STA501: Data-based Decision Making

Lecture Note 7

Question 1. A data scientist is interested in estimating the production function for AI products which is postulated to follow a Cobb-Douglas specification:

$$Y_i = exp(\beta_0) \cdot L_i^{\beta_L} \cdot K_i^{\beta_K} \cdot exp(u_i),$$

where Y_i is a measure of output of a multi-nationally funded company's AI product for firm i, L_i is labor in the country, K_i is capital stock and u_i is an unobserved term that captures technological or managerial efficiency and other external factors (e.g., weather). The parameters to be estimated are $(\beta_0, \beta_L, \beta_K)$. Taking logs,

$$lnY_i = \beta_0 + \beta_L lnL_i + \beta_K lnK_i + u_i.$$

(1) Interpret β_L and β_K. Between developed and developing countries, on average, how would β_0, β_L, and β_K be different?

(2) Assume that you have a cross-section of independent firms and that more productive firms hire less workers (labor). Explain why OLS would not provide consistent estimates for $(\beta_0, \beta_L, \beta_K)$. Would it over- or under-estimate β_L on average? Clearly explain your answer.

(3) As a data scientist representing labor union, you would like to argue that β_L is under-estimated due to endogeneity of the model. What is your strategy? Provide an argument with data scientific background.

(4) Given (3), name any possible instrumental variable, and back up your argument.

(4) Describe in detail how you would estimate the parameters of the production function using Two Stage Least Squares (2SLS). What restrictions would be necessary for this researcher to successfully use this instrumental variable in the estimation of the parameters $(\beta_0, \beta_L, \beta_K)$ and what would you need to assume about capital stock?

(5) If average wages per firm do not vary much by firm (potentially because of unionization or high mobility of the labor force), how would this affect the properties of the estimation procedure suggested in (4)? Explain your answer.

(6) Another data scientist representing the company argues that it would have been more profitable to set up an off-shore office with a cheaper labor in AI production. Given that, the data scientist claims that the labor union asks too much raise in wage upto the level that gives the executives an incentive to seriously consider cross-border operation. How would you form your counterargument? Does the new formation of the analysis can help removing necessity of 2SLS?

Question 2. In this part, we are interested in analysing whether workplace smoking bans affects the incidence of smoking. Using data on 10,000 Swiss indoor workers from 2091 to 2093 taken from "Do Workplace Smoking Bans Reduce Smoking", the following probit regression was estimated. (Find the table below.)

The dependent variable, *smoker*, is a dummy variable indicating whether a worker smokes (1=yes, 0=no) and the explanatory variables are *smkban*, a dummy variable indicating whether there is a ban on smoking in the workplace (1=yes, 0=no), the worker's *age* (in years), *gender* (male/female), *ethnicity* (black/hispanic/white) and level of *education* ($E1$=highschool dropout, $E2$=highschool graduate, $E3$=some college, $E4$=college graduate, $E5$=Master degree or above).

Probit regression — Number of obs = 10,000
LR chi2(9) = 569.63
Prob > chi2 = 0.0000
Log likelihood = -5252.3489 — Pseudo R2 = 0.0514

smoker	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
smkban	-.1517626	.0289268	-5.25	0.000	-.208458	-.0950671
female	-.1106249	.0287785	-3.84	0.000	-.1670298	-.05422
age	-.0042031	.0011748	-3.58	0.000	-.0065057	-.0019006
black	-.07969	.0525369	-1.52	0.129	-.1826604	.0232804
hispanic	-.3327039	.0476677	-6.98	0.000	-.4261308	-.2392769
E1	1.094231	.0714121	15.32	0.000	.9542663	1.234197
E2	.8518588	.0594747	14.32	0.000	.7352906	.9684271
E3	.6492566	.0606989	10.70	0.000	.530289	.7682241
E4	.2224747	.0649939	3.42	0.001	.0950891	.3498603
_cons	-.9842425	.0756055	-13.02	0.000	-1.132427	-.8360584

(1) Build an ATE model that can test whether smoking ban is an effective means of reducing smoking, assuming that you have a neighboring twin city's data without workplace smoking ban.

(2) It turned out that the twin city's data is not free and your company cannot afford the expense for acquiring the data. Without the data, build a test whether these results show that smoking in the work place has a significant effect on the incidence of smoking ban.

(3) Following (2), what is the limitation of the missing data from the twin city?

(4) Your colleague from no statistics background claim that more data is absolutely necessary to support your logic in (2). Provide your rebuttal.

(5) Explain how you can estimate the effect of the smoking ban on the probability of smoking for a 50-year old white, college graduated man. You are not expected to use your calculator, clarity of the computations required is enough.

(6) Given upcoming election, the incumbent city mayor claims that the workplace smoking ban has given a dramatic decrease in smoking and increase in public health. A contending candidate argues that given the fact that workplaces earn subsidy by reporting smoking ban, there are considerable number of companies faked in their reports. She also questions the accuracy of smoker's response whether they quit smoking during the survey. Formulate a model to test the contending candidate's argument.

Hint: You may recall that for the Probit model, we will specify

$$Pr(smoker = 1|x) = \Phi(\beta_0 + \beta_1 smkban + \beta_2 female + ... + \beta_8 E_3 + \beta_9 E_4)$$

where Φ is the standard normal CDF (cumulative distribution function).

