1

Multiple Regression and Model Building

This book focuses on the use of systematic quantitative analysis for purposes of building, refining and testing theoretical propositions in the policy and social sciences. All of the tools discussed so far – including univariate, bi-variate, and simple regression analysis – provide means to evaluate distributions and test hypotheses concerning simple relationships. Most policy and social theories, however, include multiple explanatory variables. *Multiple regression* extends the utility of simple regression by permitting the inclusion of two or more explanatory variables. This chapter discusses strategies for determining what variables to include (or exclude) in the model. As before, we use the `tbur` data.

1.1 Model Building

Model building is the process of deciding which independent variables to include in the model.[1] 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.

1.1.1 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

[1] Model building also concerns decisions about model functional form, which we address in the next chapter.

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 inclusion of variables.

1.1.2 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:

1. Improved prediction (increase in adjusted R^2)
2. Statistically and substantively significant estimated coefficients
3. Stability of model coefficients
    - Do other coefficients change when adding the new one?
    - Particularly look for sign changes for estimated coefficients.

Coefficient of Determination: R^2

R^2 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** R^2. Recall that the adjusted R^2 corrects for the inclusion of multiple independent variables; R^2 is the ratio of the explained sum of squares to the total sum of squares (ESS/TSS). The components of R^2 for an observation are illustrated in Figure 1.1. As before, for each observation Y_i, variation around the mean can be decomposed into that which is "explained" by the regression model and that which is not.

R^2 is expressed as:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{1.1}$$

However, this formulation of R^2 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 k independent variables, can increase the R^2. Adjusted R^2 penalizes the R^2 by correcting for the degrees of freedom. It is defined as:

$$\text{adjusted}R^2 = 1 - \frac{\frac{RSS}{n-k-1}}{\frac{TSS}{n-k-1}} \tag{1.2}$$

The R^2 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) %>%
  select(glbcc_risk, age, education, income, ideol) %>%
  na.omit()

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

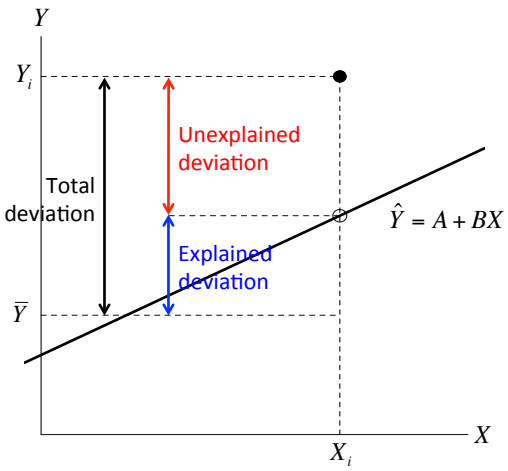

Figure 1.1

```
##
## 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)  6.161e+00  3.425e-01  17.987  < 2e-16 ***
## age         -1.557e-02  4.519e-03  -3.446  0.00058 ***
## education    2.253e-01  3.657e-02   6.160 8.58e-10 ***
## income      -5.576e-06  1.110e-06  -5.022 5.51e-07 ***
## ---
## 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: 9.815e-13

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)  1.092e+01  3.092e-01  35.326  < 2e-16 ***
## age         -4.423e-03  3.669e-03  -1.206  0.22810
## education    6.328e-02  2.994e-02   2.113  0.03468 *
## income      -2.603e-06  9.021e-07  -2.886  0.00394 **
## ideol       -1.037e+00  2.992e-02 -34.650  < 2e-16 ***
## ---
## 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: < 2.2e-16
```

As can be seen by comparing the model results, the more complex model that includes political ideology has a higher R^2 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 F-test.

F-test

The F-test is a test statistic based on the F distribution, in the same way the the t-test is based on the t distribution. The F distribution skews right and ranges between 0 and ∞. Just like the t distribution, the F distribution approaches normal as the degrees of freedom increase.[2]

F-tests are used to test for the statistical significance of the overall model fit. The null hypothesis for an F-test is that the model offers no improvement for predicting Y_i over the mean of Y, $\bar{Y}$.

The formula for the F-test is:

$$F = \frac{\frac{ESS}{k}}{\frac{RSS}{n-k-1}} \tag{1.3}$$

where k is the number of parameters and $n - k - 1$ are the degrees of freedom. Therefore, F is a ratio of the explained variance to the residual variance, correcting for the number

[2]Note that the F distribution is the square of a t-distributed variable with m degrees of freedom. The F distribution has 1 degree of freedom in the numerator and m degrees of in the denominator:

$$t_m^2 = F_{1,m}$$

of observations and parameters. The F-value is compared to the F-distribution, just like a t-distribution, to obtain a p-value. Note that the R output includes the F statistic and p value.

