Problems to go with Mastering Metrics

Steve Pischke

Chapter 1

1. Consider the following three causal questions:

   - Many firms, particularly in southern European countries, are small, and owned and run by families. Are family owned firms growing more slowly than firms with a dispersed ownership?
   - What is the effect of studying economics rather than sociology on the salaries of university graduates?
   - What is the effect of mortgage interest rates on the number of new housing starts?

   For each of these questions answer the following:

   (a) What is the outcome variable and what is the treatment?
   (b) Define the counterfactual outcomes Y_{0i} and Y_{1i}.
   (c) What plausible causal channel(s) runs directly from the treatment to the outcome?
   (d) What are possible sources of selection bias in the raw comparison of outcomes by treatment status? Which way would you expect the bias to go and why?

2. For this question we will use a dataset from a randomised experiment conducted by Marianne Bertrand and Sendhil Mullainathan, who sent 4,870 fictious resumes out to employers in response to job adverts in Boston and Chicago in 2001. The resumes differ in various attributes including the names of the applicants, and different resumes were randomly allocated to job openings. Some of the names are distinctly white sounding and some distinctly black sounding. The researchers collecting these data were interested to learn whether black sounding names obtain fewer callbacks for interviews than white names.
   Download the data set `bm.dta` from Moodle.

   (a) The data set contains two dummy variables (0-1 variables) for female (`female`) and whether the applicant has computer skills (`computerskills`). Tabulate these variables by `black`. Using the command

   ```
   tab female black, col
   ```

   will give you cross-tabulation of female and race, and display the precentages of males and females in each race group. Do gender and computer skills look balanced across race groups?

   (b) Do a similar tabulation for `education` and the number of jobs previous held (`ofjobs`). These variables take on 5 and 7 different values, respectively. Does education and the number of previous jobs look balanced across race groups?

(c) Use the `summarize` command to look at the mean and standard deviation for the variable for years of experience (`yearsexp`) separately for black and whites (using the `if` modifier). Does this variable look similar by race?

(d) What do you make of the overall results on resume characteristics? Why do we care about whether these variables look similar across the race groups?

(e) The variable of interest on the data set is the variable call, which indicates a `call` back for an interview. Do you find differences in call back rates by race?

(f) What do you conclude from the results of the Bertand and Mullainathan experiment?

3. In the last problem we looked at the experimental Bertrand and Mullainathan resume data. For this question, download the data set `cps.dta`, which comes from the responses to the monthly US Current Population Survey (CPS) in 2001, a large labour market survey. This data set contains data on 8,891 individuals living in Boston and Chicago. We want to use these data to compare the skills of real live blacks and whites, and their employment outcomes and see how they differ from the resume findings in problem set 1.

   (a) The data set contains a variable `education`, which takes on four values (high school dropouts, high school graduates, some college, and college degree and more). It also contains a category for resumes not reporting any education. Use the `education` variable to create a new dummy for resumes indicating some college or more (i.e. those in the some college category plus those in the college and more category). What fraction of respondents has at least some college education?

   (b) Carry out a t-test for whether the mean of your some college or more variable is the same for blacks and whites. You can do this with the `ttest` command. E.g. if your variable is called `somecol`, you would type

   ```
   ttest somecol, by(race)
   ```

   Report

   1. the mean of the variable for whites
   2. the mean of the variable for blacks
   3. the difference in the means
   4. the t-statistic
   5. the p-value for the null hypothesis that the two means are the same.
   6. Do you find any evidence that this variable differs for whites and blacks in the CPS?

   (c) Calculate the t-statistics for equality of the means in the years of experience (`yearsexp`). Do you find evidence that this variables differs significantly for whites and blacks?

   (d) Discuss your results from (b) and (c). Why do your conclusions for the education and experience variables differ? Why do we care about whether these variables look similar by race?

(e) Calculate the t-statistics for equality of the means in whether the individual has a job (`employed`). Do you find evidence that this variables differs significantly for whites and blacks?

(f) In the light of your results in (d) and (e), what do think you can conclude about racial discrimination in employment from the CPS data?

Chapter 2

1. A small company sells medical supplies to hospitals. Management wants to assess the efficacy of the company's advertising, and an analyst has produced the following three regressions:

$$\begin{aligned} \text{sales}_i &= -516.4 + 2.47\ \text{advertising}_i + 1.86\ \text{bonus}_i + e_i \\ \text{sales}_i &= -156.5 + 2.77\ \text{advertising}_i + e_i \\ \text{bonus}_i &= 193.5 + 0.16\ \text{advertising}_i + e_i \end{aligned}$$

where sales_i are sales in territory i (in £1,000), advertising_i is spending on advertising (in £100), and bonus_i is the amount of bonuses paid to sales people in the territory (in £100).

