State Data

Quang Duong

2017-06-26

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

```
data(state)
statedata <- cbind(data.frame(state.x77), state.abb, state.area, state.center,  state.division, state.na
```

Inspect the data set

```
str(statedata)
```

```
'data.frame':   50 obs. of  15 variables:
 $ Population    : num  3615 365 2212 2110 21198 ...
 $ Income        : num  3624 6315 4530 3378 5114 ...
 $ Illiteracy    : num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
 $ Life.Exp      : num  69 69.3 70.5 70.7 71.7 ...
 $ Murder        : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
 $ HS.Grad       : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
 $ Frost         : num  20 152 15 65 20 166 139 103 11 60 ...
 $ Area          : num  50708 566432 113417 51945 156361 ...
 $ state.abb     : Factor w/ 50 levels "AK","AL","AR",..: 2 1 4 3 5 6 7 8 9 10 ...
 $ state.area    : num  51609 589757 113909 53104 158693 ...
 $ x             : num  -86.8 -127.2 -111.6 -92.3 -119.8 ...
 $ y             : num  32.6 49.2 34.2 34.7 36.5 ...
 $ state.division: Factor w/ 9 levels "New England",..: 4 9 8 5 9 8 1 3 3 3 ...
 $ state.name    : Factor w/ 50 levels "Alabama","Alaska",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ state.region  : Factor w/ 4 levels "Northeast","South",..: 2 4 4 2 4 4 1 2 2 2 ...
```

```
summary(statedata)
```

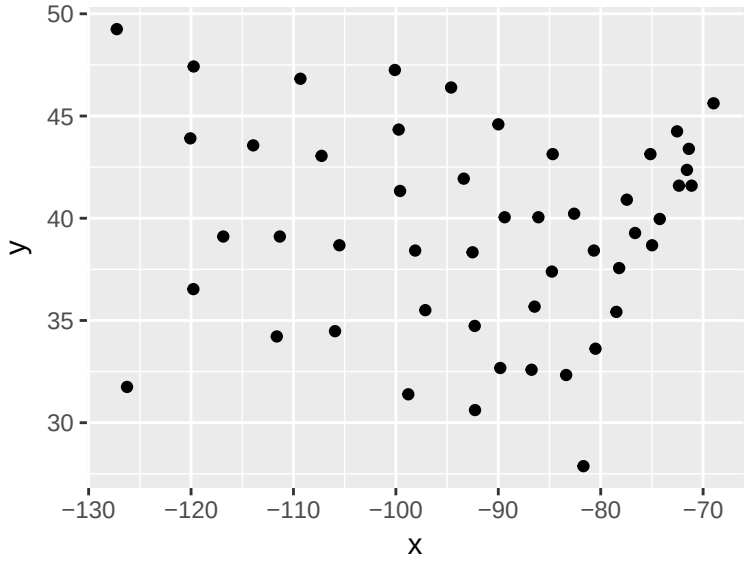
```
   Population        Income       Illiteracy       Life.Exp
 Min.   :  365   Min.   :3098   Min.   :0.500   Min.   :67.96
 1st Qu.: 1080   1st Qu.:3993   1st Qu.:0.625   1st Qu.:70.12
 Median : 2838   Median :4519   Median :0.950   Median :70.67
 Mean   : 4246   Mean   :4436   Mean   :1.170   Mean   :70.88
 3rd Qu.: 4968   3rd Qu.:4814   3rd Qu.:1.575   3rd Qu.:71.89
 Max.   :21198   Max.   :6315   Max.   :2.800   Max.   :73.60

     Murder          HS.Grad          Frost             Area
 Min.   : 1.400   Min.   :37.80   Min.   :  0.00   Min.   :  1049
 1st Qu.: 4.350   1st Qu.:48.05   1st Qu.: 66.25   1st Qu.: 36985
 Median : 6.850   Median :53.25   Median :114.50   Median : 54277
 Mean   : 7.378   Mean   :53.11   Mean   :104.46   Mean   : 70736
 3rd Qu.:10.675   3rd Qu.:59.15   3rd Qu.:139.75   3rd Qu.: 81163
```

```
 Max.   :15.100   Max.   :67.30   Max.   :188.00   Max.   :566432

  state.abb    state.area          x                y
 AK     : 1   Min.   :  1214   Min.   :-127.25   Min.   :27.87
 AL     : 1   1st Qu.: 37317   1st Qu.:-104.16   1st Qu.:35.55
 AR     : 1   Median : 56222   Median : -89.90   Median :39.62
 AZ     : 1   Mean   : 72368   Mean   : -92.46   Mean   :39.41
 CA     : 1   3rd Qu.: 83234   3rd Qu.: -78.98   3rd Qu.:43.14
 CO     : 1   Max.   :589757   Max.   : -68.98   Max.   :49.25
 (Other):44
           state.division      state.name        state.region
 South Atlantic    : 8     Alabama   : 1   Northeast    : 9
 Mountain          : 8     Alaska    : 1   South        :16
 West North Central: 7     Arizona   : 1   North Central:12
 New England       : 6     Arkansas  : 1   West         :13
 East North Central: 5     California: 1
 Pacific           : 5     Colorado  : 1
 (Other)           :11     (Other)   :44
```

Data Exploration

We begin by exploring the data. 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 look like the outline of the United States! Note that Alaska and Hawaii have had their coordinates adjusted to appear just off of the west coast.

```
library(ggplot2)
ggplot(statedata) + geom_point(aes(x,y))
```