Nested F-test

For model building we turn to the nested F-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:

$$F = \frac{\frac{ESS_1 - ESS_0}{q}}{\frac{RSS_1}{n-k-1}} \tag{1.4}$$

where q is the difference in the number of IVs between the simpler and the mode complex models. The complex model has k IVs (and estimates k parameters), and the simpler model has $k-q$ IVs (and estimates only $k-q$ parameters). ESS_1 is the explained sum of squares for the complex model. RSS_1 is the residual sum of squares for the complex model. ESS_0 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 F 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 *R*:

```
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 < 2.2e-16 ***
## ---
## 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.

1.1.3 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 X's? These can increase model complexity, reduce adjusted R^2, 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 X's? Not including important variables can fail to capture fit and can bias other estimated coefficients, particularly when the omitted X is related to both other X's and to the dependent variable Y.

Finally, remember that we are using sample data. Therefore, about 5% of the time, our sample will include random observations of X's that result in B's that meet classical hypothesis tests – resulting in a Type I error. Conversely, the B's may be important, but the sample data will randomly include observations of X 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.

1.2 Evils of Stepwise Regression

Almost all statistical software packages (including *R*) permit a number of mechanical "search strategies" for finding IVs that make a statistically significant contribution to the prediction of the model dependent variable. The most common of these is called `stepwise regression`, which may also be referred to as forward, backward (or maybe even upside down!) stepwise regression. Stepwise procedures do not require that the analyst think – you just have to designate a pool of possible IVs and let the package go to work, sifting through the IVs to identify those that (on the basis of your sample data) appear to be related to the model dependent variable. The stepwise procedures use sequential F-tests, sequentially adding variables that "improve the fit" of the mindless model until there are no more IVs that meet some threshold (usually $p < 0.05$) of statistical significance. These procedures are like mechanically wringing all of the explanation you can get for Y out of some pool of X.

You should already recognize that these kind of methods pose serious problems. First and foremost, this is an atheoretical approach to model building. But, what if you have no

theory to start with – is a stepwise approach appropriate then? No, for several reasons. If any of the candidate X variables are strongly correlated, the inclusion of the first one will "use up" some of the explanation of the second, because of the way OLS calculates partial regression coefficients. For that reason, once one of the variables is mechanically selected, the other will tend to be excluded because it will have less to contribute to Y. Perhaps more damning, stepwise approaches as highly susceptible to inclusion of spuriously related variables. Recall that we are using samples, drawn from the larger population, and that samples are subject to random variation. If the step-wise process uses the classical 0.05 cut-off for inclusion of a variable, that means that one time in twenty (in the long run) we will include a variable that meets the criterion only by random chance.[3] Recall that the classical hypothesis test requires that we specify our hypothesis in advance; step-wise processes simply rummage around within a set of potential IVs to find those that fit.

There have been notable cases in which mechanical model building has resulted in seriously problematic "findings" that have very costly implications for society. One is recounted in the PBS Frontline episode called "Currents of Fear".[4] The story concerns whether electromagnetic fields (EMFs) from technologies including high-voltage power lines cause cancer in people who are exposed. The problem was that "cancer clusters" could be identified that were proximate to the power lines, but no laboratory experiments could find a connection. But concerned citizens and activists persisted in believing there was a causal relationship. In that context, the Swedish government sponsored a very ambitious study to settle the question. Here is the text of the discussion from the Frontline program:

> ... in 1992, a landmark study appeared from Sweden. A huge investigation, it enrolled everyone living within 300 meters of Sweden's high-voltage transmission line system over a 25-year period. They went far beyond all previous studies in their efforts to measure magnetic fields, calculating the fields that the children were exposed to at the time of their cancer diagnosis and before. This study reported an apparently clear association between magnetic field exposure and childhood leukemia, with a risk ratio for the most highly exposed of nearly 4.
>
> The Swedish government announced it was investigating new policy options, including whether to move children away from schools near power lines. Surely, here was the proof that power lines were dangerous, the proof that even the physicists and biological naysayers would have to accept. But three years after the study was published, the Swedish research no longer looks so unassailable. This is a copy of the original contractor's report, which reveals the remarkable thoroughness of the Swedish team. Unlike the published article, which just summarizes part of the data, the report shows everything they did in great detail, all the things they measured and all the comparisons they made.
>
> When scientists saw how many things they had measured – nearly 800 risk ratios are in the report – they began accusing the Swedes of falling into one of the most fundamental errors in epidemiology, sometimes called the multiple comparisons fallacy.

[3] Add to that the propensity of journals to publish articles that have new and exciting findings, in the form of statistically significant modeled coefficients, and you can see that there would be a substantial risk: that of finding and promoting nonsense findings.

[4] The program was written, produced and directed by Jon Palfreman, and it was first broadcast on June 13, 1995. The full transcript can be found here: http://www.pbs.org/wgbh/pages/frontline/programs/transcripts/1319.html

So, according to the Frontline report, the Swedish EMF study regressed the incidence of nearly 800 possible cancers onto the proximity of its citizens to high-voltage power lines. In some cases, there appeared to be a positive relationship. These they reported. In other cases, there was no relationship, and in some the relationship was negative - which would seem to imply (if you were so silly as to do so) that living near the high voltage lines actually protected people from cancer. But only the positive relationships were included in the reports, leading to a false impression that the study had confirmed that proximity to high-voltage lines causes cancer. Embarrassing to the study authors, to put it mildly.

1.3 Summary

This chapter has focused on multiple regression model building. The keys to that process are understanding (a) the critical role of theory and prior research findings in model specification, and (b) the meaning of the partial regression coefficients produced by OLS. When theory is not well-developed, you can thoughtfully employ nested F-tests to evaluate whether the hypothesized inclusion of an X variable meaningfully contributes to the explanation of Y. But you should avoid reliance on mechanical model-building routines, like step-wise regression, because these can lead you down into statistical perdition. None of us want to see that happen!