Mgmt 469

Noisy Variables, Heteroskedasticity, and Grouped Data

Introduction to Noisy Variables

A variable is *noisy* if does not exactly equal the variable of interest (the one that best fits what the theory demands), or if it is mismeasured. Here are some examples:

- You want to measure the impact of product-level advertising on product sales. You have data on firms' *total* advertising budgets. To estimate product-level budgets, you divide the total budget by the number of products. Your measure of product-level advertising is noisy.

- You want to determine if inventory turnaround is faster in firms that use just-in-time (JIT) inventory techniques. You survey logistics managers to get information about inventory turnaround. The busy managers provide rough, and therefore noisy, estimates.

- Continuing this investigation of inventory turnaround, you next study whether turnaround times differ by nation. Using the survey responses, you compute the average turnaround in each nation. The number of survey respondents ranges from 75 in the U.S. to 2 in Chile. From the law of large numbers you know that the 75 U.S. firms in the sample are fairly representative of the U.S. as a whole. But you feel that the 2 Chilean responses may not accurately reflect all Chilean firms. Your measure of nation-level turnaround times is noisy, especially for nations with few sample respondents.

The first part of this note describes the implications of noisy variables and suggests possible ways to deal with them.

Implications of noisy variables

It is not always easy to determine which variables are noisy. After all, the best way to know if a variable is noisy is to compare it with an accurate measure. But if you had an accurate measure, why would you use the noisy one? It is sometimes possible to apply some statistical common sense to determine whether variables are noisy. For now, we will suppose that we know when a variable is noisy and discuss what that means for your analysis.

Noisy Dependent Variables

Here are two key facts:

1) Coefficients obtained from OLS regressions with noisy dependent variables are unbiased. This implies that your predictions are also unbiased.

2) Coefficients obtained from OLS regressions with noisy dependent variables are estimated less precisely (i.e., the standard errors increase). Thus, your predictions are less accurate.

These statements are readily confirmed. Suppose that the true model relating X to Y is:

(1) $$Y = B_0 + B_1X + \varepsilon_y$$

where ε_y is normally distributed. Suppose further that you do not have an accurate measure of Y.

Instead, you have

(2) $$Z = Y + \varepsilon_z,$$

where ε_z is a normally distributed noise term that is independent of ε_y.[1]

Substituting from (2) into (1) yields:

(3) $$Z = B_0 + B_1X + (\varepsilon_y + \varepsilon_z)$$

This is a regression equation, of course. In fact, the only difference between equations (1) and (3) is that the error term is larger in equation (3): $(\varepsilon_y + \varepsilon_z)$ versus ε_y.[2] This implies that the standard errors on B_0 and B_1 are larger when you use Z as the dependent variable. This causes the standard errors of any predictions to increase as well.

[1] In general, you do not know the precise nature of the noise. Assuming that it is normally distributed is usually a good approximation, and makes the math much easier.

[2] Recall that the sum of two normally distributed variables is also normal. Thus, $\varepsilon_x + \varepsilon_y$ is normal, so that equation (2) is a standard OLS regression model.

Noisy Predictor Variables

Things are a bit different when the predictor variables are noisy. Let's see what happens when X is noisy. Suppose that the true model is:

(4) $$Y = B_0 + B_1X + \varepsilon_y$$

Suppose that you cannot measure X with precision. Instead, you measure

(5) $$Q = X + \varepsilon_q$$

where ε_q is normally distributed and independent of ε_y.

We won't derive it here, but an important result in econometrics is that the estimated B_1 will tend towards the following value:

(6) $$\text{Estimated } B_1 = (\text{True } B_1)/(1 + \sigma^2_q/\sigma^2_y)$$

Noting that the denominator is larger than 1, we conclude that the estimate of B_1 is biased towards zero.[3] This is known as *attenuation bias*. The degree of attenuation bias depends on the relative values of σ^2_q and σ^2_y. If σ^2_q is large relative to σ^2_y (i.e., X is measured with a lot of noise relative to the regression error) then the bias can be quite large.

Most of the time, you should not be overly concerned about attenuation bias. It is inevitable that you will measure some predictor variables with error. If the measurement errors are relatively small, the bias is small as well. Moreover, if you are mainly interested in hypothesis testing, as opposed to examining magnitudes, then the bias is of the "right" type. That is, if the estimated B_1 is statistically significant when you have measurement error, then the true B_1 would be larger and it would likely be more significant if you could eliminate that error.

