**References for sessions 1 & 2 (class 1):**

* [AP] Angrist, Joshua and Jorn-Steffen Pischke (2009). Mostly Harmless Econometrics. An Empiricists Companion. Princeton University Press.
* [W] Wooldridge, Jefferey. 2009. Econometrics.

1. ***Causality***
2. Idea of potential outcomes

Consider an example. Do those admitted to hospital services may get valuable services? The answer is likely to be yes, logically. But will the data back this up? A natural approach to use data to show this is to compare the health status of those who have been to hospital and those who have not been. In the National Health Interview Survey (NHIS), individuals self-report health status as excellent (5), very good (4), good (3), fair (2), poor (1). The survey also asks a question, ‘In the last 12 months, have you stayed in a hospital over night? (Yes/No)”. Suppose we simply compare mean health status of those who go to hospital and those who do not.

|  |  |  |  |
| --- | --- | --- | --- |
| Group | Sample Size | Mean health status | Standard error |
| Hospital | 7,774 | 3.21 | 0.014 |
| No hospital | 90,049 | 3.93 | 0.003 |

Source: NHIS 2005

Taken at face value, these suggest that going to hospital makes people sicker. This is not impossible, since hospital is full of sick people and one can get an infection going to a hospital. Still, there is a larger reason why the mean health status of those going to hospital is lower: People going to hospital are less healthy to begin with. Also, even after hospitalization those who went to hospital are not as healthy as those who did not go.

More examples:

1. Effects of on the job training on wages for ex-offenders, drug addicts, and long term unemployed. Studies based on non-experimental comparisons of participants and non-participants often show that after training, the trainees earn less than comparison groups (Ashenfelter 1978, Ashenfelter and Card 1983, Lalonde 1995). Is this regression causal? If not, what is the bias here?

Answer: Those who attended training would have less ability, because of which they may earn less to begin with. In other words, in the potential outcomes framework, the wages when there is no training, is higher among non-participants compared to participants.

1. Effects of education (in years) on wages. Is this regression causal? If not, what is the bias here?

Answer: Those who got more education have higher ability than those who did not get education. Regressions based non observational data tend to overestimate the effects of education on wages.

Let’s say is health status, and is binary variable of going to hospital or not. To obtain the effect of on let’s defined potential outcomes. denote the health status of an individual who did not go to hospital, irrespective whether he/she did. Similarly, denote the health status of an individual who went to hospital, irrespective whether he/she did.

Causal effect for each person i is simply:.

Simple average difference in observational data:

1. First term is the treatment effect on the treated.
2. Second term is selection bias. It shows how if not treated, the outcome of those who end up getting treatment and those who do not are different. Compare this to: how health status if not going to the hospital are so different between those who *actually* go to hospital and those *actually* who do not.

How do we eliminate the selection bias? What if are independent. That is, going to hospital is independent of someone’s health status. Then from the above equation we get that,

E

This implies that if we randomize treatment, we can take simple differences of outcomes between treatment and control status, and obtain causal effects.

1. ***Regression fundamentals***
2. Simple regression terminologies: Let’s stick with scalar notation because the intuition is more explicit that way.

Dependent variable, independent variables, error term, slope parameter, intercept parameter. Show example regressions: yield on fertilizer, and wage on education (W, p32 and 33). What are some of the factors in u, in each of the cases?

Terminology – “run the regression of y on x”. When we say this, we always mean to estimate the parameters and in the equation – essentially the constant and the slope coefficient.

1. *Ceteris paribus* definition (equation 2.2 in w, p33).

means keeping constant everything in u. In reality, it is very hard to hold all factors in u constant because we do not even know what is in u. But if we are able to hold everything else constant, we can obtain causal effects.

Some assumptions about u.

1. E(u)=0. This is not restrictive since. You can always rearrange the constant term so that E (u)=0 (AP, p22)

Where the new error term v = , and new redefined constant term is

1. Take another assumption. E (u|x) = E(u). That is, the expected value of u does not depend on the value of x. this assumption is restrictive. Given the assumption in 3, this means the zero conditional mean assumption. E (u|x) = E(u)=0. What does E (u|x) = E(u) mean in practice?

