Violations of Gauss Markov Assumptions: Omitted Variable Bias

Econometric Methods, ECON 370

We have found that heteroskedasticity does not seem to be a really difficult problem to handle given that we have a choice of using robust standard errors, or WLS. Further, the OLS estimator remains unbiased and consistent. The inefficiency is easily handled by robust standard errors. Now we deal with the more difficult problem of correlation between ϵ and $\mathbf{x}$, that is the violation of $E(\epsilon|\mathbf{x}) = 0$. This problem where the independent variable is correlated with the errors is known the endogeneity or endogenous explanatory variables.

However, before we explore what might bring this about, a presentation of the assumption's importance is as follows;

Reminder 1: This is perhaps a better time to reiterate the importance of the Gauss Markov or OLS assumptions since we are emphasizing it violation here. What does $E(\epsilon|\mathbf{x}) = 0$ mean really? Intuitively, when the conditional expectation of the errors conditioning on the covariates or independent variables are zero, it essentially is saying the regardless of the realization of $\mathbf{x}$, ϵ would still on the average be zero. Put another way, it means that the distribution of the errors is not dependent on the values of the covariates, or ϵ is not correlated with $\mathbf{x}$. We know that when two random variables, say ϵ and one of the covariates, x_j are uncorrelated, the following holds,

$$\rho_{\epsilon,x_j} = \frac{cov(\epsilon, x_j)}{\sqrt{var(\epsilon)var(x_j)}} = 0$$

which then necessarily means that $cov(\epsilon, x_j) = 0$

Reminder 2: **Why is the** $cov(\epsilon, \times_j) = 0$ **assumption so important?**

Let the regression model be,

$$y = \alpha_0 + \alpha_1 x_1 + \epsilon$$

Then

$$cov(y, x_1) = cov(\alpha_0 + \alpha_1 x_1 + \epsilon, x)$$

$$= E\left(\alpha_0 x_1 + \alpha_1 x_1^2 + \epsilon x_1\right) - E(y)E(x_1)$$

$$= \alpha_0 E(x_1) + \alpha_1 E(x_1^2) + E(\epsilon x_1) - E(\alpha_0 + \alpha_1 x_1 + \epsilon)E(x_1)$$

$$= \alpha_0 E(x_1) + \alpha_1 E(x_1^2) + E(\epsilon x_1) - \alpha_0 E(x_1) - \alpha_1 \left[E(x_1)\right]^2 - E(\epsilon)E(x_1)$$

By the Gauss Markov assumptions that $E(\epsilon) = 0$ and $E(\epsilon x_1) = 0$, so that we get

$$cov(y, x_1) = \alpha_1 E(x_1^2) + E(\epsilon x_1) - \alpha_1 \left[E(x_1)\right]^2 - E(\epsilon)E(x_1)$$

$$= \alpha_1 var(x_1) \Rightarrow p\lim \widehat{\alpha}_1 = \frac{cov(y, x_1)}{var(x_1)} = \alpha_1$$

It is clear from here that if for any reason, $E(\epsilon) \neq 0$ and $E(\epsilon x_1) \neq 0$, due to the violations we will be examining later, the estimator would be instead,

$$cov(y, x_1) = \alpha_1 var(x_1) + cov(\epsilon, x_1)$$

$$\Rightarrow p\lim \widehat{\alpha}_1 = \frac{cov(y, x_1)}{var(x_1)} + \frac{cov(\epsilon, x_1)}{var(x_1)}$$

$$= \alpha_1 + \frac{cov(\epsilon, x_1)}{var(x_1)} \neq \alpha_1$$

Which means that when the assumption is violated, our estimators are biased, and inconsistent. As an additional note, the conditioning on the covariate does not mean that the assumption holds only for the levels value of x_j alone, but includes that same variable in other functional forms such as squared or cubed, etc..

With the above reminders we can now proceed to understand what happens, how is happened, and what are the effects.

