LECTURE NOTE 5: MULTIVARIATE MODELS

1 Introduction

This lecture note expands on the previous by exploring linear regression with multiple covariates. After reviewing the basics of multiple regression, we will discuss omitted variables bias and standard errors for functions of estimators, for example $\hat{\beta}_2 - \hat{\beta}_1$ or $\hat{\beta}_2/\hat{\beta}_1$.

2 Review of Multiple Regression

Often we will be interested in the regression of Y on many X's:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + U + \cdots + \beta_K X_K + U \\ &= \beta_0 + \textstyle\sum_{k=1}^{K} \beta_k X_k + U \end{aligned} \tag{1}$$

We say that β_k represents the relationship between X_k and Y, "controlling for" (or "conditional on," or "holding constant") all other $X_{\tilde{k}}$ with $\tilde{k} \neq k$.

In the sample, we observe $(Y_i, X_{1i}, \cdots, X_{Ki})_{i=1}^N$. The least squares minimization problem is:

$$\min_{\hat{b}_0, \cdots, \hat{b}_K} \sum_{i=1}^{N} \left(Y_i - \hat{b}_0 - \hat{b}_1 X_{1i} - \cdots - \hat{b}_K X_{Ki} \right)^2$$

For this lecture note, we will assume i.i.d. observations with heteroskedastic errors. We take the three assumptions from Lecture Note 2 and add one more to accommodate multiple regressors.

1. $E[U_i | X_{1i}, \cdots, X_{Ki}] = 0$
2. $(Y_i, X_{1i}, \cdots, X_{Ki})$ are independently and identically distributed (i.i.d.) draws from their joint distribution.
3. $X_{1i}, \cdots, X_{Ki}$ and Y_i have non-zero finite fourth moments.
4. No perfect multicollinearity. This assumption is equivalent to saying that for any $k \leq K$, the regression of any X_k on all the other X variables has an R^2 of less than 1.

Assumption (4) is new, and it implies that when the X_k's are "too" correlated with each other, we cannot estimate the regression.

3 Omitted Variables Bias

Suppose the "true" model of Y is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \qquad \text{(long regression)}$$

But we only observe (Y, X_1), and as a result, we can only estimate the following regression:

$$Y = \alpha_0 + \alpha_1 X_1 + V \qquad \text{(short regression)}$$

We would like to estimate the "true" coefficient β_1, but we can only estimate α_1. Under what conditions are they the same? To answer this question, we define a third regression:

$$X_2 = \gamma_0 + \gamma_1 X_1 + \varepsilon \qquad \text{(auxiliary regression)}$$

In all three cases, we construct the errors such that they have mean zero and are uncorrelated with the X_k's in the same regression. Now we derive how α_1 in the short regression relates to the parameters in the long and auxiliary regressions:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U \\ &= \beta_0 + \beta_1 X_1 + \beta_2(\gamma_0 + \gamma_1 X_1 + \varepsilon) + U \\ &= \underbrace{(\beta_0 + \beta_2\gamma_0)}_{\alpha_0} + \underbrace{(\beta_1 + \beta_2\gamma_1)}_{\alpha_1} X_1 + (U + \beta_2\varepsilon) \end{aligned}$$

Thus, $\alpha_1 = \beta_1 + \beta_2\gamma_1$. For α_1 to equal β_1, either β_2 or γ_1 must equal zero. If $\beta_2 = 0$, then X_2 has no predictive power (conditional on X_1), so the short regression is equivalent to the long regression. Meanwhile, $\gamma_1 = 0$ is equivalent to saying that X_1 and X_2 are uncorrelated. Thus, even if $\beta_2 \neq 0$, the OLS estimator $\hat{\alpha}_1$ is an unbiased and consistent estimator of β_1 if X_1 and X_2 are uncorrelated. If X_1 and X_2 are correlated, however, then $\hat{\alpha}_1$ is biased and inconsistent. This result is known as omitted variables bias (OVB).

One can view OVB through the lens of the error in the short regression. We can write $V = \beta_2 X_2 + U$. If $\beta_2 \neq 0$ and $\gamma_1 \neq 0$, then V is correlated with X_1, which implies a violation of least squares assumption (1).

