Midterm Review

Basics of the OLS Estimator

The goal of econometrics is to use sample data to obtain estimates of unknown population parameters. For the population model:
\[ y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \]
we can use ordinary least squares (OLS) to obtain a sample model:
\[ y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i. \]
and thus get a sample prediction:
\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \hat{\beta}_k x_{ik}. \]

The OLS estimator can be derived based on one crucial assumption (and one incidental assumption).

The incidental assumption is that \( E(u) = 0 \).
The crucial assumption is that \( E(u|x) = E(u) = 0 \) (zero conditional mean).

These assumptions allow us to derive that
\[ \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(\sum y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \]
for the simple regression. This can also be interpreted as the sample covariance between \( x \) and \( y \) divided by the sample variance of \( x \).

We can think of the OLS estimator as having minimized the sum of squared residuals (\( \hat{u} \)), where \( y_i = \hat{y}_i + \hat{u}_i \) and
\[ \sum (y_i - \bar{y})^2 = SST \]
\[ \sum (\hat{y}_i - \bar{y})^2 = SSE \]
\[ \sum \hat{u}_i^2 = SSR \]
so that \( SST = SSE + SSR \).

The \( R^2 \) is the fraction of the variation in \( y \) that is explained by the estimated model, and thus measures goodness of fit. It is defined as:
\[ R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \]
The \( R^2 \) can also be interpreted as the square of the correlation between \( y \) and \( \hat{y} \).

Unbiasedness & Omitted Variable Bias

OLS is an unbiased estimator, provide the following 4 assumption hold:
1. The population model is linear in parameters
2. We can use a random sample from the population to estimate the model
3. \( E(u|x) = E(u) = 0 \)
4. None of the \( x \)'s is constant, and there are no exact linear relationships

Omitting a variable that belongs in the model will often violate assumption 3 and can lead to omitted variable bias. We can sign the bias by deriving that if when the true population model is
\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u
\]
we estimate the simple regression, where \( \widetilde{\beta}_1 \) is the OLS estimate, then
\[
E(\widetilde{\beta}_1) = \beta_1 + \beta_2 \widetilde{\delta}
\]
where \( \widetilde{\delta} \) is the sample covariance between \( x_1 \) and \( x_2 \) divided by the sample variance of \( x_1 \). We just need to use common sense to think about what signs \( \beta_2 \) and \( \widetilde{\delta} \) are likely to have, then we can decide if \( \widetilde{\beta}_1 \) is too big (positively biased) or too small (negatively biased) relative to \( \beta_1 \). Note that asymptotic bias (consistency) can be thought of in exactly the same manner, except that technically we are thinking about \( \delta \), the population covariance divided by the population variance, instead of \( \widetilde{\delta} \).

If the omitted variable is not correlated with the variable of interest, there is no bias, since \( \widetilde{\delta} = 0 \).

**Variance of the OLS Estimator**

In order to derive inferences from our estimates, we need to know something about the variance of our OLS estimator. We first need to add the assumption of homoskedasticity to the previous 4. All 5 are collectively known as the Gauss-Markov Assumptions. This new assumption is:
\[
\text{Var}(u|x) = \sigma^2
\]
That is, that the errors have a constant variance, conditional on the \( x \)'s.

Given the Gauss-Markov Assumptions, we can derive that:
\[
\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1-R_j)}
\]
where \( \text{SST}_j \) is the total sum of squares for \( x_j \), and \( R^2_j \) is the \( R^2 \) from regressing \( x_j \) on the other \( x \)'s.

Thus, the error variance is larger when:
1. The variance of the errors is larger
2. The total sample variation in \( x_j \) is smaller
3. The relationship between \( x_j \) and the other \( x \)'s is stronger

Unfortunately, we don’t know what \( \sigma^2 \) is, so we need to estimate it as:
\[
\hat{\sigma}^2 = \frac{\text{SSR}}{(n-k-1)}
\]
Then we can substitute this into the variance formula, and take the square root to obtain the standard error of the estimate:
\[
\text{se}(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{\text{SST}_j(1-R_j)} = \frac{(\text{SSR}/df)}{\text{SST}_j(1-R_j)}
\]
**The Gauss-Markov Theorem**

The Gauss-Markov Theorem states that under the 5 Gauss-Markov assumptions, OLS is BLUE – the Best, Linear, Unbiased, Estimator.