With the above reminders we can now proceed to understand what happens, how it happened, and what are the effects of endogeneity. **Endogeneity** occurs for three reasons, we first examine them briefly, namely

1. **Omitted Variable Bias**: This bias occurs often due to a lack of data. Consider the following, we are interested in finding the following relationship

$$E\left(y|\mathbf{x},\mathbf{q}\right)$$

where just like the vector of independent variables $\mathbf{x}$, we can express the vector of other independent variables $\mathbf{q}$ as a linear relationship with respect to y, so you can think of it as us performing an OLS. Omitted Variable bias then occurs when we do not have $\mathbf{q}$, and we end up performing

$$E\left(y|\mathbf{x}\right)$$

The two expressions in fact need not even be related in any manner when we allow $\mathbf{x}$ and $\mathbf{q}$ to be correlated. Another way to think about this is the following, suppose what we want to find out is

$$y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k+\beta_qq+\epsilon$$

$$\Rightarrow y=E(y|\mathbf{x},q)+\epsilon$$

$$\Rightarrow E(\epsilon|\mathbf{x})=E(\epsilon)=0$$

But because q is unobservable, we end up performing

$$y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k+\eta$$

$$\Rightarrow y=E(y|\mathbf{x})+\eta$$

Suppose q is correlated with say x_1 by the following relationship,

$$q=\alpha_0+\alpha_1x_1+\nu$$

where $\nu\sim N\left(0,\sigma_\nu^2\right)$ and $\epsilon\sim N\left(0,\sigma_\epsilon^2\right)$. In that case,

$$\eta=\beta_qq+\epsilon=\beta_q\left(\alpha_0+\alpha_1x_1+\nu\right)+\epsilon$$

Substituting this back into the second regression equation you get,

$$y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k+\beta_q\left(\alpha_0+\alpha_1x_1+\nu\right)+\epsilon$$

$$\Rightarrow y=\left(\beta_0+\beta_q\alpha_0\right)+\left(\beta_1+\beta_q\alpha_1\right)x_1+\beta_2x_2+...+\beta_kx_k+\beta_q\nu+\epsilon$$

$$\Rightarrow y=\gamma_0+\gamma_1x_1+\beta_2x_2+...+\beta_kx_k+\beta_q\nu+\epsilon$$

Note however that although

$$E(\eta)=\beta_qE(\nu)+E(\epsilon)=0$$

it is different from $E(\eta|\mathbf{x})$ since

$$E(\eta|\mathbf{x})=\beta_qE(\nu|\mathbf{x})+E(\epsilon|\mathbf{x})\neq 0$$

Since $E(\epsilon|\mathbf{x}, q) = E(\epsilon) = 0 \neq E(\epsilon|\mathbf{x})$. Further, when you examine the coefficent,

$$E\left(\widehat{\gamma}_1\right) \neq \beta_1$$

That is the estimator for the effect of x_1 on y is biased and will be inconsistent. To be precise,

$$p\lim \widehat{\gamma}_1 = \beta_1 + \beta_q \alpha_1 = \beta_1 + \beta_q \frac{cov(q, x_1)}{var(x_1)}$$

when what we had wanted was β_1. It should also be clear that the estimate of the variance will also be incorrect, consequently compromising your inference.

Further note that

$$var(\eta) = \beta_q^2 var(\nu) + var(\epsilon) + 2\beta_q cov(\nu, \epsilon) = \beta_q^2 \sigma_\nu^2 + \sigma_\epsilon^2 + 2\beta_q \sigma_{\nu,\epsilon} \neq \sigma_\epsilon^2$$

and if we assume that the two error terms are uncorrelated,

$$var(\eta) = \beta_q^2 \sigma_\nu^2 + \sigma_\epsilon^2 = \sigma_\epsilon^2$$

2. **Measurement Error**: Suppose what we would like to measure is the effect on a dependent variable, y, due to an independent variable, x_j. However, we are unable to get a perfect measure of the independent variable, and get instead $x_j^* = x_j + \gamma$. This then introduces a bias into the estimator of the effect of x_j. Note first that we will assume $E(\gamma) = 0$, and that it is uncorrelated with the rest of the independent variable, and the error term of the original regression. Next note that based on the population regression equation