It is useful to distinguish between two types of control variables, *confounders* and *mediators*. A *causal graph*, or *directed acyclic graph*, clarifies this distinction. For starters, assume that $\beta_2 = 0$, so that X_2 is not

a relevant omitted variable. That assumption leads to the canonical causal graph for OLS:

$$\begin{array}{ccc} X_1 & \rightarrow & Y \\ & & \uparrow \\ & & U \end{array}$$

Here, X_1 and the unobservables in U both affect Y, but X_1 and U are unrelated. The regression coefficient is unbiased, and we can interpret it as a causal effect.

Enter X_2. How we treat this variable depends on whether it is a *confounder* or a *mediator*.

confounder			mediator		
X_1	$\rightarrow$	Y	X_1	$\rightarrow$	Y
$\uparrow$	$\nearrow$	$\uparrow$	$\downarrow$	$\nearrow$	$\uparrow$
X_2		U	X_2		U

When X_2 is a confounder, it affects both X_1 and Y. For example, if X_1 is a dummy for attending college and Y is income at age 30, X_2 could be a measure of pre-college academic ability, like the SAT or GPA. Academic ability might lead people to stay in school longer and earn more. In this case, if we were interested in the effect of college attendance on income, we would want to control for X_2 to avoid conflating the effect of college with the effect of academic ability. In contrast, when X_2 is a mediator, it lies on the causal path between X_1 and Y. In the college example, X_2 might be knowledge of college-level computer science. Going to college increases knowledge of computer science, which in turn raises income. In this case, we would not necessarily want to control for X_2, since it is part of the effect of college on income.

4 Inference on Linear Combinations of Coefficients

We often want to do inference on linear combinations of coefficients, for example $a\beta_1 + b\beta_2$ for some constants a and b. We can use $\hat{\beta}_1$ and $\hat{\beta}_2$ to obtain a consistent estimator for the linear combination: $a\hat{\beta}_1 + b\hat{\beta}_2 \xrightarrow{p} a\beta_1 + b\beta_2$. The formula for the variance of a linear combination of random variables implies:

11:40-45

$$\boxed{V[a\hat{\beta}_1 + b\hat{\beta}_2|\mathbf{X}] = a^2\sigma^2_{\hat{\beta}_1} + b^2\sigma^2_{\hat{\beta}_2} + 2ab\sigma_{\hat{\beta}_1,\hat{\beta}_2}}$$

write out V[\beta]

Here, $\sigma^2_{\hat{\beta}_1}$ and $\sigma^2_{\hat{\beta}_2}$ are familiar quantities: the squared standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. However, $\sigma_{\hat{\beta}_1,\hat{\beta}_2}$ is less familiar: the covariance of the two estimators. Whenever $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimated in the same sample, this covariance may be non-zero. It is zero *only* if $\hat{\beta}_1$ and $\hat{\beta}_2$ are estimated in independent samples.

Stata and R always estimate these covariances, even if the regression output only reports the standard errors for individual coefficients. You can ask Stata and R to use the estimated variances and covariances to estimate the variance of the linear combination. The Stata command is `lincom`, while the R function is `glhm()` from the `multcomp` package. Both return the standard error of the linear combination, of the square root of the variance above.

For a more concrete example, suppose we ran a Mincer-style regression in which Y was log hourly earnings, X_1 was years of education, X_2 was years of potential experience, and X_3 was years of potential experience squared. Suppose further that we were interested in predicting log hourly earnings for a worker with 12 years of education and 5 years of potential experience. The predicted value is $\hat{Y} = \hat{\beta}_0 + 12\hat{\beta}_1 + 5\hat{\beta}_2 + 25\hat{\beta}_3$. Because we are now working with four coefficient estimates instead of two, the variance of this linear combination is more complicated than the one above. But the expression still follows the same logic, summing the variances and two times the covariances:

$$\begin{aligned} V[\hat{\beta}_0 + 12\hat{\beta}_1 + 5\hat{\beta}_2 + 25\hat{\beta}_3|\mathbf{X}] = \sigma^2_{\hat{\beta}_0} + 144\sigma^2_{\hat{\beta}_1} + 25\sigma^2_{\hat{\beta}_2} + 625\sigma^2_{\hat{\beta}_3} \\ + 24\sigma_{\hat{\beta}_0,\hat{\beta}_1} + 10\sigma_{\hat{\beta}_0,\hat{\beta}_2} + 50\sigma_{\hat{\beta}_0,\hat{\beta}_3} \\ + 120\sigma_{\hat{\beta}_1,\hat{\beta}_2} + 600\sigma_{\hat{\beta}_1,\hat{\beta}_3} + 250\sigma_{\hat{\beta}_2,\hat{\beta}_3} \end{aligned}$$

