Predictive Analytics
Homework 2

Due Monday, July 23

Your submission should consist of a neatly formatted PDF report and Stata do files that document your commands.

Problem 1 - you will use the dataset HousePrice.dta. This dataset collects data on houses in Windsor, ON. Your task is to build a model estimating the effects of a variety of house characteristics on its price.

1. We will start with the simple model

$$price = \beta_0 + \beta_1 lotsize + \varepsilon$$

   Create a well-labeled graph showing a scatter plot of price against lot size, with a regression line from the above model overlaid.

2. Adjust the model in (1) to use a quadratic functional form. Create a new graph showing your new regression curve overlaid on the scatter plot. Give an intuitive explanation for why this model does a better job predicting house prices than the linear model.

3. Add the number of bedrooms and number of bathrooms to the linear model in (1). If our goal is to determine the causal impact of lot size on house price, explain why and how adding bedrooms and bathrooms to the model helps us do this. Be specific about the role of these variables and the effect they had on the estimate of β_1.

4. Add the rest of the variables in the dataset to the model and briefly discuss the interpretation of the estimates. How does the estimate of β_1 compare with (1) and (3)? Give an intuitive explanation.

5. Discuss some possible omitted variables which may still causing endogeneity in $lotsize$ in model (4). Explain why you think they may be biasing your results.

6. Construct a table showing and comparing coefficient estimates, standard errors, and R^2 from parts (1), (2), (3), and (4).

Problem 2 - Use the data in CARD.dta. This dataset contains information on wages, experiences, educational attainment, and whether people grew up near a four year college. The data contain information from 1976, when wages are recorded, as well as information from 1966, when the respondents were growing up.

1. Estimate a regression of log wages on education and controls for experience, experience squared, black, south, SMSA, SMSA status in 1966, and dummies for each of the regions in 1966, leaving one out. (After you load the dataset, type "describe" to see what each of these variables are called.) What is the coefficient on education and its standard error? Does this tell us the causal effect of education on wages? Why or why not? Explain.

2. We will use nearc4 as an instrumental variable. This variable is a dummy variable indicating whether the respondent lived near a four year college when growing up. State the assumptions necessary for nearc4 to be a valid instrument. Do these seem like plausible assumptions?

3. Run the first stage regression of education on nearc4 and all of the other controls. Does nearc4 affect educational attainment?

4. Run the second stage regression of log wages on nearc4 and all the other controls. What do you find? What does this imply?

5. Run the ivregress 2sls command in Stata. Report the coefficient on education along with its standard error. What do you estimate are the returns to education? How does this compare to your answer in part (1) and (4)?