Question 3. Let $ds10$ denote the percentage of students at SIAI receiving passing score on data science. We are interested in estimating the effect of per student spending on data science performance. A simple model is

$$ds10_i = \beta_0 + \beta_1 \log(expend_i) + \beta_2 \log(enroll_i) + \beta_3 math_i + u_i$$

where $math_i$ is the percentage of students have taken linear algebra. There are some transfer students from less sophisticated schools where quality of math education is questionable. In other words, you are faced with the fact that data is partly unavailable or measured with error on a key variable: $math$. You do have information available on a closely related variable: the students' SAT score for math, sat_i.

(1) In this question we want to use sat_i as a proxy for $math_i$ that we consider running the regression

$$ds10_i = \gamma_0 + \gamma_1 \log(expend_i) + \gamma_2 \log(enroll_i) + \gamma_3 sat_i + e_i$$

where we assume that the following relationship exists

$$math_i = \alpha_0 + \alpha_1 sat_i + v_i$$

(1-1) Briefly discuss why sat_i is a sensible proxy variable for the variable in question, $math_i$.

(1-2) Discuss the assumptions you need to make to enable consistent parameter estimates on β_1 and β_2 using your estimable equation for $ds10_i$. Will your estimates of β_1 and β_2 be unbiased as well? Prove your statements.

(2) The OLS results with and without sat_i as an explanatory variable are given by (standard errors in parentheses):

$$\widehat{ds10}_i = \underset{(26.72)}{-69.24} + \underset{(3.30)}{11.13} \log(expend_i) + \underset{(0.615)}{0.022} \log(enroll_i),$$
$$N = 428, \quad R^2 = 0.0297$$
$$\widehat{ds10}_i = \underset{(24.99)}{-23.14} + \underset{(3.04)}{7.75} \log(expend_i) - \underset{(0.58)}{1.26} \log(enroll_i) - \underset{(0.036)}{0.324} sat_i,$$
$$N = 428, \quad R^2 = 0.1893$$

Explain why the effect of expenditures on $ds10_i$ is lower in the regression where sat_i is included than where it is excluded.

(3) Provide a test for the hypothesis that we have a significant positive effect of expenditure at the 5% level of significance.

(4) A vocational school for computer coding launches a promotion that since there are nearly a perfect correlation between SAT in math and computer programming, wannabe data scientists must take the vocational school's computer programming course. Evaluate the promotion.

(5) Your research assisitant for data collection confessed that the entries for $enroll_i$ are partly made up number due to unavailability. How does this affect your estimation of γ_2? Does it affect other estimates?

(6) Another research assistant plugged out the data storage too early in copying, which resulted in 30% loss of $ds10_i$. Other variables are intact. How does this affect your statistics?

(7) Growing disbelief in your research assistants led you to hire new RAs. It turned out that both of your failing RAs are from the same student club dedicated to data science, where they spend extensive study hours together, share problem set solutions and exam information, with limited membership by GPA, and non-members are sidelined and often absent in classes. Although you are no longer willing to hire RAs from the same student club, non-members are not even responsive to your emails. Does this impact your regression? As a data scientist, can you justify your bias?

(8) A neighboring school with questionable quality teaching in data science approaches to SIAI that they have 10 times more students for the same data set. A researcher from that school claims that $N = 428$ is not a big data, but with more than 4,000 students, the bigger data can significantly raise R^2. Provide your rebuttal with pertinent differenciation between larger set of data and "BigData".

(9) Does the neighboring school's data can help you for (7), assuming the school does not have such a prestigious student club?

Question 4. (Assignment for TA Session) This question looks at a population of applicants to the Doctor of Medicine degree in 2085 who had similarly marginal credentials, which are followed up over time. Suppose that admission was offered randomly to half of the population in question. Consider the following data scientific framework:

$$\begin{aligned} Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\ P_i &= D_i P_{1i} + (1 - D_i) P_{0i}, \end{aligned} \tag{1}$$

where D_i is an indicator for person i having been admitted in 2085 to a Doctor of Medicine degree; Y_i is log annual earnings of person i in 2110; Y_{1i} is log annual earnings of person i in 2110 if they had been admitted in 2085 to a doctor of Medicine degree; Y_{0i} is log annual earnings of person i in 2110 if they had not been admitted in 2085 to a Doctor of Medicine degree; P_i is an indicator for person i working as a physician in 2110, P_{1i} is an indicator for person i working as a physician in 2110 if they had been admitted in 2085 to a Doctor of Medicine degree; and P_{0i} is an indicator for person i working as a physician in 2010 if they had not been admitted in 2085 to a Doctor of Medicine degree.

(1) Could you estimate the causal effect of admission to the Doctor of Medicine degree in 2085 on log annual earnings in 2110 for person i in this population? Explain your answer briefly.

(2) Could you estimate an average causal effect of admission to the Doctor of Medicine degree in 2085 on log annual earnings in 2110? Explain your answer briefly.

(3) Now suppose that we regress Y_i on D_i for the people for whom $P_i = 1$. Write the probability limit of the estimated coefficient on D_i and its relation to the effect $\mathbb{E}[Y_{1i} - Y_{0i} | P_{1i} = 1]$. Explain your answer.

(4) Explain intuitively (without equations) why conditioning a regression of Y_i on D_i for the people for whom $P_i = 1$ is fundamentally different from conditioning a regression of Y_i on D_i on an indicator for being born after 2062. Which approach makes more sense and why?