OLS in Matrix Form

1 The True Model

- Let X be an $n \times k$ matrix where we have observations on k independent variables for n observations. Since our model will usually contain a constant term, one of the columns in the X matrix will contain only ones. This column should be treated exactly the same as any other column in the X matrix.
- Let y be an $n \times 1$ vector of observations on the dependent variable.
- Let ϵ be an $n \times 1$ vector of disturbances or errors.
- Let β be an $k \times 1$ vector of unknown population parameters that we want to estimate.

Our statistical model will essentially look something like the following:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix}_{n\times k} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \vdots \\ \beta_n \end{bmatrix}_{k\times 1} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1}$$

This can be rewritten more simply as:

$$y = X\beta + \epsilon \tag{1}$$

This is assumed to be an accurate reflection of the real world. The model has a systematic component ($X\beta$) and a stochastic component (ϵ). Our goal is to obtain estimates of the population parameters in the β vector.

2 Criteria for Estimates

Our *estimates* of the population parameters are referred to as $\hat{\beta}$. Recall that the criteria we use for obtaining our estimates is to find the estimator $\hat{\beta}$ that minimizes the sum of squared residuals ($\sum e_i^2$ in scalar notation).[1] Why this criteria? Where does this criteria come from?

The vector of residuals e is given by:

$$e = y - X\hat{\beta} \tag{2}$$

[1] Make sure that you are always careful about distinguishing between disturbances (ϵ) that refer to things that cannot be observed and residuals (e) that can be observed. It is important to remember that $\epsilon \neq e$.

The sum of squared residuals (RSS) is $e'e$.[2]

$$\begin{bmatrix} e_1 & e_2 & \dots & \dots & e_n \end{bmatrix}_{1\times n} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} e_1 \times e_1 + e_2 \times e_2 + \ldots + e_n \times e_n \end{bmatrix}_{1\times 1} \tag{3}$$

It should be obvious that we can write the sum of squared residuals as:

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \tag{4}$$

where this development uses the fact that the transpose of a scalar is the scalar i.e. $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$.

To find the $\hat{\beta}$ that minimizes the sum of squared residuals, we need to take the derivative of Eq. 4 with respect to $\hat{\beta}$. This gives us the following equation:

$$\frac{\partial e'e}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0 \tag{5}$$

To check this is a minimum, we would take the derivative of this with respect to $\hat{\beta}$ again – this gives us $2X'X$. It is easy to see that, so long as X has full rank, this is a positive definite matrix (analogous to a positive real number) and hence a minimum.[3]

[2]It is important to note that this is very different from ee' – the variance-covariance matrix of residuals.

[3]Here is a brief overview of matrix differentiaton.

$$\frac{\partial a'b}{\partial b} = \frac{\partial b'a}{\partial b} = a \tag{6}$$

when a and b are K×1 vectors.

$$\frac{\partial b'Ab}{\partial b} = 2Ab = 2b'A \tag{7}$$

when A is any symmetric matrix. Note that you can write the derivative as either 2Ab or $2b'A$.

$$\frac{\partial 2\beta'X'y}{\partial b} = \frac{\partial 2\beta'(X'y)}{\partial b} = 2X'y \tag{8}$$

and

$$\frac{\partial \beta'X'X\beta}{\partial b} = \frac{\partial \beta'A\beta}{\partial b} = 2A\beta = 2X'X\beta \tag{9}$$

when $X'X$ is a K×K matrix. For more information, see Greene (2003, 837-841) and Gujarati (2003, 925).

From Eq. 5 we get what are called the 'normal equations'.

$$(X'X)\hat{\beta} = X'y \tag{10}$$

Two things to note about the $(X'X)$ matrix. First, it is always square since it is $k \times k$. Second, it is always symmetric.

Recall that $(X'X)$ and $X'y$ are known from our data but $\hat{\beta}$ is unknown. If the inverse of $(X'X)$ exists (i.e. $(X'X)^{-1}$), then pre-multiplying both sides by this inverse gives us the following equation:[4]

$$(X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \tag{11}$$

We know that by definition, $(X'X)^{-1}(X'X) = I$, where I in this case is a $k \times k$ identity matrix. This gives us:

$$\begin{aligned} I\hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\beta} &= (X'X)^{-1}X'y \end{aligned} \tag{12}$$

Note that we have not had to make any assumptions to get this far! Since the OLS estimators in the $\hat{\beta}$ vector are a linear combination of existing random variables (X and y), they themselves are random variables with certain straightforward properties.

