Multiple regression models

DClin Research Methods 1

Dr Christopher Wilson
Teesside University

Recap

- Thinking about more than outcomes. Designing studies to answer more **specific research questions / think about process**.
  - "Why is this happening?"
  - "What is the mechanism?"
- Thinking beyond significance testing. Using **confidence intervals and effect sizes** to interpret results.
  - "How big is the effect?"
  - "What is the range of plausible values?"
- Thinking about the relationship between variables. Modelling relationships between variables using regression.
  - "Does **Predictor Variable** (e.g. Treatment Group, Avoidance, Trait) predict **Outcome Variable** (e.g. Wellbeing, Depression, Behaviour)?"
  - "How much variance is explained by the model?"

In the coming weeks

- Thinking about more than outcomes. Designing studies to answer more **specific research questions / think about process**.
  - “Why is this happening?”
  - “What is the mechanism?”

- We will learn more about modelling our data to address these questions.

- Remember that analysis alone cannot answer these questions. We need to design our studies to address these questions based on theory.

Overview

- Multiple regression
- Hierarchical regression

What type of research question?

Research scenario

- Imagine we are interested in factors that predict depression in students
- In terms of demographics, age and gender have been shown to be important
- Previous research has shown that depression is associated with loneliness and stress
- More recent research has shown that self-esteem might also be a factor

Research question: Does self-esteem predict depression?

There are several different ways we could approach this:

1. We could focus on self-esteem and depression in isolation (i.e., simple regression)

2. We could include the other variables as covariates in a single model (i.e., multiple regression)
3. We could use a hierarchical approach, where we enter the variables in stages, to see the variance explained by self-esteem, above and beyond what is explained by the other variables (i.e., hierarchical regression)

Multiple regression

Multiple regression

- We will use the multiple regression approach to answer our research question

There are some considerations:

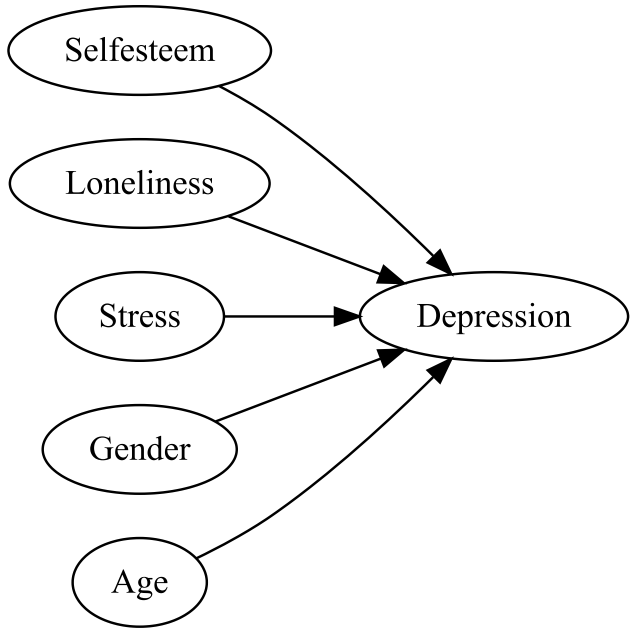
- More variables means a larger sample size is required (power analysis)
- There is an additional assumption called *multicollinearity*. This means that the predictor variables should not be too highly correlated with each other

Running the analysis

- The process is the same as what we have done previously. The additional predictors are added to the model.

```
model1 <- lm(data = data, lm(depression ~ age + gender + stress + lonelines
```

Visualising this model



Testing multicollinearity

- We can test for multicollinearity using the mctest() function, from the mctest package

```r
library(mctest)

mctest(model1, type= "b")
```