[3] If the true value of B_1 is positive, the computer will report an estimate of B_1 that is a smaller positive number. Similarly, if the true value is negative, the computer will report a smaller (in magnitude) negative number.

Heteroscedasticity

A key assumption of OLS regression is that the errors for all observations are distributed identically. In other words, you expect the model to give equally precise predictions for all observations. Any variation in the errors must be completely random. You are likely to be violating this assumption if the *magnitude of the residuals* is correlated with some factor Z.[4] Any Z will do, whether it is in your model or not. For example, your residual may be large in magnitude whenever Z is large, and your residual may be small in magnitude whenever Z is small. If this occurs, then the underlying error is likely to be correlated with Z, implying that you have *heteroscedasticity*. When you have heteroscedasticity, *your standard errors are reported incorrectly*. As a result, your results are not convincing.

There are two ways to cope with heteroscedasticity. One involves a common situation where the magnitude of the errors is proportional to some factor Z. The second involves more complex relationships between the errors and one or more Z's. Fortunately, both have solutions that are easy to implement in Stata.

Using Weighted Least Squares to Correct Heteroscedasticity

An important technique known as *weighted least squares* (WLS) fixes heteroscedasticity when the magnitude of the errors is proportional to some factor Z. WLS works by giving the most credence (the most weight) to the observations that are expected to have the smallest errors. To illustrate why it makes sense to do this, suppose that you are trying to find the slope of a line. You run a regression and discover that residuals tend to be systematically larger for some values

[4] Remember – the error is the ε in the underlying model. The residual is the difference between the actual and predicted values. The two are not the same, due to the randomness of the process that generates your data. Even so, the residual is your best estimate of the actual error.

of Z than for others. Perhaps the residual is relatively large in magnitude when Z<10 and relatively small when Z>10. This is blatant heteroscedasticity.

It is apparent that your model is much more accurate when Z>10. You might be tempted to throw out the observations for which Z<10. But this data is not completely useless – it is just not as accurate. A better solution is to use all the data, but to place more weight on the observations for which Z>10. This is what WLS is all about.

To implement WLS, you must identify some factor, like Z in the above example, that is correlated with the model's accuracy. This will be your *weighting factor*. A good weighting factor is often difficult to find. Moreover, the choice of a factor is very dependent on the application. As a result, there are few hard and fast rules about when to weight and what to use as a weight.

There is one important class of applications for which the weighting factor is easy to identify and essential to use. This occurs *whenever the LHS variable is drawn from individual survey data that is aggregated up to the market level*. That is a rich sentence, with lots of content. So let's break it down.

1) You need to have *survey data*.

2) The survey data was used to construct the *LHS variable*.

3) The LHS variable was computed by *aggregating individual responses* to create a *market level* "mean".

If all three conditions hold (and they often do), then WLS is indicated.

The following example should make things clearer. Suppose that you are studying determinants of television viewing in different cities. You survey lots of viewers in lots of cities to find out about their viewing habits. Your unit of analysis is going to be the city, so you

compute city-wide average viewing levels. In some cities, you may have just one or two responses. In others, you have 50 or 100 responses. Simple statistics tells you that in those cities with 50-100 responses, the city-wide averages you compute are probably pretty close to the actual averages for those cities (assuming you have a representative sample.) In those cities with only one or two responses, however, the averages you compute may be very different from the citywide averages.

Because the sample sizes are rather small in many cities, your LHS variable – estimated city-wide viewership – is noisy. But there is something predictable about the magnitude of the noise. A bit of statistics will show that if n_i is the number of respondents in city i, and e_i is the regression residual for city i, then the magnitude of e_i is proportional to $1/\sqrt{n_i}$. This is heteroscedasticity – e_i is systematically related to some factor (in this case n_i).

To eliminate heteroscedasticity, you must ask the computer to pay less attention to those cities with fewer respondents. Specifically, you should weight each observation by $\sqrt{n}$. If you do this, and then run OLS regression, your results will be just fine. Weighting by $\sqrt{n}$ means that you *multiply each and every value in your data set by* $\sqrt{n}$ before running the regression.

