

ECON 5340 Class Notes

Chapter 8. Specification Analysis, Model Selection and Data Problems

1 Introduction

Most of this chapter is concerned with choosing the correct regression model. That is, how do you choose between competing models and if you get it wrong, what are the consequences. The last part of my notes, deals with several different practical problems that may occur in the data.

2 Specification Analysis

2.1 Omission of Relevant Variables

Suppose that the "true" regression model is

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon \quad (1)$$

where X_1 is a $(n \times k_1)$ matrix and X_2 is a $(n \times k_2)$ matrix. Now assume that the researcher mistakenly estimates the following

$$Y = X_1\beta_1 + \epsilon. \quad (2)$$

The least squares estimate of β_1 is

$$\begin{aligned} b_1 &= (X_1'X_1)^{-1}X_1'Y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \epsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\epsilon. \end{aligned}$$

Taking expectations then gives

$$E(b_1) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2.$$

This implies that b_1 is a biased estimator of β_1 unless

1. $\beta_2 = 0$, which means that equation (2) was the "true" model and X_2 was not really relevant or
2. X_1 and X_2 are orthogonal.

Neither of these are likely to be true, so omitting relevant variables produces biased estimates of the coefficients. Although b_1 is biased, its variance will not be larger (and is likely to be smaller) than the LS estimator for β_1 when X_2 is included (call this estimator $b_{1,2}$). These two variances are

$$\begin{aligned} \text{var}(b_1) &= \sigma^2(X_1'X_1)^{-1} \\ \text{var}(b_{1,2}) &= \sigma^2(X_1'M_2X_1)^{-1} = \sigma^2(X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1)^{-1} \end{aligned}$$

where M_2 is the "residual maker" matrix for X_2 . Note, however, that the estimates of $\text{var}(b_1)$ and $\text{var}(b_{1,2})$ may not reflect this ordering because s^2 is a biased estimator of σ^2 when excluding X_2 from the model.

2.2 Pretest Estimators

At least on a mean-square error basis, it is not clear which estimator is better: b_1 or $b_{1,2}$. A third (and quite popular) choice is the so-called pretest estimator, call it b_1^* . This estimator is a mix of the previous two. First, you estimate model (1) and then perform a statistical test to see if X_2 belongs in the model. If you reject the null (X_2 does matter), then you settle on $b_{1,2}$. Otherwise, you choose b_1 . Using an F test, we can write

$$E(b_1^*) = E(b_1) \Pr(F < F_c) + b_{1,2} \Pr(F > F_c) \neq \beta_1.$$

Therefore, b_1^* is a biased estimator unless the F test is designed to always reject the null hypothesis (size $\simeq 1$). The variance of b_1^* is non-trivial to calculate. The gauss example below performs a Monte Carlo experiment to see which of these three estimators performs better on a mean-square error basis.

2.2.1 Gauss Example. MSE Comparison of a Pretest Estimator.

For this experiment, we let

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i$$

and examine the mean square error of three estimators of β_2 : b_2 , $b_{2,3}$ and b_2^* . For given values of the independent variables, we then draw 2000 different samples, each of size ($n = 50$). See Gauss example 8.1 for more details.

2.3 Inclusion of Irrelevant Variables

Now assume that the "true" regression model is

$$Y = X_1\beta_1 + \epsilon$$

and the researcher mistakenly estimates

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon.$$

As shown earlier, the estimator for β_1 from the latter model is

$$\begin{aligned} b_{1.2} &= (X_1' M_2 X_1)^{-1} X_1' M_2 Y \\ &= (X_1' M_2 X_1)^{-1} X_1' M_2 (X_1\beta_1 + \epsilon) \\ &= \beta_1 + (X_1' M_2 X_1)^{-1} X_1' M_2 \epsilon \end{aligned}$$

which is clearly unbiased. However, as shown above, there is a cost involved with including the unnecessary regressors X_2 . The variance of $b_{1.2}$ is inflated relative to the correct estimator b_1 .

3 Choosing Between Nonnested Models

Sometimes we want to choose between two models that are not nested. For example, maybe we want to distinguish between model #1, $Y = X\beta + \epsilon$, and model #2, $Y = Z\gamma + \epsilon$. Assuming X and Z each have a variable not included in the other, neither model can be written as a special case of the other. No simple t or F test can reject one model in favor of the other. One solution to this problem is to artificially nest the two models in the compound model

$$Y = (1 - \alpha)X\beta + \alpha Z\gamma + \epsilon$$

where $0 \leq \alpha \leq 1$.

