STATE DATA

We often take data for granted. However, one of the hardest parts about analyzing a problem you're interested in can be to find good data to answer the questions you want to ask. As you're learning R, though, there are many datasets that R has built in that you can take advantage of.

In this problem, we will be examining the "state" dataset, which has data from the 1970s on all fifty US states. For each state, the dataset includes 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.

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

```
data(state)

statedata = cbind(data.frame(state.x77), state.abb, state.area, state.center,  state.division, state.name, state.region)
```

Inspect the data set using the command: str(statedata)

For more information about this data set, type ?state in the R console.

PROBLEM 1.1 - DATA EXPLORATION (1 point possible)

We begin by exploring the data by examining the latitude and longitude of each state. Plot all of the states' centers with latitude on the y axis (the "y" variable in our dataset) and longitude on the x axis (the "x" variable in our dataset). The shape of the plot should be the familiar outline of the United States! Note that Alaska and Hawaii have had their coordinates adjusted to appear just off of the west coast.

In the R command you used to generate this plot, which variable name did you use as the first argument?

- statedata$y
- statedata$x
- I used a different variable name.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 1.2 - DATA EXPLORATION (1 point possible)

Using the tapply command, determine which region of the US (West, North Central, South, or Northeast) has the highest average high school graduation rate of all the states in the region:

- West
- North Central
- South

- ○ Northeast

Show Answer *You have used 0 of 1 submissions*

PROBLEM 1.3 - DATA EXPLORATION (1 point possible)

Now, make a boxplot of the murder rate by region (for more information about creating boxplots in R, type ?boxplot in your console).

Which region has the highest median murder rate?

- ○ Northeast
- ○ South
- ○ North Central
- ○ West

Show Answer *You have used 0 of 1 submissions*

PROBLEM 1.4 - DATA EXPLORATION (1 point possible)

You should see that there is an outlier in the Northeast region of the boxplot you just generated. Which state does this correspond to? (Hint: There are many ways to find the answer to this question, but one way is to use the subset command to only look at the Northeast data.)

- ○ Delaware
- ○ Rhode Island
- ○ Maine
- ○ New York

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.1 - PREDICTING LIFE EXPECTANCY - AN INITIAL MODEL (1 point possible)

We would like to build a model to predict life expectancy by state using the state statistics we have in our dataset.

Build the model with all potential variables included (Population, Income, Illiteracy, Murder, HS.Grad, Frost, and Area). Note that you should use the variable "Area" in your model, NOT the variable "state.area".

What is coefficient for income?

Show Answer *You have used 0 of 3 submissions*

PROBLEM 2.2 - PREDICTING LIFE EXPECTANCY - AN INITIAL MODEL (1 point possible)

Call the coefficient for income x (the answer to Problem 2.1). What is the interpretation of the coefficient x?

- ○ For a one unit increase in income, predicted life expectancy increases by |x|

- For a one unit increase in income, predicted life expectancy decreases by |x|
- For a one unit increase in predicted life expectancy, income decreases by |x|
- For a one unit increase in predicted life expectancy, income increases by |x|

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.3 - PREDICTING LIFE EXPECTANCY - AN INITIAL MODEL (1 point possible)

Now plot a graph of life expectancy vs. income using the command:

plot(statedata$Income, statedata$Life.Exp)

Visually observe the plot. What appears to be the relationship?

- Life expectancy is somewhat positively correlated with income.
- Life expectancy is somewhat negatively correlated with income.
- Life expectancy is not correlated with income.

Show Answer *You have used 0 of 1 submissions*

PROBLEM 2.4 - PREDICTING LIFE EXPECTANCY - AN INITIAL MODEL (1 point possible)

The model we built does not display the relationship we saw from the plot of life expectancy vs. income. Which of the following explanations seems the most reasonable?

- Income is not related to life expectancy.
- Multicollinearity

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.1 - PREDICTING LIFE EXPECTANCY - REFINING THE MODEL AND ANALYZING PREDICTIONS

(1 point possible)
Recall that we discussed the principle of simplicity: that is, a model with fewer variables is preferable to a model with many unnnecessary variables. Experiment with removing independent variables from the original model. Remember to use the significance of the coefficients to decide which variables to remove (remove the one with the largest "p-value" first, or the one with the "t value" closest to zero), and to remove them one at a time (this is called "backwards variable selection"). This is important due to multicollinearity issues - removing one insignificant variable may make another previously insignificant variable become significant.

You should be able to find a good model with only 4 independent variables, instead of the original 7. Which variables does this model contain?

- Income, HS.Grad, Frost, Murder
- HS.Grad, Population, Income, Frost
- Frost, Murder, HS.Grad, Illiteracy
- Population, Murder, Frost, HS.Grad

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.2 - PREDICTING LIFE EXPECTANCY - REFINING THE MODEL AND ANALYZING PREDICTIONS

(1 point possible)
Removing insignificant variables changes the Multiple R-squared value of the model. By looking at the summary output for both the initial model (all independent variables) and the simplified model (only 4 independent variables) and using what you learned in class, which of the following correctly explains the change in the Multiple R-squared value?

- We expect the "Multiple R-squared" value of the simplified model to be slightly worse than that of the initial model. It can't be better than the "Multiple R-squared" value of the initial model.
- We expect the "Multiple R-squared" value of the simplified model to be slightly better than that of the initial model. It can't be worse than the "Multiple R-squared" value of the initial model.
- We expect the "Multiple R-squared" of the simplified model to be about the same as the intial model (we have no way of knowing if it will be slightly worse or slightly better than the Multiple R-squared of the intial model).

Show Answer *You have used 0 of 2 submissions*

PROBLEM 3.3 - PREDICTING LIFE EXPECTANCY - REFINING THE MODEL AND ANALYZING PREDICTIONS

(2 points possible)
Using the simplified 4 variable model that we created, we'll now take a look at how our predictions compare to the actual values.

Take a look at the vector of predictions by using the predict function (since we are just looking at predictions on the training set, you don't need to pass a "newdata" argument to the predict function).

Which state do we predict to have the lowest life expectancy? (Hint: use the sort function)

- South Carolina
- Mississippi
- Alabama
- Georgia

Which state actually has the lowest life expectancy? (Hint: use the which.min function)

- South Carolina
- Mississippi
- Alabama
- Georgia

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.4 - PREDICTING LIFE EXPECTANCY - REFINING THE MODEL AND ANALYZING PREDICTIONS

(2 points possible)
Which state do we predict to have the highest life expectancy?

- Massachusetts
- Maine
- Washington
- Hawaii

Which state actually has the highest life expectancy?

- Massachusetts
- Maine

- Washington
- Hawaii

Show Answer *You have used 0 of 1 submissions*

PROBLEM 3.5 - PREDICTING LIFE EXPECTANCY - REFINING THE MODEL AND ANALYZING PREDICTIONS

(2 points possible)
Take a look at the vector of residuals (the difference between the predicted and actual values).

For which state do we make the smallest absolute error?

- Maine
- Florida
- Indiana
- Illinois

For which state do we make the largest absolute error?

- Hawaii
- Maine
- Texas
- South Carolina

Show Answer *You have used 0 of 1 submissions*

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

Show Discussion

New Post

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