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_j x_j^* + ... + \beta_k x_k + \epsilon$$

but with measurement error, it becomes

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + ... + \beta_j \left(x_j + \gamma\right) + ... + \beta_k x_k + \epsilon$$

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + ... + \beta_j x_j + ... + \beta_k x_k + \beta_j \gamma + \epsilon$$

$$\Rightarrow y = E(y|\mathbf{x}) + \beta_j \gamma + \epsilon$$

Next note the following; by assumption $cov(x_j^*, \gamma) = 0$, but this means that x_j will be correlated with γ, and the conditional expectation of errors being zero assumption in OLS would be violated. To see the former,

$$cov(x_j, \gamma) = E(x_j \gamma) - E(x_j)E(\gamma) = E(x_j^* \gamma + \gamma^2) = E(x_j^* \gamma) + E(\gamma^2) = \sigma_\gamma^2$$

$$cov(x_j, \gamma|\mathbf{x}) = E(x_j \gamma|\mathbf{x}) + E(x_j|\mathbf{x})E(\gamma|\mathbf{x})$$

$$= E(x_j^* \gamma + \gamma^2|\mathbf{x})$$

$$= E(x_j^* \gamma|\mathbf{x}) + E(\gamma^2|\mathbf{x}) = \sigma_\gamma^2 \neq 0$$

Let $\eta = \beta_j\gamma + \epsilon$, then the significance of the above is that,

$$E(\eta|\mathbf{x}) = E(\beta_j\gamma + \epsilon|\mathbf{x}) = \beta_j E(\gamma|\mathbf{x}) \neq 0$$

Clearly violating the Gauss Markov assumption. To see that, continue with our example,

$$cov(x_j, \eta|\mathbf{x}) = \sigma_\gamma^2 = E(x_j\gamma|\mathbf{x}) - E(x_j|\mathbf{x})E(\gamma|\mathbf{x})$$

$$\Rightarrow E(\gamma|\mathbf{x}) = \frac{E(x_j\gamma|\mathbf{x}) - \sigma_\gamma^2}{E(x_j)}$$

$$cov(x_j, \gamma|\mathbf{x}) = \sigma_\gamma^2 = E(x_j\gamma|\mathbf{x}) - E(x_j|\mathbf{x})E(\gamma|\mathbf{x})$$

$$\Rightarrow E(\gamma|\mathbf{x}) = \frac{E(x_j\gamma|\mathbf{x}) - \sigma_\gamma^2}{E(x_j)}$$

3. **Simultaneity**: This occurs when some independent variables are determined along with the dependent variable. Put another way, it may be possible that some of the independent variable is dependent on the dependent variable. In that case, those independent variables will be correlated with the error term. For example, does the relaxing of divorce laws increase incidences of divorce, or does the increase incidence of divorce forced the relaxing of divorce laws. We will examine this problem much later. The problem principally arises as a result of the fact that although what we need is for the independent variable to arise first, thereby causing variation in the dependent variable, the data we have typically are such that both y, and $\mathbf{x}$ are generated simultaneously.

We will now examine the problem arising from Misspecification Errors.

