Multiple Regression

Advanced Psychological Research Methods

Dr Christopher Wilson

Questions from last week's session?

Submit your attendance

Attendance code: 9456

http://bit.ly/APRM22

Note from last week:

What do we do if our assumptions are violated?

- Normalilty: transformation or bootstrapping
- Linearity: Consider alternatives such as non-linear regression or polynomial approaches
- Homogeneity of variance or influential cases: Robust regression can reduce standard errors

Overview

- What is multiple regression?
- Assumptions of multiple regression
- Sample size in regression
- Using categorical predictors in R
- Testing all predictors at once
  - Interpreting the output of Multiple Regression
- Hierarchical regression
- Stepwise regression

What is multiple regression?

- An extension of simple regression
- Same format as simple regression but adding each predictor:

$$Y = b_1 X_1 + b_2 X_2 + b_0$$

(The constant can be referred to in the equation as **c** or **b0**)

What are the assumptions of Multiple Regression?

- They are primarily the same as simple regression
- The additional assumption of no **multicollinearity** (due to having multiple predictors)
  - i.e. predictors should not be highly correlated

What is multicollinearity?

- Multicollinearity = predictors correlated highly with each other.
- This is not good because:
  - It makes it difficult to determine the role of individual predictors
  - Increases the error of the model (higher standard errors)
  - Difficult to identify significant predictors - wider confidence interval

Testing multicollinearity

```r
## use the mctest package

library(mctest)

m1 <- lm(aggression_level ~ treatment_group + treatment_duration + trust_score, data=regressi

mctest(m1)
```

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


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.9229         0
Farrar Chi-Square:         7.7960         0
Red Indicator:             0.1547         0
Sum of Lambda Inverse:     3.1728         0
Theil's Method:           -0.8800         0
Condition Number:         13.6549         0
```

- The format of *mctest()* is:

  mctest(model)

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)

Sample size for multiple regression

- Is based on the number of predictors
- More predictors = more participants needed
- **Do a power analysis**
- Loose “rule of thumb” = 10-15 participants per predictor

Approaches to multiple regression: All predictors at once #1

Research question: Do a client’s treatment duration and treatment group predict aggression level?

```r
model1 <- lm(data = regression_data, aggression_level ~ treatment_duration + treatment_group)
```

- Here we are including all of the predictors at the same time
- Note that we are using a plus sign + between each predictor
    - This means that no interactions will be tested

Using categorical predictors in R

- Treatment group is a categorical (also called “nominal” or “factor”) variable
- No special “dummy coding” is required in R to use categorical predictors in regression
- R will use the first group as the reference category and test whether being in another group shows a significant difference
- R chooses the reference group based on numerical value or alphabetical order
- If you want you can change the reference category or “force” it using the relevel function:

```r
regression_data$treatment_group <- relevel(regression_data$treatment_group, ref = "therapy1")
```

Reviewing the output

```r
summary(model1)
```

```
Call:
lm(formula = aggression_level ~ treatment_duration + treatment_group,
    data = regression_data)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9468 -1.1104  0.0205  0.9621  3.4481

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             11.58713    0.77331  14.984  < 2e-16 ***
treatment_duration      -0.66024    0.07119  -9.274 4.96e-15 ***
treatment_grouptherapy2  0.85032    0.30449   2.793   0.0063 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpreting the output

- Multiple R^2 = Total variance in outcome that is explained by the model
- p-value = Statistical significance of the model
- Coefficients = Contribution of each predictor to the model
  - Pr = Significance of the individual predictor
  - Estimate = Change in the outcome level that occurs when the predictor increases by 1 unit of measurement

Approaches to multiple regression: All predictors at once #2

Research questions: - Do a client’s treatment duration and treatment group predict aggression level - Do the predictors interact?

```r
model2 <- lm(data = regression_data, aggression_level ~ treatment_duration * treatment_group)
```

- Here we are including all of the predictors at the same time
- Note that we are using an asterisk * between each predictor
  - This means that interactions will be tested

Reviewing the output

```r
summary(model2) %>% coefficients
```

```
                                           Estimate Std. Error    t value
(Intercept)                              12.3529190  1.1006127 11.2236751
treatment_duration                       -0.7334435  0.1033086 -7.0995381
treatment_grouptherapy2                  -0.5615517  1.4753596 -0.3806202
treatment_duration:treatment_grouptherapy2  0.1394649  0.1425977  0.9780305
                                             Pr(>|t|)
(Intercept)                              3.599000e-19
treatment_duration                       2.166226e-10
treatment_grouptherapy2                  7.043260e-01
treatment_duration:treatment_grouptherapy2 3.305175e-01
```