(a) Why is the coefficient on advertising different in the first two regressions? Show how the coefficient in the second regression relates to the one in the first using the information provided.

(b) Is either of the regressions likely to provide a good indication of the causal effect of advertising spending on sales? Why or why not?

2. Suppose a doctor assigns treatment T_i solely on the basis of three factors: age of the patient, blood pressure, and blood sugar level. Can you estimate the following regression equation

$$Y_i = \alpha + \rho T_i + \beta_1 \text{Age}_i + \beta_2 (\text{Blood pressure})_i + \beta_3 (\text{Blood sugar})_i + e_i$$

to get the causal effect of treatment on the outcome Y_i? Why or why not?

(a) A specific condition can be treated either with a traditional treatment, or with a new method. Denote the new treatment by the dummy variable T_i . The condition and general health of a patient is diagnosed with multiple measures, which are collected in a score S_i. Patients with higher scores are more ill, and they are more likely to receive the new treatment. But assignment differs somewhat from patient to patient (e.g. different doctors use different (implicit) thresholds to assign the treatment). In order to understand the treatment assignment better, a researcher estimates the following equation:

$$T_i = \underset{(0.113)}{-0.456} + \underset{(0.009)}{0.035} S_i - \underset{(0.016)}{0.035} (\text{Female})_i - \underset{(0.028)}{0.004} (\text{Age})_i + \underset{(0.039)}{0.003}\, (\text{Age})_i^2 /100 + e_i.$$

The researcher then goes on to estimate

$$Y_i = \underset{(0.225)}{0.006} - \underset{(0.572)}{1.211} T_i + \underset{(0.165)}{0.798} S_i - \underset{(0.221)}{0.562} (\text{Female})_i - \underset{(0.191)}{0.766} (\text{Age})_i + \underset{(0.138)}{0.481} (\text{Age})_i^2 /100 + e_i$$

where Y_i is a measure of health status 3 months after the initial treatment, and a lower value of Y_i denotes better health. What is the rationale for including the regressor for female in the outcome equation? What is the rationale for including the regressors for age and age square in the outcome equation?

(b) Another researcher is worried that the outcome regression in (b) may not identify the causal effect of treatment. That researcher notices a variable which measures the number of days the patient was hospitalized after treatment $(\text{Days})_i$ and includes it in the regression:

$$\begin{aligned} Y_i \;=\; & \underset{(0.233)}{-1.347} - \underset{(0.549)}{0.625} T_i + \underset{(0.277)}{0.423} S_i + \underset{(0.023)}{0.111} (\text{Days})_i - \underset{(0.239)}{0.538} (\text{Female})_i \\ & \underset{(0.203)}{-0.815} (\text{Age})_i + \underset{(0.142)}{0.465} (\text{Age})_i^2 /100 + e_i. \end{aligned}$$

Do you think the regression in (b) or (c) is more likely to estimate the causal effect of treatment? Explain why.

Chapter 3

1. Download the data ajr.dta from the Moodle website. The data set contains per capita income in 1995 as well as a number of other variables for 62 non-European countries. The data have been collected by Acemoglu, Johnson, and Robinson (AJR), who, like many other economists, believe that rich countries are rich primarily because they have "institutions" which are more conducive to growth. "Institutions" refers to wide set of political and economic arrangements, including democracy versus autocratic rule, the security of property rights, the enforcement of law and contracts, the efficiency of the bureaucracy versus corruption, etc. In this question we want to assess the particular hypothesis tested by AJR that the protection of property rights is conducive to growth, and hence should be correlated with the level of contemporary per capita income (you must have been growing a lot in the past to be rich now). The data set contains a variable `risk`, indicating the protection of property rights (with larger values indicating more protection–I know that's crazy but I haven't changed their coding). The log of GDP per capita is called `loggdp`.

   (a) Run an OLS regression of `loggdp` on `risk`. Comment on your result.

   (b) Why might you be worried about interpreting the effect of property rights or expropriation risk on GDP per capita causally? Explain.