1 Misspecification

Our regression model suffers from **Functional Form Misspecification** when it does not account for the relationship between the dependent and independent variables correctly. An example: suppose the true relationship between the contract salary that a baseball player receives is as follows,

$$log(salary) = \beta_0 + \beta_1 RBI + \beta_2 Homers + \beta_3 FieldError + \beta_4 PitcherSal + \beta_5 Homers^2 + e$$

Where the square on homers (home runs) is largely due to the idea that there is only so much one player can do, given that baseball is a team sport. Then performing a regression as follows,

$$log(salary) = \beta_0 + \beta_1 RBI + \beta_2 Homers + \beta_3 FieldError + \beta_4 PitcherSal + e$$

consitutes a functional form misspecification, and as noted above, this leads to biased estimators, particularly for all the parameters (Can you see why? Show it to yourself). Of course the magnitude of the bias is dependent on the magnitude of the concavity, or β_5 and the correlation between $Homers^2$ and all the other variables (recalling correlation between independent variables are common). Note further that with this incorrect formulation, we cannot estimate the effect of Home Runs on salary.

Functional Form Misspecification is not confined however to just this one type of misspecification. It my also be a result of us ignoring the interaction effects between the independent variables. This type of

error is ultimately a result of our carelessness as researchers (we're all just humans), that is the problem can easily be solved since we do have the data on the missing variables. Further, we already know we can easily test for misspecified functional form since we can always use the F Test to test whether restrictions are correct. This highlights the prudence in the practice of always adding quadratic terms (especially when a priori expectations tells us we should).

However, you should keep in mind that significant quadratic terms can also signals of broader problems with functional form such as whether to use log values or levels of the variables in question. We can of course use log and add quadratics to detect the appropriate functional form.

1.1 RESET as a General Test for Functional Form Misspecification

Regression Specification Error Test (RESET) is a useful test to test for functional form misspecification. The idea behind the test is as follows: If we believe that a linear function is the correct function form, all quadratics or other polynomials would be statistically insignificant. This means that we perform a F Test or the Lagrange Multiplier Test of the hypothesis that the functional form is linear. However just like the White Test for Heteroskedasticity takes away to much degrees of freedom if the model in question has a lot of independent variables we can use the squared and cubed and ... predicted values of the dependent variable. That is in the k variable case, the alternate or restricted model would be,

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 \widehat{y}^2 + \delta_2 \widehat{y}^3 + e$$

The F statistic would then be asymptotically distributed as a $F_{2,n-(k+1)-2}$, while the χ^2 statistic would be asymptotically distributed as a χ^2_2.

To make sure that the model does not yield contrary results in log you could perform the above in log. A caveat as noted in your text about conclusions that may be drawn, note that it is possible that the RESET test reject quadratics in levels but not in log. That does not mean that we should use the linear regression in log. In fact, RESET test does not provide any guidance on what is the correct specification. All it says is whether your linear specification is correct or not. Further note that in log you would have already included concavity.

1.2 Tests against Nonnested Alternatives

Sometimes the functional forms on the independent variables are nonnested, in which case our usual F tests cannot be used. An example is when we wish to test log values of the independent variables, against the levels.

There are two approaches:

1. We can always perform a full model where we include both the levels and log values, and test each set of parameters in turn. Consider the simple regression case, then two possibilities of functional form are

$$y = \beta_0 + \beta_1 x_1 + e$$

and

$$y = \alpha_0 + \alpha_1 \log(x_1) + \nu$$

then a comprehensive model would be

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 \log(x_1) + \nu$$

So that to test the levels model, we would have as $H_0 : \gamma_2 = 0$, while the logs we would have $H_0 : \gamma_1 = 0$. This approach is due to Mizon and Richard (See the reference in your text).

2. Another approach was suggested by Davidson and MacKinnon (again see your text for the reference if you're interested). Consider the same example as above. The rationale of the test is as follows; if the level specification is correct, that the predicted values of the model in logs would be statistically insignificant in the levels model. Similarly if the logs model is correct, the fitted values from the levels would be statistically insignificant as a covariate in the logs regression.

   Let the predicted values from the *log*s model be $\widehat{\widehat{y}}$, and that from the levels model be $\widehat{y}$, then to test the levels specification, perform regression

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 \widehat{\widehat{y}} + e$$

   The null hypothesis is $H_0 : \beta_2 = 0$, which means that all you need to do is a two sided t test. You can do the same for the log model.

