Mgmt 469

Model Specification: Choosing the Right Variables for the Right Hand Side

Even if you have only a handful of predictor variables to choose from, there are infinitely many ways to specify the right hand side of a regression. How do you decide what variables to include? The most important consideration when selecting a variable is its *theoretical relevance*. A lot of things can go wrong when you add variables willy-nilly without a reasonably sound theoretical foundation (also known as "regressing faster than you think.") Of course, the definition of "reasonably sound" is a bit murky, and you can do just as much harm by excluding too many variables. This note spells out the tradeoffs involved in model specification.

The Kitchen Sink

You will undoubtedly come across "kitchen sink" regressions that include dozens of variables. This is often an indication that the researcher was brain dead, throwing in every available predictor variable, rather than thinking about what actually belongs. You can imagine that if completely different predictors had been available, the researcher would have used those instead. And who knows what the researcher would have done if there were thousands of predictors in the data set? (Not to mention the possibilities for exponents and interactions!)

A little trial and error is perfectly okay. After all, sometimes a problem is so new that we don't have any theory to go on. Or sometimes we know we want a certain category of variables (e.g., some measure of education) but we do not know the best way to measure it (E.g., "percent of population with a college education" versus "percent of population with an advanced degree".) Even so, do try to resist the temptation to include every variable at your disposal. Kitchen sink regressions can reduce regression precision and even give misleading results.

I use the term *junk variable* to describe a variable that is included in the regression only because you have it in your data set, not because of its theoretical relevance. You already know one practical reason to keep junk variables out of your regressions: adding arbitrary variables uses up valuable *degrees of freedom* (df's). This reduces the precision (i.e., increases the standard errors) of the estimates of all the valid predictor variables. This "unwanted imprecision" effect is especially pronounced when you do not have a lot of observations.

Here are some valuable rules of thumb:

1) Use no more than one predictor for every 5 observations if you have a good predictive model (most predictors significant).

2) You no more than one predictor for every 10 observations if you have a weaker model (few predictors significant) or you are experimenting with a lot of junk variables.

3) You can cut yourself some slack if you have categorical variables. Treat each included category as a half of a normal predictor.

More Reasons to Keep Out the Junk

There are at least three other potential problems that may arise when you introduce junk variables even when you have sufficient df's to get significant findings:

1) Junk variables may be statistically significant due to random chance. Suppose your acceptable significance level is .05. If you introduce ten junk variables, there is about a 40 percent chance that at least one will be significant, just due to random chance.[1] If you don't know what is junk and what is not, you will often find yourself claiming that junk variables really matter. If someone tries to reproduce your findings using different data, they will usually be unable to reproduce your junk result. Your shoddy methods are then exposed for all to see.

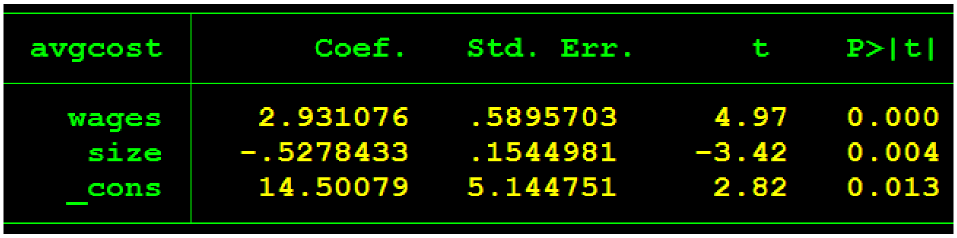
[1] The odds of all ten randomly constructed variables failing to achieve the .05 significance threshold is $1\text{-}.95^{10}$ = .40

2) A junk variable that is correlated with another valid predictor may, by sheer luck, also have a strong correlation with the LHS variable. This could make the valid predictor appear insignificant, and you may toss it out of the model. (This is related to multicollinearity, which I will discuss later in this note.)   The bigger the kitchen sink, the better the chance that this happens. The bottom line: when you add junk variables, “stuff” happens. “Stuff” is not good.

3) Adding some variables to your model can affect how you interpret the coefficients on others. This occurs when one RHS variable is, itself, a function of another. This is not as serious a problem as (1) and (2), but does require you to be careful when you describe your findings. The next section shows you how the interpretation of one variable can change as you add others.