Here is why this works. This works because if you multiply everything by $\sqrt{n}$, then the error term for each observation is also multiplied by $\sqrt{n}$. This, in turn, implies that the squared errors are multiplied by n. Recall that OLS works hardest to fit the observations that contribute the most to the sum of squared errors. By multiplying the squared errors by $\sqrt{n}$, you force the computer to do a good job of fitting the observations with the largest n's, which is exactly what you want.

Stata has a built-in command to estimate WLS. Suppose you have survey data on television viewing that you aggregate to the city level (e.g., you have data for Chicago, Detroit, etc.) You also have city-level data on income and education, and you want to see how these affect viewing. The variable n represents the number of survey respondents in each city.

If you suspect that you ought to be using WLS, where the weighting factor is some variable **n**, you can test for this by typing

regress tvviewing percapincome schooling

hettest n

If you find that the test statistic is significant, then you should use WLS by typing:

regress tvviewing percapincome schooling [w=n][5]

Stata will do the rest. Note that when you run WLS, coefficients may change.

Review: A Guide to Weighting

1) You may want to put more weight on some observations than others.

2) This is certainly the case if the errors are systematically smaller for some observations; these observations deserve more weight. This occurs when you aggregate survey data, for example.

3) You can check for the need to do WLS by correlating the absolute value of the residuals with the potential weighting factor n, or, better still, doing **hettest n**.

4) WLS multiplies the LHS and RHS by $\sqrt{n}$, where n is the weighting factor. This weights the squared errors by n, which is what you want.

5) It is easy to perform WLS in Stata – just add **[w=n]** at the end of your regression statement.

[5] Note the use of brackets [] rather than parentheses ().

Testing for Heteroscedasticity, and the "White-washing" Solution

WLS is not appropriate for most regressions. But you still may have heteroscedasticity. Stata provides a simple way of testing for heteroscedasticity. For example, you can type:

regress tvviewing percapincome schooling

hettest

(This is the same as before, but you use "hettest" instead of "hettest n.") Stata performs the Cook-Weisberg test for heteroscedasticity. If the test statistic is significant, then you need to correct the OLS standard errors.

Statistician Halbert White discovered the proper formula for adjusting standard errors in 1980. Many statistical packages, including Stata, feature the White correction. To estimate White-corrected standard errors, just run your OLS regression as follows:

regress tvviewing percapincome schooling, robust

The coefficients are identical to the OLS coefficients, but your standard errors are now correct. Some econometricians call this technique "White-washing".

***Summarizing Heteroscedasticity*:**

1) You have heteroscedasticity if the magnitude of the standard errors is correlated with some unmeasured factor.

2) Heteroscedasticity biases the standard errors; the coefficients are unbiased, however.

3) A common source of heteroscedasticity is the use of aggregated survey data. This can be corrected by using WLS.

4) You can test for heteroscedasticity using **hettest** command in Stata.

5) The **,robust** option corrects the standard errors in heteroscedastic OLS regressions.

6) WLS is preferred to **,robust** whenever the former is indicated.

Grouped data

Another critical assumption of OLS is that all the observations are independent. This assumption is frequently violated in practice. A prime example is regression with *grouped* data. For example, you may run a regression of profits for firms in a variety of industries. It seems plausible that profits will be correlated for firms within any given industry.

Here is a more extreme example (to be mimicked in class, so be careful!) Suppose you want to know if redheads are more popular than brunettes. You have two friends named John and Paul. John is a brunette and Paul has red hair. At 1pm, you poll the class to see how many classmates like John more than Paul. You find that 45 prefer John and 15 prefer Paul. You repeat this poll at 1:10, 1:20, etc. .

Your data looks as follows

Name	Hair color	Popularity
John	B	45
Paul	R	15
John	B	46*
Paul	R	15
John	B	46
Paul	R	15
John	B	46
Paul	R	15
John	B	46
Paul	R	14*

*One student arrived later and another student left class early to go to a job interview.

Given these 10 observations, you regress popularity on an indicator variable for hair color, where Hair=1 if the hair is brown, and Hair=0 if red. The result is $B_{hair} \approx 30.5$ and this coefficient is statistically significant, thanks to the 10 observations and the apparent 9 degrees of freedom.