   (c) AJR suggested using the mortality of European colonial settlers as an instrument for property rights protection. Their argument is that European imperial powers set up different institutions in various countries depending on whether they decided to settle there (as in the USA, Argentina, or Australia) or whether they decided simply to exploit the natural resources of the colony (as in many African countries). Some colonies had conditions more conducive to the settlement than others, and these are measured by `logmort0`, the log of European settler mortality (measured mostly in the 1800s).

1. What conditions need to be satisfied for settler mortality to be a valid instrument for property rights? Discuss whether each of these is likely to hold in this context. Which of these conditions can you check in the data?
2. Estimate the first stage equation and explain what you find.
3. Estimate the reduced form and explain what you find.
4. Run the IV regression of `loggdp` on `risk` using `logmort0` as instrument. Comment on your result and compare them to your results in (a). Is the differences you find explained by the biases you discussed in (b)?
5. Construct the IV estimate from your results in part 2. and 3.

(d) A critic of these regressions is worried that the current level of GDP is correlated with the disease environment in a country, which in turn will be correlated with European settler mortality in the 1800s.

1. If the critic is right, what is the consequence of this for your IV results you obtained in part (c)?
2. The critic therefore proposes to include another regressor, the current incidence of malaria in the country. Explain why this is a potential solution to the problem identified by the critic.
3. Repeat the first stage including malaria as a regressor. Compare your result to that from part (c) and comment on your findings.
4. Repeat your OLS and IV regressions `loggdp` on `risk` including `malaria` as a regressor. Compare your results to those from parts (a) and (c) and comment on your findings.

(e) Other economists dispute the institutions view of economic development. For example, Jeffrey Sachs argues that geographic conditions are mostly responsible for underdevelopment. Run a "horserace" between the AJR view and the Sachs view by including instrumented `risk` in your regression, controlling for malaria, and also add the absolute value of latitude (`latitude`, proxying for general climate and soil conditions), minimum monthly rainfall (`rainmin`) and mean temperature (`meantemp`) to your regression.

1. Can we simply include the variables `latitude`, `rainmin`, and `meantemp` in the regression and interpret it causally? Explain.
2. What do you conclude from this regression about the AJR versus the Sachs view of the world?

2. A bank offers a week long management training programme to all its loan officers at the end of the second year in their job. Participation in the programme is voluntary. The bank is interested in knowing whether participation in the programme makes it more likely that a loan officer is promoted to branch manager. You have data on all the bank's loan officers, their participation in the programme, and whether they have been promoted between years three and five of being in their job. You run a regression for the promotion decision on a constant and a dummy for participation in the training programme.

(a) Suppose the bank encourages participation in the training programme by sending a personal letter from the CEO to selected employees who have been recommended by their supervisor as showing particular potential for a position in management. Would it be useful to use the CEO letter as an instrument for participation in the training programme? Explain why or why not.

(b) Suppose the bank encourages participation in the training programme by sending a personal letter from the CEO to a randomly chosen set of employees. Would it be useful to use this CEO letter as an instrument for participation in the training programme? Explain why or why not.

Chapter 5

1. Download the dataset called "minwage.dta". It contains data collected by David Card and Alan Krueger on fast food restaurants in New Jersey (NJ) and eastern Pennsylvania (PA) during two interview waves in March and November/December of 1992. On April 1, 1992 New Jersey raised its minimum wage from \$4.25 to \$5.05. The minimum in Pennsylvania remained at the federal level of \$4.25. Use this data to analyze the impact of the minimum wage increase in New Jersey on employment in the fast food industry.

   Throughout, variable names with a trailing "2" refer to the second (Nov./Dec.) wave of the data, the same names without any number refer to the corresponding variable from the March wave. "fte" and "fte2" are full time equivalent employment, it is the sum of the number of full time employees and one half the number of part time employees, excluding managers; "dfte" refers to the change in full time equivalent employment between the second and first interview (fte2 - fte); "dw" refers to the change in the starting wage between the second and first interview, and "sample" is dummy variable which is 1 if both wage and employment data are available in both the first and second interview wave, and 0 otherwise. I want you to do the following analysis for the part of the data with sample equal to 1. If you don't specify this, Stata will make calculations with the full set of available observations for each variables, so you may not be comparing the same set of restaurants between March and November, or you may compare wages and employment for different restaurants.

   (a) Calculate the average starting wage (`wage_st`) separately for restaurants in NJ and in PA, both for each interview wave.

