Economics 50/HKS SUP 135

Gregory Bruich, Ph.D.
Dept. of Economics, Harvard University

Lab 6: Predicting Social Mobility using Trees and Regression
[R Labs]

Methods/concepts: decision trees, prediction, in-sample vs. out of sample

LAB DESCRIPTION

This is the first of three labs on prediction policy questions. In this lab, you will predict upward mobility using *decision trees* and *multivariable regression*. The measure of upward mobility that we will focus on is **Statistic 1: Absolute Mobility at the 25th Percentile** in each commuting zone (**kfr_pooled_pooled_p25**). For more details on the variables included in these data, see Table 1.

The focus of this lab will be on the *concepts* and not the coding. In the R labs, we will start from starter scripts that can either be run on your computer or on the FAS On Demand server.

QUESTIONS

1. Why do we split our data into "test" and "training" datasets for prediction applications?

2. Now turn to the .R starter script and the mobility.dta data set. Just like in Lab 1 where we randomly assigned observations to treatment and control groups, in this week's lab we will randomly assign half of the data into a "test" and half into a "training" subsamples as follows:

    a. First, set the "seed" using your Harvard University ID number (`set.seed(12345)` in R), which will make your simulation reproducible, but different from your classmates' simulations. Then assign to each observation a random number drawn uniformly between 0 and 1.

    b. Generate a new variable *train_flag* that equals 1 ("training sample") if the number generated in part a is greater than or equal to 0.5, and otherwise equals 0 when the number is less than 0.5 ("test sample"). How many observations are in the treatment group? How many are in your control group?

3. Then subset the mobility.dta data set to create two new data frames:

    `train` is the training dataset containing observations where *train_flag=1*

    `test` is the test dataset containing observations where *train_flag=0*

4. Now we will use linear regression to predict upward mobility.
    a. Start by estimating a regression of **kfr_pooled_pooled_p25** on at least three predictor variables in the training data. Please choose at least three, but you can choose more than three if you want. The three that you choose should not include my two predictors: 'bowl_per_capita' and 'singleparent_share1990'. Pick your own!
    b. Use the estimated coefficients from the regression to predict **kfr_pooled_pooled_p25** for Milwaukee, WI, which has `cz == 24100`. For example, in a regression using bowling alleys per 100,000 residents and single parent families as predictors, we could simply use the estimated regression coefficients and the fact that Milwaukee, WI has 5.72 bowling alleys per 100,000 residents and 22.6% single parent families in 1990 to obtain the predicted value. What is the prediction error for Milwaukee?

c. Obtain predictions for the **training data** and create a new variable called `y_train_predictions_ols`.
d. Obtain predictions for the **test data** and create a new variable called `y_test_predictions_ols`.
e. Calculate the **root mean squared prediction error** in the training data and the test data
f. Compare the prediction error in the test vs. the training data. Which is higher?

5. Next we will use a decision tree to predict upward mobility.
   a. Estimate a decision tree to predict **kfr_pooled_pooled_p25** using the same predictor variables in the training data that you used for the regression.
   b. Visualize your decision tree in "tree form" and include your image in your solutions. Use the graphical representation of the decision tree to predict **kfr_pooled_pooled_p25** for Milwaukee, WI. What is the prediction error for Milwaukee?
   c. Obtain predictions for the **training data** and create a new variable called `y_train_predictions_tree`.
   d. Obtain predictions for the entire **test data** and create a new variable called `y_test_predictions_tree`.
   e. Calculate the **root mean squared prediction error** in the training data and the test data
   f. Compare the prediction error in the test vs. the training data. Which is higher?

6. To conclude this week's lab, we will illustrate the overfit problem. The key issue in using a decision tree to make predictions is choosing how big of a tree you want to grow. How many splits in the tree? Or in other words, the depth of tree.

   Decision trees have a tendency to overfit the training data. By growing a bigger and bigger tree, we can drive the in-sample prediction error down to zero. The tree that minimizes the in-sample error would be a tree where each observation is in its own leaf. But that large decision tree is not likely to do well when trying to make an out of sample prediction. As with regression, it's possible to fit our existing data perfectly but have terrible predictions for new data.

   To show this, fit a tree in R using `rpart()` with maximum depth `maxdepth = 30`, complexity parameter `cp = 0`, minimum number of observations in each leaf `minbucket = 1`, and the minimum number of observations in a leaf for a split to be attempted `minsplit = 1`. Calculate the **root mean squared prediction error** in the training data and the test data.

7. Which of the three models – the linear regression, the small decision tree, or the big decision tree – performs best on the training sample? What about the test sample?

8. The files to submit for this lab are:

   a. Your well annotated .R/.rmd file replicating all your analyses above (with enough comments that a principal investigator on a research project would be able to follow and understand what each step of the code is doing). Please submit this to Gradescope.

   b. A PDF version of the solutions to the above questions. For graphs, save them as .png files and insert them into the document. Please submit this file to the same gradescope assignment as the .R/.rmd file. (Please do not submit a word document: we can only read PDFs in gradescope. Using R Markdown is never required; but if you have chosen to use it, you can *knit* the file to generate the PDF).

