

GB704 - Homework 2
Due : September 28

Problem 1 (50pts)

Download dataset data1D2D.xlsx from my webpage.

1. Regress $\ln(\text{spending})$ on gender. Does gender have a significant effect on spending according to that regression?
2. Now regress $\ln(\text{spending})$ on income and gender. What happens to the significance of gender? Explain in a sentence or two what caused this change (if any).
3. Now we want to forecast what a new male customer with income \$150,000 and age 40 is going to spend using a regression model with $\ln(\text{spending})$ on the left-hand side. Possible explanatory variables are age , age^2 , $income$, $income^2$ and $gender$. Choose the model that you feel is best for that purpose (explain why you chose it) and use that model to forecast the spending of our new customer.
4. (Nearest neighbors forecasting) Perform the same forecasting exercise as in part 3 by averaging the spending of the 10 closest consumers in our existing data. Use the same notion of distance as in class.
5. (Cross-validation) Estimate your preferred model after leaving the first 500 observations out. Use the resulting model to forecast the 500 observations you left out. Plot observed outcomes vs predicted outcomes. Fit a line through those dots and interpret the resulting R^2 in **one** sentence.