13.1: Model Building

Model building is the process of deciding which independent variables to include in the model.[22] For our purposes, when deciding which variables to include, theory and findings from the extant literature should be the most prominent guides. Apart from theory, however, this chapter examines empirical strategies that can help determine if the addition of new variables improves overall model fit. In general, when adding a variable, check for: a) improved prediction based on empirical indicators, b) statistically and substantively significant estimated coefficients, and c) stability of model coefficients—do other coefficients change when adding the new one – particularly look for sign changes.

13.1.1 Theory and Hypotheses

The most important guidance for deciding whether a variable (or variables) should be included in your model is provided by theory and prior research. Simply put, knowing the literature on your topic is vital to knowing what variables are important. You should be able to articulate a clear theoretical reason for including each variable in your model. In those cases where you don't have much theoretical guidance, however, you should use model *parsimony*, which is a function of simplicity and model fit, as your guide. You can focus on whether the inclusion of a variable improves model fit. In the next section, we will explore several empirical indicators that can be used to evaluate the appropriateness of variable inclusion.

13.1.2 Empirical Indicators

When building a model, it is best to start with a few IV's and then begin adding other variables. However, when adding a variable, check for:

- Improved prediction (increase in adjusted R2R2)
- Statistically and substantively significant estimated coefficients
- Stability of model coefficients
- Do other coefficients change when adding the new one?
- Particularly look for sign changes for estimated coefficients.

Coefficient of Determination: R2R2

R2R2 was previously discussed within the context of simple regression. The extension to multiple regression is straightforward, except that multiple regression leads us to place greater weight on the use of the **adjusted R2R2**. Recall that the adjusted R2R2 corrects for the inclusion of multiple independent variables; R2R2 is the ratio of the explained sum of squares to the total sum of squares (*ESS/TSS*).

R2R2 is expressed as:

R2=1−RSSTSS(13.1)(13.1)R2=1−RSSTSS

However, this formulation of R2R2 is insensitive to the complexity of the model and the degrees of freedom provided by your data. This means that an increase in the number of kk independent variables, can increase the R2R2. Adjusted R2R2 penalizes the R2R2 by correcting for the degrees of freedom. It is defined as:

adjustedR2=1−RSSn−k−1TSSn−k−1(13.2)(13.2)adjustedR2=1−RSSn−k−1TSSn−k−1

The R2R2 of two models can be compared, as illustrated by the following example. The first (simpler) model consists of basic demographics (age, education, and income) as predictors of climate change risk. The second (more complex) model adds the variable measuring political ideology to the explanation.

```
ds.temp <- filter(ds) %>%
  dplyr::select(glbcc_risk, age, education, income, ideol) %>%
  na.omit()

ols1 <- lm(glbcc_risk ~ age + education + income, data = ds.temp)
summary(ols1)
```

```
##
## Call:
## lm(formula = glbcc_risk ~ age + education + income, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9189 -2.0546  0.0828  2.5823  5.1908
##
## Coefficients:
##                Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)  6.160506689  0.342491831  17.987 < 0.0000000000000002 ***
## age         -0.015571138  0.004519107  -3.446              0.00058 ***
## education    0.225285858  0.036572082   6.160       0.000000000858 ***
## income      -0.000005576  0.000001110  -5.022       0.000000551452 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.008 on 2268 degrees of freedom
## Multiple R-squared:  0.02565,    Adjusted R-squared:  0.02437
## F-statistic: 19.91 on 3 and 2268 DF,  p-value: 0.0000000000009815
```

```
ols2 <- lm(glbcc_risk ~ age + education + income + ideol, data = ds.temp)
summary(ols2)
```

```
##
## Call:
## lm(formula = glbcc_risk ~ age + education + income + ideol, data = ds.temp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7991 -1.6654  0.2246  1.4437  6.5968
##
## Coefficients:
##                  Estimate    Std. Error t value             Pr(>|t|)
## (Intercept) 10.9232861851  0.3092149750  35.326 < 0.0000000000000002 ***
## age         -0.0044231931  0.0036688855  -1.206              0.22810
## education    0.0632823391  0.0299443094   2.113              0.03468 *
## income      -0.0000026033  0.0000009021  -2.886              0.00394 **
## ideol       -1.0366154295  0.0299166747 -34.650 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.433 on 2267 degrees of freedom
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.3619
## F-statistic:   323 on 4 and 2267 DF,  p-value: < 0.00000000000000022
```

As can be seen by comparing the model results, the more complex model that includes political ideology has a higher R2R2 than does the simpler model. This indicates that the more complex model explains a greater fraction of the variance in perceived risks of climate change. However, we don't know if this improvement is statistically significant. In order to determine whether the more complex model adds significantly to the explanation of perceive risks, we can utilize the FF-test.

