

ECON 282 Midterm Assignment

(1) Find or compile a dataset to use for this project, looking for one that is relatively fat, i.e. has a large number of variables per observation. If you are interested in sports, team or player performance data should work well. Many large datasets are available at the websites listed below. Be warned that it may take some data wrangling to get them into R.

www.data.gov

www.census.gov

research.stlouisfed.org

data.worldbank.org

www.nsf.gov/statistics/data.cfm

azure.microsoft.com/en-us/documentation/articles/machine-learning-use-sample-datasets/

datahub.io

aws.amazon.com/datasets

(2) Describe the dataset in words. Choose a dependent variable and explain what you are interested in about it.

(3) Use the apriori algorithm and theory to build a regression model for your dependent variable. Describe what you learned from your regression model.

This project will not be graded on what you find, but rather on how carefully you perform each step and how well you describe what you are doing. Make sure to use the techniques we discussed in lectures and readings. It is not necessary to use every single topic as if you were completing a checklist, but it is important to proceed thoughtfully and justify each step. When justifying "theory" steps, include a citation unless your step is just applying common sense.

Your printed midterm should be no more than 5 pages long, and no more than 2 pages of that should be your own text (the remainder R output). Please cut and paste from R into a word processor. Please use an 11-point monospaced font like Courier or Monaco for the R commands and output, and a different 11-point text font for your commentary.

It is okay to get help on this assignment as long as you discuss who gave you the help and what they said in your commentary. QAC tutoring hours can be found at:

www.wesleyan.edu/qac/tutoring.html