STAT115: Introduction to Biostatistics

University of Otago

Ōtākou Whakaihu Waka

Lecture 24: Multiple Linear Regression

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.

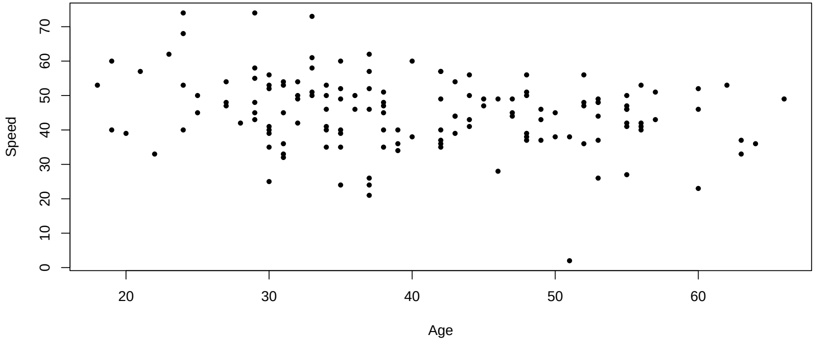
Neurocognitive scores: data

- 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: speed $= \beta_0 + \beta_1$ age $+ \varepsilon$
  - Score in the speed of processing test: outcome variable y
  - Age of participant: predictor variable x
- If we take $y =$ speed and $x =$ 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

```r
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

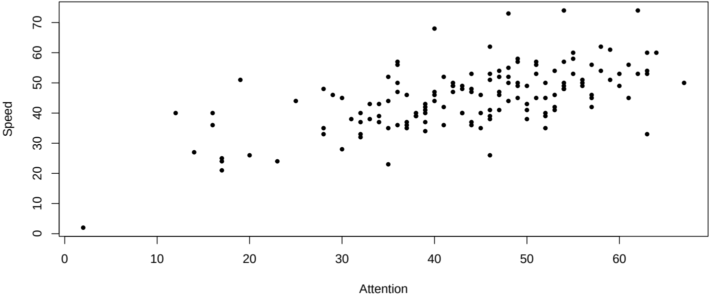
```r
confint(m_neuro)
```

```
##                 2.5 %   97.5 %
## (Intercept) 47.8914 60.40223
## age         -0.3736 -0.07447
```

- We are 95% confident that the increase in expected speed score is between -0.3736 and -0.0745 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 0.0745 and 0.3736 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 attention score, 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) 26.2216 37.1107
## age         -0.3418 -0.1499
## attention    0.4473  0.6226
```

- Interpreting the confidence interval for β_2
  - We are 90% confident that the average speed score will increase by between 0.4473 and 0.6226 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

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

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

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

```
##     fit   lwr   upr
## 1 48.58 47.14 50.02
```

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

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

```
##     fit   lwr   upr
## 1 48.58 33.11 64.05
```

Model assumptions

- The multiple linear regression model is

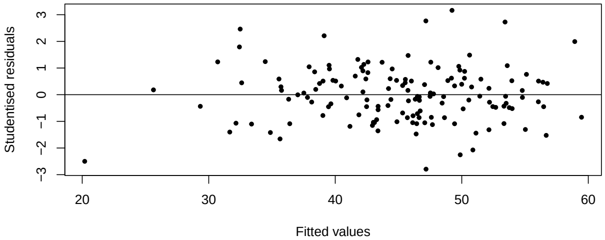
$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\mu_y} + \varepsilon$$

- 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^2_ε ('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
- 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