There are however problems with both of the test.

1. The tests may tell you all the models are good, or none at all. That is it may not provide much guidance. If all models are good, or more than 1, than we can always use R^2 as a guide to picking the models. Often times, alot depends on the researcher as well, how would she like to interpret the results, in levels or as percentages. Also, if the magnitudes suggested by all models are similar, then it does not really matter which models we should use.

2. Just because one of the models is rejected does not imply the other(s) which may not be rejected is(are) correct.

2 When Omitted Variable Bias is due to Unobservable Variables

As we had noted, the functional form has a "easy "solution, discover the correct functional form through testing more general functional forms.

We also noted that the more sinister problem arises when the omitted variable arises because the data is simply not collected or could not be collected. This problem makes our variable(s) of interest biased, inconsistent and OLS inefficient.

2.1 Using Proxy Variables for Unobserved Explanatory Variables

We suggested that perhaps we could substitute a variable that is very correlated with the missing variable with a **proxy variable** (such as substituting educational attainment say number of years of education or grades in mathematics for IQ in a wage regression). Conceptually, this seems like a easy one, but we have to be sure it works, and when it does not work. Let have somthing more concrete, suppose the population regression is actually of the following form,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + \epsilon$$

However, the third variable, x_3^* is missing or unobservable. However, we do observe x_3 which is very correlated with x_3^* both theoretically, and through our a priori believes. Their relationship is of the following form;

$$x_3^* = \delta_0 + \delta_3 x_3 + \nu$$

There are two key assumptions that need to be made for this to work;

1. The error term in ϵ must be uncorrelated with the independent variable, x_1, x_2 and x_3^* which is not new. We also need now in addition that x_3 the proxy we use is also uncorrelated with ϵ. The last assumption simply implies that if we have x_3^* is unnecessary in the regression. In short,

$$E(\epsilon|x_1, x_2, x_3^*, x_3) = 0$$

2. That ν is uncorrelated with x_1, x_2 and x_3^*, as well as x_3 itself.

$$E(\nu|x_1, x_2, x_3^*, x_3) = 0$$

Together, they imply that;

$$E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) = \delta_0 + \delta_3 x_3$$

The first equality implies that once x_3 is controlled for, the expected value of x_3^* will not be affected by x_1,and x_2,since it will not contain any additional information. Put another way, once we have controlled for the effect x_3 has on x_3^*, there is no additional effect from x_1 and x_2 that could explain the variation in x_3^*.

It should become obvious now how the assumptions work to the favour of the idea of using a proxy variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + \epsilon$$

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\delta_0 + \delta_3 x_3 + \nu) + \epsilon$$

$$\Rightarrow y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3\delta_3 x_3 + (\beta_3\nu + \epsilon)$$

Let $(\beta_0 + \beta_3\delta_0) = \alpha_0$, $\beta_3\delta_3 = \alpha_3$ and $\eta = \beta_3\nu + \epsilon$, then the above can be written as,

$$\Rightarrow y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + \eta$$

Note that $E(\nu|x_1, x_2, x_3) = \beta_3 E(\nu|x_1, x_2, x_3) + E(\epsilon|x_1, x_2, x_3) = 0$. However, it is clear that neither the intercept nor the parameter of interest, β_3 would be unbiased, since $(\beta_0 + \beta_3\delta_0) \neq \beta_0$, and $\beta_3\delta_3 \neq \beta_3$.