3.1 J Test

The J test of Davidson and MacKinnon (1981) is designed to test whether or not $\alpha = 0$ (model #1) or $\alpha = 1$ (model #2). Normally, we would just estimate α and run a simple t test. The problem is that α is not identified – it is nothing more than an arbitrary scaling of β and γ . The J test solves this problem using

the following procedure:

1. Estimate γ by a LS regression of Y on Z .
2. Estimate β and α by a LS regression of Y on X and $Z\hat{\gamma}$.
3. Using $\hat{\alpha}$, carry out an asymptotic t test of the null hypothesis $H_0: \alpha = 0$.
4. Reverse the order in (1) - (3) and test the null hypothesis $H_0: \alpha = 1$.

Unfortunately, in finite samples there are four possible outcomes – reject both nulls, fail to reject both nulls and reject one of each of the nulls.

4 Model Selection Criteria

The J test is implicitly designed to distinguish between models based on goodness-of-fit within the sample. An example of a less sophisticated model-selection criterion would be to see which model produced a greater R^2 . The problem with these approaches is that what works well in-sample, may not work so well out-of-sample. In this case, we need a penalty for over-parameterizing a model. Here are some options:

1. Choose explanatory variables to maximize $\bar{R}^2 = 1 - [(n - 1)/(n - k)]R^2$ or alternatively minimize s^2 .
2. Choose explanatory variables to minimize, Akaike Information Criterion (AIC), $\ln(e'e/n) + 2k/n$.
3. Choose explanatory variables to minimize, Schwarz Criterion, $\ln(e'e/n) + k \ln(n)/n$.

These three criteria will tend to produce increasingly parsimonious models, as the penalty for additional explanatory variables increases.

5 Data Problems

This section is an eclectic collection of practical data problems.

5.1 Multicollinearity

There are two types of multicollinearity (MC): perfect and imperfect. Perfect MC violates the Classical assumption that the X matrix is of full rank, in which case OLS cannot be calculated. This section deals with imperfect MC between the explanatory variables, in which case OLS can be calculated.

5.1.1 Properties of the OLS Estimator

Given that imperfect MC does not violate any of the Classical assumptions, we know that the Gauss-Markov theorem still holds and b is the best linear unbiased estimator. This is a surprising result to some, but it simply means that given the multicollinear regressors, there is no better way than OLS to estimate the population parameters. Of course, all else equal, having less multicollinear regressors would produce more reliable estimates (smaller standard errors), but that is not an option.

5.1.2 Detection

The first two procedures to detect MC involve using simple correlations and variance inflation factors (VIFs).

1. Simple Correlation Coefficients. The easiest method to detect MC is to print out a matrix of simple, pairwise correlation coefficients between the explanatory variables and look for values close to one in absolute value (say, greater than 0.8 in magnitude).
2. Variance Inflation Factors. The problem with pairwise correlation coefficients is that it can miss more sophisticated forms of multicollinearity that involve multiple explanatory variables. VIFs are calculated according to

$$VIF(\hat{\beta}_j) = (1 - R_j^2)^{-1}$$

where R_j^2 is the coefficient of determination for a regression of the j th explanatory variable on all other explanatory variables. It is interpreted as the amount $var(\hat{\beta}_j)$ is inflated relative to the case of no MC.

Another approach to detecting MC is the diagnostic approach of looking for MC in the OLS results. Under severe MC, OLS properties include

1. Small changes in the data (e.g., eliminating a single observation or variable) can cause large changes in the $\hat{\beta}$ s.
2. High R^2 s and low ts .
3. Unexpected signs on the $\hat{\beta}$ s (of course this could also be caused by an inappropriate theory so be cautious).

5.1.3 Solutions

There are many ways to handle MC and none of the potential solutions are uniformly the best. Here are some options:

1. Do nothing. Recall that OLS is still BLUE.
2. Transform the data. Taking ratios, linear combinations or first-differences of the explanatory variables can often reduce MC.
3. Drop variables. This is probably the most common solution. Many researchers use economic theory, common sense and initial regression results to choose variables to drop. You need to be very careful, however, to not drop a relevant variable because it will bias all the remaining estimates.
4. Mechanical approaches. Routines such as ridge regressions and principal components are options but are not widely accepted by the discipline.