When RHS Variables are Functions of Each Other

Suppose you want to know if firm size affects costs. (We will flesh out this example in class.) You have firm-level data. The dependent variable is unit **cost**. The predictors are **size** and **wages.** Here is the regression result (I will run this regression in class):



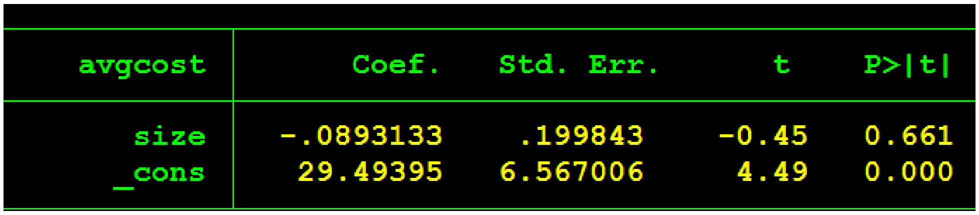
```
 avgcost |      Coef.   Std. Err.      t    P>|t|
---------+-------------------------------------------
   wages |   2.931076   .5895703     4.97   0.000
    size |  -.5278433   .1544981    -3.42   0.004
   _cons |   14.50079   5.144751     2.82   0.013
```

The results seem to show that once we control for wages, there are economies of scale – larger firms have lower average costs. Surprisingly, this does not imply that larger firms in this data set have a cost advantage. The reason is that wages are a potential function of size. In other words, our control variable is a candidate to be a LHS variable. As a result, it is difficult to fully interpret the regression without doing further analysis.

If we want to more fully understand the determinants of costs, we should run a simpler regression that is stripped of any predictors that might themselves be used as dependent variables. In this case, we run:

regress cost size

I will run a regression like this in the classroom. Here is the output:



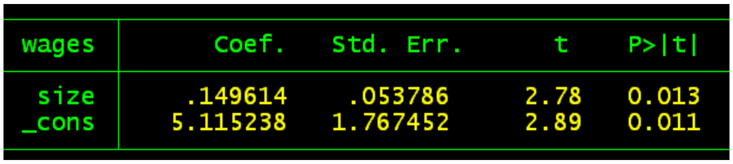
```
     avgcost |      Coef.   Std. Err.      t    P>|t|
-------------+--------------------------------------
        size |  -.0893133    .199843   -0.45   0.661
       _cons |   29.49395   6.567006    4.49   0.000
```

The coefficient on size is close to zero – i.e., there do not appear to be scale economies in this simple regression.

We have now run two regressions and obtained two seemingly conflicting results. Are there scale economies? Which equation should you use?

…Pause for you to take in the drama…

This is a good time for a little story telling. Size may affect costs for many reasons. One effect, which we might call the *direct effect*, is simple economies of scale. For example, larger firms may have lower average fixed costs. Another effect, which we can call the *indirect effect*, is through the potential effect of size on wages and the resulting effect of wages on costs. (The following regression shows that size affects wages):



```
 wages |     Coef.   Std. Err.      t    P>|t|
-------+--------------------------------------
  size |   .149614    .053786    2.78   0.013
 _cons |  5.115238   1.767452    2.89   0.011
```

If you **regress cost size**, then the coefficient on **size** picks up both the direct and indirect effect of size on costs. In our data, the overall direct plus indirect effect is estimated to be -.089 and not statistically different from 0. This occurs because the direct and indirect effects of size observed in the initial regression offset each other.

So here is a story that is consistent with all of our regressions:

1) Larger firms in our data pay higher wages.

2) Larger firms have some other offsetting cost advantages

3) Overall, larger firms have comparable costs to smaller firms.

We might have missed these nuances if we had examined only the model that included both variables.

We can also see how this works mathematically. (Note: I will omit the error terms from these equations for simplicity. The conclusions are *almost* exactly correct, close enough to make the point.) Suppose that the following equations are correct:

(1) $\text{Cost} = B_0 + B_1\text{Wage} + B_2\text{Size}$

(2) $\text{Wage} = C_0 + C_1\text{Size}$

In other words, wages and size affect costs, and size affects wages.

