LECTURE NOTE 4: HETEROSKEDASTICITY AND DEPENDENCE

1 Introduction

The classical approach to the OLS estimator assumes homoskedastic and independent (or uncorrelated) errors. This lecture is about what happens when the data do not meet these assumptions. For simplicity, we continue to deal with a bivariate model here and will generalize to a multivariate model next time.

2 Variance of the OLS Estimator

In this section, we discuss the main topic of the lecture: inference for the OLS estimator. We first derive a general expression for the estimator's variance and then discuss how to estimate it under different assumptions.

To start, let's come up with a slightly different expression for $\hat{\beta}_1$. You'll see in a bit why it's useful.

$$\begin{aligned}
\hat{\beta}_1 &= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \\
&= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(\beta_0 + \beta_1 X_i + U_i - \beta_0 - \beta_1 \bar{X} - \bar{U}\right) \\
&= \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(\beta_1 \left(X_i - \bar{X}\right) + \left(U_i - \bar{U}\right)\right) \\
&= \beta_1 + \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right)\left(U_i - \bar{U}\right) \\
&= \beta_1 + \frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right) U_i
\end{aligned}$$

This expression implies that $E[\hat{\beta}_1] = \beta_1$, thus confirming the unbiasedness of the OLS estimator. Also, the expression $\frac{\sum_i (X_i - \bar{X}) U_i}{\sum_i (X_i - \bar{X})^2}$ converges in probability to zero, so $\hat{\beta}_1$ is a consistent estimator of β_1.

For now, let's still assume that X_i is non-random, as in the Gauss-Markov setup. We'll relax this assumption below. Since X_i is non-random, we can treat all of the terms involving X_i as constant terms. Recall that if c_k is a series of constant terms and Z_k is a series of random variables, then $V\left[\sum_k c_k Z_k\right] =$

$\sum_k c_k^2 V[Z_k] + \sum_k \sum_{l \neq k} c_k c_l cov(Z_k, Z_l)$. Thus, if we take the variance of the estimator above, we obtain:

$$\begin{aligned} V\left[\hat{\beta}_1\right] &= V[\beta_1] + V\left[\frac{1}{\sum_i (X_i - \bar{X})^2} \sum_i \left(X_i - \bar{X}\right) U_i\right] \\ &= \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \left[\sum_i \left(X_i - \bar{X}\right)^2 V[U_i] + \sum_i \sum_{j \neq i} \left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right) cov\left(U_i, U_j\right)\right] \end{aligned} \quad (1)$$

We will use this equation to derive estimators for the variance (and therefore also the standard error) of the OLS estimator under different sets of assumptions.

2.1 Classical Model

At this point, we have not used any of Gauss and Markov's three assumptions. Equation (1) is a general result that holds under any assumptions about the variances and covariances of the errors. We *always* place some restrictions on these variances and covariances. In Gauss and Markov's case, these restrictions are that $V[U_i] = \sigma^2$ and $cov(U_i, U_j) = 0$ for all $i \neq j$. These restrictions imply:

$$\begin{aligned} V\left[\hat{\beta}_1\right] &= \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \sum_i \left(X_i - \bar{X}\right)^2 \sigma^2 \\ &= \frac{\sigma^2}{\sum_i \left(X_i - \bar{X}\right)^2} \\ &= \frac{\sigma^2}{NV[X]} \end{aligned}$$

We have an estimator for the denominator of this expression, but we still need to estimate σ^2. An unbiased and consistent estimator for σ^2 is:

$$s^2 = \hat{\sigma}^2 = \frac{1}{N-2} \sum_i U_i^2$$

This expression is very similar to the variance estimator we developed in Lecture Note 2. First, because the mean of U_i is zero, $\left(U_i - \bar{U}\right)^2 = U_i^2$. Second, in both cases, we divide the sum of squared deviations by the sample size minus the *degrees of freedom* adjustment: the number of parameters we needed to estimate along the way. When we estimate the variance of U_i, we need to estimate both β_0 and β_1, leaving us with $N - 2$ degrees of freedom. In large samples, these adjustments will not matter much; an estimator that used N as the denominator would still be consistent. We call s^2 the *standard error of the regression*.

