**Lab 5: Correlation, Regression**

Geog 4300/6300-Shannon | Fall 2016

Value: 20 points

[Next time: Add descriptive stats for each variable and use in model interpretation. Replace q. 2? Or roll that into q1?]

For this lab, you’ll be using data for the average date of the last spring frost for several sites in the Southeast. This data is available on ELC as “Lab 5-DLSF”. The fields in this file are as follows:

* Station: Monitoring station
* State: State where the monitoring station is located
* AvgDLSpF: Average date of the last spring frost—days after Jan 1st.
* Latitude, longitude, and elevation: self explanatory
* DisttoCoast: Distance from the ocean coastline

1. ***(3 points)*** Test the correlation between the following variables using R’s rcorr command, found in the Hmisc package. You can select just the continuous variables (the last three listed above, columns 4-8 of the data) using this command in R:

DLSF.subset<-Lab.5.DLSF[,4:8]

To do this, you’ll also have to test whether these variables are normally distributed or not and use the appropriate test (Pearson’s r or Spearman’s rho). You’ll also need to make sure your data are in a matrix format—see the in class example.

Copy and paste your correlation and significance matrices. Explain how you determined which test to use and interpret the results these matrices show, focusing on both the magnitude and direction of the correlation.

1. ***(3 points)*** Visualize these correlations using the corrgram package in R. Create a two panel plot, different visualizations for the upper and lower parts of the correlogram. Pick whichever methods (panel.pie, panel.pts, etc.) of plotting on the upper and lower halves best show the correlations in the data. Use the help documentation for the function to see your visualization options, or see this page for more information: <http://www.statmethods.net/advgraphs/correlograms.html>.  
     
   Along with your visualization, briefly describe what you think each panel tells us (if anything) in addition to the correlation matrix from question 1.
2. ***(5 points)*** Do a univariate regression finding the effect of each of elevation, distance to coast, latitude, and longitude on the date of the last spring frost. Create a table like the one below that summarizes the results of each model, and summarize what these models tell you in a paragraph or so.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Independent var. | Beta | St. Error | p value | Model R2 |
| Elevation |  |  |  |  |
| Distance to coast |  |  |  |  |
| Latitude |  |  |  |  |
| Longitude |  |  |  |  |

1. ***(2 points)*** Create a scatterplot of one pair of variables from question 3 with the fit line included (using the abline command from your model in question 3). Copy and paste that plot in your response. Interpret it in light of the results you outlined in question 3, focusing on the direction and strength of the association between the variables.
2. ***(7 points)*** Conduct a multivariate regression with date of last frost as your dependent variable and the others as independent variables. Copy and paste your model output into your response. Interpret the magnitude, direction, and significance of each coefficient. How does the R2 of this model compare to the ones in question 3, and how would you interpret that difference? Are there any potential issues with this model for issues of multicollinearity or heteroskedasticity? Do the residuals indicate any problems?