ES 207 Spring 2016

Homework Lab 3: Pre-Regression

Updated: February, 2016

Homework Exercise (3 Points total)

These exercises rely on the lab data found in the resource section of UCMCROPS, `riparian_survey.csv` which you worked with in the last lab.

Your document should have the following sections, and provide written explanations formatted in RMarkdown that explains your code, output and graphics in the following format:

NAME
CLASS
DATE

Homework Assignment 3

Objective Statement: [What are you trying to accomplish?]

Methods: [In general terms, what analyses are you doing?]
- **Data**: [What are the data and where did they come from?]
- **Code**: [In specific terms, what is the code that was used to conduct the analysis?]

Results: [What do the results show? Numerical evidence and graphic evidence are required.]

Discussion: [What do the results mean?]

Limitations: [What are the limitations, caveats, and assumptions of the analysis?]

OBJECTIVE

You have been asked to analyze a dataset of field measurements and observations taken at various project sites throughout northern California that are intended to estimate aboveground carbon stocks in riparian areas. Some of these project sites are likely more productive than others, and you are being asked to determine which sites have more carbon stocks than the others. While carbon (C) is a function of biomass (roughly 50%), we cannot measure biomass directly without destructive sampling. We can, however, estimate volumes based on a few structural dimensions. As an exercise in data analysis, you decided that you could extend the report from last week (Homework 2) that tested the assumption that the project sites are independent of each other in the frequency of trees present in the sample plots. There were a bunch of different species, so you decided that a subsample of the most frequently occurring genera and selected the top four most frequent observations (i.e., *Populus*, *Salix*, *Quercus*, *Acer*).

To extend your knowledge on this issue, and in order to make reasonable estimates of standing carbon, you decide to develop a linear model that relates tree height as a function of its diameter (at breast height) or dbh. However, you have a hunch this linear model (i.e., regression) could vary by project site or by genus. In this exercise, you build a new report showing which, if any, statistical differences can be determined for the general function of `riparian$ht ~ riparian$dbh` as determined by project site and/or genus.

Step 1 - Exploratory Data Analysis

EDA Refresher

Load your output from last homework (refer to step 5 of homework 2). Re-examine columns (variables) using `str(), head(), tail()`. This should refresh your memory on these data. Take a moment to examine *ProjCode* and *Genus* using `levels()`. Develop exploratory data analysis graphs, using `boxplot()`, to show how height varies by project site and genus, separately, and via their interaction using `":"` within the formula. For example `boxplot(rip$ht ~ rip$ProjCode:rip$Genus)`. What trends do you observe? Similarly, look at the histograms and density plots of tree heights and diameters. Are these (approximately) normally distributed?

Step 2 Develop a Linear Model

Linear models or `lm()` include regressions of various types. In this case, we develop a basic linear model that relates tree height in centimeters to diameter at breast height in cm.

```
rip$htcm <- rip$ht * 100 #note correct column names from data.frame
riplm <- lm(rip$htcm ~ rip$dbh)
summary(riplm)
```

Interpret the results of the linear model regression (ie., the coefficient of determination). Is this a decent fit? `Plot()` the variables and then use `abline()` to show the predicted response. *Advanced users are challenged to develop a model with a zero intercept, as in reality trees start growing at zero cm height.*

Step 3 - Outliers and Removal

Outliers often confound environmental data analysis. These are often from errors in data collection or transcription, and occasionally are just random, freak events in nature. Their inclusion in datasets used to derive generalized models can reduce predictive power. We want to identify and remove outliers.

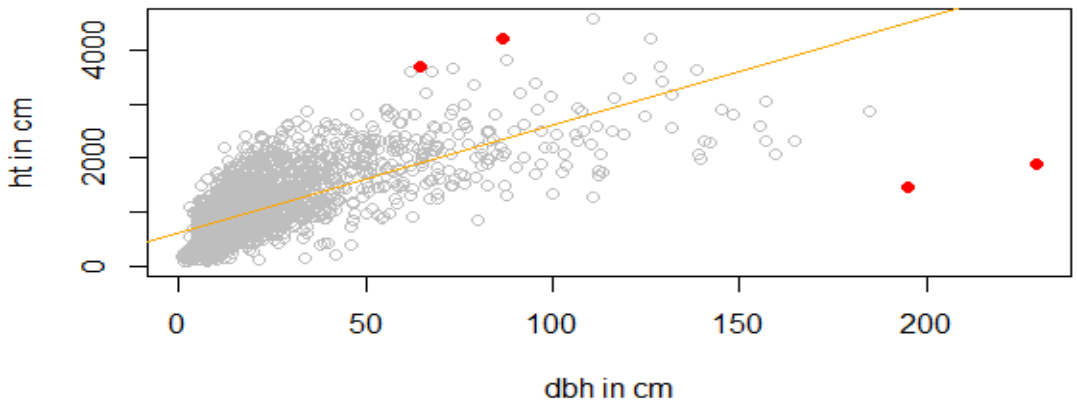
Using the results of the Bonferroni `outlierTest()` from the `car` package, create a table of outlier records from your linear model.

Use the `points()` function to add the outlier points to the plot in red. The following is a guide on how to extract the information from the outlier results:

```
# The outlierTest results come in the form of a table which has row #
"names" for each row. These coincide with the index number
#from our original data.frame. We need to identify those rows
#in our plot. The following is one way to do it using a for loop

#set an obj to the outlierTest
ol<-outlierTest(mylm)
ol.ids <- as.integer(names(ol$rstudent)) #rstudent is first obj

#create for loop
for (i in 1:length(ol.ids){
  #print(rip.ol.ids[id])
  r<-ol.ids[i]
  points(rip$dbh[r], rip$[r], col="red",pch=19)
}
#Do these look like outliers?
#Can you do this in a more elegant way?
```



Remove the outliers from your riparian dataframe using functions from last homework, such as `subset()`, `x[which(),]`, etc. and create a new linear model without these records, summarize and plot as before. Did the model improve?

Step 4 - Create and Evaluate Alternative Models

From our EDA, we know that both project site and genus likely have some influence on the relationship between tree height and diameter. We also know that the sample values are not likely to be normally distributed.

Create a global model that natural log transforms the sample values (ie `log(x)`) and compare to your first one. Is this a better representation?

Create competing linear models for each Project Site and for each Genus. Use `summary()` to examine the adjusted R^2 values to determine which one you think is the best model. Which model do you think should be used for your evaluation and projection of C stocks? Using a

panel layout, create a 2 x 4 plots that have Project Code and Genus linear model plots with predicted responses.

Step 5 - Create Master Scatter Plot

Create a single plot where each Genus is represented by a different symbol and each Project Code is represented by a different color. Overlay the global model fit line with log transformed axes. Lastly, label the fit equation on the plot, along with the coefficient of determination.

Step 6 - Commit to Git

Data

- `riparian_cleaned.csv` from previous Lab 2

Keywords

`lm(), outlierTest()`