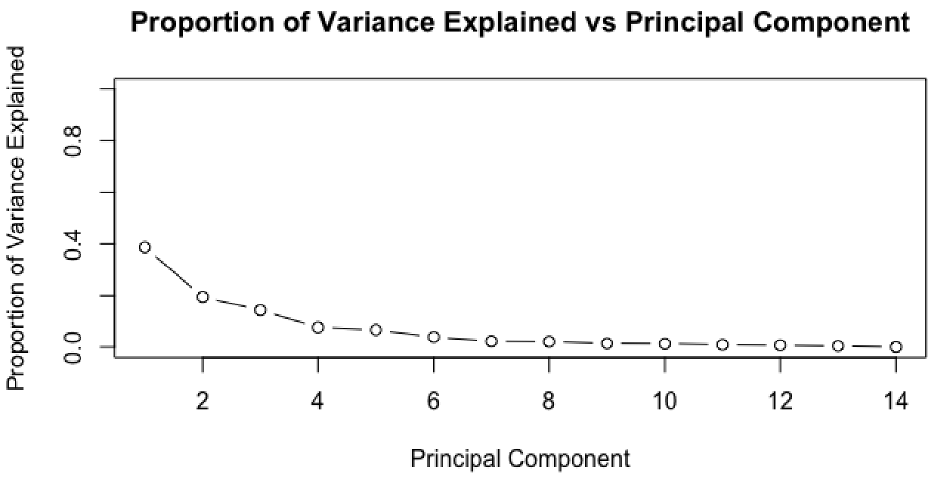
**Question 9.1**
Using the same crime data set as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that 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 unscale the coefficients (i.e., do the scaling calculation in reverse!)

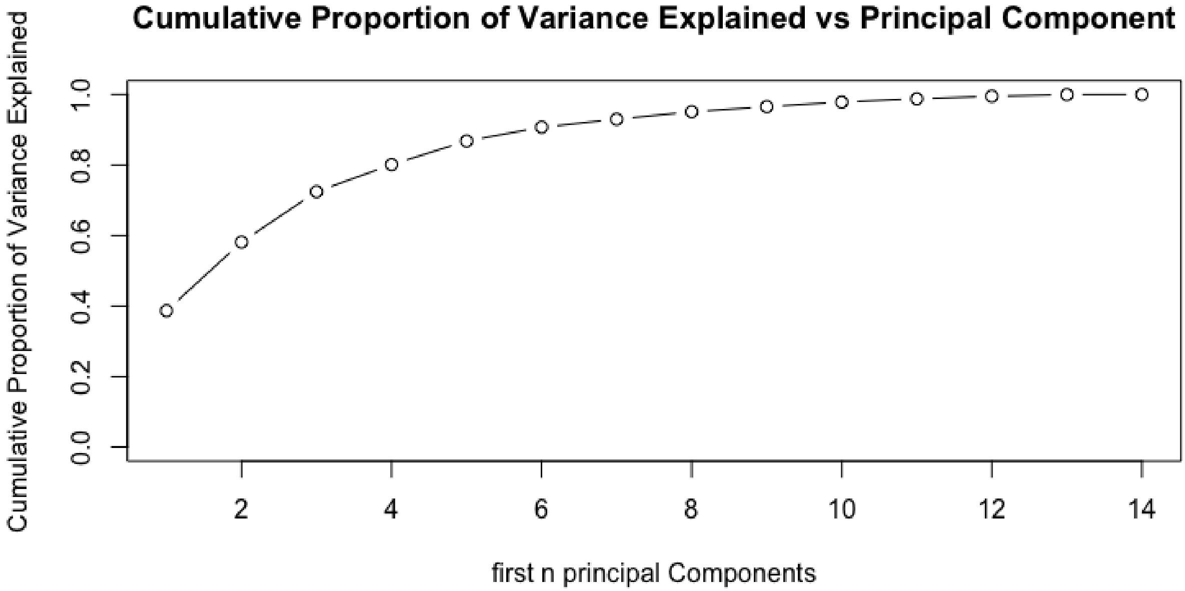
(Please note that my R code for this question is contained in the file 'homework_6_Q9.1_.R', and if you wish to run the code, you will need to change line 5 of my R code to your local directory which contains the 'temps.txt' data. I have also included the cod from homework 5 in the file 'homework_5_Q8.2_.R')

**Response**
To solve this problem, I implemented *principal component analysis* via R's built-in *prcomp* function and then created a regression model on the principal components via R's built-in *lm* function. To begin my analysis, I read in the us crime data and saved it as a variable called *crime_data*. *crime_data* contains 47 data points and 15 predictor variables as well as 1 response variable called *Crime*. One of the predictor variables in *crime_data,* called *So*, is a binary variable. Since *principal component analysis* does not work on data sets which contain categorical and binary variables, I first removed this column from the data set prior to my implementation of PCA.

