# State Data (OPTIONAL)

**IMPORTANT NOTE: This problem is optional, and will not count towards your grade. We have created this problem to give you extra practice with the topics covered in this unit.**

## State Data (OPTIONAL)

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)

If you can't access the state dataset in R, here is a CSV file with the same data that you can load into R using the read.csv function: statedata.csv

After you have loaded the data into R, inspect the data set using the command: str(statedata)

This dataset has 50 observations (one for each US state) and the following 15 variables:

- **Population** - the population estimate of the state in 1975
- **Income** - per capita income in 1974
- **Illiteracy** - illiteracy rates in 1970, as a percent of the population
- **Life.Exp** - the life expectancy in years of residents of the state in 1970
- **Murder** - the murder and non-negligent manslaughter rate per 100,000 population in 1976
- **HS.Grad** - percent of high-school graduates in 1970
- **Frost** - the mean number of days with minimum temperature below freezing from 1931–1960 in the capital or a large city of the state
- **Area** - the land area (in square miles) of the state
- **state.abb** - a 2-letter abreviation for each state
- **state.area** - the area of each state, in square miles
- **x** - the longitude of the center of the state
- **y** - the latitude of the center of the state
- **state.division** - the division each state belongs to (New England, Middle Atlantic, South Atlantic, East South Central, West South Central, East North Central, West North Central, Mountain, or Pacific)
- **state.name** - the full names of each state
- **state.region** - the region each state belong to (Northeast, South, North Central, or West)

---

## Problem 1.1 - Data Exploration

0 points possible (ungraded)

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.

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.

Explanation
To generate the described plot, you should type plot(statedata$x, statedata$y) in your R console. The first variable here is statedata$x.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 1.2 - Data Exploration

0 points possible (ungraded)
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

Explanation
You can find the average high school graduation rate of all states in each of the regions by typing the following command in your R console:
tapply(statedata$HS.Grad, statedata$state.region, mean)
The highest value is for the West region.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 1.3 - Data Exploration

0 points possible (ungraded)
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

Calculator

North Central

West

Explanation
To generate the boxplot, you should type boxplot(statedata$Murder ~ statedata$state.region) in your R console. You can see that the region with the highest median murder rate (the one with the highest solid line in the box) is the South.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 1.4 - Data Exploration

0 points possible (ungraded)
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 ✔

Explanation
The correct answer is New York. If you first use the subset command:
NortheastData = subset(statedata, state.region == "Northeast")
You can then look at NortheastData$Murder together with NortheastData$state.abb to identify the outlier.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 2.1 - Predicting Life Expectancy - An Initial Model

0 points possible (ungraded)
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 the coefficient for "Income" in your linear regression model?

Answer: -0.0000218

Calculator

Explanation
You can build the linear regression model with the following command:
LinReg = lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area, data=statedata)
Then, to find the coefficient for income, you can look at the summary of the regression with summary(LinReg).

Submit You have used 0 of 3 attempts

Answers are displayed within the problem

## Problem 2.2 - Predicting Life Expectancy - An Initial Model

0 points possible (ungraded)
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|

Explanation
If we increase income by one unit, then our model's prediction will increase by the coefficient of income, x. Because x is negative, this is the same as predicted life expectancy decreasing by |x|.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 2.3 - Predicting Life Expectancy - An Initial Model

0 points possible (ungraded)
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.

Explanation
Although the point in the lower right hand corner of the plot appears to be an outlier, we observe a positive linear relationship in the plot.

Submit You have used 0 of 1 attempt

Calculator

Answers are displayed within the problem

## Problem 2.4 - Predicting Life Expectancy - An Initial Model

0 points possible (ungraded)
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 ✓

Explanation
Although income is an insignificant variable in the model, this does not mean that there is no association between income and life expectancy. However, in the presence of all of the other variables, income does not add statistically significant explanatory power to the model. This means that multicollinearity is probably the issue.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 3.1 - Predicting Life Expectancy - Refining the Model and Analyzing Predictions

0 points possible (ungraded)
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 ✓