Another common example arises when we want to test whether $\beta_1 = \beta_2$. This test is equivalent to asking whether $\beta_1 - \beta_2 = 0$. So we can estimate $V[\hat{\beta}_1 - \hat{\beta}_2|\mathbf{X}] = \sigma^2_{\hat{\beta}_1} + \sigma^2_{\hat{\beta}_2} - 2\sigma_{\hat{\beta}_1,\hat{\beta}_2}$ and take the square root to obtain the standard error. Then we can form the t-statistic:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\hat{V}[\hat{\beta}_1 - \hat{\beta}_2|\mathbf{X}]}}$$

explain when can set cov=0 and use SE^2

to test the hypothesis that $\beta_1 = \beta_2$.

5 Inference on Non-Linear Functions of Coefficients

What if we want to test hypotheses or construct confidence intervals for a non-linear function of coefficients? For instance, we might want to know about the ratio of two coefficients, e.g. $\frac{\beta_2}{\beta_1}$, or about the inverse of a coefficient, e.g. $\frac{1}{\beta_1}$. We will compute the standard error for a non-linear function of coefficients using the *delta method*. say 1st-order Taylor approximation, then draw graph, then go back to V[a*beta1 + b*beta2]

Suppose we have a multivariate model with parameters $\beta_0, \beta_1, \cdots, \beta_K$ want to know about some nonlinear function of the parameters $g(\beta_0, \beta_1, \cdots, \beta_K)$. We can again use $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K$ to obtain a consistent estimator for the value of $g(\cdot)$: $g(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K) \xrightarrow{p} g(\beta_0, \beta_1, \cdots, \beta_K)$. But how do we estimate the variance of

$g(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K)$? An exact expression for this variance is very difficult to find (and may be impossible). However, we can *approximate* the correct answer by using a tool from calculus, the Taylor approximation. Specifically, we take a first-order Taylor approximation of $g(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_K)$ near the true value $g(\beta_0, \beta_1, \cdots, \beta_K)$. In practice, that means we evaluate the first derivatives of $g(\cdot)$ at $\beta_0, \beta_1, \cdots, \beta_K$ and treat these derivatives as scalars in a linear combination:

$$g(\hat{\beta}_0, \hat{\beta}_1, \cdots) \approx g(\beta_0, \beta_1, \cdots) + g_{\beta_0}(\beta_0, \beta_1, \cdots)\{\hat{\beta}_0 - \beta_0\} + g_{\beta_1}(\beta_0, \beta_1, \cdots)\{\hat{\beta}_1 - \beta_1\} + \cdots$$

where $g_{\beta_k}(\beta_0, \beta_1, \cdots) = \partial g(\beta_0, \beta_1, \cdots)/\partial \beta_k$. Taking the variance of this linear approximation is straightforward; we simply use the formula for the variance of a linear combination. It turns out that as the sample size grows large, the distribution of the linear approximation converges to the distribution of $g(\hat{\beta}_0, \hat{\beta}_1, \cdots)$. The central limit theorem implies that this large-sample distribution is normal, so we can do inference as usual, using standard normal critical values.