**Statistical Testing – t-tests**

Now that we have a standard error for our OLS estimate, we can use it for statistical testing. Unless we have large samples, and can rely on asymptotic normality, we must add yet another assumption to the Gauss-Markov assumptions. The new assumption is:

\[ u \sim \text{Normal}(0, \sigma^2) \]

That is, that the errors are normally distributed.

Now since \( \left( \hat{\beta}_j - \beta_j \right) / \text{se}(\hat{\beta}_j) \sim t_{n-k-1} \) we can perform a t-test for whether to reject a null hypothesis about one parameter. This test can be one-sided or two-sided, although most standard packages (including Stata) assume a two-sided test. Usually, our null hypothesis will be

\[ H_0: \beta_j = 0, \]

but a more general specification is \( H_0: \beta_j = a \), where \( a \) is a constant.

Since OLS is unbiased, if our null hypothesis is true, then the distribution of \( \hat{\beta}_j \) will be centered on \( a \), since then \( E(\hat{\beta}_j) = \beta_j \) by definition of unbiased. Thus, we can form a t statistic that is just a standardized version of \( \beta_j \) and which has a \( t \) distribution as noted above. The idea is that if our estimate is “too far” into the tail of the hypothesized distribution, then our hypothesis is likely wrong, and we’ll reject it.

**Classical Statistical Testing**

Classical statistical testing chooses a significance level, \( \alpha \), first – say 5% – meaning there will only be a 5% chance of rejecting the null if it is really true. That means we will reject the null if our estimate is beyond 5% of the distribution. In a one-sided test (\( H_1: \beta_j > a \), or \( H_1: \beta_j < a \)) we set a critical value based on leaving all 5% in one tail. In a two-sided test (\( H_1: \beta_j \neq a \)) we set a critical value based on leaving 2.5% in each tail. A table of the \( t \)-distribution will allow us to pick these critical values. Having picked our critical value, \( c \), then:

- reject the null if \( |t| > c \)
- fail to reject if \( |t| < c \)

We can also use these critical values to form confidence intervals (95% in this case) as

\[ \hat{\beta}_j \pm c \times \text{se}(\hat{\beta}_j) \]

**Inference Using p-values**

Alternatively, we can simply calculate where in the distribution our estimate is. This p-value represents the answer to the question,

“what is the smallest significance level at which the null would be rejected?”
It gives us the probability that we would observe the $t$ statistic we did, if the null were actually true (i.e. if the distribution were really centered on $a$).

To test a linear combination, say
\[ H_0: \beta_1 = \beta_2, \]
we can always use algebra to rearrange the model to directly estimate the hypothesis of interest,
\[ H_0: \theta = \beta_1 - \beta_2 = 0, \]
using the standard error printed by our statistical package.

**Statistical Testing – Multiple Linear Restrictions**

In order to test multiple linear restrictions, or exclusion restrictions, we have to perform an $F$ test. The idea is to see if the change in SSR from restricting the model is “big enough” to reject the null that these restrictions are true. We form an $F$ statistic as
\[
F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}
\]
or, equivalently, as
\[
F = \frac{(R^2_r - R^2_{ur})/q}{(1 - R^2_r)/(n-k-1)}
\]
where $r$ stands for restricted, $ur$ stands for unrestricted and $q = df_r - df_{ur}$ (i.e. the # of restrictions). Note that if the restricted and unrestricted models have different dependent variables, we must use the first form.

Just as with a $t$-test, we can either choose a significance level ahead of time, or calculate and report a p-value.

An alternative test, valid only when our sample size is large, is an LM statistic. To form an LM statistic we have to estimate the restricted model, then run an auxiliary regression of the residuals on all of the $x$’s in the unrestricted model. The LM statistic is calculated as $n*R^2$ from this auxiliary regression. The LM statistic $\sim \chi^2_q$. As with the a $t$-test, or an $F$-test, we can either choose a significance level ahead of time, or calculate and report a p-value.