3 Properties of the OLS Estimators

The primary property of OLS estimators is that they satisfy the criteria of minimizing the sum of squared residuals. However, there are other properties. These properties do not depend on any assumptions - they will always be true so long as we compute them in the manner just shown.

Recall the normal form equations from earlier in Eq. 10.

$$(X'X)\hat{\beta} = X'y \tag{13}$$

Now substitute in $y = X\hat{\beta} + e$ to get

$$\begin{aligned} (X'X)\hat{\beta} &= X'(X\hat{\beta} + e) \\ (X'X)\hat{\beta} &= (X'X)\hat{\beta} + X'e \\ X'e &= 0 \end{aligned} \tag{14}$$

[4]The inverse of $(X'X)$ may not exist. If this is the case, then this matrix is called non-invertible or singular and is said to be of less than full rank. There are two possible reasons why this matrix might be non-invertible. One, based on a trivial theorem about rank, is that $n < k$ i.e. we have more independent variables than observations. This is unlikely to be a problem for us in practice. The other is that one or more of the independent variables are a *linear* combination of the other variables i.e. *perfect* multicollinearity.

What does $X'e$ look like?

$$\begin{bmatrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} X_{11} \times e_1 + X_{12} \times e_2 + \ldots + X_{1n} \times e_n \\ X_{21} \times e_1 + X_{22} \times e_2 + \ldots + X_{2n} \times e_n \\ \vdots \\ \vdots \\ X_{k1} \times e_1 + X_{k2} \times e_2 + \ldots + X_{kn} \times e_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \tag{15}$$

From $X'e = 0$, we can derive a number of properties.

1. **The observed values of X are uncorrelated with the residuals.**

   $X'e = 0$ implies that for every column x_k of X, $x_k'e = 0$. In other words, each regressor has zero sample correlation with the residuals. Note that this does not mean that X is uncorrelated with the disturbances; we'll have to assume this.

   If our regression includes a constant, then the following properties also hold.

2. **The sum of the residuals is zero.**

   If there is a constant, then the first column in X (i.e. X_1) will be a column of ones. This means that for the first element in the $X'e$ vector (i.e. $X_{11} \times e_1 + X_{12} \times e_2 + \ldots + X_{1n} \times e_n$) to be zero, it must be the case that $\sum e_i = 0$.

3. **The sample mean of the residuals is zero.**

   This follows straightforwardly from the previous property i.e. $\overline{e} = \frac{\sum e_i}{n} = 0$.

4. **The regression hyperplane passes through the means of the observed values ($\overline{X}$ and $\overline{y}$).**

   This follows from the fact that $\overline{e} = 0$. Recall that $e = y - X\hat{\beta}$. Dividing by the number of observations, we get $\overline{e} = \overline{y} - \overline{x}\hat{\beta} = 0$. This implies that $\overline{y} = \overline{x}\hat{\beta}$. This shows that the regression hyperplane goes through the point of means of the data.

5. **The predicted values of y are uncorrelated with the residuals.**

   The predicted values of y are equal to $X\hat{\beta}$ i.e. $\hat{y} = X\hat{\beta}$. From this we have

   $$\hat{y}'e = (X\hat{\beta})'e = b'X'e = 0 \tag{16}$$

   This last development takes account of the fact that $X'e = 0$.

6. **The mean of the predicted Y's for the sample will equal the mean of the observed Y's i.e. $\overline{\hat{y}} = \overline{y}$.**

These properties always hold true. You should be careful not to infer anything from the residuals about the disturbances. For example, you cannot infer that the sum of the disturbances is zero or that the mean of the disturbances is zero just because this is true of the residuals - this is true of the residuals just because we decided to minimize the sum of squared residuals.

