Final Exam

Econometrics

Elizabeth Goodwin

12/13/2021

1

The R^2 of a linear regression measures the amount the variance of the dependent variable that can be explained by the independent variable. The importance of this estimator can vary significantly. An example where it can be important is in a well designed RCT, where endogeneity is mostly eliminated. If we isolate the causal impact of the independent variable, the R^2 can represent the amount of variance applying the treatment can explain. An example where it would not matter is a poorly designed correlation. If we, say, measure the college wage premium as a simple correlation, the R^2 is less useful as a confounding variable, such as innate ability, may actually explain said change in variation.

2

a

If we used the lasso instead of OLS, we would no longer have unbiased coefficient estimates. The lasso is designed to trade biased coefficients for lower variation, and is primarily useful for use cases where you wish to most accurately predict $\hat{Y}$. The lasso intentionally biases the coeffients to achieve this.

b

It would be justified to use the lasso in situations where the primary objective is prediction. A good example of this would be something such as credit default risk, where the goal of the model is simply to predict who is likely to default. If you want to know why people default, the model would be far less informative. For simple prediction problems, however, trading unbiasedness for lower variance may be worth it. Examples where you are trying to explain the causal reasons why something happened, for example, should not be using the lasso. If you are trying to explain the causal effect of going to college on wages, you are looking to figure out $\hat{B}$, or the impact of college. Adding bias to the model would destroy your ability to do this type of causal inference, and the use of the lasso would be unjustified.

3

Selection bias is when the composition of groups being studied is nonrandom and distorts the results. Going back to examples from before, if harder working or more intelligent people choose to go to college, and earn higher wages, that may not mean that college caused those higher wages, as they may have earned those higher wages either way. This means that NATE differs from ATE, as the simple difference in wages

between those with and without a college degree is significant, but does not represent the average effect of the treatment. The decision to going college is not random, and without that randomization of who receives the treatment, it is difficult to tell if the treatment explains the average difference. For ATE and NATE to be the same, the untreated group must actually represent the unobserved counter factual of if those going to college chose not to go, and vice versa. If the groups differ in relevant ways, that may not be true.

4

a

In this situation, the local average treatment effect would only measure a very specific subgroup of people who would not have gone to college without being selected by the lottery, but would if being selected. This does not include those who would not go either way, so those with GPA/SAT scores below the cuttoff, those who do not wish to go to college, or those who cannot for any other reason. The instrument would not effect college attendance upon this group. Another group it would not observe would be those who would go to college either way. So this includes the group with "exceptional merit", who gets accepted without the randomization, and groups that would have gone to a a private university or a university in a different state even if they were not selected at random. The last group that it does not measure is defiers, or those who would go if not selected, but would go if not selected. This group also biases the end result, and must be assumed to not exist. So in total, it measures the effect of college upon wages for those who have a GPA above the cutoff, would not go to college otherwise, and when chosen would go to college.

b

If this was an RDD, it would be a fuzzy RDD. This is because not all of those chosen would go, so it cannot represent a jump of 0 to 1. Some of those chosen at random would have gone to college either way, and some may change thier mind after being accepted as well. It would only jump the proportion of people going, and and therefore be fuzzy. For it to be sharp, it would have to change it from 0 to 1. If there were no noncompliers, and everyone selected would go to definitely going, it would be sharp.

c

In this context, a few things would have to be true for RDD to work. Most important is the continuity assumption, or essentially that without the treatment, the outcomes would not have jumped. As we cannot observe the counterfactual reality of those past the cutoff without the treatment or those before the cutoff with the treatment, we must assume that with the jump removed, the trend would be continuous. Continuity removes omitted variable bias, as it assumes that the treatment is what is causing the cutoff. Another assumption is that the values around the cutoff must not be manipulated, or essentially changed in because knowledge of its existence. Also, the people at either side of the cutoff inside the bandwidth are assumed to be exchangeable

d.

In the context of this question, this means the following. First, it means that without being treated, the jump in wages around the cutoff would not exist. Essentially, the treatment must be what is causing the change. Very importantly for this case, there must not be manipulation. If the exact GPA or SAT cutoff is public, hard working people (that may naturally earn more anyways) that are right below the cutoff might intentionally work harder than otherwise to marginally improve their score to just above the cutoff. In addition, whose who know they are above the cutoff may work less hark knowing marginal improvement above that score may not effect their chances. If this is correlated with income, it could distort the results.

This also impacts the changeability criteria, as those right below and above the cutoff would not be the same in the case of manipulation, as the manipulators may be systematically different in ways that impact future earnings than those right below the cutoff. To test for this, you could run a density test. Essentially, if the distribution of people around the cutoff breaks from a normal distribution, that could be a sign of manipulation. Without manipulation, you would expect the density of those before and after to follow the existing distribution, but if that is not the case around the break, that may be a sign of bias.