This is all very abstract, so let's illustrate the ideas with an example. Suppose we were interested in racial disparities in education, so we ran a regression of years of education on a binary variable for white racial identity:

$$years_i = \beta_0 + \beta_1 white_i + u_i$$

The coefficient β_1 measures the absolute difference in average education between whites and non-whites. But suppose we were interested in the *ratio* of average education between whites and non-whites:

$$\frac{\mu_{white_i=1}}{\mu_{white_i=0}} = \frac{\beta_0 + \beta_1}{\beta_0}$$

The delta method allows us to compute the variance of this ratio. The function g is:

$$g(\beta_0, \beta_1) = \frac{\beta_0 + \beta_1}{\beta_0}$$

The first derivatives of $g(\cdot)$ are:

$$g_{\beta_0}(\beta_0, \beta_1) = -\frac{\beta_1}{\beta_0^2} \quad \text{and} \quad g_{\beta_1}(\beta_0, \beta_1) = \frac{1}{\beta_0}$$

So we can estimate the variance of $g(\hat{\beta}_0, \hat{\beta}_1)$ as:

$$\hat{V}\left[g(\hat{\beta}_0, \hat{\beta}_1)|\mathbf{X}\right] \approx \left(\frac{\hat{\beta}_1}{\hat{\beta}_0^2}\right)^2 \hat{\sigma}^2_{\hat{\beta}_0} + \left(\frac{1}{\hat{\beta}_0}\right)^2 \hat{\sigma}^2_{\hat{\beta}_1} - 2\left(\frac{\hat{\beta}_1}{\hat{\beta}_0^2}\right)\left(\frac{1}{\hat{\beta}_0}\right)\hat{\sigma}_{\hat{\beta}_0,\hat{\beta}_1}$$

$$= \frac{\hat{\beta}_1^2}{\hat{\beta}_0^4}\hat{\sigma}^2_{\hat{\beta}_0} + \frac{1}{\hat{\beta}_0^2}\hat{\sigma}^2_{\hat{\beta}_1} - 2\frac{\hat{\beta}_1}{\hat{\beta}_0^3}\hat{\sigma}_{\hat{\beta}_0,\hat{\beta}_1}$$

As usual, the standard error is the square root of this quantity. You can implement the delta method in Stata using the `nlcom` command and in R using the `deltaMethod()` function from the `car` package. If the function $g(\cdot)$ is linear, then the delta method reduces to the linear combination method discussed in Section 4. So to keep our coding as simple as possible, we will use `deltaMethod()` for both linear and nonlinear combinations.

6 Interactions

Problem Sets 2 and 3 will ask about interaction terms, and students often express confusion. Here we will review how to interpret and work with interaction terms in the context of a specific example. Suppose we are public health researchers interested in the determinants of body mass index (weight/height^2). In a simple random sample of the US population, we regress BMI on age and dummies for male sex, white racial identity, and residence in the South:

$$BMI_i = \beta_0 + \beta_1 age_i + \beta_2 male_i + \beta_3 white_i + \beta_4 south_i + u_i \tag{2}$$

To run this regression in R using `feols()`, we would type:

```
feols(bmi ~ age + male + white + south, data = df, vcov = 'robust')
```

where `df` is the name of the data frame containing the survey data. Suppose we find that the coefficients on `male` and `south` are both positive, but we want to know if the gender gap in BMI is significantly different from the regional gap (i.e., is $\beta_2 = \beta_4$?). That is equivalent to asking whether the coefficient on `male` minus the coefficient on `south` is significantly different from zero (i.e., is $\beta_2 - \beta_4 = 0$?). We can run this test using `deltaMethod()`, since the delta method embeds linear combinations. We assign the name `model1` to the model above and then run `deltaMethod()`.

```
model1 <- feols(bmi ~ age + male + white + south, data = df, vcov = 'robust')
deltaMethod(model1, "male - south")
```

R will then report an estimate, standard error and 95% confidence interval for $\beta_2 - \beta_4$. If we wish to formally test the null hypothesis that $\beta_2 = \beta_4$, we can add:

```
deltaMethod(model1, "male - south", rhs=0)
```

Now suppose we want to know whether the accumulation of BMI with age is different between the South

and non-South. One way to do so is to run separate regressions by South and non-South:

$$BMI_i = \beta_0 + \beta_1^S age_i + \beta_2^S male_i + \beta_3^S white_i + u_i \text{ if } i \text{ lives in the South} \tag{3}$$

$$BMI_i = \beta_0 + \beta_1^N age_i + \beta_2^N male_i + \beta_3^N white_i + u_i \text{ if } i \text{ lives in the non-South} \tag{4}$$