Note that we know nothing about $\hat{\beta}$ except that it satisfies all of the properties discussed above. We need to make some assumptions about the true model in order to make any inferences regarding β (the true population parameters) from $\hat{\beta}$ (our estimator of the true parameters). Recall that $\hat{\beta}$ comes from our sample, but we want to learn about the true parameters.

4 The Gauss-Markov Assumptions

1. $y = X\beta + \epsilon$

   This assumption states that there is a linear relationship between y and X.

2. X is an $n \times k$ matrix of full rank.

   This assumption states that there is no perfect multicollinearity. In other words, the columns of X are linearly independent. This assumption is known as the identification condition.

3. $E[\epsilon|X] = 0$

$$E\begin{bmatrix} \epsilon_1|X \\ \epsilon_2|X \\ \vdots \\ \epsilon_n|X \end{bmatrix} = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{17}$$

   This assumption - the zero conditional mean assumption - states that the *disturbances* average out to 0 for any value of X. Put differently, no observations of the independent variables convey any information about the expected value of the disturbance. The assumption implies that $E(y) = X\beta$. This is important since it essentially says that we get the mean function right.

4. $E(\epsilon\epsilon'|X) = \sigma^2 I$

   This captures the familiar assumption of homoskedasticity and no autocorrelation. To see why, start with the following:

$$E(\epsilon\epsilon'|X) = E\begin{bmatrix} \epsilon_1|X \\ \epsilon_2|X \\ \vdots \\ \epsilon_n|X \end{bmatrix} \begin{bmatrix} \epsilon_1|X & \epsilon_2|X & \dots & \epsilon_n|X \end{bmatrix} \tag{18}$$

which is the same as:

$$E(\epsilon\epsilon'|X) = E\begin{bmatrix} \epsilon_1^2|X & \epsilon_1\epsilon_2|X & \dots & \epsilon_1\epsilon_n|X \\ \epsilon_2\epsilon_1|X & \epsilon_2^2|X & \dots & \epsilon_2\epsilon_n|X \\ \vdots & \vdots & \vdots & \vdots \\ \epsilon_n\epsilon_1|X & \epsilon_n\epsilon_2|X & \dots & \epsilon_n^2|X \end{bmatrix} \tag{19}$$

which is the same as:

$$E(\epsilon\epsilon'|X) = \begin{bmatrix} E[\epsilon_1^2|X] & E[\epsilon_1\epsilon_2|X] & \dots & E[\epsilon_1\epsilon_n|X] \\ E[\epsilon_2\epsilon_1|X] & E[\epsilon_2^2|X] & \dots & E[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|X] & E[\epsilon_n\epsilon_2|X] & \dots & E[\epsilon_n^2|X] \end{bmatrix} \tag{20}$$

The assumption of homoskedasticity states that the variance of ϵ_i is the same (σ^2) for all i i.e. $\text{var}[\epsilon_i|X] = \sigma^2 \ \forall \ i$. The assumption of no autocorrelation (uncorrelated errors) means that $\text{cov}(\epsilon_i, \epsilon_j|X) = 0 \ \forall \ i \neq j$ i.e. knowing something about the disturbance term for one observation tells us nothing about the disturbance term for any other observation. With these assumptions, we have:

$$E(\epsilon\epsilon'|X) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \tag{21}$$

Finally, this can be rewritten as:

$$E(\epsilon\epsilon'|X) = \sigma^2\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I \tag{22}$$

Disturbances that meet the two assumptions of homoskedasticity and no autocorrelation are referred to as spherical disturbances. We can compactly write the Gauss-Markov assumptions about the disturbances as:

$$\Omega = \sigma^2 I \tag{23}$$

where Ω is the variance-covariance matrix of the disturbances i.e. $\Omega = E[\epsilon\epsilon']$.

5. X may be fixed or random, but must be generated by a mechanism that is unrelated to ϵ.

6. $\epsilon|X \sim N[0, \sigma^2 I]$

   This assumption is not actually required for the Gauss-Markov Theorem. However, we often assume it to make hypothesis testing easier. The Central Limit Theorem is typically evoked to justify this assumption.