Once I removed the categorical variable *So*, I began my analysis. I used R's *prcomp* function to conduct principal component analysis on the remaining variables in the *crime_data* data set. It should also be mentioned that I took care of scaling the variables directly within the *prcomp* function by setting *scale* equal to *TRUE.* To analyze the results generated by implementing R's *prcomp* function I plotted the proportion of variance explained by each of the principal components, as may be seen in the figure below:

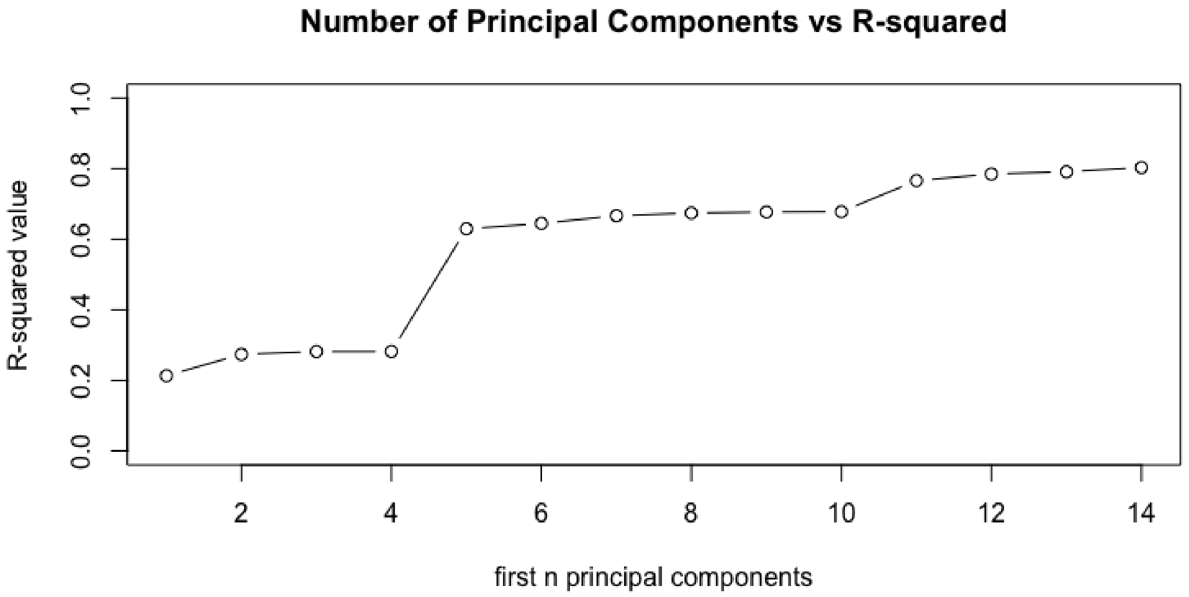


As may be observed in the figure above, R's *prcomp* function automatically sorts the principal components in descending order based upon the amount of variance explained by each principal component. To further analyze the total proportion of variance that is explained by the *prcomp* function, I next plotted the cumulative sum of the variance explained by each of the first *n* principal components. This may be seen in the following figure:



As may be expected, we see a large increase in the proportion variance explained due to the inclusion of the first several principal components, while we see a very small increase in the proportion of variance explained associated with the inclusion of the last several principal components. This indicates that we likely should not include all of the principal components in our final model, as this may lead to problems of overfitting and heteroscedasticity due to strong correlations amongst the predictor variables. Additionally, when we observe the figure above, we may note that just the first *4* of the *14* principal components explain approximately *80%* of the variance in the model.

