Summary of Individual Analysis

Summary of Individual Analysis

Multivariate analysis has been an eye opening learning experience throughout the quarter. In fact, one of the most interesting aspects that has been an continuing learning theme is transforming a data set's weakness into strengths to extrapolate latent details in the data. A perfect example of this was learning to use PCA/CFA in combination with linear models in order to improve performance and gain further meaning. Our project is a great example of applying what I've learned about multivariate analysis. This project allowed me to combine multiple techniques that compliment each other. In addition, using PCA with modeling allowed me to have a better understanding how data points can relate to each other and finally being able to use geometric calculations to judge potential importance. As a result, my process can be broken down into three sections, initial analysis, individual analysis and comparisons.

Initial Analysis

The initial analysis involved merging and cleaning data, correlation visualizations and variable transformations. More cleaning was conducted by other members of the group while I performed transformations such as standardizing units (i.e. lbs to kg) and removing incomplete observations. After clean up, I created the correlation plots and stacked histogram visualizations on every variable showing the original, logged and squared root values in order to determining if any of the variables could be become more normal. Ultimately, none of the variables truly benefited from any of the transformations since many of the measurements were based on a predetermined scale. This meant that many of the variable transformations would not be appropriate if performed on the data set.

Individual Analysis

After the initial analysis we each conducted our own analysis on the data to compare methodologies (in terms of finding the best model). This involved PCA/CFA, OLS and regularization models. My modeling was based on sparser data set, I omitted all the blanks at read, while Les had more robust set where she used 3rd party data to fill it any potential blanks in the data. This was done in order to compare the sets to see if the extrapolated missing values would be drastically different from the sparser data set For the first portion of the analysis I ran PCA in 2 different iterations. Each iteration involved comparing the results with its scaled|unscaled and rotated | rotated counterparts.

1. No variables excluded

   Unscaled PCA held the entire variability in PC1, This would create uninterpretable results. Fortunately, using scaled PCA allowed the variability to become more evenly spread out. More importantly, I was able to use the variance value of 1 as a breaking point for determining the number of components.

2. Excluding at CF = 0.90

   Using a correlation test with a confidence level of 90%, I was able to eliminate very high correlated variables. From the original 29 numerical variables reduced to only 4 variables.

As expected, using all the variables resulted in difficult interpretations. The scaled results, surprisingly are harder to interpret than that of the unscaled.

The regularization models were next. This was done in number of ways. First, OLS models were created this includes step wise AIC for modeling and feature selection. Then, was conducting the regularization models. Using cross validation and randomly splitting the data into training and test sets (80/20) I was able to compare the different models and judge for over fitting. The regularization models and outputs were stored in tables using a sequence of alpha from 0 to 1 by increments of 0.05 (21 models) and subsequently comparing the results using both calculated lambdas (lambda.min and lambda 1SE). This meant a total of 42 models for each iteration. The root mean squared errors were then calculated and the model with the smallest diagnostic was then used for residual analysis. This same process was repeated only using the PCA scores that were calculated earlier.

Residual analysis was conducted at 3 levels. At each level, a number of observations were removed that had calculated standardized residuals greater than 3.5. A total of 15 influential outlines were found using this methodology and 16 influential outliers were found when using the PCA scores in regularization. Without repeating outputs, the 15 residuals were did not seem to have any similarities except that their actual cupper points were rated low relative to other observations that had similar scores. The biggest ones were residuals over 10. Three observations (Kenya, Indonesia, and Taiwan) had high predictor scores, but had actual cupper points of less than 6. Compared to all of the models, each predicted these observations to be much higher. Ultimatly, it seems we are missing variables in the dataset that do seem to explain those observations very well.

Comparisons

Compared to Les's analysis there was not much difference in the models, ultimately deciding that using more observations was better since the 3rd party data did not seem to vary drastically. As a result, for the final report as well as the presentation we opted to use her PCA analysis coupled with my linear and regularization models. The results flowed very nicely together as the raw data and the PCA analysis complemented each others interpretation. For example, the first PCA component was composed of the same variables as my best linear model and regularization model using the raw data. This is evidence of the significance and potential importance for predicting the variability in cupper points.

Final Thoughts

I think the most important aspect that could be drawn from my contribution to the project would have to the residual analysis and visualizations. The models, when compared without removing the residuals outliers, had more than 10% less in calculated captured variance (R^2). Removing the outliers helped stabilize the model to reveal less inflated/deflated beta coefficients in both regularization and OLS.