5 The Gauss-Markov Theorem

The Gauss-Markov Theorem states that, conditional on assumptions 1-5, there will be no other linear and unbiased estimator of the β coefficients that has a smaller sampling variance. In other words, the OLS estimator is the Best Linear, Unbiased and Efficient estimator (BLUE). How do we know this?

Proof that $\hat{\beta}$ is an unbiased estimator of β.

We know from earlier that $\hat{\beta} = (X'X)^{-1}X'y$ and that $y = X\beta + \epsilon$. This means that

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \epsilon) \\ \hat{\beta} &= \beta + (X'X)^{-1}X'\epsilon \end{aligned} \tag{24}$$

since $(X'X)^{-1}X'X = I$. This shows immediately that OLS is unbiased so long as either (i) X is fixed (non-stochastic) so that we have:

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\ &= \beta + (X'X)^{-1}X'E[\epsilon] \end{aligned} \tag{25}$$

where $E[\epsilon] = 0$ by assumption or (ii) X is stochastic but independent of ϵ so that we have:

$$\begin{aligned} E[\hat{\beta}] &= E[\beta] + E[(X'X)^{-1}X'\epsilon] \\ &= \beta + (X'X)^{-1}E[X'\epsilon] \end{aligned} \tag{26}$$

where $E(X'\epsilon) = 0$.

Proof that $\hat{\beta}$ is a linear estimator of β.

From Eq. 24, we have:

$$\hat{\beta} = \beta + (X'X)^{-1}X'\epsilon \tag{27}$$

Since we can write $\hat{\beta} = \beta + A\epsilon$ where $A = (X'X)^{-1}X'$, we can see that $\hat{\beta}$ is a linear function of the disturbances. By the definition that we use, this makes it a linear estimator (See Greene (2003, 45).

Proof that $\hat{\beta}$ has minimal variance among all linear and unbiased estimators

See Greene (2003, 46-47).

6 The Variance-Covariance Matrix of the OLS Estimates

We can derive the variance-covariance matrix of the OLS estimator, $\hat{\beta}$.

$$\begin{aligned} E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] &= E[((X'X)^{-1}X'\epsilon)((X'X)^{-1}X'\epsilon)'] \\ &= E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] \end{aligned} \tag{28}$$

where we take advantage of the fact that $(AB)' = B'A'$ i.e. we can rewrite $(X'X)^{-1}X'\epsilon$ as $\epsilon'X(X'X)^{-1}$. If we assume that X is non-stochastic, we get:

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1} \tag{29}$$

From Eq. 22, we have $E[\epsilon\epsilon'] = \sigma^2 I$. Thus, we have:

$$\begin{aligned} E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \\ &= \sigma^2 I(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{30}$$

We estimate σ^2 with $\hat{\sigma}^2$, where:

$$\hat{\sigma}^2 = \frac{e'e}{n-k} \tag{31}$$

To see the derivation of this, see Greene (2003, 49).

What does the variance-covariance matrix of the OLS estimator look like?

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'] = \begin{bmatrix} \text{var}(\hat{\beta}_1) & \text{cov}(\hat{\beta}_1,\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_1,\hat{\beta}_k) \\ \text{cov}(\hat{\beta}_2,\hat{\beta}_1) & \text{var}(\hat{\beta}_2) & \dots & \text{cov}(\hat{\beta}_2,\hat{\beta}_k) \\ \vdots & \vdots & \vdots & \vdots \\ \text{cov}(\hat{\beta}_k,\hat{\beta}_1) & \text{cov}(\hat{\beta}_k,\hat{\beta}_2) & \dots & \text{var}(\hat{\beta}_k) \end{bmatrix} \tag{32}$$

As you can see, the standard errors of the $\hat{\beta}$ are given by the square root of the elements along the main diagonal of this matrix.