- We get additional information in the coefficients table about the interaction between variables
  - e.g. does the interaction between level of trust and treatment duration predict the outcome (aggression level)?
- We can see from the output that none of the interactions are significant

Hierarchical multiple regression: Theory driven “blocks” of variables

- It might be the case that we have previous research or theory to guide how we run the analysis
- For example, we might know that treatment duration and therapy group are likely to predict the outcome
- We might want to check whether client’s level of trust in the clinician has any **additional** impact on our ability to predict the outcome (aggression level)

Hierarchical multiple regression: Theory driven “blocks” of variables

- To do this, we run three regression models
  - Model 1: treatment duration and therapy group
  - Model 2: treatment duration and therapy group and trust score
- We then compare the two regression models to see if:
  - Model 2 is better than Model 1

Hierarchical multiple regression: Running and comparing 2 models

```r
## run regression using the same method as above
model1 <- lm(data = regression_data, aggression_level ~ treatment_duration + treatment_group)
model2 <- lm(data = regression_data, aggression_level ~ treatment_duration + treatment_group

## use the aov() command to compare the models
anova(model1,model2)
```

```
Analysis of Variance Table

Model 1: aggression_level ~ treatment_duration + treatment_group
Model 2: aggression_level ~ treatment_duration + treatment_group + trust_score
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     97
2     96 217.86  1     0.399 0.1757   0.676
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We can see that:
    - Model 2 (treatment duration, treatment group and trust score) shows no significant change compared to Model 1

Stepwise multiple regression: computational selection of predictors

- Stepwise multiple regression is controversial because:
    - The computer selects which predictors to include based on Akaike information criterion (AIC)
        - This is a calculation of the quality of statistical models when they are compared to each other

What's the problem?

- This selection is not based on any underlying theory or understanding of the real-life relationship between the variables

Stepwise multiple regression: loading the MASS package and run the full model

1. **install and load the MASS package**
2. **run a regression model with all of the variables**
3. use the *stepAIC()* command on the full model to run stepwise regression
4. View the best model

```r
library(MASS)

# Run the full model
full.model <- lm(data = regression_data, aggression_level ~ treatment_duration + treatment_gr
```

Stepwise multiple regression: Use stepAIC( ) with options

- **Trace** *(TRUE or FALSE)*: do we want to see the steps that were involved in selecting the best model ?
- **Direction** *("forward", "backward" or "both")*:
  - start with no variables and add them *(forward)*
  - start with all variables and subtract them *(backward)*
  - use both approaches *(both)*

```r
# Run stepwise
step.model <- stepAIC(full.model, direction = "both", trace = TRUE)
```

```
Start:  AIC=85.87
aggression_level ~ treatment_duration + treatment_group + trust_score

                     Df Sum of Sq    RSS     AIC
- trust_score         1     0.399 218.26  84.052
<none>                            217.86  85.869
- treatment_group     1    17.877 235.74  91.755
- treatment_duration  1   188.709 406.57 146.259

Step:  AIC=84.05
```

```
aggression_level ~ treatment_duration + treatment_group

                     Df Sum of Sq    RSS    AIC
<none>                            218.26  84.052
+ trust_score         1     0.399 217.86  85.869
```

Stepwise multiple regression: Display the best model

1. install and load the MASS package

2. run a regression model with all of the variables

3. **use the *stepAIC()* command on the full model to run stepwise regression**

4. **View best model**

```
#view the stepwise output
summary(step.model)
```

```
Call:
lm(formula = aggression_level ~ treatment_duration + treatment_group,
    data = regression_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9468 -1.1104  0.0205  0.9621  3.4481 

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)    
(Intercept)             11.58713    0.77331  14.984  < 2e-16 ***
treatment_duration      -0.66024    0.07119  -9.274 4.96e-15 ***
treatment_grouptherapy2  0.85032    0.30449   2.793   0.0063 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Summary

- Multiple regression is an extension of simple regression
- We need to check the same assumptions + multicolinearity
- When entering multiple predictors:
  - Heirarchical: we have a theoretical basis for the models
  - Stepwise: the computer selects the best model
- Comparing multiple models using Akaike information criterion (AIC)

Questions?

Submit your attendance

Attendance code: 9456

http://bit.ly/APRM22