By plugging equation (2) into equation (1), we get:

$\text{Cost} = B_0 + B_1(C_0 + C_1\text{Size}) + B_2\text{Size}$, or

(3) $\text{Cost} = (B_0 + B_1C_0) + (B_1C_1 + B_2)\text{Size}$

Let's think about these equations as regression equations.

- Equation (1) corresponds to the regression: **regress cost wage size**. This regression will report the coefficient on **size** to be B_2. This is the direct effect of size on cost.
- Equation (3) roughly corresponds to the regression **regress cost size**. This regression reports the coefficient on **size** to be $B_1C_1 + B_2$. This includes the direct (B_2) and indirect (B_1C_1) effects.

It is valid to estimate both equations. Just be careful how you interpret the results. If you could only choose one regression to relate size and cost, use the simpler one (without wages). Your interpretation will be correct, if somewhat incomplete.

Action

When you perform a regression, you hope there is enough information in the data to precisely figure out how changes in X affect Y. To get an intuitive grasp of how much information is in your data, think of each observation of X and Y as an *experiment*. If X does not vary much from one experiment to the next, then there is not much information in the data and it will be difficult to determine with any precision how changes in X affect Y.

It follows that good predictors have *action* – they move around a lot from observation to observation. You should always examine each key predictor for action, for example by computing the range and the standard error. You should also plot your dependent variable against each key predictor. The extreme values of the predictors are likely to drive the regression. Does Y vary much as the predictor moves from its lowest to highest value? This plot should foreshadow the regression results (bearing in mind that simple two-way plots mask the effects of control variables.)

Action and Multicollinearity

It is time to address the over-hyped problem of multicollinearity. Suppose you have two predictors, X and Z, and a dependent variable Y. When you examine the data, you see that X, Y, and Z all move together. (I.e., they have high correlations.) You are now quite certain that either X or Z affects Y. Perhaps both do. But you cannot be sure which one matters more. Unfortunately, the computer may also be unable to sort this out. This is multicollinearity.

Let us use the concept of action to better understand multicollinearity. If X and Z are highly correlated, then their “experiments” are not independent. This makes it difficult to determine which one is causing the associated movements in Y. As a result, if you include both in the regression, *the computer will report large standard errors* around their estimated coefficients because it cannot with confidence figure out which predictor really matters.

An immediate implication is that it is possible to get a high R^2 without having any significant predictors! Taken together X and Z give good predictions of Y, but the computer can’t be sure which one is really responsible, so R^2 is high even though significance levels are poor. In other cases, the computer may report a large positive coefficient on one of the correlated predictors and a large negative coefficient on the other. This “sign flipping” often arises when the two variables are essentially identical and the computer uses slight differences between them to fit a few outliers.

There is no test statistic for multicollinearity. There is no particular level of correlation or any other measure that indicates that you have a definite problem. In fact, very high correlations between predictors are not necessarily indicative of multicollinearity and should not automatically deter you from adding both to a model. Consider a model with 1000 observations in which predictor variables X and Z have a correlation of .90. Roughly speaking, X and Z move together 90 percent of the time and move independently 100 times. These 100 “independent experiments” could be enough for your computer to determine with some precision the effects of each predictor on Y. (Of course, a smaller correlation would mean more “independent experiments” and even more precise estimates.) The more observations you have, the more experiments you have. This means that you can tolerate higher correlations among predictors as your sample size increases.

By construction, categorical dummy variables (e.g., indicators for Winter, Spring, and Summer) are negatively correlated, but they do not normally introduce multicollinearity. They are merely a convenient way to break down the action in the predictor (seasonality). There is usually ample action for the computer to estimate their independent effects (provided there are enough observations for each category.)

Finally, note that multicollinearity can be hidden among 3 or more predictors. But the symptoms will be the same.

Signs of multicollinearity

Although there is no definitive test for multicollinearity, there are some symptoms to watch for:

1) You find that the two or more correlated variables have insignificant coefficients when entered jointly in the regression, but each has a significant coefficient when entered one at a time.

2) An F-test shows that two correlated variables add to the predictive power of the model, even though neither has a significant coefficient.

3) Variables have the same sign when entered independently, but have opposite signs when entered together.