5

To answer the question of the impact of the ADA on wages, you can use a differences in differences approach. To test for this, we first have to test the trends of existing data. The theory is, essentially, that disabled and non disabled workers, even if being employed at different rates, would follow the same employment trend. To put this in a regression, you could do something like this. This is a regression with the dependent variable being Employment Rate, a constant term(for both disabled and non disabled), and the other terms being interaction terms between employment rate trends and an indicator for the year. The most important terms, however, are the interaction between year and disability status. The coefficient there represents the difference per year in the employment rate between those disabled and those not. The beta hats with multiple numbers represent the coefficients for each year, as it interacts a year indicator with each beta hat. Disability would be a dummy variable with 1 = disabled and 0 meaning not disabled.

$$\widehat{EmpRate} = \hat{\beta}_0 + \hat{\beta}_1 D + \hat{\beta}_{2-22} \times Year + \hat{\beta}_{22-42} D \times Year$$

You can also do the same thing, but for income as well. You would want to log income though, to show a percent change in income instead of an absolute one. $\hat{\beta}_{22-42}$ represents the difference in percent income per year between disabled and nondisabled workers.

$$\widehat{LogWage} = \hat{\beta}_0 + \hat{\beta}_1 D + \hat{\beta}_{2-22} \times Year + \hat{\beta}_{22-42} D \times Year$$

For this DiD model to be valid, the parallel trends assumption must not be violated. You can attempt to justify this assumption by looking at previous trends. If the law was not in place before 1992, you can look at the coefficients of the interaction term between disability status and year in both regressions before the law was announced. If the coefficient is statistically significantly different before this point, the trends are different. You also may want to look specifically at the years right before the law, as if they started to diverge before the law was passed, it could indicate a different underlying trend in the data creating a different trend. If the interaction between those years and disability is not significantly different from 0, there is no significant difference in trends beforehand that we can measure. We cannot prove anything for certain, but we are unable to disprove it, which is the best we can really do given we can never know the counterfactual.

6

a

The bias you would be seeing in this case is omitted variable bias. Essentially, this means your model has endogeneity issues, and there is likely a missing variable from the model. If those relevant variables were added in, the correlation would not exist. It missing means that the model is not accurately showing the effect of x on y, but the effect of x on y *and* the indirect effect of the omitted variable on y, through x.

b

i.

In this situation, Z passes the relevance test, but not the exogeneity test. As Z is also correlated with the error term, it is facing the same problems as what it is trying to measure, and is an inproper instrument.

ii.

The term describing the bias is endogeneity. You may also say it does not pass the exclusion criteria, or the failure of exogeneity.

iii.

The OLS estimator is less biased than that of the IV estimator if a weak instrument magnifies the bias to a level above the bias of OLS. Essentially the equation below is the bias of 2SLS in this situation. If it is greater than the bias of the OLS variable, as shown below, 2SLS is more biased than OLS. Weak instruments mean a low covariance in the denominator, which magnifies the bias significantly.

$$\frac{Cov(Z,\epsilon)}{Cov(Z,X)} > Cov(X,\epsilon)$$

iv.

For a fixed non-zero value of $Cov(Z,\epsilon)$, weak instruments can make the problem of bias very bad. One issue of weak instruments is simply that it makes the error larger in general, as it will by nature magnify the standard error. In addition to this, the bias being magnified can even make a less biased instrument bias 2SLS more than normal OLS. This means you must be able to justify the exogeneity criteria quite well for weak instruments, as any failure in exogeneity can be very bad.

7

This argument is simply incorrect. The primary reason is it misses the entire point of instruments. If Z is correlated with X, and X is correlated with Y, you would expect Z to be correlated with Y as well. The point of an instrument is to use an instrument to measure the effect of X on Y through Z, to avoid endogeneity issues of using X itself. If you add X and Z together in one regression, so conditioning on X, you are not just conditioning on X, but also the unobserved variable U that is mediated through X. If you could just condition on X to see the causal effect of X on Y, and use that to test Z, you wouldn't need Z in the first place. The regression of Y on both X and Z would have omitted variable bias, and you would have to be able to condition on X, Z, and U, but U is unobserved, and you are using IV to get around the problem of U.

8

a

```
gen robcap = (robberies / (population/1000) )

gen bacap = (assaults + batteries) / (population/1000)

sum population

gen popstand = (population - 857.2294)/479.5102

encode community, gen(enc_c)
```

b

```
eststo: reg robcap popstand i.enc_c bus_stops i.intersection i.majorstreet pct_black ///
pct_hisp pct_hh_w_public_assist pct_residential pct_industrial pct_commercial, vce(robust)
```

In Table 1, the first regression. Population has a significant negative correlation with robberies per caita, pct residential had a significant negative correlation with robberies per capita. Percent industrial had a significant negative correlation with crime, but less significant than before, with pretty large standard errors. Percent commercial had a huge and very significant positive correlation with crime, with a much greater magnitude than the others.