      1. Calculate the difference in the average wages between the second and first interviews.
      2. Now calculate the difference between NJ and PA of the time differences just obtained.
      3. What is the interpretation of such a difference-in-differences estimate of the wage effect? Under what conditions does this provide a valid estimate of the of the minimum wage increase on wages in the fast food industry?
      4. Interpret your finding.

   (b) Repeat the same exercise as in (a) for full time equivalent employment. What is the impact of the minimum wage increase on relative employment in NJ restaurants?

(c) Difference-in-difference estimates can also be calculated from the regression

$$Y_{i,s,t} = \alpha + \beta TREAT_{i,s} + \gamma POST_t + \delta_{rDD}\left(TREAT_{i,s} * POST_t\right) + e_{i,s,t},$$

where $Y_{i,s,t}$ is employment in restaurant i in state s and period t, $TREAT_{i,s}$ is an indicator for the treatment area (NJ or low wage restaurants in NJ), $POST_t$ is an indicator for the treatment period (Nov/Dec) and $TREAT_{i,s} * POST_t$, is the interaction of these two dummies. Note that this regression uses the data for individual restaurants i, unlike in the lecture, where we worked with state/period averages of banks or deaths. Here we leave the averaging to the regression.

1. Write the equation separately for March and Nov/Dec and show that the DD model for two periods ($t = 1, 2$) can be estimated as

$$Y_{i,s,2} - Y_{i,s,1} = \gamma + \delta_{rDD} TREAT_{i,s} + e_{i,s,2} - e_{i,s,1}. \quad (1)$$

2. What are the regression DD estimates on wages and employment using this regression? How do they compare to the results you found in (a) and (b)?
3. The regression allows you to control for other factors. Repeat the regressions, entering a dummy variable for whether the restaurant is company owned ("co_owned," as compared to franchised) and three dummy variables for three of the four chains in the dataset (Burger King, KFC, Roy Rogers, and Wendy's; you will have to construct the dummies from the variable "chain") or use `i.chain`.
4. Do your results change when you enter restaurant specific covariates? Would you have expected the results to change? Explain why or why not.

(d) An alternative to comparing NJ and PA restaurants is comparing restaurants within NJ which have high and low wages before the minimum wage increase. Restrict your sample to restaurants in NJ.

1. Would you expect the DD assumptions to be satisfied more easily for the within NJ comparison than for the NJ - PA comparison?
2. Construct a variable for those restaurants paying starting wages of less than \$5.00 before the minimum wage increase. Use the regression to obtain a DD estimate of the employment and wage effects of the minimum wage increase. What is the relative impact of the minimum wage on starting wages and employment within NJ?
3. How do your witin NJ estimates compare to those obtained in part (c) for the NJ - PA comparison?

(e) You can create a variable for those restaurants paying starting wages of less than \$5.00 *in PA* in the initial period. There is no minimum wage forcing those restaurants to pay more in the second period but there may be general wage growth.

1. Now run a regression of changes in employment and wages just for PA using this new variable for low paying restaurants in PA. How do your results differ from those just for NJ?
2. Carry out a statistical test of the hypothesis that the coefficient on the low wage dummy is the same in NJ and in PA.

3. Why is this a check on how well the methodology is doing in uncovering effects of the minimum wage increase? What do you conclude?

2. You are interested in the impact of a new fertiliser on the yield of wheat. You have the following data for 253 farms over 7 years:

yield	crop yield on farm i in year t
capital	the amount of capital equipment used on farm i in year t in constant 2000 £
labour	the number of farm workers on farm i in year t
rain	annual rainfall on farm i in year t in mm
fertiliser	fertiliser applied by farm i in year t in 1,000 kg
size	number of hectares available for planting on farm i

You don't have any data on fixed farm attributes like ability of the farm manager or soil quality.

(a) Suppose you knew that fertiliser was randomly assigned to farms in some years and not other years. How would you specify an econometric model to estimate the causal effect of fertiliser on crop yield? Be precise about your model, the variables you use, and your method of estimation and statistical inference, and explain why you make these choices.

(b) Continue with the assumption that fertiliser is randomly assigned to farms and years. How would you carry out a statistical test that fertiliser has the same impact on yield on big farms (> 30 hectares) as on small farms?

(c) Suppose that the use of the new fertiliser is chosen by farmers rather than randomly assigned. Nobody uses fertiliser in the first year the data is available and different farmers start using fertiliser in different years. How would you go about estimating the causal effect of the new fertiliser now? Which additional assumptions do you need to make compared to the case in a)?