lab_6

Lindsey Greenhill

3/31/2022

Question 1

We need to split our data into test and training datasets so we can develop a model using the training set and then test that model on the test set and see how well the model performs with out of sample predictions. This is a way to test if the model is overfit to the data used to develop it.

Question 2

There are 378 observations in the training set (treatament) and 363 in the test set (control).

Question 3

Question 4

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Thu, Mar 31, 2022 - 14:40:20

Table 1:

	Dependent variable:	
	kfr_pooled_pooled_p25	
	(1)	(2)
bowl_per_capita	0.369***	0.361***
	(0.035)	(0.035)
singleparent_share1990	−65.535***	−65.237***
	(4.246)	(4.276)
frac_coll_plus2000	7.717***	
	(2.979)	
Constant	52.840***	54.239***
	(1.058)	(0.916)
Observations	378	378
R^2	0.596	0.589
Adjusted R^2	0.593	0.587
Residual Std. Error	3.781 (df = 374)	3.809 (df = 375)
F Statistic	184.155*** (df = 3; 374)	268.786*** (df = 2; 375)
Note:		*p<0.1; **p<0.05; ***p<0.01

Part a

The table above shows the regression results for the two variable regression in the starter code and a modified regression with 3 variables.

Part b (check this)

- Using theregression coefficients from the 3 variable model, we can predict the upward mobility rate in Milwaukee, WI using the equation y = 52.84 + .369bowl_per_capita -65.535singleparent_share1990 + 7.717frac_coll_plus2000. Using this equation, we can predict that Milwaukee has an upward mobility rate of **42.0697047**.
- To calculate the prediction error, we subtract the prediction calculated above from the actual value of Milwaukee's kfr_pooled_pooled_25 variable (which is 38.88789). The prediction error = **3.1818147**

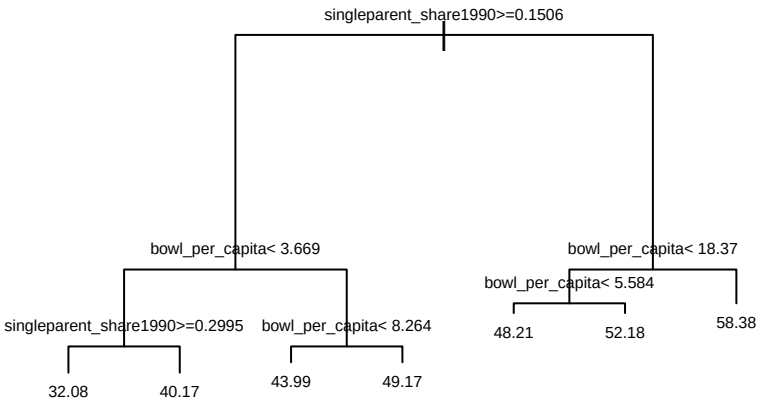
Part c, d, e, f

- The root mean squared prediction error for the test data = 3.966274.
- The root mean squared prediction error for the train data = 3.7604983.
- The rmspe for the test data is higher than the rmspe for the train data.

Question 5

Part b

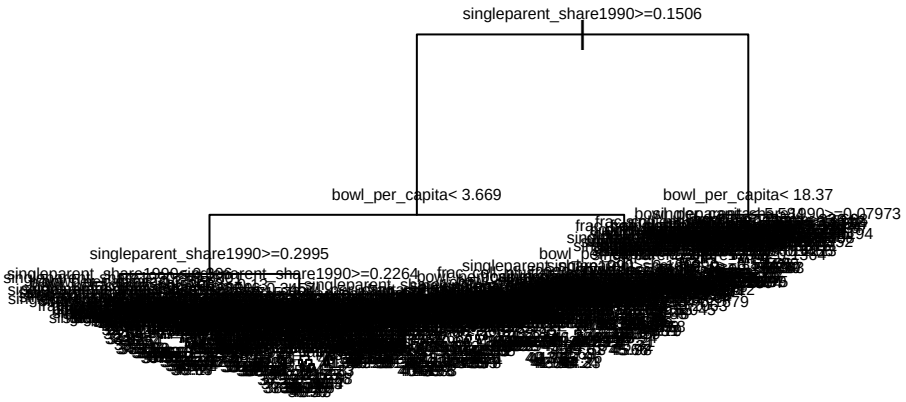
To get our answer for Milwaulkee, we can go down the tree and follow the branches based on the values for different variables. Single parent share is > .15, so we go to the right of the tree, then, bowl_per_capita is less than 18.37 and greater than 5.584, so we end up at 52.18 as a prediction. The actual kfr_pooled_pooled_p25 = 38.88. As such, the prediction error = 52.18 - 38.88 = 13.3



Parts c, d, e, f

- The rmspe for the test data = 4.3012301
- The rmspe for the train data = 3.5943684
- The rmspe is higher for the test data than for the training data.

Question 6



- The rmspe for the test data = 4.8290559
- The rmspe for the training data = 0.
- Obviously, the rmspe is larger for the test data than the training data.

Question 7

- Training sample: When comparing the rmspe for the three models on the training data set, The big decision tree performs best followed by the small decision tree and the regression, respectively.
- Test sampleL when comparing the rmspe for the three models on the test data set,