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

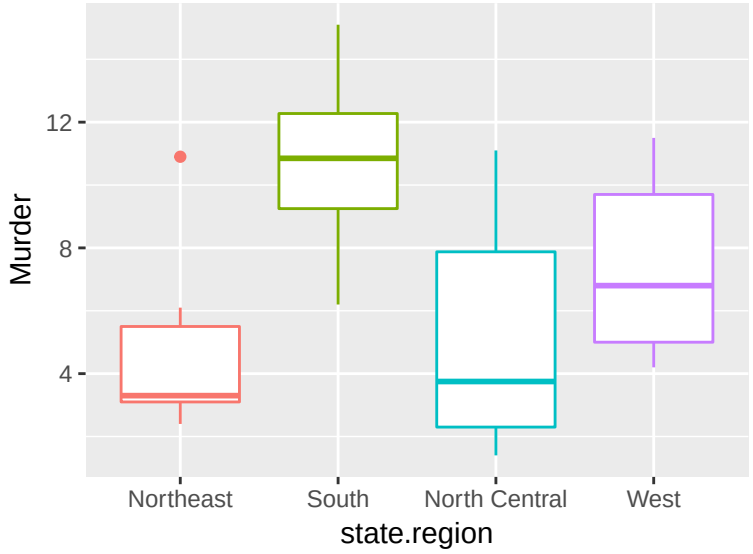
```
tapply(statedata$HS.Grad, statedata$state.region, mean)
```

```
    Northeast         South North Central          West
     53.96667      44.34375      54.51667      62.00000
```

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?

```
ggplot(statedata) + geom_boxplot(aes(state.region, Murder, color=state.region)) + guides(color=FALSE)
```



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.)

```
subset(statedata, state.region=='Northeast')['Murder']
```

```
              Murder
Connecticut      3.1
Maine            2.7
Massachusetts    3.3
New Hampshire    3.3
New Jersey       5.2
New York        10.9
Pennsylvania     6.1
Rhode Island     2.4
Vermont          5.5
```

Predicting Life Expectancy - Initial Model

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".

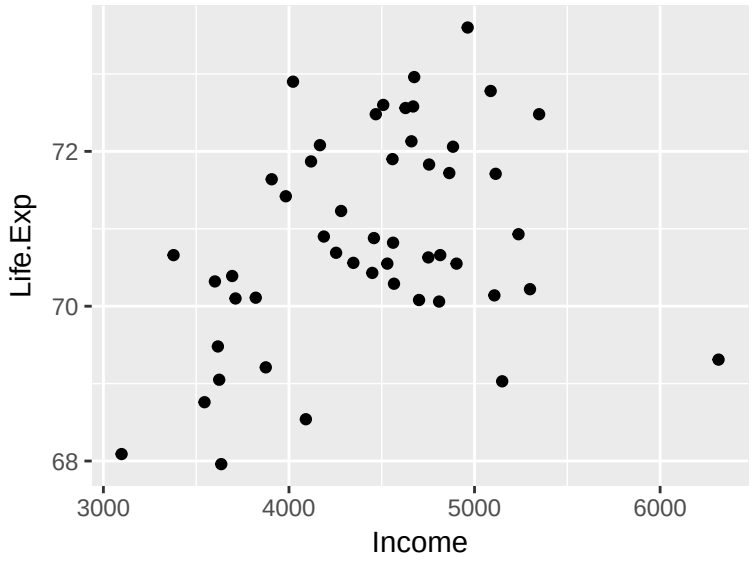
```
attach(statedata)
lifeReg <- lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area, data = sta
summary(lifeReg)
```

```
Call:
```

```
lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +
    HS.Grad + Frost + Area, data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.48895 -0.51232 -0.02747  0.57002  1.49447

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
Population   5.180e-05  2.919e-05   1.775   0.0832 .
Income      -2.180e-05  2.444e-04  -0.089   0.9293
Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
Area        -7.383e-08  1.668e-06  -0.044   0.9649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7448 on 42 degrees of freedom
Multiple R-squared:  0.7362,    Adjusted R-squared:  0.6922
F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

Now plot a graph of life expectancy vs. income.

```
ggplot(statedata) + geom_point(aes(Income, Life.Exp))
```



Visually observe the plot. It is appear that life expectancy is somewhat positively correlated with income. The model we built does not display the relationship we saw from the plot of life expectancy vs. income. Multicollinearity might be an reasonable explanation for this fact.

Predicting Life Expectancy - Refining the Model and Analyzing Predictions

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.

```
lifeReg2 <- lm(Life.Exp ~ Population + Murder + HS.Grad + Frost, data = statedata)
summary(lifeReg2)
```

```
Call:
lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
    data = statedata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47095 -0.53464 -0.03701  0.57621  1.50683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
Population   5.014e-05  2.512e-05   1.996  0.05201 .
Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7197 on 45 degrees of freedom
Multiple R-squared:  0.736, Adjusted R-squared:  0.7126
F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

Removing insignificant variables changes the Multiple R-squared value of the model. 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.

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). Observe the difference between our prediction and the actual values.

```
lifePredic <- predict(lifeReg2)
statedata$state.name[which.min(lifePredic)]
statedata$state.name[which.min(statedata$Life.Exp)]
```

```
[1] Alabama
50 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
[1] South Carolina
50 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```
lifePredic <- predict(lifeReg2)
statedata$state.name[which.max(lifePredic)]
statedata$state.name[which.max(statedata$Life.Exp)]
```

```
[1] Washington
50 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
[1] Hawaii
```

```
50 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```
statedata$state.name[which.min(abs(lifeReg2$residuals))]
statedata$state.name[which.max(abs(lifeReg2$residuals))]
```

```
[1] Indiana
50 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
[1] Hawaii
50 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```