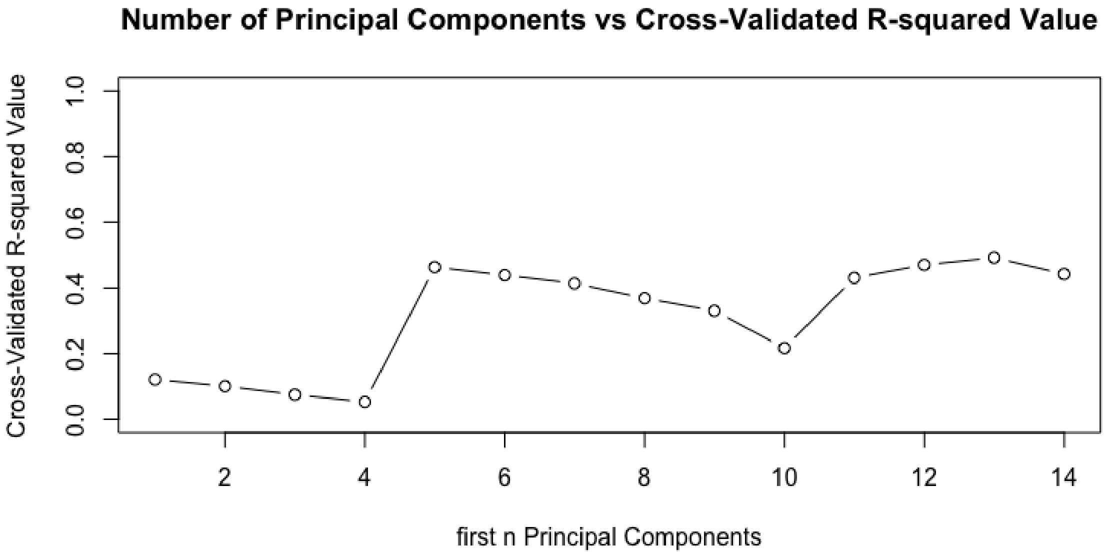
To select the number of principal components to include in our model, I next found the $R^2$ value associated with including each of the first *n* principal components in our regression model. To do this, I implemented a for-loop in R which created a regression model based upon including each of the first *n* principal components, and then found the $R^2$ value associated with each of those 14 models. The results of this analysis may be seen in the following figure:



By observing the figure above, it is not surprising to find that there is a trend in which the *R*$^2$ value increases as the number of principal components included in the model increases. However, to create the best model, we do not want to simply maximize the *R*$^2$ value. Problems of overfitting and heteroscedasticity are often associated with regression models that include an excessive number of predictor variables, and therefore model quality does not always increase along with the *R*$^2$ value as more predictor variables are included. Of particular importance to us is the large jump in the *R*$^2$ value when moving from 4 principal components to 5 principal components. Beyond 5 principal components there is a relatively minimal increase in *R*$^2$ value. This makes the choice of a creating a regression model based upon the first 5 principal components a compelling choice. To further verify that including only the first *5* principal components is indeed the correct choice, I found the *R*$^2$ and *adjusted R*$^2$ value associated with the model which includes the first *5* principal components. I found an *R*$^2$ value of *0.6297* and an *adjusted R*$^2$ value of *0.5845*. The *R*$^2$ value is always greater than or equal to the *adjusted R*$^2$ value, but a difference this minimal is a good sign, since a large difference between the *R*$^2$ and *adjusted R*$^2$ values indicates that overfitting is likely occurring in the model. Therefore, based upon these above considerations, I chose to include only the first *5* principal components in my regression model.

While the above *R*$^2$ and *adjusted R*$^2$ values give an indication of the model's performance, creating and testing a model on the same set of data will generally lead to an overly optimistic estimation of the model's true performance when generalized independent data sets. To analyze how the model will perform when applied to independent data sets, we must test the model's performance on a different set of data than the data set which was used to create the model. To do this, I used *k-fold* cross validation via R's built-in *cv.lm* function. Based upon the size of the *crime_data* data set, I chose to use *5-fold* cross validation. *10-fold* cross validation is generally considered to be standard number of fold, but since *crime_data* only includes *47* observations, the choice of a smaller number of folds is appropriate here.

After implement cross validation, I again chose to find the *$R^2$* value associated with including each of the first *n* principal components in order to ensure that I had truly selected the appropriate number of principal components to include in the regression model. To do this, I implemented a for-loop in R in which 5-fold cross-validation was used to train and independently test the performance of the regression model based upon including each of the first *n* principal components. I next found the *$R^2$* value associated with each of the 14 models created. The results of this analysis may be seen in the following figure:



As may be observed in the figure above, the greatest *$R^2$* value occurs when the number of principal components is equal to *5.* This confirms our choice of including the first 5 principal components in the regression model was the correct choice. The fact that we see a decrease in the *$R^2$* value in the models which include more than the first *5* principal components is evidence that overfitting is occurring in these models, since you would generally expect to see the *$R^2$* value increase as the number of predictors increases. The *$R^2$* value associated with the PCA model which includes the first five principal components is equal to *0.4635*. Not surprisingly, this value is smaller than the *$R^2$* value of *0.6297* which we found when for the regression model which included the same 5 principal components, but did not utilize cross-validation. However, while the cross-validated *$R^2$* value of *0.4635* may seem relatively small, this number is still indicative of a fairly robust, yet simple regression model, thanks to the fact that we are only including 5 principal components as predictor variables. Thus, we can feel confident in the predictive power of this model.