① The mctest() function takes a model as an argument. The type argument specifies the type of tests to run. The b argument specifies that we want to test both overall and individual predictor multicollinearity.

```r
library(mctest)

mctest(model1, type= "b")
```

```
Call:
omcdiag(mod = mod, Inter = Inter, detr = detr, red = red, conf = conf,
    theil = theil, cn = cn)


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.3331         0
Farrar Chi-Square:       216.0184         1
Red Indicator:             0.3329         0
Sum of Lambda Inverse:     7.6888         0
Theil's Method:           -1.9998         0
Condition Number:        103.6436         1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test


===================================

Call:
imcdiag(mod = mod, method = method, corr = FALSE, vif = vif,
    tol = tol, conf = conf, cvif = cvif, ind1 = ind1, ind2 = ind2,
    leamer = leamer, all = all)


All Individual Multicollinearity Diagnostics Result

              VIF    TOL      Wi      Fi Leamer    CVIF Klein   IND1   IND2
age        1.0263 0.9744  1.2811  1.7169 0.9871 -0.1753     0 0.0200 0.0893
gender1    1.0242 0.9764  1.1791  1.5802 0.9881 -0.1750     0 0.0200 0.0824
stress     1.6061 0.6226 29.5495 39.6014 0.7891 -0.2744     0 0.0128 1.3164
loneliness 1.9477 0.5134 46.2012 61.9175 0.7165 -0.3328     0 0.0105 1.6972
selfesteem 2.0844 0.4797 52.8658 70.8492 0.6926 -0.3561     0 0.0098 1.8147

1 --> COLLINEARITY is detected by the test
```

```
0 --> COLLINEARITY is not detected by the test

age , gender1 , coefficient(s) are non-significant may be due to
multicollinearity

R-square of y on all x: 0.8583

* use method argument to check which regressors may be the reason of
collinearity
===================================
```

Higher variance inflation factors (VIFs) indicate higher multicollinearity. A VIF of 5 or more is considered problematic.

Interpreting the output of mctest()

- The output of the mctest() function is several different tests of multicollinearity
- We need to review them as a whole and make a judgement about whether multicollinearity is a problem
- If there seems to be a problem, we would need to look into the data to see which of the variables are highly correlated

What to do if multicollinearity exists:

- Remove some of the highly correlated predictors
- Linearly combine some predictors.
- Perform an analysis designed for highly correlated variables (e.g. = PCA or partial least squares regression)

Remember, we also need to test the other assumptions

- There is another package called gvlma that provides diagnostics for all of the assumptions of linear regression.
- We can use this in combination with our diagnostic plots (from last week)

```
library(gvlma)

gvlma(model1)
```

Viewing the output of gvlma()

```r
library(gvlma)

gvlma(model1)
```

```
Call:
lm(formula = lm(depression ~ age + gender + stress + loneliness +
    selfesteem), data = data)

Coefficients:
(Intercept)          age      gender1       stress   loneliness   selfesteem
  -15.00538     -0.03025     -0.03374      0.10901      0.25165      0.54264


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = model1)

                    Value p-value                   Decision
Global Stat        8.6719 0.06984    Assumptions acceptable.
Skewness           3.6750 0.05524    Assumptions acceptable.
Kurtosis           0.3004 0.58365    Assumptions acceptable.
Link Function      0.4887 0.48450    Assumptions acceptable.
Heteroscedasticity 4.2079 0.04024 Assumptions NOT satisfied!
```

- Global Statistic: Test of the overall model.
- Link function test: Is this relationship linear?

Looking at the output of multiple regression

```
1 summary(model1)
```

```
Call:
lm(formula = lm(depression ~ age + gender + stress + loneliness +
    selfesteem), data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74865 -0.57635 -0.00253  0.52990  2.20394

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.00538    1.94985  -7.696 6.97e-13 ***
age          -0.03025    0.05233  -0.578    0.564
gender1      -0.03374    0.11160  -0.302    0.763
stress        0.10901    0.02179   5.002 1.27e-06 ***
loneliness    0.25165    0.03354   7.502 2.20e-12 ***
selfesteem    0.54264    0.03553  15.273  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7797 on 194 degrees of freedom
Multiple R-squared:  0.8583,	Adjusted R-squared:  0.8547
F-statistic:   235 on 5 and 194 DF,  p-value: < 2.2e-16
```

Interpreting the output of multiple regression

- First we look at the overall model significance and R^2 values. These tell us whether the model is significant and how much variance is explained by the model.
- If the overall model is significant, we look at the individual predictors. We look at the significance of the predictors and the coefficient values (Estimate).

Hierarchical regression

Hierarchical regression - research scenario

- Imagine we are interested in factors that predict depression in students
- In terms of demographics, age and gender have been shown to be important
- Previous research has shown that depression is associated with loneliness and stress
- More recent research has shown that self-esteem might also be a factor