Take the wage on education regression. In this let u be ability. Can we say that E (u|8) = E (u|16)? In reality, ability levels may be higher for those better educated, or those with higher ability may choose to get more education (W p35).

More examples: (1) . Does this satisfy assumption E (u|x) = E(u)? What is correlation between u and attendance? (2) . What are the factors in u? Are they likely correlated with education? What is the sign? (3) Suppose you are interested in measuring the effect of hours of preparation on cat score. You got a grant to implement am RCT. How will you randomize this? Suppose you do not have that grant but only have observational data. Then if you run the following regression:. What are the factors in u? Do they have positive or negative correlation with hours?

1. Population regression function (PRF)

, because E(u)=0

Here, one unit increase in x increases *expected value* of y by (REMEMER, not y *itself*). Given E(u) =0, we can redefine as y = , or systematic part non-systematic-part. (see W, figure 2.1 p36)

1. ***OLS estimator ()***
2. What is “least square”-estimator intuitively?

Choosing and such that we minimizing the sum of square of residuals .In other words, should be as small as possible. If you solve for the first order conditions for this with respect to and , you get two equations in two unknowns and you can obtain and . Figure 2.4 in W, p40 will be useful.

1. Formula for beta

Simple regression formulas

-

Note that the denominator has to be positive for to be defined. If you divide numerator and denominator by n-1, is simply the ratio of sample covariance between x and y divided by sample variance of x.

Matrix notation:

. The first order condition is 2\* E()) = 0. Solving the first order condition, you get

1. Recall the assumptions of OLS (Wooldridge p52)
2. Model is linear in parameters (y=ax+b is linear in parameters, but y=xa/zb is not; Note logy=a logx+b or y = a + b are still linear in parameters.)
3. The data are a random sample of the population (i=1…n, random observations)
4. Sample variation in explanatory variable (V(x) ≠ 0)
5. Zero conditional mean, E(u|x)=E(u)=0
6. The residuals have constant variance (Homoskedasticity). V() = for all i.

Assumptions i- iv make the OLS estimator unbiased; v makes it have lowest variance is unbiased if i-iv is satisfied.

1. Note that =

Numerator is

In this, note that first term is 0, and second term is equivalent to . Numerator becomes,

E () = () = (second term is zero, E(u|x)=0 implies E(ux)=0)

Show the regression of lunchprg on math10 (data from MEAP93, CD in W). lunchprg denotes the percentage of students in the school eligible for a school lunch program. math10 is the proportion of tenth graders obtaining a passing score in standardized math exam. This simple regression produces a negative sign on lunchprg. Does that make sense? Or could it be that biased? What can be in u that is correlated with x that is causing this bias? Answer: Biased because u contains school and student level characteristics, such as poverty rate which is correlated with lunch program eligibility.

1. Variance of with homoskedasticity (for single equation non-matrix proof, see W, p59).

V () = V (

V () = =

Intuition: The reduction of , the error variance can reduce ’s variance. In regressions, adding control variables which have explanatory power reduces and hence V(. The denominator indicates that if the variation in x is low, the variance of is high.

1. How to estimate error, u and error variance, V(u) in practice? u cannot be observed, but residual - its estimate -, can be observed () (W, p60). .

cannot be observed in practice. But can be. We cannot estimate this as because it’s biased. The formula for unbiased variance, based on the loss of two degrees of freedom given in W, p61.

1. The standard error of uses the formula for (W, p61-62)

se () =

1. R2 and sum of squares formulas:

Given that ; we can rewrite this as (for proof, see W, p46):

Which is Total sum of squares (TSS) = explained sum of squares (SSE) + residual sum of squares (SSR)

R2 = SSE/TSS

R2 tells us what proportion of the dependent variable is explained by independent variables.

