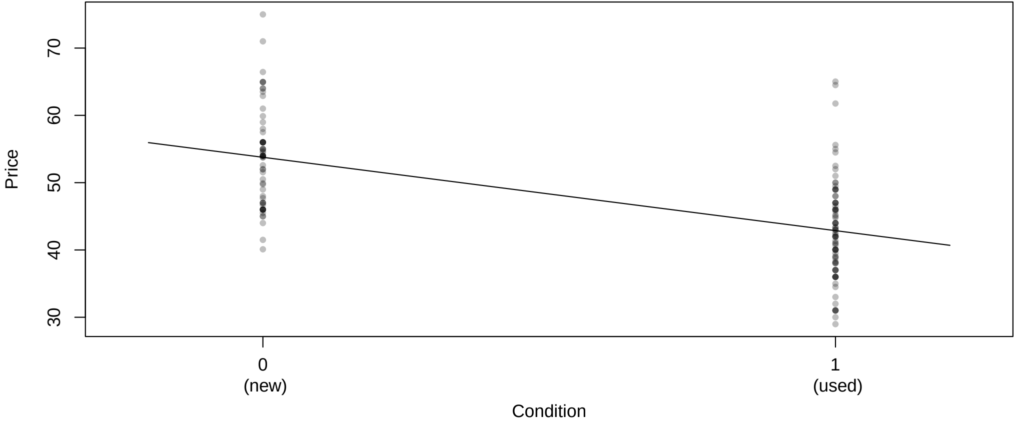
Visualisation: Mario Kart



Dummy (or indicator) variables

- The boxplot suggests a way forward
- Relabel (or encode) the condition variable to take numeric values
  - One level takes the value 0 (new)
  - Other level takes the value 1 (used)
- That is, our predictor variable x is
  - 0 if `cond = new`
  - 1 if `cond = used`
- Referred to as a dummy (or indicator) variable
- We now have a quantitative variable and can fit a regression model

Another visualisation: fitted regression



Regression model

- The mean response from a linear regression model: $\mu_y = \beta_0 + \beta_1 x$
  - The mean response when $x = 0$ (condition = new)

$$\mu_y = \beta_0 + \beta_1 x = \beta_0 + \beta_1 \times 0 = \beta_0$$

  - The mean response when $x = 1$ (condition = used)

$$\mu_y = \beta_0 + \beta_1 x = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

- β_0 is the mean response when $x = 0$
  - β_0 is the expected price when the game is new
- β_1 is the difference in mean response for $x = 1$ compared to $x = 0$
  - β_1 is the difference in the expected price between used and new games

Fitting the model in R

- To fit the model in R we could obtain the dummy variable ourselves
  - We don't have to
  - We will let R do it for us
- We make use of the data type `factor` in R
  - Used to represent categorical data
- When using a factor in R it automatically includes a dummy variable for us
  - Value 0: level that comes first in alphabet (for us this is `new`)
  - Value 1: other level (for us this is `used`)
    - This order can be changed: no reason to change it in this course
- We make `cond` a factor variable using `as.factor`

```
mario$cond = as.factor(mario$cond) # cond is now a factor variable
```

Fitting the model in R

```r
m_mario = lm(price ~ cond, data = mario)
summary(m_mario)
```

```
##
## Call:
## lm(formula = price ~ cond, data = mario)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.891  -5.831   0.129   4.129  22.149
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    53.77       0.96   56.03  < 2e-16 ***
## condused      -10.90       1.26   -8.66  1.1e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.37 on 139 degrees of freedom
## Multiple R-squared:  0.351, Adjusted R-squared:  0.346
## F-statistic:   75 on 1 and 139 DF,  p-value: 1.06e-14
```

Mario Kart: interpretation

- The fitted model is

$$\hat{y} = 53.77 - 10.9\,x, \qquad \text{or}$$

$$\widehat{\text{price}} = 53.77 - 10.9\,\text{used}$$

- The estimated expected price for new games is $\hat{\beta}_0 = 53.77$
- The estimated change in expected price for used games (compared to new games) is $\hat{\beta}_1 = -10.9$
  - We could refer to this as an estimated decrease in expected price of 10.9
- Using what we learned for linear regression:
  - We can find confidence intervals for β_1 (or β_0): see below
  - We can conduct hypothesis tests for β_1

Comparison with t.test

- Comparing linear regression (with dummy variable) to the model with two independent groups we find:
  - The parameter $\beta_0 = \mu_1$, the mean of the first group
  - The parameter $\beta_1 = \mu_2 - \mu_1$, the difference in means between the groups
- Regression model assumes equal variance: both groups have the same variance
- The independent group model allowed the two groups to have different variances
  - We can assume both groups have same variance when using `t.test`
    - Next slide
  - We can extend regression model to have different variance
    - Actually quite difficult

Comparison with t.test

- To use `t.test` we find the two groups

```
new = subset(mario, cond == "new")
used = subset(mario, cond == "used")
```

