Multiple Regression

Christopher Wilson

Advanced Psychological Research Methods

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:

$$Y = b_1X_1 + b_2X_2 + b_0$$

(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

```
## use the mctest package
# install.packages('mctest')o
library(mctest)


mctest(cbind(regression_data$treatment_duration,regression_data$treatment_gro
up,regression_data$trust_score),
        regression_data$aggression_level)

##
## Call:
## omcdiag(x = x, y = y, 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:         16.0564         0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
```

- The format of *mctest()* is:

  mctest(predictors, outcome)

- In the above example we used the *cbind()* function to bind 3 columns of data together (the predictors)

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?

```
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:

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

Reviewing the output

```
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
##
## Residual standard error: 1.5 on 97 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:  0.5107
## F-statistic: 52.67 on 2 and 97 DF,  p-value: 3.267e-16
```

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?

```
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

```
summary(model2) %>% coefficients

##                                              Estimate Std. Error
## (Intercept)                                12.3529190  1.1006127
## treatment_duration                         -0.7334435  0.1033086
## treatment_grouptherapy2                    -0.5615517  1.4753596
## treatment_duration:treatment_grouptherapy2  0.1394649  0.1425977
##                                               t value     Pr(>|t|)
## (Intercept)                                11.2236751 3.599000e-19
## treatment_duration                         -7.0995381 2.166226e-10
## treatment_grouptherapy2                    -0.3806202 7.043260e-01
## treatment_duration:treatment_grouptherapy2  0.9780305 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)
- To do this, we run three regression models
  - Model 0: the constant (baseline)
  - 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 1 is better than Model 0 (the constant)
  - Model 2 is better than Model 1

Hierarchical multiple regression: Running and comparing 2 models

```
## run regression using the same method as above
model0 <- lm(data = regression_data, aggression_level ~ 1)
model1 <- lm(data = regression_data, aggression_level ~ treatment_duration +
treatment_group)
model2 <- lm(data = regression_data, aggression_level ~ treatment_duration +
treatment_group + trust_score)

## use the aov() command to compare the models
anova(model0,model1,model2)

## Analysis of Variance Table
##
## Model 1: aggression_level ~ 1
## Model 2: aggression_level ~ treatment_duration + treatment_group
## Model 3: aggression_level ~ treatment_duration + treatment_group +
trust_score
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     99 455.27
## 2     97 218.26  2   237.013 52.2195 4.507e-16 ***
## 3     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 1 (treatment duration and treatment group) is significant relative to the constant (Model 0)
  - 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

```
library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

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

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)*

```
# 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
## - treatment_group     1    17.547 235.81  89.785
## - treatment_duration  1   193.515 411.78 145.531
```

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
##
## Residual standard error: 1.5 on 97 degrees of freedom
## Multiple R-squared:  0.5206, Adjusted R-squared:  0.5107
## F-statistic: 52.67 on 2 and 97 DF,  p-value: 3.267e-16
```

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)