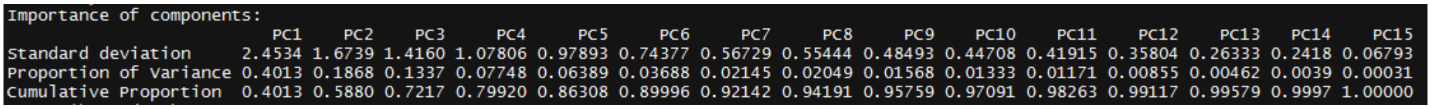
Bryson Cook
HW6

Question 9.1
Using the same crime data set as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your solution to Question 8.2. You can use the R function `prcomp` for PCA. Note that to first scale the data, you can include `scale. = TRUE` to scale as part of the PCA function. Don't forget that, to make a prediction for the new city, you'll need to un-scale the coefficients (i.e., do the scaling calculation in reverse!)

First, I brought in the crime data and ran the R-function prcomp() on the predictors, with scaling turned on. A summary of the output is shown below in Figure 1. From the summary we can see the Proportion of Variance for each Principle Component, or how much of the data each factor explains.



```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12    PC13   PC14    PC15
Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377 0.56729 0.55444 0.48493 0.44708 0.41915 0.35804 0.26333 0.2418 0.06793
Proportion of Variance 0.4013 0.1868 0.1337 0.07748 0.06389 0.03688 0.02145 0.02049 0.01568 0.01333 0.01171 0.00855 0.00462 0.0039 0.00031
Cumulative Proportion  0.4013 0.5880 0.7217 0.79920 0.86308 0.89996 0.92142 0.94191 0.95759 0.97091 0.98263 0.99117 0.99579 0.9997 1.00000
```

Figure 1. Summary of PCA Analysis

To get an understanding of the basic structure of the first two PC's I made a biplot of their Eigen vectors, which is shown in Figure 2. From the biplot we see that PC1 is a most likely a function of Wealth, Ineq, M, and So and that PC2 is most likely a function of Time, Pop, M.F, and L.F. These are easily seen because these are most parallel to the axes and thus have the largest variances in those particular scales.

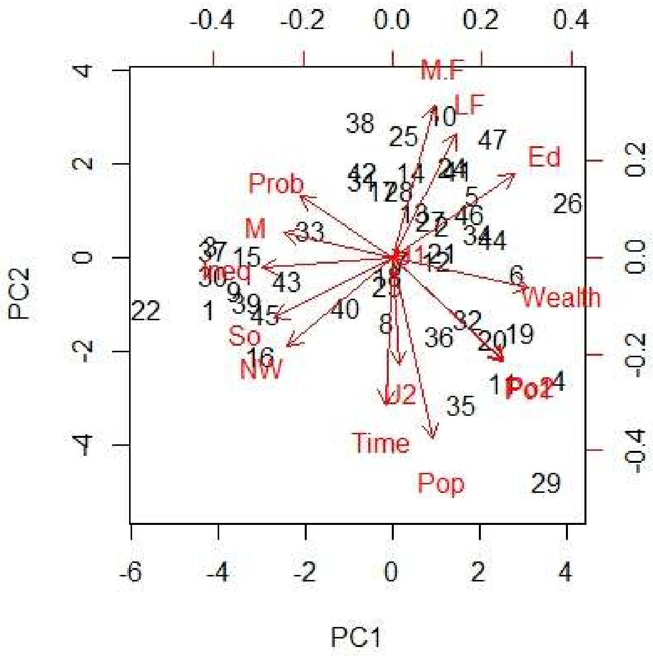


Figure 2. Biplot of PC1 and PC2

I then created a scree plot to choose the number of PC's to use in my regression model, which is shown in Figure 3. From the scree plot and plotting Proportion of Variances vs the PC number, we can visually see a diminishing return. I will choose the first 7 PC's, which from the summary accounts for 92.1% of the values. Other values can be chosen, but to me this is the point after which the curve flattens significantly.

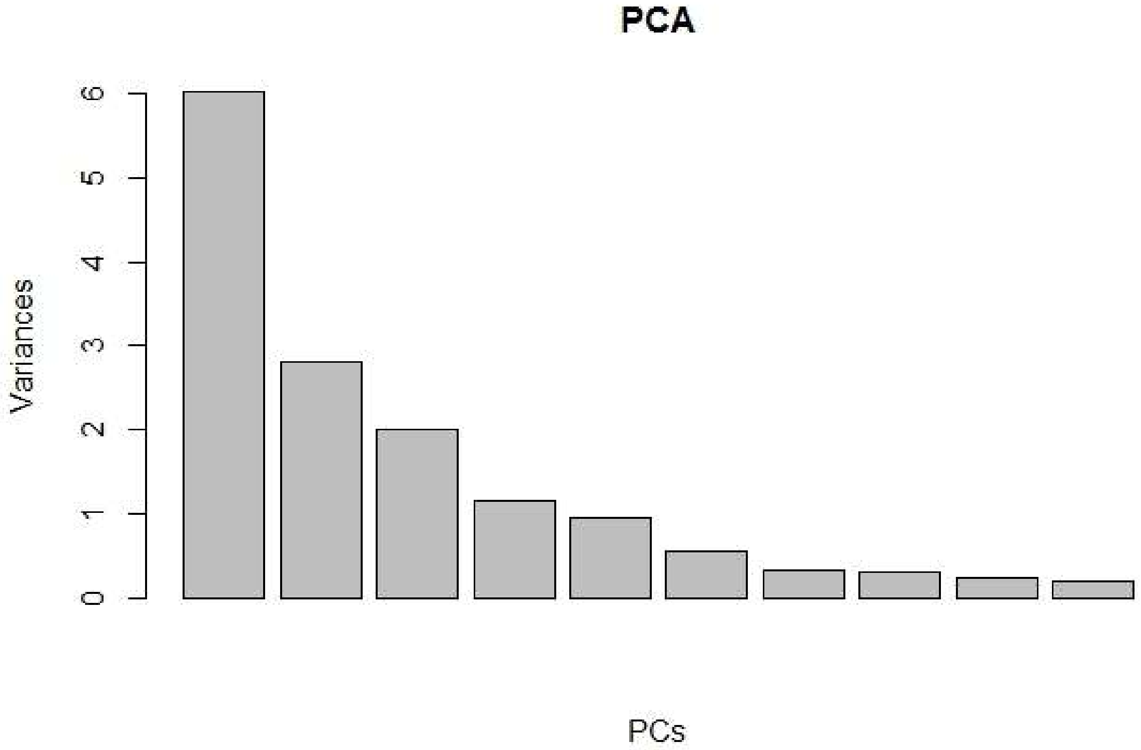


Figure 3. Scree plot of PC's Variances

Running the lm() function on the chosen PC's, un-scaling the coefficients, and inserting the new city's data yields a predicted crime rate of 1230 and an adjusted R^2 value of 0.6322. The final model from question 8.2 of HW5, which used the formula Crime ~ M + Ed + Po1 + U2 + Ineq + Prob, had a predicted crime rate of 1304 and an adjusted R^2 value of 0.766, showing that perhaps there is just too much correlation in the data for the PCA model to overcome and that removing multiple predictors as in the last homework yields a better model. Using cross validation, we calculate an R^2 of 0.456 for the PCA, again using the top 7 PC's, vs 0.413, for the model using all 15 predictors. However, this is when using all 15 variables. When reducing the number of predictors to the above formula, we saw a R^2 of 0.638, which again shows that removing predictors in linear regression is superior, at least in this case.