We are now equipped to derive an asymptotic (large-sample) estimator for the variance of $\hat{\beta}_1$ under the Gauss-Markov assumptions:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{s^2}{N \hat{\sigma}^2_X}$$

As in Lecture Note 2, the central limit theorem guarantees that as the sample size grows, $\hat{\beta}_1$ approaches a normal distribution with mean β_1 and variance $\hat{\sigma}^2_{\hat{\beta}_1}$. The standard error of $\hat{\beta}_1$ is $SE\left(\hat{\beta}_1\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$, and the t-statistic $t = \frac{\hat{\beta}_1 - \beta_1^0}{SE\left(\hat{\beta}_1\right)}$ has a standard normal distribution for the null hypothesis that $\beta_1 = \beta_1^0$. You can estimate the classical OLS model in Stata using `regress` and in R using `lm()` or `fixest::feols()`.

2.2 Normal Linear Model

At this point, we have said nothing about the distribution of $\hat{\beta}_1$ in small samples. We can obtain the *exact* sampling distribution of $\hat{\beta}_1$, holding for any sample size, if we make an additional assumption about the distribution of the error U_i. If, in addition to assumptions (1)-(3) above, U_i is normally distributed, then the t-statistic $t = \frac{\hat{\beta}_1 - \beta_1^0}{SE\left(\hat{\beta}_1\right)}$ is distributed according to a Student's t-distribution with $N-2$ degrees of freedom. We will not derive this result in class, but it can come in handy when the sample size is too small to justify the premise that $N \to \infty$. In fact, Stata and R automatically compute regression p-values and confidence intervals using the $t(N-2)$ distribution. In large samples, the adjustment does not matter.

2.3 Random X's

The preceding results assumed that X_i was non-random, which may strike you as implausible for many situations. Fortunately, if we condition on the X_i's, the results still hold when X_i is random. Thus, the Gauss-Markov assumptions become:

1. $E[U_i|X_1, X_2, \cdots, X_N] = 0$ for all i
2. $V[U_i|X_1, X_2, \cdots, X_N] = \sigma^2$ for all i
3. $cov(U_i, U_j|X_1, X_2, \cdots, X_N) = 0$ for all $i \neq j$

and the normality assumption in the normal linear model becomes $U_i|X_1, X_2, \cdots, X_N \sim \mathcal{N}(0, \sigma^2)$. The results also become conditional on the X_i's: for instance, $E[\hat{\beta}_1|X_1, \cdots, X_N]$ and $V[\hat{\beta}_1|X_1, \cdots, X_N]$. But aside from this more cumbersome notation, we introduce no new complications by allowing X_i to be random.

2.4 Allowing for Heteroskedasticity

In practice, researchers rarely have good justification for assuming that the error terms have constant variance, i.e. that the errors are *homoskedastic*. Fortunately, an alternative estimator of the standard error of $\hat{\beta}_1$ allows for heteroskedasticity: $V[U_i|X_i]$ may vary with i. This is the workhorse regression method of empirical work in economics and other quantitative social sciences. For this method, we make three alternative assumptions:

1. $E[U_i|X_i] = 0$
2. (X_i, Y_i) are independently and identically distributed (i.i.d.) draws from their joint distribution.

3. X_i and Y_i have non-zero finite fourth moments, meaning that large outliers are unlikely.

Assumption (1) is identical to before. Assumption (2) is a slightly stricter and more intuitive version of our previous assumption that $cov(U_i, U_j|X_i) = 0$ for all $i \neq j$. The covariance assumption is a statistical condition that bears no obvious relation to the real world. In contrast, the i.i.d. assumption basically implies that the observations in our sample are unconnected but comparable. Assumption (3) is a technicality that is necessary for proofs we won't do in this class. The basic idea is that if extreme outliers are likely, $\hat{\beta}_1$ will not converge in distribution to a normal distribution.

Note that we have left Gauss-Markov world. Our allowance for heteroskedasticity implies that the OLS estimator may not be efficient. It is still a "linear unbiased estimator," but it is not generally "best." Nevertheless, we would still like to estimate its variance.