1. ***Multiple linear regression model***
2. For a *ceteris paribus* interpretation, the more variables we control for, the better. Multiple regressions also have more predictive power. For these reasons, we often run multiple regression models. But there is a limit to adding variables, because variance of β may get affected. We will see this soon.
3. How can we get a ceteris paribus effect here (W, p73). Total differentiation of y gives,

When x2 is held constant (ceteris paribus),

Key point here is that, we are able to control for x2 (in other words, maintain x2 constant), when we want to examine the effect of change in x1 on change in y.

What does that mean intuitively? Consider a model where wage depends on education and experience. Suppose, we do not add experience in the model (may be because we do not measure it).

Here does not provide the ceteris paribus effect because experience also affects wages and is sitting in the error term u. We cannot hold experience constant even though experience changes as education changes. But if we add that as a separate variable, we can effectively hold that constant and obtain ceteris paribus effect.

We have university student level data (GPA1.dta from W CD): *colGPA* is college Grade Point Average. *hsGPA* is high school Grade Point Average. *ACT* is 'achievement' score. Regression results of colGPA on hsGPA and ACT gives us the following relationship.

1. How do you interpret the coefficients in the first regression?
2. What does the difference in the coefficient of ACT between first and second regression signify?
3. Consider a multiple regression with k variables ,

OLS estimator in multiple regression is: (equation 3.1.3 in AP p35)

is the residual from the regression of on all other explanatory variables. This formula is much more intuitive than the matrix formula. It tells us that each slope coefficient in a multivariate model is the slope coefficient of a bivatiate model on that corresponding regressor after partialing out all other covariates. A theorem proves this – called the Frisch–Waugh–Lovell (FWL) theorem, named after the econometricians who worked on it.

Proof: Substitute into the above equation to see that the right hand side becomes . We will see many examples based on this theorem in the hands-on class.

1. Relationship between simple and multiple regression coefficients. Suppose we have a simple regression model as below:

Consider a multiple regression model below:

Relationship between variables in simple regression and multiple regression models W, p76).

where is slope coefficient of on . = are equal if or if and ) are uncorrelated in the sample.

Example: Load GPA1.dta. Compare the following results with the regression of on and ACT (earlier)

Correlation between ACT and hsGPA is 0.346 which is non-trivial (meaning, is high). But the coefficient on ACT is low (0.0094), which is probably why 0.482 is not that different from the multiple regression coefficient of hsGPA, which is 0.453.

INTUITION AND PRACTICAL RULES: From the above formula, what are the consequences of including an irreverent variable?

Let’s say a model has two explanatory variables x1 and x2:

Now we add x3.

How does including x3 in the model, which is irrelevant () affect other coefficients? From point 4, we know that this does not really affect the value of or . But note that it will affect their variance (W, p84) – we will see this below.

1. Omitted variable bias: What are the consequences of omitting a relevant variable in a model (say x3), which has x1 and x2?

Say, the “true” model is:

+

(this is parallel to where we included both x1 and x2.)

But we somehow end up estimating the model below, excluding ability:

+

(this is parallel to where we included only x1)

What are the consequences for if we miss x2?

We already know that . Take expectation on both sides. E( = E(.

Here, because they are coefficients from the full regression including x1 and x2.

Consequently, E( = ; Bias = E( - = . Because this bias arises from an omitted variable (x2), it is called the omitted variable bias.

There are two reasons why where there in bias.

1. First, ≠0; that is, x2 significantly explains y.
2. Second, x1 and x2 are correlated.

What is the sign of the bias? It depends on the sign of both and x2.

|  |  |  |
| --- | --- | --- |
|  | Corr(x1,x2) > 0 | Corr(x1,x2) < 0 |
| >0 | Positive bias | Negative bias |
| <0 | Negative bias | Positive bias |

INTUITION AND PRACTICAL RULES:Now let’s connect this table to examples we saw already.

1. Effect of education on wages (W p86) – suppose we miss adding ability. Here >0 and Corr(x1,x2) > 0. Those with higher ability are more likely to get higher years of education. Bias is positive, so we overestimate the effects of education on wages if we do not include ability.
2. Effect of “training” on wages (AP p16) - here again, suppose we miss adding ability to the model. Here >0 and Corr(x1,x2) < 0. Those with lower ability are more likely to enrol in training. Bias is negative, so we are likely to underestimate the effects of training on wages, which is what non-experimental studies found.
3. Variance of OLS estimator (W p89, equation 3.51).

is the from regressing on all other x-variables. is the total sampling variation in . .