Nonetheless, the important issue is that the other parameters of β_1 and β_2 are unbiased and consistent, especially if what we are most concerned with is the other variables. If you can not see it, just think, if a typically regression has variable correlated with each other, the omission of one due to it being unobservable, would have its effect hidden within the error term. This would effectively bias all the estimators. Here, by assumption of the effectiveness of the proxy, we sacrifice a few parameters, but gain in the unbiasedness of the other estimated parameters. As your text book suggests, sometimes, rather fortuitously, a good proxy may provide a neater interpretation than if you had used the actual variable. Considering the IQ and education analogy, is not education an easier way in interpret ability's effect on wages, or even mathematics grades! Wouldn't it boost your interest in all things quantitative if you had known that a one grade increase, or 5 mark increase could have landed you a $200,000 job as opposed to a dead end $60,000 job even if you are a return to lander. Think about the acreage differential!

However, if the unobserved variable x_3^* is correlated with every variable in the regression, all estimators would still be biased. (Show yourself this will be true). However, in general recognizing a problem, and attempting to fix it would give you better traction in your empirical work, then if you simply ignore it.

2.2 Using Lagged Dependent Variables

The ideas so far pertaining to omission of variables has been illustrated with the background that we are using individual cross sectional data structures. It is very often that we could work with aggregated data, such as passing rate in mathematics across provinces, or unemployment rate across provinces, etc. If studies are of such aggregated nature, empirical work typically collect the data across several years together, forming what we have noted is a pooled cross sectional dataset.

Under those situations our concerns with endogeneity, a good proxy may not be obvious particularly when you are dealing with the economy at large, considering the myriad of possibilities. When we pool several cross sectional data sets together, we create a historical perspective of phenomenon. One possible manner in which we could account of omitted variables is to include **lagged** (of the previous period) observations of the dependent variable so that we account of all the historical occurrences that might have caused current differential in the dependent variables that are difficult to account for otherwise. In other words, we use a catchall in hopes that it covers everything we don't see, and pray that our parameter estimates will become unbiased with the inclusion of this **lagged dependent variable** on the right hand side of the regression equation.

Consider the simple example; let y be the academic achievement of a university measured by a weighted measure of number of publications. Let the independent variable include variables ranking of degree granting institution of the average academic in the institution x_1, proportion of money devoted to research endeavours x_2, proportion of faculty devoted to research x_3, average entry grade point average (weighted by importance of subject) of students x_4, proportion of graduates who enter top graduate programs x_5, and indicator of whether the school has a graduate program or otherwise x_6. This model could be structure as a OLS regression equation (we should definitely include province effects (i.e. provincial indicator variables),

$$y_{i,t} = \beta_0 + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + ... + \beta_6 x_{6,i,t} + \epsilon$$

where i denotes the observation, and t denotes the time from which the observation was drawn, i.e. which cross section. Despite the seeming completeness of the independent variables accounted for, there is always a heritage problem where we have a problem putting a finger on the issue at hand that might cause any deviations, some of which might be totally exogenous to the institution and yet stuck on through time. This then suggests that we might include the lagged variable for y as a covariate. Let this lagged covariate be denoted by $y_{i,t-1}$ so that the regression becomes

$$y_{i,t} = \beta_0 + \beta_1 x_{1,i,t} + \beta_2 x_{2,i,t} + ... + \beta_6 x_{6,i,t} + \beta_7 y_{i,t-1} + \epsilon$$

In all truth, in such a regression, you can imagine what you would be concerned with is the parameter for all the x variables. But we do expect that β_7 to be positive based on the idea, the good institution beget good genes and it rolls on in its achievement. What the variable essentially takes away the effect a Harvard or Yale has that a smaller institution does not, that we can't put a finger on, not that it is of substantive importance. What it might assist with is making the other covariates unbiased or probably less biased (hopefully!). It also makes interpretation far easier, since we can say holding histories the same, how would an increase in emphasis on research emphasis in terms of less teaching time might go towards raising the academic achievement of a university.

You should read the short section in your book on **A Different Slant on Multiple Regression** which essentially highlights that even if you know the bias still exists, you could "reduce the rigour "of your empirical study and accept the fact that your empirical work has flaws, but you could interpret the results still, and the result is still of substantive value.