4) You might ask Stata to compute the "variance inflation factor" or VIF.[2] After running your regression, type **vif**. There is no threshold vif score, but you *may* have a problem if a predictor variable has a vif in excess of about 20 or the average vif is "substantially higher" than 1 (perhaps 5 or higher). Note: the VIF test is invalid when applied to exponents, fixed effect dummies, or interactions. These types of variables inflate the VIF statistic but do not normally introduce multicollinearity.

If you have problematic multicollinearity, you have several options:

1) Keep all the variables; remember that your coefficients are unbiased but not precise. Your model can still be highly predictive. This is not recommended when you have sign-flipping coefficients on key predictors; the predictive benefits are likely to be miniscule.

2) Throw out one of the offending variables. Acknowledge that you may experience omitted variable bias (see below) that limits the interpretation of the included variable. If the offending variables are merely control variables (rather than the main RHS variables of interest), this is a safe approach.

3) Create a *composite score* – a single measure that captures the information in the correlated variables. For example, if both variables are measured on the same scale, you could simply compute the average of the two variables. I am a big fan of composite scores. They help you avoid problems with interpretation while adding more information to the RHS than you would have with just a single predictor.

A final word: Most, but not all, diagnoses of multicollinearity are false positives. Remember that as long as you have more than 100 or 200 observations, predictor variables can have

[2] For what it is worth, I know of no economics research paper that reports the VIF. This is probably because there

correlations of as much as .7 or .8 without causing problems. To minimize the risk of multicollinearity, avoid including predictors that are nearly identical, or capture nearly identical concepts, regardless of sample size. As an alternative, try them one at a time or create a *composite variable*.[3] Multicollinearity is the result of the researcher indiscriminately throwing variables onto the RHS. A bit of prevention should avoid the problem.

Omitted variable bias

There are so many reasons to be parsimonious when choosing RHS variables that you may be tempted to run regressions with just one predictor variable. It is time to put things in perspective and remember why we add control variables. Adding theoretically sound control variables to the RHS has two virtues:

1) It improves the predictive power of the model, and, in the process, improves the precision of your estimates.

2) Excluding relevant variables can bias the coefficients on the included variables. In other words, the computer reports values that are systematically higher or lower than the actual values, due to an *omitted variable bias*.

It is useful to examine the mathematics that underlies the omitted variable bias. This will look familiar – it is quite similar to the math showing what happens when one predictor is a function of another.

Suppose that the true economic relationship that determines the dependent variable Y is

(4) $Y = B_0 + B_xX + B_zZ + \varepsilon$

(In our in-class example, Y = income, X = schooling, and Z = health status.)

are other, more intuitive ways to detect multicollinearity and the VIF test is itself not definitive.

We may or may not realize that this is the true relationship. In any event, suppose we only have data on Y and X. We regress Y on X:

regress Y X

The computer will spit out a coefficient on X. But is this an unbiased estimate of B_x? We can answer this question after a bit of math.:

Let us suppose that the statistical relationship between X and Z is:

(5) $Z = C_0 + C_xX + \varepsilon_z$.[4]

(This is a very general statement and allows for any degree of correlation between X and Z.) Substitute equation (5) into equation (4) to obtain:

(6) $Y = B_0 + B_xX + B_z(C_0 + C_xX + \varepsilon_z) + \varepsilon_y$.

Gathering terms together gives us:

(7) $Y = [B_0 + B_zC_0] + [B_x + B_zC_x]X + [\varepsilon_y + B_z \varepsilon_z]$

Equation (7) looks exactly like a regression equation of Y on X.:

· The first term in braces $[B_0 + B_zC_0]$ is the intercept

· The second term $[B_x + B_zC_x]$ is the slope

· The third term $[\varepsilon_y + B_z \varepsilon_z]$ is the error.

In fact, this is the regression equation that the computer estimates when we run **regress Y X**.

We clearly have a problem. The intercept and slope coefficients that the computer reports are not estimates of B_0 and B_x. Instead, they are estimates of $B_0 + B_zC_0$ and $B_x + B_zC_x$.

[3] I will discuss composite variables in class.

[4] Note that no causality is necessary here. A simple statistical correlation will suffice to cause the problem.

The problem of omitted variable bias is summarized in the following table[5].