FF-test

The FF-test is a test statistic based on the FF distribution, in the same way the the tt-test is based on the tt distribution. The FF distribution skews right and ranges between 00 and ∞∞. Just like the tt distribution, the FF distribution approaches normal as the degrees of freedom increase.^[Note that the FF distribution is the square of a tt-distributed variable with mm degrees of freedom. The FF distribution has 11 degree of freedom in the numerator and mm degrees of in the denominator:t2m=F1,mtm2=F1,m

FF-tests are used to test for the statistical significance of the overall model fit. The null hypothesis for an FF-test is that the model offers no improvement for predicting YiYi over the mean of YY, ¯YY¯.

The formula for the FF-test is:

F=ESSkRSSn−k−1(13.3)(13.3)F=ESSkRSSn−k−1

where kk is the number of parameters and n−k−1n−k−1 are the degrees of freedom. Therefore, FF is a ratio of the explained variance to the residual variance, correcting for the number of observations and parameters. The FF-value is compared to the FF-distribution, just like a tt-distribution, to obtain a pp-value. Note that the R output includes the FF statistic and pp value.

Nested FF-test

For model building we turn to the nested FF-test, which tests whether a more complex model (with more IVs) adds to the explanatory power over a simpler model (with fewer IVs). To find out, we calculate an F-statistic for the model improvement:

F=ESS1−ESS0qRSS1n−k−1(13.4)(13.4)F=ESS1−ESS0qRSS1n−k−1

where qq is the difference in the number of IVs between the simpler and the more complex models. The complex model has kk IVs (and estimates kk parameters), and the simpler model has k−qk−q IVs (and estimates only k−qk−q parameters). ESS1ESS1 is the explained sum of squares for the complex model. RSS1RSS1 is the residual sum of squares for the complex model. ESS0ESS0 is the explained sum of squares for the simpler model. So the nested-F represents the ratio of the additional explanation per added IV, over the residual sum of squares divided by the model degrees of freedom.

We can use R , to calculate the FF statistic based on our previous example.

```
TSS <- sum((ds.temp$glbcc_risk-mean(ds.temp$glbcc_risk))^2)
TSS
```

```
## [1] 21059.86
```

```
RSS.mod1 <- sum(ols1$residuals^2)
RSS.mod1
```

```
## [1] 20519.57
```

```
ESS.mod1 <- TSS-RSS.mod1
ESS.mod1
```

```
## [1] 540.2891
```

```
RSS.mod2 <- sum(ols2$residuals^2)
RSS.mod2
```

```
## [1] 13414.89
```

```
ESS.mod2 <- TSS-RSS.mod2
ESS.mod2
```

```
## [1] 7644.965
```

```
F <- ((ESS.mod2 - ESS.mod1)/1)/(RSS.mod2/(length(ds.temp$glbcc_risk)-4-1))
F
```

```
## [1] 1200.629
```

Or, you can simply use the `anova` function in RR:

```
anova(ols1,ols2)
```

```
## Analysis of Variance Table
##
## Model 1: glbcc_risk ~ age + education + income
## Model 2: glbcc_risk ~ age + education + income + ideol
##   Res.Df   RSS Df Sum of Sq      F               Pr(>F)
## 1   2268 20520
## 2   2267 13415  1    7104.7 1200.6 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown using both approaches, the inclusion of ideology significantly improves model fit.

13.1.3 Risks in Model Building

As is true of most things in life, there are risks to consider when building statistical models. First, are you including irrelevant XX's? These can increase model complexity, reduce adjusted R2R2, and increase model variability across samples. Remember that you should have a theoretical basis for inclusion of all of the variables in your model.

Second, are you omitting relevant XX's? Not including important variables can fail to capture fit and can bias other estimated coefficients, particularly when the omitted XX is related to both other XX's and to the dependent variable YY.

Finally, remember that we are using sample data. Therefore, about 5% of the time, our sample will include random observations of XX's that result in BB's that meet classical hypothesis tests – resulting in a Type I error. Conversely, the BB's may be important, but the sample data will randomly include observations of XX that result in estimated parameters that do not meet the classical statistical tests – resulting in a Type II error. That's why we rely on theory, prior hypotheses, and replication.