1. Large error variances increases . We saw this before in the simple regression model as well.
2. Multi-collinearity can decrease precision. That is, higher the , higher the .

INTUITION AND PRACTICAL RULES: Consider an important scenario here. Suppose our model contains x1, x2, and x3. x2 and x3 are highly correlated. But x1 is not correlated to either x2 or x3. Remember then that the variance of is not affected because . You may actually benefit from adding x2 and x3 if it has explanatory power, and you only care about .

1. Variance versus bias.

Suppose you have an extra explanatory variable you are considering add to your model. The choice of whether you want to include that depends on the trade-off between bias and variance. Suppose the true model is (model 1):

Say we leave out x2 (model 2):

Then, if is clearly biased. So if bias is the criterion, we prefer over . Now let’s bring variance into the picture. . Unless x1 and x2 are uncorrelated, . is less precise compared to .

INTUITION AND SOME PRACTICAL RULES

If x1 and x2 are uncorrelated

If , definitely worthwhile including x2 since this reduces error variance (without causing multicollinearity).

If x1 and x2 are correlated

1. If , and are unbiased. . Prefer model 2.
2. If , is biased and is unbiased. In terms of variance of rror variance () is now lower because of better explanatory power, but the denominator also increases because of multi-collinearity ( >0 ); we do not know which effect dominates. Unclear about . Not clear indication of whether we prefer model 1 or 2. But since variance decreases with large sample size, we would benefit from adding x2 if sample size is large.

TILL NOW, END OF CLASS 1 ON 20TH DECEMBER 2016

1. Standard error of OLS estimator is based on residual variance, which is below:

1. If we assume that u is (i) normally distributed with (ii) mean 0 and (iii) variance . Note that we already assumed (ii) and (iii) before.

u ~ N(0,)

Then we can show that,

Subtracting mean and dividing by standard deviation, gives us a standard normal distribution:

Replacing σ by in , gives . This gives a t-distribution rather than Normal:

Definition of t-value to test H0: against H1: (two-sided test):

We do not reject the null hypothesis if the calculated , the critical value chosen based on the level of significance (α). α is commonly chosen as 0.05 and the critical values are read off the t-table based on the level of significance.

INTUITION AND PRACTICAL RULES: The language of hypothesis testing needs to be cautious (W, p118). We prefer to use the language “we fail to reject H0 at the x% level”, rather than saying “we accept H0”. This makes sense because our both H0=-1 and H0=-0.95, can be “accepted” statistically. Clearly both cannot be true.

INTUITION AND PRACTICAL RULES: It is important to distinguish between economic significance and statistical significance (W, p118). could be significantly different from zero (a) because its value is high, or (b) because the se is low. We need to distinguish these two, and understand based on the context whether the magnitude of coefficient is meaningfully large at all.

1. Interpretation of coefficients in log functional forms (W, p158)

log y =

is the % change in y when there is a % change in x.

1. Model with quadratic variables (p161)

y =

1. Model with interaction terms (W, p164)

y =

1. INTUITION AND PRACTICAL RULES: How much to value R2. Is the regression valuable if R2 is low? Remember nothing in a OLS regression model says that R2 should be above a particular value. Low R2 is does not imply that x and factors in u are correlated. Remember, an Randomized Control Trial(RCT) can produce unbiased estimates of treatment but can still give a low R2. So if your aim is unbiased estimation of a parameter (in answering a specific question), rather than model predictions, do not worry about R2 being low.

In APPLA.dta, the key explanatory variables were set experimentally. The outcome variable is “number of ecologically friendly apples demanded” (*ecolbs*). Each family were randomly assigned a price pair - price of regular apple (*regprc*) and price of ecolabel apples (*ecoprc*). The regression of *ecolbs* on *regprc* and *regprc* give unbiased estimates. But look at the R2 from this regression. .0364, which is very low.