6.1 Hypothesis Testing

Recall Assumption 6 from earlier, which stated that $\epsilon|X \sim N[0,\sigma^2 I]$. I had stated that this assumption was not necessary for the Gauss-Markov Theorem but was crucial for testing inferences about $\hat{\beta}$. Why? Without this assumption, we know nothing about the distribution of $\hat{\beta}$. How does this assumption about the distribution of the disturbances tell us anything about the distribution of $\hat{\beta}$? Well, we just saw in Eq. 27 that the OLS estimator is just a linear function of the disturbances. By assuming that the disturbances have a multivariate normal distribution i.e.

$$\epsilon \sim N[0,\sigma^2 I] \tag{33}$$

we are also saying that the OLS estimator is also distributed multivariate normal i.e.

$$\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}] \tag{34}$$

but where the mean is β and the variance is $\sigma^2(X'X)^{-1}$. It is this that allows us to conduct the normal hypothesis tests that we are familiar with.

7 Robust (Huber of White) Standard Errors

Recall from Eq. 29 that we have:

$$\begin{aligned} \text{var} - \text{cov}(\hat{\beta}) &= (X'X)^{-1}X'E[\epsilon\epsilon']X(X'X)^{-1} \\ &= (X'X)^{-1}(X'\Omega X)(X'X)^{-1} \end{aligned} \tag{35}$$

This helps us to make sense of White's heteroskedasticity consistent standard errors.[5]

Recall that heteroskedasticity does not cause problems for estimating the coefficients; it only causes problems for getting the 'correct' standard errors. We can compute $\hat{\beta}$ without making any assumptions about the disturbances i.e. $\hat{\beta}_{OLS} = (X'X)^{-1}X'y$. However, to get the results of the Gauss Markov Theorem (things like $E[\hat{\beta}] = \beta$ etc.) and to be able to conduct hypothesis tests ($\hat{\beta} \sim N[\beta, \sigma^2(X'X)^{-1}]$), we need to make assumptions about the disturbances. One of the assumptions is that $E[ee'] = \sigma^2 I$. This assumption includes the assumption of homoskedasticity – $\text{var}[\epsilon_i|X] = \sigma^2 \ \forall \ i$. However, it is not always the case that the variance will be the same for all observations i.e. we have σ_i^2 instead of σ^2. Basically, there may be many reasons why we are better at predicting some observations than others. Recall the variance-covariance matrix of the disturbance terms from earlier:

$$E(\epsilon\epsilon'|X) = \Omega = \begin{bmatrix} E[\epsilon_1^2|X] & E[\epsilon_1\epsilon_2|X] & \dots & E[\epsilon_1\epsilon_n|X] \\ E[\epsilon_2\epsilon_1|X] & E[\epsilon_2^2|X] & \dots & E[\epsilon_2\epsilon_n|X] \\ \vdots & \vdots & \vdots & \vdots \\ E[\epsilon_n\epsilon_1|X] & E[\epsilon_n\epsilon_2|X] & \dots & E[\epsilon_n^2|X] \end{bmatrix} \tag{36}$$

If we retain the assumption of no autocorrelation, this can be rewritten as:

$$E(\epsilon\epsilon'|X) = \Omega = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix} \tag{37}$$

Basically, the main diagonal contains n variances of ϵ_i. The assumption of homoskedasticity states that each of these n variances are the same i.e. $\sigma_i^2 = \sigma^2$. But this is not always an appropriate

[5] As we'll see later in the semester, it also helps us make sense of Beck and Katz's panel-corrected standard errors.

assumption to make. Our OLS standard errors will be incorrect insofar as:

$$X'E[\epsilon\epsilon']X \neq \sigma^2(X'X) \tag{38}$$

Note that our OLS standard errors may be too big or too small. So, what can we do if we suspect that there is heteroskedasticity?

Essentially, there are two options.

1. **Weighted Least Squares:** To solve the problem, we just need to find something that is proportional to the variance. We might not know the variance for each observation, but if we know something about where it comes from, then we might know something that is proportional to it. In effect, we try to model the variance. Note that this only solves the problem of heteroskedasticity if we assume that we have modelled the variance correctly - we never know if this is true or not.

2. **Robust standard errors** (White 1980): This method treats heteroskedasticity as a nuisance rather than something to be modelled.