**Alternative Models and their Interpretation**

We can transform the variables in our model to fit nonlinear relationships between $x$ and $y$, but we need to be careful to think about the interpretation. For example, when using an interaction term, one cannot just look at the coefficient on the main effect. For a model of the form
\[
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u,
\]
to describe the effect of $x_1$ on $y$, you must look at
\[
\beta_1 + \beta_3 x_2
\]
since this is the derivative of $y$ with respect to $x_1$.

The adjusted $R^2$ is one way to think about comparing the fit of models with the same dependent variable. It adjusts for the fact that the $R^2$ will always rise when adding more variables. It is:
\[ R^2 = 1 - \frac{[SSR / df]}{SST /(n-1)} = 1 - \frac{\hat{\sigma}^2}{SST /(n-1)} \]

We can use our estimated model to predict \( E(y|x_i = c_i) \), that is for a given set of \( x \)'s, as
\[
\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + ... + \hat{\beta}_k c_k
\]
Since this is just a linear combination of our estimates, we can always just rearrange our model to estimate \( \hat{y}^0 \) and \( se(\hat{y}^0) \) directly. To also take into account the variance in \( y \) that comes from the error term, we need to take into account \( \sigma \) as well. We can obtain the standard error of the prediction as \( \sqrt{se(\hat{y}^0)^2 + \hat{\sigma}^2} \).

**Special Issues with log Dependent Variables**

When the dependent variable is in logs, there are a few extra things to remember besides interpreting coefficients in percentage terms. First, taking the exponent of the predicted \( \ln(y) \) will underestimate the predicted \( y \). It needs to be scaled up by \( E(\exp(u)) \). If errors are normal, this just means scaling up by \( \exp(\hat{\sigma}^2 / 2) \).

If the errors are not normal, \( E(\exp(u)) \) can be estimated using an auxiliary regression. After getting predicted \( \ln(y) \) and taking the exponent to get \( \hat{y}_c \), run the following auxiliary regression:
\[
y = \phi \hat{y}_c + u.
\]
Note that this regression does not include an intercept! The estimated coefficient, \( \hat{\phi} \), is the estimate of \( E(\exp(u)) \) that can be used to scale up \( \hat{y}_c \).

When comparing 2 models, one of which uses \( y \) and one of which uses \( \ln(y) \) as the dependent variable, you can’t just compare the \( R^2 \). However, there is something you can do using the just described auxiliary regression. Find the sample correlation between \( y \) and the predicted \( y \) from the auxiliary regression. Square this and compare it to the \( R^2 \) from the regression in levels. This makes sense because the \( R^2 \) from the levels regression can be interpreted as the squared correlation between \( y \) and the predicted \( y \) from the levels model, which is now being compared to the squared correlation between \( y \) and the predicted \( y \) from the log model.

**Using Dummy Variables**

Dummy variables are derived from categorical variables – you need one less dummy variable than there are categories. A dummy independent variable can be thought of as shifting the intercept, that is as estimating a separate intercept for each group. Interacting dummy variables with a continuous variable can be thought of as changing the slope, that is as estimating a different slope for each group. Interacting dummy variables subdivides your groups, but is still just estimating a separate intercept for each subgroup.

**The Chow Test**
If you think two groups should have different intercepts and slopes, you could estimate one model with a complete set of dummy variables and interactions. Then you could do an F-test on the exclusion restrictions for the dummy variables and interactions. An equivalent test is the Chow test, that requires running the model without the dummy variables and interactions (the restricted model) for the whole sample, and then separately for each group. This allows you to form the following F statistic:

\[
F = \left( \frac{SSR - (SSR_1 + SSR_2)}{SSR_1 + SSR_2} \right) \times \frac{n - 2(k + 1)}{k + 1}
\]

This is really just a standard F-test, since the sum of the SSR from estimating the model separately for each group is the same as the SSR from estimating the unrestricted model, there are \( k+1 \) restrictions (the group dummy, plus the group dummy times each of the \( k \) different \( x \) variables), and \( n - 2(k + 1) \) degrees of freedom.

**Linear Probability Models**

When the binary variable is the dependent variable, we call that a linear probability model. The linear probability model will violate the assumption of homoskedasticity, but is often a good place to start.