Do you conclude that people with red hair are more popular? Of course not. The reason you get a significant coefficient on hair color is that *you do not have 10 independent* observations. You have 2 observations that are each repeated 5 times. The computer has no reason to know this, thinks that you have lots of experiments, and computes the standard errors accordingly. This is an extreme example of *groupiness* in the data. If you do not account for the "groupiness" of your data, you will overstate the true degrees of freedom in your model and the reported standard errors will be artificially small. You run a great risk of tricking yourself into thinking that you have significant findings when in reality you do not.

One way to deal with grouped data is to estimate fixed effects. In fixed effects models, the computer ignores within-group variation when estimating the coefficients. Thus, only across-group variation matters. (To determine the effect of hair color in the prior example, either John or Paul would need to change theirs from brunette to red, or vice versa.)

There are times when you do not want to estimate fixed effects models. This is especially possible if you do not have much within-group variation. For example, suppose you want to study the effect of market demographics on yogurt sales. The demographics of the communities surrounding the stores will change little over time. If you include store fixed effects, you will not have sufficient within-store variation. You will have to omit the store dummies and rely on across-store variation. (You now run a heightened risk of omitted variable bias, of course, but if you have a rich set of demographics, this risk is minimized.)

Suppose you go ahead and omit the store fixed effects. It is now likely that the standard errors across observations within each store are no longer independent. You have grouped data and if you don't account for it, your standard errors will be biased.

The technique for adjusting the standard errors to account for groupiness is preprogrammed into Stata. Continuing the example, if you have data on the income of each store's local community, you could estimate the following regression:

regress sales1 price1 promo1 income

To correct the standard errors for possible groupiness, just use the **cluster** subcommand in Stata. In this case, the groupiness comes from the variable store, so you type:

regress sales1 price1 promo1 income, cluster(store)

Coping with Grouped Data

1) Use common sense as a guide to determine if your data falls naturally into groups. As an alternative, examine the error terms for observations within specific groups. Are they systematically positive or negative? If so, then you may not have independent observations. As a result, your standard errors are too small.

2) You can estimate a fixed effects model to avoid the resulting bias in the standard errors. But you will be unable to examine the effects of variables that vary only across groups.

3) If you want to preserve inter-group action but avoid biased standard errors, use the **,cluster(groupname)** option in Stata.

An unexpected problem (math optional)

Suppose that your initial model is:

$$Y = B_0 + B_1X + \varepsilon_y.$$

You decide that you want to divide both Y and X by some other variable V. An example might be when you express both variables in per capita amounts, where V is the size of the population. If you divide Y by V, then you must divide the right hand side by V to keep the equation correct. This means that you are effectively regressing

$$Y/V = B_0/V + (B_1X)/V + \varepsilon_y/V.$$

Note that the error term is now clearly larger when V is smaller — that is, when the dependent variable and independent variable are smaller. This is blatant heteroscedasticity.

One excuse for keeping the model this way is that the underlying model is, in fact,

$$Y/V = B_0 + (B_1X)/V + \varepsilon_y$$

OLS appears to be safe here. If you think this is the correct model, you are almost safe. A word of caution is still necessary.

Suppose the true model is

$$Y/V = B_0 + (B_1X)/V + \varepsilon_y$$

but you do not have a precise measure of V. Instead, you have $U = V + \varepsilon_u$. Thus, you are actually regressing:

$$Y/(V + \varepsilon_u) = B_0 + (B_1X)/(V+ \varepsilon_u) + \varepsilon_y$$

Note that whenever ε_u is positive, both the dependent variable ($Y/(V + \varepsilon_u)$) and the predictor variable ($X/(V+ \varepsilon_u)$) in the regression are smaller in magnitude than the corresponding variables in the true model. Similarly, if ε_u is negative, both variables are larger than they are supposed to be. This implies that the two variables move together in the data, not because the variables are

causally related, but because of noisy measurement of V. This will bias upwards the estimate of B_1— it is more positive than the true B_1.

This bias emerges whenever you divide the dependent and predictor variable by the same variable, *and the divisor is a noisy variable*. Many empirical researchers feel that such bias is inevitable, and suggest that you restate the regression in such a way as to avoid dividing both the LHS and RHS by the same variable. I generally side with this skeptical group, although I think it important to determine if the divisor does or does not accurately measure the theoretical construct. For example, I am less worried about dividing the LHS and RHS by population (to obtain per capita values) than I am about dividing by other variables that might be measured with considerable noise (or be noisy measures of the underlying theoretical construct.)