STATE DATA REVISITED

We will be revisiting the "state" dataset from Week 2. Recall that this dataset has, for each of the fifty U.S. states, the population, per capita income, illiteracy rate, murder rate, high school graduation rate, average number of frost days, area, latitude and longitude, division the state belongs to, region the state belongs to, and two-letter abbreviation. This dataset comes from the U.S. Department of Commerce, Bureau of the Census.

Load the dataset into R and convert it to a data frame by running the following two commands in R:

```
data(state)
statedata = data.frame(state.x77)
```

Inspect the data set using the command:

```
str(statedata)
```

We will try to build a model for life expectancy using regression trees, and employ cross-validation to improve our tree's performance.

PROBLEM 1.1 - LINEAR REGRESSION MODELS (1 point possible)

Let's recreate the **linear regression** models we made in the previous homework question. First, predict *Life.Exp* using all of the other variables as the independent variables (*Population, Income, Illiteracy, Murder, HS.Grad, Frost, Area*). Use the entire dataset to build the model.

What is the **adjusted** R-squared of the model?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.2 - LINEAR REGRESSION MODELS (1 point possible)

Calculate the sum of squared errors (SSE) between the predicted life expectancies using this model and the actual life expectancies:

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.3 - LINEAR REGRESSION MODELS (1 point possible)

Build a second **linear regression** model using just *Population, Murder, Frost, and HS.Grad* as independent variables (the best 4 variable model from the previous homework). What is the **adjusted** R-squared for this model?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.4 - LINEAR REGRESSION MODELS (1 point possible)

Calculate the sum of squared errors again, using this reduced model:

Show Answer *You have used 0 of 3 submissions*

PROBLEM 1.5 - LINEAR REGRESSION MODELS (1 point possible)

Which of the following is correct?

- ○ Trying different combinations of variables in linear regression is like trying different numbers of splits in a tree - this controls the complexity of the model.
- ○ Using many variables in a linear regression is **always** better than using just a few.
- ○ The variables we removed were uncorrelated with *Life.Exp*

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.1 - CART MODELS (1 point possible)

Let's now build a **CART model** to predict *Life.Exp* using all of the other variables as independent variables (*Population, Income, Illiteracy, Murder, HS.Grad, Frost, Area*). We'll use the default *minbucket* parameter, so don't add the *minbucket* argument. Remember that in this problem we are not as interested in *predicting* life expectancies for new observations as we are understanding how they relate to the other variables we have, so we'll use all of the data to build our model. You shouldn't use the method="class" argument since this is a regression tree.

Plot the tree. Which of these variables appear in the tree?

- ☐ Population
- ☐ Murder
- ☐ Frost
- ☐ HS.Grad
- ☐ Area

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.2 - CART MODELS (1 point possible)

Use the regression tree you just built to predict life expectancies (using the predict function), and calculate the sum-of-squared-errors (SSE) like you did for linear regression. What is the SSE?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.3 - CART MODELS (1 point possible)

The error is higher than for the linear regression models. One reason might be that we haven't made the tree big enough. Set the *minbucket* parameter to 5, and recreate the tree.

Which variables appear in this new tree?

- ☐ Population
- ☐ Murder
- ☐ Frost
- ☐ HS.Grad
- ☐ Area

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.4 - CART MODELS (1 point possible)

Do you think the default minbucket parameter is smaller or larger than 5 based on the tree that was built?

- ○ Smaller
- ○ Larger

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.5 - CART MODELS (1 point possible)

What is the SSE of this tree?

This is much closer to the linear regression model's error. By changing the parameters we have improved the fit of our model.

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.6 - CART MODELS (1 point possible)

Can we do even better? Create a tree that predicts *Life.Exp* using **only** *Area*, with the *minbucket* parameter to 1. What is the SSE of this newest tree?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.7 - CART MODELS (1 point possible)

This is the lowest error we have seen so far. What would be the best interpretation of this result?

- ○ Trees are much better than linear regression for this problem because they can capture nonlinearities that linear regression misses.
- ○ We can build almost perfect models given the right parameters, even if they violate our intuition of what a good model should be.
- ○ Area is obviously a very meaningful predictor of life expectancy, given we were able to get such low error using just Area as our independent variable.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.1 - CROSS-VALIDATION (1 point possible)

Adjusting the variables included in a linear regression model is a form of model tuning. In Problem 1 we showed that by removing variables in our linear regression model (tuning the model), we were able to maintain the fit of the model while using a simpler model. A rule of thumb is that simpler models are more interpretable and generalizeable. We will now tune our regression tree to see if we can improve the fit of our tree while keeping it as simple as possible.

Load the *caret* library, and set the seed to 111. Set up the controls exactly like we did in the lecture (10-fold cross-validation) with *cp* varying over the range 0.01 to 0.50 in increments of 0.01. Use the *train* function to determine the best *cp* value. What value of cp does the train function recommend? (Remember that the train function tells you to pick the largest value of cp with the lowest error when there are ties, and explains this at the bottom of the output.)

Show Answer *You have used 0 of 4 submissions*

PROBLEM 3.2 - CROSS-VALIDATION (2 points possible)

Create a tree with this value of *cp*. You'll notice that this is actually quite similar to the first tree we created with the initial model. Interpret the tree: we predict the life expectancy to be 70 if the murder rate is greater than or equal to

and is less than

Show Answer *You have used 0 of 4 submissions*

PROBLEM 3.3 - CROSS-VALIDATION (1 point possible)

Calculate the SSE of this tree:

Show Answer *You have used 0 of 3 submissions*

PROBLEM 3.4 - CROSS-VALIDATION (1 point possible)

Recall the first tree (default parameters), second tree (minbucket = 5), and the third tree (selected with cross validation) we made. Given what you have learned about cross-validation, which of the three models would you expect to be better if we did use it for prediction on a test set? For this question, suppose we had actually set aside a few observations (states) in a test set, and we want to make predictions on those states.

- The first model
- The second model
- The model we just made with the "best" cp

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.5 - CROSS-VALIDATION (1 point possible)

At the end of Part 2 we made a very complex tree using just Area. Use *train* with the same parameters as before but just using Area as an independent variable to find the best cp value (set the seed to 111 first). Then build a new tree using just Area and this value of cp.

How many splits does the tree have?

Show Answer *You have used 0 of 4 submissions*

PROBLEM 3.6 - CROSS-VALIDATION (2 points possible)

The lower left leaf (or bucket) corresponds to the lowest predicted Life.Exp, (70). Observations in this leaf correspond to states with area greater than

and area less than

Show Answer *You have used 0 of 4 submissions*

PROBLEM 3.7 - CROSS-VALIDATION (1 point possible)

We have simplified the previous "Area tree" considerably by using cross-validation. Calculate the SSE of the cross-validated "Area tree", and select the correct statements:

- ☐ The best model in this whole question is the first "Area tree" because it had the lowest SSE.
- ☐ The Area variable is not as predictive as Murder rate.
- ☐ Cross-validation is intended to decrease the SSE for a model on the training data, compared to a tree that isn't cross-validated.
- ☐ Cross-validation will always improve the SSE of a model on unseen data, compared to a tree that isn't cross-validated.

Show Answer *You have used 0 of 2 submissions*

Please remember not to ask for or post complete answers to homework questions in this discussion forum.

Show Discussion

New Post

MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.

(https://twitter.com/edXOnline)

(https://plus.google.com/108235383044095082

(http://youtube.com/user/edxonline)