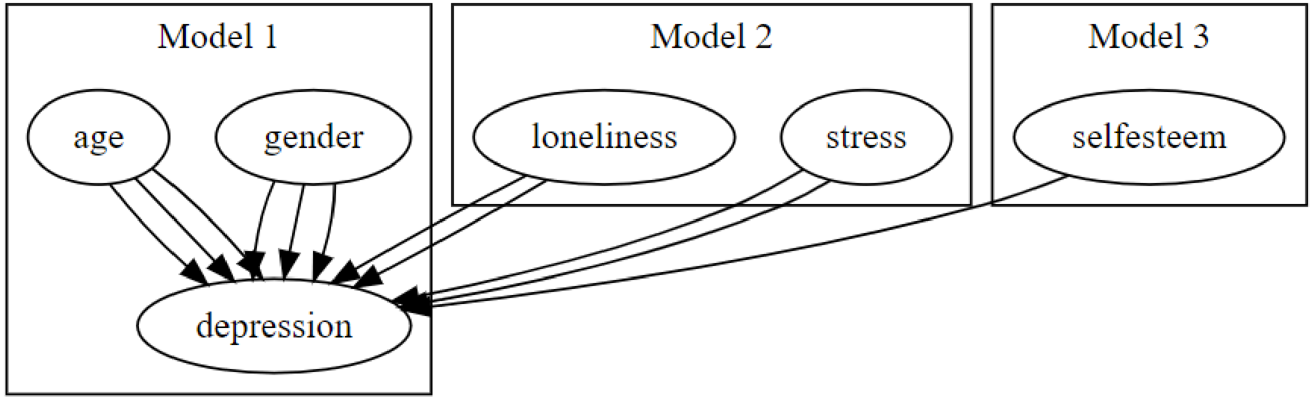
Using a hierarchical approach

- We could use a hierarchical approach, where we enter the variables in stages.
- This involves running several regression models, each with a different set of predictors
- We can then compare the models to see how much variance is explained by each set of predictors

Which models would we test?

- This depends on our understanding of the variables
- For example, we might do the following:
  - Model 1: Demographics
  - Model 2: Demographics + stress + loneliness
  - Model 3: Demographics + stress + loneliness + self-esteem

Visualising models



Running the analysis

```
model0 <- lm(depression ~ 1, data = data) ①

model1 <- lm(depression ~ age + gender, data = data) ②

model2 <- lm(depression~ age + gender + stress + loneliness, data = data) ③

model3 <- lm(depression ~ age + gender + stress + loneliness + selfesteem, ④
```

① Model 0 is the null model. It is a model with no predictors. It is used as a baseline for comparison.

② Model 1 is the first model. It includes the demographic variables.

③ Model 2 is the second model. It includes the demographic variables, stress and loneliness.

④ Model 3 is the third model. It includes the demographic variables, stress, loneliness and self-esteem.

Comparing the models

- We can compare the models using the anova() function

```
1  anova(model0, model1, model2, model3)
```

```
Analysis of Variance Table

Model 1: depression ~ 1
Model 2: depression ~ age + gender
Model 3: depression ~ age + gender + stress + loneliness
Model 4: depression ~ age + gender + stress + loneliness + selfesteem
  Res.Df    RSS Df Sum of Sq        F  Pr(>F)
1    199 832.35
2    197 826.62  2      5.73   4.7127 0.01003 *
3    195 259.75  2    566.88 466.2325 < 2e-16 ***
4    194 117.94  1    141.81 233.2590 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparing the models

- The anova() function compares the change in R^2 between the models
- If the change in R^2 is significant, then the model is significantly better than the previous model
- This way, we can see the value of each set of predictors

What is the intercept-only model?

- The intercept-only model is the null model.
- It is a model with no predictors.
- It is used as a baseline for comparison, so we can see if the the first set of predictors (i.e., demographics) explain more variance than the null model.

Assessing the quality of the models

- Significance is not the only thing we should look at when comparing models
- There are several measures of model fit that we can use to assess the quality of regression models
- AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are two commonly used measures
- Both are assessing the same thing: the trade-off between model fit and model complexity

Assessing the quality of the models

- We can use the built-in AIC() and BIC() functions to calculate these measures

```
AIC(model0, model1, model2, model3)
```

```
       df      AIC
model0  2 856.7637
model1  4 859.3821
model2  6 631.8528
model3  7 475.9464
```

```
BIC(model0, model1, model2, model3)
```

```
       df      BIC
model0  2 863.3604
model1  4 872.5754
model2  6 651.6427
model3  7 499.0346
```

Lower values indicate better model fit (i.e., the model explains more variance), taking into account the number of predictors.

Models and hypothesis testing

What do the models tell us about our research question?

- We were interested in whether self-esteem predicts depression
- The models tell us that self-esteem explains a significant amount of variance in depression, above and beyond:
  - Demographics (age and gender)
  - stress + loneliness (identified in previous research)
- This gives us a clear indication that self-esteem is an important predictor of depression
- The final model is the best model in terms of R^2 and model fit, suggesting that all of the predictors are important

Final points

- Model building should be theory driven
- We should have a clear rationale for the predictors we include
- We should have a clear rationale for the order in which we enter the predictors, based on previous research