5.2 Measurement Error

Many economic variables are measured with error. For example, the consumer price index is calculated from a sample of prices across many metropolitan areas and tends to miss new goods, often fails to account for improvements in existing goods, and doesn't fully recognize consumers ability to substitute toward cheaper goods. Survey data are also often measured with error as respondents misstate their true behavior or characteristics. Let's consider two types of measurement error.

5.2.1 Measurement Error in the Dependent Variable

Assume the true model is

$$y_i^* = \beta_1 + \beta_2 x_i + \epsilon_i, \tag{3}$$

where y_i^* represents the true and unobserved value of the dependent variable. The researcher, unfortunately, is endowed with $y_i = y_i^* + \mu_i$, a noisy measure of y_i^* . Rewriting (3) gives

$$y_i = \beta_1 + \beta_2 x_i + (\epsilon_i + \mu_i).$$

Therefore, as long as μ_i is i.i.d. and uncorrelated with x_i , the OLS estimates of the β s will be BLUE.

5.2.2 Measurement Error in the Independent Variables

Now assume the true model is

$$y_i = \beta_1 + \beta_2 x_i^* + \epsilon_i, \quad (4)$$

where x_i^* represents the true and unobserved value of the dependent variable. The researcher, unfortunately, is endowed with $x_i = x_i^* + \mu_i$, a noisy measure of x_i^* . Rewriting (4) gives

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + (\epsilon_i - \beta_2 \mu_i) \\ &= \beta_1 + \beta_2 x_i + \epsilon_i^*. \end{aligned}$$

It is clear that the $\text{corr}(x_i, \epsilon_i^*) \neq 0$, which violates a Classical assumption and will result in biased and inconsistent estimates of β_2 . In fact,

$$\text{cov}(x_i, \epsilon_i^*) = \text{cov}(x_i^* + \mu_i, \epsilon_i - \beta_2 \mu_i) = -\beta_2 \sigma_\mu^2$$

and the inconsistency in b_2 , measuring the variables in their deviation-from-the-mean form, is given by

$$\text{plim}(b_2) = \text{plim}\left(\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}\right) = \text{plim}\left(\frac{\frac{1}{n} \sum_{i=1}^n (x_i^* + \mu_i)(\beta_2 x_i^* + \epsilon_i)}{\frac{1}{n} \sum_{i=1}^n (x_i^* + \mu_i)^2}\right).$$

Using Slutsky's theorem and $Q^* = \text{plim}(\frac{1}{n} \sum_{i=1}^n x_i^{*2})$ we can show that

$$\text{plim}(b_2) = \frac{\beta_2 Q^*}{Q^* + \sigma_\mu^2}$$

so if $\sigma_\mu^2 > 0$, b_2 is downwardly inconsistent (in magnitude). This matches the fact that $\text{corr}(x_i, \epsilon_i^*) = -\beta_2 \sigma_\mu^2 < 0$ when $\beta_2 > 0$ and $\text{corr}(x_i, \epsilon_i^*) = -\beta_2 \sigma_\mu^2 > 0$ when $\beta_2 < 0$, which causes b_2 to be biased toward zero. Signing the bias is much more complicated in a multivariate setting. Finally, the typical solution is instrumental variables estimation, that is, find a proxy variable for x_i that is not correlated with the measurement error μ_i .

5.3 Missing Observations

A third practical problem with economic data is missing observations (i.e., "holes" in your dataset). This is a common occurrence in survey data as people refuse to answer questions. If observations for certain

questions are missing there are several options.

1. Eliminate the entire row (entire observation) from the dataset. There are two problems with this approach. First, missing observations are often not random, so eliminating them will produce a sample that is not representative of the population (e.g., maybe old people are reluctant to state their age). Second, this often leaves you with too few remaining observations.
2. Replace the missing value with the sample mean. If the entire row of the X matrix is missing, this is no different than entirely eliminating the observation. Furthermore, if missing values are systematically related to X , the sample mean may not be an representative estimate of the true value of X .
3. Dummy variable approach. Create a new dummy variable for each variable that has missing observations (provided they are missing in different rows) and add the dummies to the X matrix. In this fashion, the researcher is using all the available observations on an explanatory variable in calculating the corresponding coefficient. One downside is that like (1) and (2) above, it assumes that the observations are missing at random, which is not always the case.
4. Sophisticated interpolation. There are several available routines that allow one to use in-sample and out-of-sample information to make a more sophisticated (than the unconditional mean) guess at the missing value. Little is known about the property of these estimators, and what is known, typically comes from simulation exercises in special contexts.