Problem Set 6

Due December 2nd, 6:30 pm

1. Instrumental variable inequalities

Give an example in which all the instrumental variable inequalities hold and another example in which not all the instrumental variable inequalities hold. You only need to specify the joint distribution of (Z, D, Y) with binary Z and D.

2. Numerical equivalence of the indirect least squares and two-stage least squares estimators (Stat 260 only)

Consider the following canonical regression with one endogenous regressor D and one instrumental variable Z:

$$\begin{aligned} Y &= \tau D + X\beta + \varepsilon_1, \\ D &= \alpha Z + X\gamma + \varepsilon_2, \end{aligned}$$

where $Y, D, Z, \varepsilon_1, \varepsilon_2$ are $n \times 1$ vectors, and X is an $n \times p$ matrix. If D is endogenous, then the OLS fit for the first equation gives a biased estimator for τ. Instead, we can use either the indirect least squares or the two-stage least squares estimator.

The indirect least squares estimator has three steps: first, fit the OLS of Y on Z and X and get the coefficient of Z, denoted by $\hat{\theta}_Y$; second, fit the OLS of D on Z and X and get the coefficient of Z, denoted by $\hat{\theta}_D$; third, the indirect least squares estimator is the ratio $\hat{\tau}_{\text{ils}} = \hat{\theta}_Y / \hat{\theta}_D$.

The two-stage least squares estimator has two steps: first, fit the OLS of D on Z and X and obtain the fitted vector $\hat{D}$; second, fit the OLS of Y on $\hat{D}$ and X and obtain the coefficient of $\hat{D}$, denoted by $\hat{\tau}_{\text{tsls}}$.

Many textbooks claim that $\hat{\tau}_{\text{ils}} = \hat{\tau}_{\text{tsls}}$ without a formal proof. Note that this is a linear algebra fact without assuming any modeling assumptions. We can verify this using the following simple numerical example.

```
> n = 10^5
> u = rnorm(n)
> v = rnorm(n)
> x = matrix(rnorm(n*2), n, 2)
> z = rnorm(n)
> d = z + as.vector(x%*%c(1, 2)) + u
> y = d + as.vector(x%*%c(1, -1)) + u + v
> summary(lm(y ~ d + x))$coef[2]
[1] 1.500528
> summary(lm(y ~ z + x))$coef[2]/summary(lm(d ~ z + x))$coef[2]
[1] 1.001864
> dhat = lm(d ~ z + x)$fitted.values
> summary(lm(y ~ dhat + x))$coef[2]
[1] 1.001864
```

Now prove that $\hat{\tau}_{\text{ils}} = \hat{\tau}_{\text{tsls}}$.

3. Data analysis: a job training program (Schochet et al. 2008)

`jobtraining.rtf` contains the description of the data files `X.csv` and `Y.csv`.

`X.csv` contains the pretreatment covariates; you can view the sampling weight variable `wgt` as a covariate too. It is generally difficult to deal with sampling weights. Many previous analyses made this simplification. Conduct analyses with and without covariates.

`Y.csv` contains the sampling weight, treatment assigned, treatment received, and many post-treatment variables. Therefore, this data contains many outcomes depending on your questions of interest. The data also have many complications. First, some outcomes are missing. Second, unemployed individuals do not have wages or incomes. Third, the outcomes are repeatedly observed over time. When you do the data analysis, please give details about your choice of the questions of interest and estimators.

4. Data analysis: Card (1993)

Card (1993) used the geographic variation in college proximity to estimate the return to schooling. The treatment is education and the outcome is log wage. The instrumental variable is the college proximity. `card.csv` is a slightly modified version of the data from Card's homepage with detailed explanations of the variable names in `code_bk.txt`.

You can dichotomize the education variable and infer the LATE. How do you choose the threshold for the dichotomization? Are the results sensitive to your choice? Conduct an analysis with covariates.

Without dichotomizing the education variable, you can use the standard two-stage least squares estimator. Compare the results to the above analysis.

What if the return of education is nonlinear? Do the data provide information for us to test the nonlinearity?

5. Data analysis: Efron and Feldman (1991)

Efron and Feldman (1991) was one of the early studies dealing with noncomppliance under the potential outcomes framework. The original randomized experiment, the Lipid Research Clinics Coronary Primary Prevention Trial (LRC-CPPT), was designed to evaluate the effect of the drug cholestyramine on cholesterol levels. In the dataset `EF.csv`, the first column contains the binary indicators for treatment and control, the second column contains the proportions of the nominal cholestyramine dose actually taken, the last three columns are cholesterol levels. Note that the individuals did not know whether they were assigned to cholestyramine or to the placebo, but differences in adverse side effects could induce differences in compliance behavior by treatment status. All individuals were assigned the same nominal dose of the drug or placebo, for the same time period. Column 3, C_3, was taken prior to a communication about the benefits of a lowcholesterol diet, Column 4, C_4, was taken after this suggestion, but prior to the random assignment to cholestyramine or placebo, and Column 5, C_5, an average of post-randomization cholesterol readings, averaged over two-month readings for a period of time averaging 7.3 years for all the individuals in the study. Efron and Feldman (1991) used the change in cholesterol level as the final outcome of interest, defined as $C_5 - 0.25C_3 - 0.75C_4$. The original paper contains more detailed descriptions.

This dataset is more complicated than the noncompliance problem discussed in class. You can analyze it based on your understanding of the problem, but you need to justify your choice of method. There is no gold-standard solution for this problem.

REFERENCES

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. Technical report, National Bureau of Economic Research.

Efron, B. and Feldman, D. (1991). Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–17.

Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does job corps work? impact findings from the national job corps study. *American Economic Review*, 98(5):1864–86.