- We then use `t.test` with option `var.equal = TRUE`

```
t_mario = t.test(used$price, new$price, var.equal = TRUE)
t_mario
##
##  Two Sample t-test
##
## data:  used$price and new$price
## t = -9, df = 139, p-value = 1e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.39  -8.41
## sample estimates:
## mean of x mean of y
##      42.9      53.8
```

Comparison with t.test

- The confidence interval for $\mu_{\text{used}} - \mu_{\text{new}}$ from `t.test`

```
t_mario$conf.int

## [1] -13.387540  -8.411621
## attr(,"conf.level")
## [1] 0.95
```

- The confidence interval for β_1 when using linear regression

```
confint(m_mario, parm = 2) # parm = 2 gives CI for 2nd parameter only

##              2.5 %    97.5 %
## condused -13.38754 -8.411621
```

- They are identical!

Categorical variable: more than 2 groups

- We may be interested in categorical predictor variables with more than two groups, e.g.
  - Prioritised ethnicity (assigned to one ethnic group, even if they identify with multiple ethnicities, based on a predefined order of priority)
  - Highest education level attained (primary, high school, undergraduate, postgraduate)
  - Fertilizer (in agricultural trial)
  - Drug (control, drug A, drug B)
  - etc
- How can we extend the approach above for categorical predictors with more than two groups?

Example

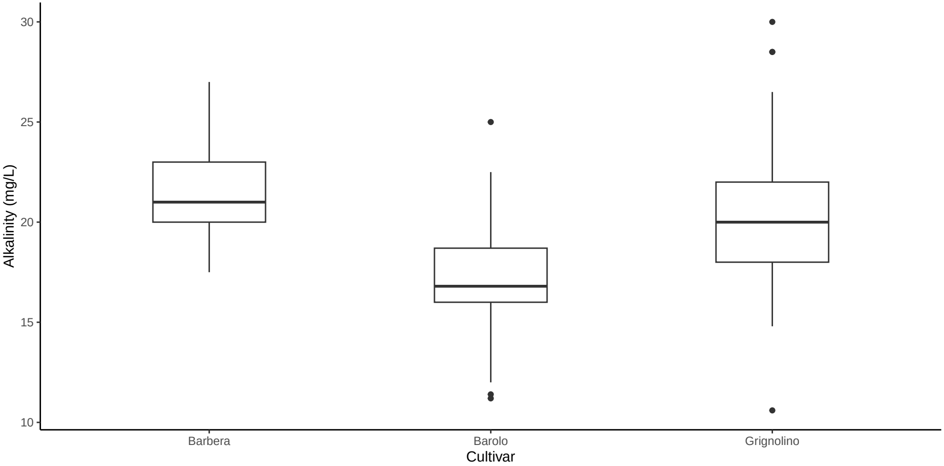
- Data on chemical composition of Italian wines
  - Three cultivars: barbera, barolo, grignolino
- We will focus on the alkalinity of the wine (measured in mg/L)
- Import the data

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

- Look at the data

```
head(wine)
##   cultivar alkalinity
## 1   Barolo       15.6
## 2   Barolo       18.6
## 3   Barolo       16.0
## 4   Barolo       18.0
## 5   Barolo       16.8
## 6   Barolo       16.0
```

Visualise the data



Statistical model: categorical predictor with K levels

- We can extend the independent group model we have seen earlier
  - Outcome variable in group 1 is normally distributed with mean μ_1 and variance σ^2
  - Outcome variable in group 2 is normally distributed with mean μ_2 and variance σ^2
  - $\ldots$
  - Outcome variable in group K is normally distributed with mean μ_K and variance σ^2
- Assume the variance is the same for all groups
- This is called an ANOVA (analysis of variance) model
  - More precisely, it is a one-way ANOVA model
- Again, this model is a special case of a linear regression
  - STAT 210 explores (and exploits) the connection in more detail

Big picture: what do we want to know

- What do we want to know: how do the mean outcome differ between groups?
  - We could look at pairwise differences in the means
    - Is there a difference in the mean alkalinity between Barbera and Grignolino
  - This approach is unreliable, particularly when there are a lot of groups (large K)
    - End up making many comparisons: with 10 groups there are 45 pairwise comparisons
    - Increased chance of finding a difference, even if there is no difference in the population
    - Look at this more in the next lecture, and later in course

Hypothesis test

- Start with a slightly different question: does the mean outcome from any group differ from the mean outcome in the other groups?
  - Is there a difference in the mean alkalinity among any of the cultivars?
- We can express this as a set of hypotheses
  - $\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_K$
  - H_A : at least one mean is different
- Develop a hypothesis test to simultaneously compare the mean of all groups
  - Next lecture

Summary

- Categorical predictor variables
- Include them in a linear regression
  - Dummy (indicator) variables
  - Relabel the two groups as 0/1
- Equivalence of linear regression (with categorical predictor) and difference in two means (independent groups)
- Introduced categorical variables with more than two groups
  - ANOVA model