We do not include $south_i$ in the regession because it does not vary within the regional sub-samples. We are interested in $\hat{\beta}_1^S - \hat{\beta}_1^N$. Because we have a simple random sample, the South and non-South are independent subsamples, so we can compute the standard error of $\hat{\beta}_1^S - \hat{\beta}_1^N$ "by hand" as

$$SE\left[\hat{\beta}_1^S - \hat{\beta}_1^N\right] = \sqrt{SE\left[\hat{\beta}_1^S\right]^2 + SE\left[\hat{\beta}_1^N\right]^2} \tag{5}$$

Note that we could **not** take this approach when comparing the coefficients on $male_i$ and $south_i$ above. Those coefficients were estimated from the same sample, so we cannot assume that they have no covariance, and based on the regression output alone, we do not know their covariance. The `lincom` or `nlcom` commands in Stata and the `glhm()` or `deltaMethod()` functions in R access an estimate of the covariance that is not reported with the usual regression output.

An quicker way to assess differences in age slopes between regions is with an interaction term. To distinguish from the β's above, I will use α's to refer to the parameters of the interacted model. If we want to **exactly** match the estimates of equations (2) and (3), we need to interact the $south_i$ dummy with **every** other covariate. That's because equations (2) and (3) allow **all** of the coefficients to vary by region. If we only interacted $south_i$ with age_i, then we would effectively be assuming that the coefficients on $male_i$ and $white_i$ are the same in the South and non-South, i.e. $\beta_2^S = \beta_2^N$ and $\beta_3^S = \beta_3^N$. This assumption would also change the estimated coefficients on age. The fully interacted model is:

$$\begin{aligned} BMI_i =& \alpha_0 + \alpha_1 age_i + \alpha_2 male_i + \alpha_3 white_i + \alpha_4 south_i \\ &+ \alpha_5 south_i \times age_i + \alpha_6 south_i \times male_i + \alpha_7 south_i \times white_i + u_i \end{aligned}$$

To see how the interaction terms work, let's compute the predicted value for someone living in the non-South. For such a person, $south_i = 0$, so $\alpha_4, \alpha_5, \alpha_6$, and α_7 are all multiplied by 0, dropping out:

$$\widehat{BMI}_i^N = \alpha_0 + \alpha_1 age_i + \alpha_2 male_i + \alpha_3 white_i$$

Now let's compute the predicted value for someone living in the South. In this case, $south_i = 1$, so $\alpha_4, \alpha_5, \alpha_6$,

and α_7 all "turn on." We have:

$$\begin{aligned}\widehat{BMI}_i^S =& \alpha_0 + \alpha_1 age_i + \alpha_2 male_i + \alpha_3 white_i + \alpha_4 + \alpha_5 age_i + \alpha_6 male_i + \alpha_7 white_i \\ =& (\alpha_0 + \alpha_4) + (\alpha_1 + \alpha_5) age_i + (\alpha_2 + \alpha_6) male_i + (\alpha_3 + \alpha_7) white_i\end{aligned}$$

where the second line collects terms from the first line. Now we can see that the "main effect" of each covariate (i.e., the uninteracted term) is the same as coefficient from the non-South-only regression, while sum of the the "main effect" and "interaction effect" is the same as the coefficient from the South-only regression. For example, $\beta_1^N = \alpha_1$ and $\beta_1^S = \alpha_1 + \alpha_5$. Thus, $\alpha_5 = \beta_1^S - \beta_1^N$, and the standard error should be the same as the one we computed "by hand" using equation (4).

The preceding paragraphs were all about testing for a **difference** in the age slope between regions. What if we want to know about the **ratio** of slopes? Then we would want to compute:

$$\widehat{\text{ratio}} = \frac{\hat{\alpha}_1 + \hat{\alpha}_5}{\hat{\alpha}_1}$$

To compute the standard error, construct confidence intervals, etc., we can use the delta method. First, define the interaction terms:

```
df <- df %>% mutate(ageXsouth=age*south, maleXsouth=male*south, whiteXsouth=white*south)
```

Then run the model and delta method:

```
model2 <- feols(bmi ~ age + male + white + south + ageXsouth + maleXsouth + whiteXsouth,
                data = df, vcov = 'robust')
deltaMethod(model1, "(age + ageXsouth)/age")
```

R also has syntax for generating interaction terms within the regression function. See the example code.