c

```
eststo: reg bacap popstand i.enc_c bus_stops i.intersection i.majorstreet pct_black ///
pct_hisp pct_hh_w_public_assist pct_residential pct_industrial pct_commercial, vce(robust)
```

Results, shown in regression 2 of Table 1, were overall pretty similar to before, with a few very notable changes. With batteries and assaults, the negative correlation with population was greater. In addition, the residential correlation was slightly smaller. Most interestingly, the negative correlation with industrial areas is much, much greater and very statistically significant. And the coefficient with percent commercial, which was already big, nearly triples in size.

d

```
eststo: ivregress 2sls robcap bus_stops i.enc_c i.intersection i.majorstreet ///
pct_black pct_hisp pct_hh_w_public_assist (popstand pct_residential pct_industrial ///
pct_commercial = pct_low_dens_zoned pct_medium_dens_zoned pct_high_dens_zoned ///
pct_highest_den_zoned ct_commercial_zoned pct_industrial_zoned ///
pct_residential_zoned), vce(robust)

eststo: ivregress 2sls bacap bus_stops i.enc_c i.intersection i.majorstreet pct_black ///
pct_hisp pct_hh_w_public_assist (popstand pct_residential pct_industrial ///
pct_commercial = pct_low_dens_zoned pct_medium_dens_zoned pct_high_dens_zoned ///
pct_highest_den_zoned pct_commercial_zoned pct_industrial_zoned ///
pct_residential_zoned), vce(robust)
```

After being instrumented, the regressions change significantly. First off, many of the results are just far less signficiant, and many we can't draw conclusions from at all. These include population, percent industrial, and percent commercial. There was a very significant negative correlation between percent residential and crime, however, and it was larger in the batteries and assaults regression (4). In addition, the hidden variable representing multifamily both end up with a significant positive correlation with both types of crimes. In previous regressions, they have had a insigificant or slightly negative correlation, but not here. With Assults and batteries in particular it is very significant.

e

```
eststo: ivregress 2sls robcap bus_stops i.enc_c i.intersection i.majorstreet pct_black ///
pct_hisp pct_hh_w_public_assist (popstand pct_residential pct_industrial pct_commercial = ///
pct_low_dens_zoned pct_medium_dens_zoned pct_high_dens_zoned pct_highest_den_zoned  ///
pct_commercial_zoned pct_industrial_zoned pct_residential_zoned pct_low_dens_zoned ///
  c.pct_low_dens_zoned#c.pct_residential_zoned c.pct_medium_dens_zoned#c.pct_residential_zoned///
c.pct_high_dens_zoned#c.pct_residential_zoned c.pct_highest_den_zoned#c.pct_residential_zoned), vce(robu

eststo: ivregress 2sls bacap bus_stops i.enc_c i.intersection i.majorstreet pct_black ///
pct_hisp pct_hh_w_public_assist (popstand pct_residential pct_industrial pct_commercial = ///
pct_low_dens_zoned pct_medium_dens_zoned pct_high_dens_zoned pct_highest_den_zoned  ///
pct_commercial_zoned pct_industrial_zoned pct_residential_zoned pct_low_dens_zoned ///
  c.pct_low_dens_zoned#c.pct_residential_zoned c.pct_medium_dens_zoned#c.pct_residential_zoned ///
c.pct_high_dens_zoned#c.pct_residential_zoned c.pct_highest_den_zoned#c.pct_residential_zoned), vce(robu
```

We might do this because the density of the zoning may not refer to high density zoning in areas where people live, ie residential areas. I did not have time to add all the interactions but i added them with residential which should be informative and help adjust for specifically high density residential areas in particular. The values did not really seem to change very much at all, however. At least the significant values. Seen in section 6 and 7 of Table 1.

Table 1: Regression Table (Land Use Only)

	(1)	(2)	(3)	(4)	(5)	(6)
popstand	-0.660***	-1.674***	-0.114	1.850	-0.168	1.551
	(0.0665)	(0.168)	(0.583)	(1.271)	(0.511)	(1.124)
pct_residential	-0.790***	-0.674**	-2.695***	-5.575***	-2.672***	-5.694***
	(0.144)	(0.261)	(0.648)	(1.239)	(0.618)	(1.181)
pct_industrial	-0.588*	-4.543***	-2.432	-1.095	-2.573	-2.114
	(0.289)	(0.611)	(2.408)	(5.253)	(2.153)	(4.721)
pct_commercial	2.341***	6.111***	1.153	0.988	1.137	0.796
	(0.310)	(0.643)	(0.774)	(1.593)	(0.741)	(1.528)
_cons	0.184	-1.941*	2.550***	5.430***	2.494***	5.349***
	(0.373)	(0.914)	(0.640)	(1.369)	(0.639)	(1.364)
N	19330	19330	19330	19330	19330	19330
adj. R^2	0.220	0.322	0.207	0.272	0.207	0.277

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$