**Heteroskedasticity Robust Standard Errors**

When the assumption of homoskedasticity is violated, OLS remains unbiased and consistent, but the standard errors are biased and we cannot rely on the usual t statistics, F statistics and LM statistics for drawing inferences. Essentially, in deriving the variance of the OLS estimator, we relied on the fact that \( \sigma^2 \) was a constant to pull it outside the summation signs. Instead, now we will have to substitute \( \hat{u}_i^2 \) for \( \sigma^2 \) in our variance formulas. This will provide us with a consistent estimator of the variance. In the multiple regression case, the estimator of \( \text{Var}(\hat{\beta}_j) \) is:

\[
\sum \frac{\hat{r}_i^2 \hat{u}_i^2}{SST_j^2},
\]

where \( \hat{r}_i \) is the \( i \)th residual from regressing \( x_j \) on all other independent variables, and \( SST_j \) is the sum of squared residuals from this regression. Luckily, Stata will calculate robust standard errors for us. Note that in small samples, this correction will still not be sufficient.

**Heteroskedasticity Robust LM Statistic**

In the presence of heteroskedasticity, an alternative version of the LM statistic must be used. To obtain this statistic, begin by saving the residuals from the restricted model. Then regress each of the excluded variables on all of the included variables, saving each set of residuals \( \hat{r}_1, \hat{r}_2, \ldots, \hat{r}_q \). Interact the original residuals with each of these residuals to form \( q \) new variables \( \hat{r}_j \hat{u} \). Regress a variable equal to 1 on all of these \( q \) new variables, and no intercept. The LM statistic is \( n - \text{SSR} \) from this regression.

**The Breusch-Pagan Test**
The Breusch-Pagan test simply tests for whether the error variance (estimated by $\hat{\sigma}^2$) is a linear function of the $x$’s in the model. Thus you just regress the residuals squared on all of the $x$’s and do either a standard F or standard LM test for the joint significance of all of the $x$’s.

The White Test

The White test is more general as it allows the error variance to be a nonlinear function of the $x$’s in the model. Thus, it adds squares and cross-products of all the $x$’s. Again a standard test for exclusion restrictions is carried out. With more than a few $x$’s, this regression can get out of hand, so there is an alternative form of this test. It recognizes that the predicted $y$ from the model is a function of all of the $x$’s, and thus just adds nonlinear functions of it. So, this test requires regression the squared residuals on $\hat{y}$ and $\hat{y}^2$, testing for the joint significance of these 2 variables.

**When the Form of Heteroskedasticity is Known**

If the exact form of the heteroskedasticity is known, more efficient estimates can be obtained by transforming the equation. For example, if $\sigma^2 = h(x, \sigma^2)$, where $h(x)$ is a known function of the $x$’s, we can just divide through the whole equation by the square root of $h(x)$. An equivalent procedure, known as weighted least squares, allows us to obtain the same result without having to do the transformation. We simply weight by $1/h(x)$.

**Feasible Generalized Least Squares (GLS)**

When we don’t know the exact form, we can assume that $h$ is a flexible function of $x$, and proceed as follows. Run the original model, saving the residuals. Take the log of the squared residuals and regress that on all of the $x$’s. Use $1/\exp$(predicted value) as the weight.

**Specification Tests**

Ramsey’s RESET tests for whether there should be nonlinear functions of $x$ included in the model. Simply include $\hat{y}^2$ and $\hat{y}^3$ to the model and test for their joint significance. If you reject the null that they can be excluded, then additional interactions and quadratics should be included.

The Davidson-MacKinnon test can be used to compare nonnested alternative models. Simply include $\hat{y}$ from the first model as a regressor in the other model and test for significance. If you reject the null that it can be excluded, then the first model is preferred. Unfortunately, you could do this test the other way around and either accept or reject both models.

**Measurement Error**

Measurement error in $y$ is no problem if the error is uncorrelated with both the $x$’s and $u$. The effect of measurement error in $x$ depends on the assumption about the correlation between the error and the reported $x$. If uncorrelated, no problem. If correlated, then have classical errors-in-variables and expect attenuation bias – estimates that are biased toward zero.