How do robust standard errors work? We never observe disturbances (ϵ) but we do observe residuals (e). While each individual residual (e_i) is not going to be a very good estimator of the corresponding disturbance (ϵ_i), White (1980) showed that $X'ee'X$ is a consistent (but not unbiased) estimator of $X'E[\epsilon\epsilon']X$.[6]

Thus, the variance-covariance matrix of the coefficient vector from the White estimator is:

$$\text{var} - \text{cov}(\hat{\beta}) = (X'X)^{-1}X'ee'X(X'X)^{-1} \tag{39}$$

rather than:

$$\begin{aligned} \text{var} - \text{cov}(\hat{\beta}) &= X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1} \\ &= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} \end{aligned} \tag{40}$$

from the normal OLS estimator.

White (1980) suggested that we could test for the presence of heteroskedasticity by examining the extent to which the OLS estimator diverges from his own estimator. White's test is to regress the squared residuals (e_i^2) on the terms in $X'X$ i.e. on the squares and the cross-products of the independent variables. If the R^2 exceeds a critical value ($nR^2 \sim \chi_k^2$), then heteroskedasticity causes problems. At that point use the White estimator (assuming your sample is sufficiently large). Neal Beck suggests that, by and large, using the White estimator can do little harm and some good.

[6] It is worth remembering that $X'ee'X$ is a consistent (but not unbiased) estimator of $X'E[\epsilon\epsilon']X$ since this means that robust standard errors are only appropriate when the sample is relatively large (say, greater than 100 degrees of freedom).

8 Partitioned Regression and the Frisch-Waugh-Lovell Theorem

Imagine that our true model is:

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon \tag{41}$$

In other words, there are two sets of independent variables. For example, X_1 might contain some independent variables (perhaps also the constant) whereas X_2 contains some other independent variables. The point is that X_1 and X_2 need not be two variables only. We will estimate:

$$y = X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + e \tag{42}$$

Say, we wanted to isolate the coefficients associated with X_2 i.e. $\hat{\beta}_2$. The normal form equations will be:[7]

$$\begin{array}{l}(1)\\(2)\end{array}\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} \tag{43}$$

First, let's solve for $\hat{\beta}_1$.

$$\begin{aligned} (X_1'X_1)\hat{\beta}_1 + (X_1'X_2)\hat{\beta}_2 &= X_1'y \\ (X_1'X_1)\hat{\beta}_1 &= X_1'y - (X_1'X_2)\hat{\beta}_2 \\ \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 \\ \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2) \end{aligned} \tag{44}$$

8.1 Omitted Variable Bias

The solution shown in Eq. 44 is the set of OLS coefficients in the regression of y on X_1, i.e, $(X_1'X_1)^{-1}X_1'y$, minus a correction vector $(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2$. This correction vector is the equation for omitted variable bias. The first part of the correction vector up to $\hat{\beta}_2$, i.e. $(X_1'X_1)^{-1}X_1'X_2$, is just the regression of the variables in X_2 done separately and then put together into a matrix on all the variables in X_1. This will only be zero if the variables in X_1 are linearly unrelated (uncorrelated or orthogonal) to the variables in X_2. The correction vector will also be zero if $\hat{\beta}_2 = 0$ i.e. if X_2 variables have no impact on y. Thus, you can ignore all potential omitted variables that are either (i) unrelated to the included variables or (ii) unrelated to the dependent variable. Any omitted variables that do not meet these conditions will change your estimates of $\hat{\beta}_1$ if they were to be included.

Greene (2003, 148) writes the omitted variable formula slightly differently. He has

$$E[b_1] = \beta_1 + P_{1.2}\beta_2 \tag{45}$$

where $P_{1.2} = (X_1'X_1)^{-1}X_1'X_2$, where b_1 is the coefficient vector of a regression omitting the X_2

[7] To see this, compare with Eq. 10.

matrix, and β_1 and β_2 are the true coefficient vectors from a full regression including both X_1 and X_2.

8.2 The Residual Maker and the Hat Matrix

Before going any further, I introduce some useful matrices. Note that:

$$\begin{aligned} e &= y - X\hat{\beta} \\ &= y - X(X'X)^{-1}X'y \\ &= (I - X(X'X)^{-1}X')y \\ &= My \end{aligned} \tag{46}$$

where M is called the residual maker since it makes residuals out of y. M is a square matrix and is idempotent. A matrix A is idempotent if $A^2 = AA = A$.

$$\begin{aligned} MM &= (I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X') \\ &= I^2 - 2X(X'X)^{-1}X' + X(X'X)^{-1}X'X(X'X)^{-1}X' \\ &= I - 2X(X'X)^{-1}X' + X(X'X)^{-1}X' \\ &= I - X(X'X)^{-1}X' \\ &= M \end{aligned} \tag{47}$$

This will prove useful. The M matrix also has the properties that $MX = 0$ and $Me = e$.