DATA DESCRIPTION, FILE: mobility.dta

The data consist of $N = 741$ Commuting Zones. Commuting zones are geographical aggregations of counties that are similar to metro areas but cover the entire U.S., including rural areas. Commuting zones are meant to consist of local labor markets where people both live and work. For more details on the construction of the variables included in this data set, please see Chetty, Raj, John Friedman, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2018. "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility." NBER Working Paper No. 25147.

TABLE 1
Variable Definitions

	Variable (1)	Description (2)	Obs. (3)	Mean (4)	St. Dev. (5)	Min (6)	Max (7)
1	*cz*	Five-digit 1990 commuter zone code	741	n/a	n/a	n/a	n/a
2	*cz_name*	String variable consisting of the name of the commuting zone.	741	n/a	n/a	n/a	n/a
3	*kfr_pooled_pooled_p25*	Absolute Mobility at the 25th Percentile	741	42.99	5.994	23.33	66.63
4	*bowl_per_capita*	Bowling Alleys per 100,000 residents	741	3.928	5.661	0	70.50
5	*singleparent_share1990*	Share of Single-Headed Households with Children 1990	741	0.197	0.0503	0.0433	0.441
6	*singleparent_share2000*	Share of Single-Headed Households with Children 2000	741	0.268	0.0563	0.105	0.547
7	*singleparent_share2010*	*Share of Single-Headed Households with Children 2006-2010 ACS*	741	0.315	0.0687	0.109	0.573
8	*hhinc_mean2000*	Mean Household Income 2000	741	65,137	12,755	38,817	122,288
9	*mean_commutetime2000*	Average Commute Time of Working Adults in 2000	741	21.97	4.548	7.383	40.24
10	*frac_coll_plus2000*	Fraction of Residents w/ a College Degree or More in 2000	741	0.181	0.0639	0.0488	0.481
11	*frac_coll_plus2010*	Fraction of Residents w/ a College Degree or More in 2006-2010 ACS	741	0.204	0.0695	0.0764	0.481
12	*foreign_share2010*	Share of Population Born Outside the U.S. in 2006-2010 ACS	741	0.0517	0.0622	0.000817	0.722
13	*med_hhinc1990*	Median Household Income in 1990	741	24,973	6,371	12,097	51,112

14	*med_hhinc2016*	Median Household Income in 2016	741	48,983	10,936	26,645	103,043
15	*poor_share2010*	Share Below Poverty Line 2006-2010 ACS	741	0.160	0.0541	0.0500	0.442
16	*poor_share2000*	Share Below Poverty Line 2000	741	0.145	0.0569	0.0540	0.460
17	*poor_share1990*	Share Below Poverty Line 1990	741	0.165	0.0703	0.0515	0.505
18	*share_white2010*	Share White 2010	741	0.758	0.197	0.0286	0.986
19	*share_black2010*	Share Black 2010	741	0.0846	0.123	0.00151	0.697
20	*share_hisp2010*	Share Hispanic 2010	741	0.0999	0.145	0.00249	0.957
21	*share_asian2010*	Share Asian 2010	741	0.0132	0.0328	0.000414	0.428
22	*share_black2000*	Share Black 2000	741	0.0773	0.118	0	0.646
23	*share_white2000*	Share White 2000	741	0.795	0.188	0.0365	0.991
24	*share_hisp2000*	Share Hispanic 2000	741	0.0755	0.133	0.00186	0.948
25	*share_asian2000*	Share Asian 2000	741	0.0107	0.0313	0.000244	0.455
26	*gsmn_math_g3_2013*	Average School District Level Standardized Test Scores in 3rd Grade in 2013	741	3.190	0.652	-0.661	4.960
27	*rent_twobed2015*	Average Rent for Two-Bedroom Apartment in 2015	741	704.3	185.9	336.0	1,652
28	*traveltime15_2010*	Share of Working Adults w/ Commute Time of 15 Minutes Or Less in 2006-2010 ACS	741	0.450	0.143	0.152	0.991
29	*emp2000*	Employment Rate 2000	741	0.578	0.0639	0.323	0.756
30	*mail_return_rate2010*	Census Form Rate Return Rate 2010	741	79.82	5.205	47.80	88.98
31	*popdensity2010*	Population Density (per square mile) in 2010	741	109.4	284.0	0.106	5,636
32	*popdensity2000*	Population Density (per square mile) in 2000	741	101.3	271.0	0.0833	5,506
33	*job_growth_1990_2010*	Job Growth Rate 1990-2010	741	13.54	21.56	-36.52	148.2
34	*ann_avg_job_growth_2004_2013*	Average Annual Job Growth Rate 2004-2013	741	-0.001	0.0134	-0.0827	0.107
35	*job_density_2013*	Job Density (in square miles) in 2013	741	50.28	133.6	0.0425	2,595