## Summary

In order to use the above model in a predictive manner, we must first define the model in terms of the original variables, rather than the principal components. To do this, I calculated the original intercept and coefficients in R by unscaling the values (essentially doing the scaling process in reverse), and found that the regression model associated with the first 5 principal components may be described in terms of the original variables as follows:

$$\begin{gathered} Crime = -5726 + 51M + 12.6Ed + 39.2Po1 + 39.5Po1 + 1885LF + 35M.F + 1.66Pop + \ldots \\ \ldots + 10.1NW + 213U1 + 38.2U2 + 0.0306Wealth + 6.42Ineq - 1214Prob + 3.79Time \end{gathered}$$

Based upon the above formula, we may now use the model in a predictive way in order to forecast the *Crime* in a new city. In this assignment we are asked to apply this model to a new city defined by the following predictor variables, *M = 14.0  So = 0  Ed = 10.0  Po1 = 12.0 Po2 = 15.5  LF = 0.640  M.F = 94.0 Pop = 150  NW = 1.1  U1 = 0.120  U2 = 3.6 Wealth = 3200 Ineq = 20.1 Prob = 0.04 Time = 39.0*. When we use the above model to forecast the Crime within this new city we find that the forecasted value for *Crime* is equal to *1443*. This is a reasonable value that falls approximately in the middle of values that we see associated with *Crime* in the *crime_data* data set. The values of *Crime* range from *342* to *1993*, thus this is further evidence that we have created a meaningful model with regard to forecasting ability, and that our choice to include 5 principal components in the model was likely the optimal choice.

As a final step, we are asked to compare the quality of this model to the linear model we created in the previous homework in question 8.2 . In question 8.2 I determined that the ideal linear model to predict *Crime* in the *crime_data* data set is given by the following equation:

$$Crime = -5040.5 + 105*M + 196.5*Ed + 115*Po1 + 89.4*U2 + 67.7*Ineq - 3801.8*Prob$$

The details of how I determined this was the optimal linear model are left out here since this was discussed in the previous homework. However, I will compare the important metrics of this model with the PCA regression model created in this assignment, to determine which was better.

The PCA regression model created in this assignment has an $R^2$ value of *0.6297* and an *adjusted* $R^2$ value of *0.5845*, when cross-validation is not implemented. These are both strong values, and the minimal difference $R^2$ value and *adjusted* $R^2$ value indicates that minimal overfitting is exhibited by this model. The linear model from the previous assignment had an $R^2$ value of *0.7659* and an *adjusted* $R^2$ value of *0.7307* associated with it. Both of these values are greater than the corresponding values associated with the PCA regression model, which indicates that the linear model is perhaps better. However, the greater difference between the $R^2$ and *adjusted* $R^2$ values in the linear model indicates that it is more likely that some overfitting may be present in the linear model.

To test which model is definitively better, I implemented *5-fold* cross validation and found the $R^2$ value associated with each model. When cross validation was utilized, I found that PCA regression model had an $R^2$ value of *0.4635* . I found that the $R^2$ value associated with the cross-validated linear model was equal to *0.6384* . This indicates that the linear model may in fact be a better model than the regression model created via principal component analysis. However, given the above metrics outlined in this paper, I feel confident that the PCA regression model also has competent forecasting ability. Further, when both models were applied to the same new city defined by predictor variables, $M = 14.0 \quad So = 0 \quad Ed = 10.0 \quad Po1 = 12.0\ Po2 = 15.5 \quad LF = 0.640 \quad M.F = 94.0\ Pop = 150 \quad NW = 1.1 \quad U1 = 0.120 \quad U2 = 3.6\ Wealth = 3200\ Ineq = 20.1\ Prob = 0.04\ Time = 39.0$, both models produce a response that falls within a reasonable range. The linear model forecasts that the *Crime* in this city will be equal to *1304* , while the PCA regression model forecasts that *Crime* will be equal to *1443*. Thus, while it appears the linear model may be slightly better, further analysis on independent data sets would be necessary to verify that it is definitively the better choice, since both models have fairly similar metrics.