Parameter of interest	You want to estimate	The computer reports
Intercept	B_0	$B_0 + B_zC_0$
Coefficient on X	B_x	$B_x + B_zC_x$

We want an estimate of B_x. Unfortunately, the coefficient reported by the computer is $B_x + B_zC_x$. The term B_zC_x represents the bias.

Omitted variable bias appears when a variable on the RHS must do "double duty". The coefficient on the included variable includes its direct effect on Y, as well as the indirect effect of the omitted variable that happens to be correlated with it.

It is sometimes helpful if you can determine the direction of the bias. For example, suppose:

· You believe that X and Z are positively correlated ($C_x > 0$), and

· You believe that Z is positively related to Y ($B_z > 0$).

Then you should conclude that the estimate that the computer reports for B_x will be more positive than the correct value (because $B_zC_x > 0$). Continuing our example, if schooling and health status are positively correlated and health status has an independent effect on earnings, then a regression that omits health status will overstate the effect of schooling on earnings.

[5] Here is a more precise statement of the bias. Suppose that you want to estimate B_x but omit data on Z. The coefficient on X will equal $B_x + B_z\text{Cov}(X,Z)/\text{Var}(X)$, where Cov(X,Z) is the covariance between X and Z and Var(X) is the variance of X. This follows from the formula for the parameter C_x.

Coping with Omitted Variable Bias

It might seem that omitted variable bias plagues every regression. After all, it is impossible to get data on all the factors that affect the dependent variable. To some extent this is true and this is why we always think about possible biases in our regressions. Fortunately, omitted variable bias is usually a manageable problem, for three reasons.

1) Omitting variables results in biased coefficients only if the omitted variables are correlated with included variables *and* are important in their own right. If either of these conditions fails to hold, there is no bias.

2) Even if there is omitted variable bias, it may be possible to determine the direction of the bias. This will allow us to state that the reported coefficients are either upper or lower bounds on the actual effects.

3) Thinking about the omitted variable bias forces us to carefully identify the correct economic model and do a better job of variable selection in the first place.

Endogeneity bias

A RHS variable is said to be *endogenous* if it is correlated with the error in the original model.[6] In general, one can not readily interpret the coefficients on endogenous predictors. They may be biased and/or it may be impossible to draw conclusions about causality.

We have already encountered an important example of endogeneity bias – omitted variables. Any omitted variable is, by definition, part of the error term (or, in econometrician's parlance, part of the *unobservables.*) If that omitted variable is correlated with a predictor variable, it follows that the error term is correlated with the predictor variable. We will soon learn an important technique for dealing with this type of endogeneity bias – *fixed effects models.*

There are two other ways that endogeneity often arises. First, the RHS variable may be a function of Y (as opposed to being a cause of Y). For example, recall the sales and advertising model from the previous note. We normally assume that advertising affects sales. But suppose that firms increase their advertising *in anticipation of* changes in demand. In this case, sales affects advertising; advertising is therefore endogenous. In general, if you are unsure whether X causes Y or Y causes X, your regression suffers from *simultaneity bias*. It is impossible to determine the direction of causality from OLS regression. We will discuss a possible solution to simultaneity bias called *instrumental variables regression* in a later lecture.

A third source of endogeneity is *measurement error*. This occurs when a RHS variable is imprecisely measured. Some of the measurement error necessarily becomes part of the regression residual, causing the measured X to be correlated with the residual. We will discuss measurement error in a later lecture as well.

[6] RHS variables are uncorrelated with the *residual* of the regression, by construction.

All forms of endogeneity -- omitted variable bias, simultaneity bias, and measurement error -- will bias the coefficient on the affected RHS variable and any other variables that are correlated with the affected variable. This is why it is crucial to determine whether your model may suffer from endogeneity bias.

A sensible approach to regression modeling

The following approach balances concerns about kitchen sink modeling and omitted variable bias. You could do worse than to follow these steps:

1) Always begin with a "core" set of predictors that have theoretical relevance, as well as any predictors whose effects you are specifically interested in. You may estimate a "quick and dirty" OLS model at this time.

2) Finalize model specification issues (e.g., log vs. linear – to be discussed in a later note)

3) Add additional predictors that you think might be relevant. You can add them one at a time or one "category" at a time (see next section). Check for the robustness of your initial findings.