A related matrix is the hat matrix (H) which makes $\hat{y}$ our of y. Note that:

$$\hat{y} = y - e = [I - M]y = Hy \tag{48}$$

where:

$$H = X(X'X)^{-1}X' \tag{49}$$

Greene refers to this matrix as P, but he is the only one that does this.

8.3 Frisch-Waugh-Lovell Theorem

So far we have solved for $\hat{\beta}_1$.

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2) \tag{50}$$

Now we insert this into (2) of Eq. 43. This gives us

$$\begin{aligned}
X_2'y &= X_2'X_1(X_1'X_1)^{-1}X_1'y - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 + X_2'X_2\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= X_2'X_2\hat{\beta}_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= [X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2]\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= [(X_2' - X_2'X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
(X_2' - X_2'X_1(X_1'X_1)^{-1}X_1')y &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
X_2'(I - X_1(X_1'X_1)^{-1}X_1')y &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]\hat{\beta}_2 \\
\hat{\beta}_2 &= [X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2]^{-1}X_2'(I - X_1(X_1'X_1)^{-1}X_1')y \\
&= (X_2'M_1X_2)^{-1}(X_2'M_1y) \qquad (51)
\end{aligned}$$

Recall that M is the residual maker. In this case, M_1 makes residuals for regressions on the X_1 variables: M_1y is the vector of residuals from regressing y on the X_1 variables and M_1X_2 is the matrix made up of the column by column residuals of regressing each variable (column) in X_2 on all the variables in X_1.

Because M is both idempotent and symmetric, we can rewrite Eq. 51 as

$$\hat{\beta}_2 = (X_2^{*'}X_2)^{-1}X_2^{*'}y^* \qquad (52)$$

where $X_2^* = M_1X_2$ and $y^* = M_1y$.

From this it is easy to see that $\hat{\beta}_2$ can be obtained from regressing y^* on X_2^* (you'll get good at spotting regressions i.e. equations of the $(X'X)^{-1}X'y$ form. The starred variables are just the residuals of the variables (y or X_2) after regressing them on the X_1 variables.

This leads to the **Frisch-Waugh-Lovell Theorem**: In the OLS regression of vector y on two sets of variables, X_1 and X_2, the subvector $\hat{\beta}_2$ is the set of coefficients obtained when the residuals from a regression of y on X_1 alone are regressed on the set of residuals obtained when each column of X_2 is regressed on X_1.

We'll come back to the FWL Theorem when we look at fixed effects models.

8.4 Example

Imagine we have the following model.

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \epsilon \qquad (53)$$

If we regressed Y on X_1, X_2, and X_3, we would get $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$. We could get these estimators differently. Say we partitioned the variables into (i) X_1 and (ii) X_2 and X_3.

Step 1: regress Y on X_1 and obtain residuals (e1) i.e. M_1y.

Step 2: regress X_2 on X_1 and obtain residuals (e2) i.e. first column of M_1X_2.
Step 3: regress X_3 on X_1 and obtain residuals (e3) i.e. second column of M_1X_2.
Step 4: regress e1 on e2 and e3 i.e. regress M_1y on M_1X_2.
Step 5: the coefficient on e2 will be $\hat{\beta}_2$ and the coefficient on e3 will be $\hat{\beta}_3$.

Steps 2 and 3 are called partialing out or netting out the effect of X_1. For this reason, the coefficients in multiple regression are often called partial regression coefficients. This is what it means to say we are holding the X_1 variables constant in the regression.

So the difference between regressing Y on both X_1 and X_2 instead of on just X_2 is that in the first case we first regress both the dependent variables and all the X_2 variables separately on X_1 and then regress the residuals on each other, but in the second case we just regress y on the X_2 variables.