Let's rewrite equation (1) conditional on the X_i's:

$$V\left[\hat{\beta}_1|X_1, X_2, \cdots, X_N\right] = \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \left[\sum_i \left(X_i - \bar{X}\right)^2 V\left[U_i\right] + \sum_i \sum_{j \neq i} \left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right) cov\left(U_i, U_j\right)\right]$$

Independence of observations still implies that the covariance terms are zero, but $V\left[U_i\right]$ is no longer guaranteed to equal a constant σ. Thus, we need to generalize our estimator for the variance of $\hat{\beta}_1$ to allow each U_i to have a different variance. This *heteroskedasticity-robust estimator* for $V\left[\hat{\beta}_1|X_1, X_2, \cdots, X_N\right]$ is:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{N} \frac{\frac{1}{N-2}\sum_i \left(X_i - \bar{X}\right)^2 U_i^2}{\left[\frac{1}{N}\sum_i (X_i - \bar{X})^2\right]^2}$$

and $SE\left(\hat{\beta}_1|X_1, X_2, \cdots, X_N\right) = \sqrt{\hat{\sigma}^2_{\hat{\beta}_1}}$. Note that if U_i *are* homoskedastic, this estimator reduces to the classical estimator we derived in Section 3.3. So if the Gauss-Markov assumptions hold exactly, we'll still get the correct result, even if we use this alternative estimator that does not impose homoskedasticity. In Stata, implementation is quite easy, with the `robust` option for the `regress` command. In R, you cannot directly obtain heteroskedasticity-robust standard errors using `lm()`, so we will use `fixest::feols()` with the option `vcov = 'hetero'`.

2.5 Allowing for Dependence

In all of the above approaches to OLS, we have assumed that the error terms were either mutually independent or mutually uncorrelated. Here, we will discuss inference under weaker independence assumptions. This tool is useful in two common situations, both of which involve violations of the independence assumption:

1. The sample has a clustered design. For example, sometimes a survey randomly samples entire house-

holds instead of individuals. In this case, the households can be treated as mutually independent, but individuals from the same households cannot. Another example is a survey that randomly samples individuals but follows them over time. In this case, the individuals can be treated as mutually independent, but observations from the same individual in different time periods cannot.

2. A treatment is assigned at the group level. For example, many education experiments vary treatments (e.g., a teacher training program) at the classroom level. In this case, the classes can be treated as mutually independent, but the students within them cannot.

Again recall Equation (3):

$$V\left[\hat{\beta}_1\right] = \left(\frac{1}{\sum_i (X_i - \bar{X})^2}\right)^2 \left[\sum_i \left(X_i - \bar{X}\right)^2 V\left[U_i\right] + \sum_i \sum_{j \neq i} \left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right) cov\left(U_i, U_j\right)\right]$$

In Section 2.4, we maintained the assumption that the covariance terms are zero, but we allowed $V[U_i]$ to differ across i. Now we will also allow some of the covariance terms to be non-zero too. Specifically, we will allow observation i's error to be correlated with the errors of any other observations in i's cluster, but we still assume that observation i's error is uncorrelated with the errors of any observations outside i's cluster. Because we are allowing for violations of Gauss-Markov assumptions (2)-(3) (homoskedasticity and independence), we can no longer expect that the OLS estimator is efficient. But we still want to know its variance.

The *heteroskedasticity- and cluster-robust estimator* of the variance of $\hat{\beta}_1$ is:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{N} \frac{\frac{1}{N-2} \sum_i^N \left\{\left(X_i - \bar{X}\right)^2 U_i^2 + \sum_{j \in \mathcal{J}_i} \left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right) U_i U_j\right\}}{\left[\frac{1}{N} \sum_i^N (X_i - \bar{X})^2\right]^2}$$

where $\mathcal{J}_i$ is the set of all other observations in i's cluster. Compare this expression with the expression under heteroskedasticity and i.i.d., and see if you can spot the similarities. The cluster-robust standard error can be larger or smaller than the standard robust standard error. If the error terms from observations within each cluster are positively correlated, the cluster-robust standard error will be larger, and if they are negatively correlated, it will be smaller. Researchers tend to be more concerned about a positive intra-cluster correlation, which if ignored would lead us to overstate the significance of our results. In Stata, you can cluster standard errors on a variable `clustvar` by typing `cluster(clustvar)` after the comma in the `regress` command. In R, you can use `fixest::feols()` with the option `vcov = ~clustvar`.