4) When adding predictors, you should keep all the original predictors in the model, even if they were not significant. Remember, omitted variable bias can cause significant predictors to appear to be insignificant. By adding more variables, your key predictors may become significant.

5) At this point, you should know your robust findings. That is the main goal of your research.

6) If you insist on producing a "final model", then you should remove those additional predictors that were not remotely significant.

7) You can also remove core predictors if they remain insignificant and you need degrees of freedom. If you are not taxed for degrees of freedom, you may want to keep your core variables, if only to paint the entire picture for your audience.

Coping with "Groups" of Variables

Many predictor variables can be neatly categorized into groups: seasons, competitors' prices, consumer demographics, the 50 states. You will often want to determine whether a group of predictors belongs in the regression. ("Does seasonality matter?" "Do competitors' prices matter?") Analysts often examine the coefficients of each predictor in a group individually and keep those that are significant. *This is a mistake,* as I will now explain.

Suppose you are studying the demand for automobiles and you want to know whether there are differences in the level of demand across the 50 states. You include 49 dummy variables. Even if state location really does not matter, you can expect coefficients on 2 or 3 state dummies to be statistically significant just due to random chance. Upon observing those 2 or 3 significant dummies, you have two choices:

1) Conclude that those two states really are different from the omitted state;

2) Determine whether the results are consistent with the null hypothesis that there are really are *no differences* among the states

The latter approach is correct. After all, some state has to have the highest sales, even if the difference is totally random. The fact that some state really does have the highest sales is no reason to reject the null hypothesis that states don't matter.

The correct way to examine a group of variables is to perform a *partial F-test,* also called called a *Chow test*, which compares the predictive power of the model with and without all the variables under consideration. If the group of variables does not collectively add predictive power, then you cannot reject the null that the group is irrelevant.

The Chow (or partial-F) Test

Recall the formula for R^2:

$$R^2 = 1\text{-SSE/SST}.$$

The value of adding another predictor variable is that it reduces SSE. Adding any predictor variable, even one chosen at random, will decrease the SSE. A Chow/partial-F test determines whether adding a group of variables decreases the SSE by more than would be expected if you added a group of randomly constructed variables.

Here is how it works. Suppose you start with a model that excludes one or more variables from the right hand side. You then add some additional variables to the model. Naturally, the SSE will decrease. The Chow test is an F-test that determines whether the decrease in the SSE is larger than would have been expected due to random chance, given the number of added predictors and the degrees of freedom.

Stata performs the F-test for us. Suppose you want to know whether the variables X2 and X3 should be added to a model that includes variable X1. Simply type:

```
regress Y X1 X2 X3
test X2 X3
```

Stata will test to determine if the simultaneous inclusion of X2 and X3 in the regression generates a statistically significant decrease in the SSE.

Working with the Results of the Chow Test

1) Make sure you perform the test on the entire theoretically-linked group. Do not do *ex post data picking* and choose from among that group. (But do think carefully *ex ante* about whether variables really belong in the same group.)

2) If the test is insignificant, throw out the entire group. There is no reason to reject the null hypothesis that the group of variables were, in effect, no better than randomly constructed variables.

3) If the test is significant, you have a choice. You can keep the entire group, or if you are starved for degrees of freedom, you may throw out the insignificant members of the group. Be careful when using dummy variables – significance is only based on comparisons with the omitted category! If you throw out a specific category, then that category will be lumped in with the omitted category.

Some students worry about how to define a group of variables. In the yogurt project, should you lump together all prices or just competitor prices? The answer is that you should come to the analyses with specific hypotheses. Each hypothesis pertains to one or more variables. This will guide your groupings. For example, you are probably willing to claim that your yogurt’s own price will have an effect on your own sales. Thus, it is okay to examine this coefficient in isolation. On the other hand, you may wish to test whether “competitors’ prices” matter. If this is your hypothesis, then the set of three competitors’ prices is a group.

My advice is simple – try to identify one group of variables for each theoretical concept. If you are testing whether there is seasonality, test your season variables as a group. If you are testing whether demographics affect Internet demand, test the demographic variables as a group. But try as best as you can to develop your theories before you look at your results.