STAT 4400 / 5400

Homework #4

Spring 2022

Due: Mar. 8 by midnight

(1) Discrete probability simulation: suppose that a basketball player has a 60% chance of making a shot, and he keeps taking shots until he misses two in a row. Also assume his shots are independent (so that each shot has 60% probability of success, no matter what happened before).

(a) Write an R function to simulate this process.

(b) Put the R function in a loop to simulate the process 1000 times. Use the simulation to estimate the mean, standard deviation, and distribution of the total number of shots that the player will take.

(c) Using your simulations, make a scatterplot of the number of shots the player will take and the proportion of shots that are successes.

(2) Summarizing inferences and predictions using simulation: Problem 3 in Homework #3 used a Tobit model to fit a regression with an outcome that had mixed discrete and continuous data. Revisit these data and build a two-step model: (1) logistic regression for zero earnings versus positive earnings, and (2) linear regression for level of earnings given earnings are positive.

Compare predictions that result from this model with those from the regression model in Problem 3, Homework #3.

(3) Fitting the wrong model: suppose you have 100 data points that arose from the following model: $y = 3 + 0.1x_1 + 0.5x_2 + \epsilon$, with errors having a t distribution with mean 0, scale 5, and 4 degrees of freedom. We shall explore the implications of fitting a standard linear regression to these data.

(a) Simulate data from this model. For simplicity, suppose the values of x_1 are simply the integers from 1 to 100, and that the values of x_2 are random and equally likely to be 0 or 1. Fit a linear regression (with normal errors) to these data and see if the 68% confidence intervals for the regression coefficients (for each, the estimates ± 1 standard error) cover the true values.

(b) Put the above step in a loop and repeat 1000 times. Calculate the confidence coverage for the 68% intervals for each of the three coefficients in the model.

(c) Repeat this simulation, but instead fit the model using t errors.

(4) The dataset "sesame.dta" contains data from an experiment in which a randomly selected group of children was encouraged to watch the television program Sesame Street and the randomly selected control group was not. The file "sesame.vars.doc" describes the dataset in detail.

(a) The goal of the experiment was to estimate the effect on child cognitive development of watching more Sesame Street. In the experiment, encouragement but not actual watching was randomized. Briefly explain why you think this was done. (Hint: think of practical as well as statistical reasons.)

(b) Suppose that the investigators instead had decided to test the effectiveness of the program simply by examining how test scores changed from before the intervention to after. What assumption would be required for this to be an appropriate causal inference? Use data on just the control group from this study to examine how realistic this assumption would have been.

(5) Instrumental variables: come up with a hypothetical example in which it would be appropriate to estimate treatment effects using an instrumental variables strategy. For simplicity, stick to an example with a binary instrument and binary treatment variable.

(a) Simulate data for this imaginary example if all the assumptions are met. Estimate the local average treatment effect for the data by dividing the intent-to-treat effect by the percentage of compliers. Show that two-stage least squares yields the same point estimate.

(b) Now simulate data in which the exclusion restriction is not met (so, for instance, those whose treatment level is left unaffected by the instrument have a treatment effect of half the magnitude of the compliers) but the instrument is strong (say, 80% of the population are compliers), and see how far off your estimate is.

(c) Finally, simulate data in which the exclusion restriction is violated in the same way, but where the instrument is weak (only 20% of the population are compliers), and see how far off your estimate is.