If the X_i's are constant within cluster each i, then an alternative to the cluster-robust standard error is to take cluster-level averages of Y_i and then to run a cluster-level regression. If we index the clusters by $k = 1, 2, \cdots, K$, then we can represent this cluster-level regression by: $\bar{Y}_k = \beta_0 + \beta_1 X_k + \bar{U}_k$. In this regression,

as in the cluster-robust case of the individual-level regression, we are only imposing independence across k. This approach typically yields conservative p-values, however, because it effectively assumes that we get *no* independent information from each i within cluster k.

3 Weighting for Efficiency

Recall the weighted least squares (WLS) estimator from Lecture Note 3:

$$\hat{\beta}_1^{WLS} = \frac{\sum_i w_i(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_i w_i(X_i - \bar{X})^2}$$

In Lecture Note 3, we applied this estimator to try to restore representativeness in an unequal probability sample.

Another use of weighted least squares is to reduce an estimator's variance (leading to smaller standard errors). This application is useful when the errors are heteroskedastic, and the form of heteroskedasticity is known. As noted above, heteroskedasticity is a violation of the Gauss-Markov assumptions, implying that the classical OLS estimator may not be efficient (or, in the language of Gauss and Markov, "best"). In this case, heteroskedasticity-robust standard errors are correct, in the sense that they consistently estimate the square root of the estimator's variance, but other regression estimators may have smaller variance.

To restore efficiency, we can weight by the inverse of each observation's residual variance, $w_i = \frac{1}{V[U_i]}$, which gives more weight to observations with smaller variance. The weighting and heteroskedasticity exactly offset each other, so that homoskedasticity is effectively restored, and the weighted least squares estimator is the best linear unbiased estimator. Of course, knowledge of $V[U_i]$ is rare, so opportunities for this application of weighted least squares are rare as well.

An example of known heteroskedasticity occurs when individuals i are in groups g, each with group size N_g. (The groups might be cities, states, countries, classrooms, etc.) Suppose that the individual-level model is $Y_{ig} = \beta_0 + \beta_1 X_{ig} + U_{ig}$, with $V[U_{ig}|X_{ig}] = \sigma^2$. So the individual-level model is homoskedastic. But suppose also that we only observe group level averages, $\bar{Y}_g = \frac{1}{N_g}\sum_{i=1}^{N_g} Y_{ig}$ and $\bar{X}_g = \frac{1}{N_g}\sum_{i=1}^{N_g} X_{ig}$, so that we can only run the group-level regression $\bar{Y}_g = \beta_0 + \beta_1 \bar{X}_g + \bar{U}_g$. This group-level regression is similar to the regression at the end of Section 2.5 above. Note that $V[\bar{U}_g|\bar{X}_g] = \frac{\sigma^2}{N_g}$. If group sizes vary across groups, then by using group averages, we have introduced heteroskedasticity. If we compute standard errors assuming homoskedasticity, our standard errors will be incorrect. We can fix this problem by computing heteroskedasticity-robust standard errors, and these will correctly capture the variance an our regression estimator, but our regression estimator is inefficient since we no longer meet the Gauss-Markov assumptions. If we weight the regression using $w_g = \frac{1}{V[\bar{U}_g|\bar{X}_g]} = \frac{N_g}{\sigma^2}$, we restore homoskedasticity and return to the efficient world of Gauss-Markov. In fact,

since the denominator of w_g is the same for all g, we can just weight groups by N_g. This weighting scheme is intuitive; we give more weight to groups with more observations and therefore more precisely-estimated means.

Let's take stock of the two reasons we have used WLS: *efficiency* and *representativeness*. We use WLS for efficiency when we have a known form of heteroskedasticity. In this case, WLS reduces the variance of our estimator but should not change its expectation or probability limit. We use WLS for representativeness when we are working with data from a survey with a complex design. In this case, WLS may change the expectation or probability limit of our estimator.