UCLA Anderson School of Management

Prescriptive Models and Data Analytics
Problem Set #2

1 Hospital admission & quality of service

Download `health_data.csv` and load it into python. These are data from hospital admissions for coronary artery bypass graft (CABG) in the UK. Among other things, you observe whether the patient died after the surgery (coded up as `patient_died_dummy`), which hospital the patient visited (`hospital_id`), and a series of patient characteristics such as gender and age.

Question 1. Start by regressing the patient-died dummy variable on a set of hospital dummies

(a) Based on the regression output, interpret the coefficients on the constant term and the dummy for hospital D.

(b) What is the difference between the mortality rates at hospitals D and E (use the regression output to derive this)?

Causal interpretation (or lack thereof)

Question 2. Continue to use the hospital data in this question, but only use data for patients that visited either hospital A or B. Regress mortality on an intercept and a dummy for whether the patient visited hospital B.

(a) Explain why the difference in mortality rate implied by this regression cannot be interpreted as the causal effect of visiting a different hospital (i.e., the change in risk of dying when moving a patient from hospital A to B cannot be inferred from this regression).

(b) Do you think difference in mortality between hospitals are over- or under-estimated? Think about what type of patients go to which type of hospital.

(c) What are potential control variables that you might want to include in the regression, in order to obtain a causal estimate (or at least get closer to a causal estimate)? Run such a regression with suitable controls and interpret the change in the coefficient on the hospital B dummy. Explain why you included the specific set of variables.

2 Demand estimation

The dataset `demand_data.csv` contains data on sales and prices at a set of ice-cream vendors measured over 52 weeks. All ice-cream at a given store is always priced the same, so there is only one price variable. However, different vendors charge different prices and most vendors vary their prices throughout the year.

Question 1. Load `demand_data.csv` into Python. For vendor 1, run a regression of sales on price and also a regression of sales on price and a summer dummy (make sure your regression selects only the 52 weeks of data for vendor 1). Use the omitted variable bias formula to explain why the price coefficient changes when the summer dummy is also included in the regression.

Question 2. Repeat the two regressions that you just ran in question 1, but now use data only for vendor 2. In the case of the regression with the summer dummy, you should find that there might be multicollinearity problems. Why does this happen?

Question 3. Suppose that one of the vendors did not systematically charge higher or lower prices in summer. If you were to repeat the analysis you just did for vendors 1 and 2, what would you expect to happen to the price coefficient estimate and its precision in the two regressions with and without the summer dummy?