Explanation
We would eliminate the variable "Area" first (since it has the highest p-value, or probability, with a value of 0.9649), by adjusting our lm command to the following:
LinReg = lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost, data=statedata)
Looking at summary(LinReg) now, we would choose to eliminate "Illiteracy" since it now has the highest p-value of 0.9340, using the following command:
LinReg = lm(Life.Exp ~ Population + Income + Murder + HS.Grad + Frost, data=statedata)
Looking at summary(LinReg) again, we would next choose to eliminate "Income", since it has a p-value of 0.9153. This gives the following four variable model:

Calculator

LinReg = lm(Life.Exp ~ Population + Murder + HS.Grad + Frost, data=statedata)
This model with 4 variables is a good model. However, we can see that the variable "Population" is not quite significant. In practice, it would be up to you whether or not to keep the variable "Population" or eliminate it for a 3-variable model. Population does not add much statistical significance in the presence of murder, high school graduation rate, and frost days. However, for the remainder of this question, we will analyze the 4-variable model.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 3.2 - Predicting Life Expectancy - Refining the Model and Analyzing Predictions

0 points possible (ungraded)
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).

Explanation
When we remove insignificant variables, the "Multiple R-squared" will always be worse, but only slightly worse. This is due to the nature of a linear regression model. It is always possible for the regression model to make a coefficient zero, which would be the same as removing the variable from the model. The fact that the coefficient is not zero in the intial model means it must be helping the R-squared value, even if it is only a very small improvement. So when we force the variable to be removed, it will decrease the R-squared a little bit. However, this small decrease is worth it to have a simpler model.
On the contrary, when we remove insignificant variables, the "Adjusted R-squred" will frequently be better. This value accounts for the complexity of the model, and thus tends to increase as insignificant variables are removed, and decrease as insignificant variables are added.

Submit You have used 0 of 2 attempts

Answers are displayed within the problem

## Problem 3.3 - Predicting Life Expectancy - Refining the Model and Analyzing Predictions

0 points possible (ungraded)
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).

Calculator

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

- South Carolina
- Mississippi
- Alabama ✔
- Georgia

Explanation
If your simplified 4-variable model is called "LinReg", you can answer this question by typing
sort(predict(LinReg))
in your R console. The first state listed has the lowest predicted life expectancy, which is Alabama.

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

- South Carolina ✔
- Mississippi
- Alabama
- Georgia

Explanation
You can find the row number of the state with the lowest life expectancy by typing
which.min(statedata$Life.Exp) into your R console. This returns 40. The 40th state name in the vector statedata$state.name is South Carolina.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 3.4 - Predicting Life Expectancy - Refining the Model and Analyzing Predictions

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

- Massachusetts
- Maine
- Washington ✔
- Hawaii

Explanation
If your simplified 4-variable model is called "LinReg", you can answer this question by typing
"sort(predict(LinReg))" in your R console. The last state listed has the highest predicted life expectancy, \

Calculator

is Washington.

Which state actually has the highest life expectancy?

- Massachusetts
- Maine
- Washington
- Hawaii ✔

Explanation
You can find the row number of the state with the highest life expectancy by typing which.max(statedata$Life.Exp) into your R console. This returns 11. The 11th state name in the vector statedata$state.name is Hawaii.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

## Problem 3.5 - Predicting Life Expectancy - Refining the Model and Analyzing Predictions

0 points possible (ungraded)
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

Explanation
You can look at the sorted list of absolute errors by typing
sort(abs(model$residuals))
into your R console (where "model" is the name of your model). Alternatively, you can compute the residuals manually by typing
sort(abs(statedata$Life.Exp - predict(model)))
in your R console. The smallest absolute error is for Indiana.

For which state do we make the largest absolute error?

- Hawaii ✔
- Maine
- Texas

Calculator

South Carolina

Explanation
You can look at the sorted list of absolute errors by typing
sort(abs(model$residuals))
into your R console (where "model" is the name of your model). Alternatively, you can compute the residuals manually by typing
sort(abs(statedata$Life.Exp - predict(model)))
in your R console. The largest absolute error is for Hawaii.

Submit You have used 0 of 1 attempt

Answers are displayed within the problem

< Previous Next >

Calculator