GTx: ISYE6501x - Homework 4

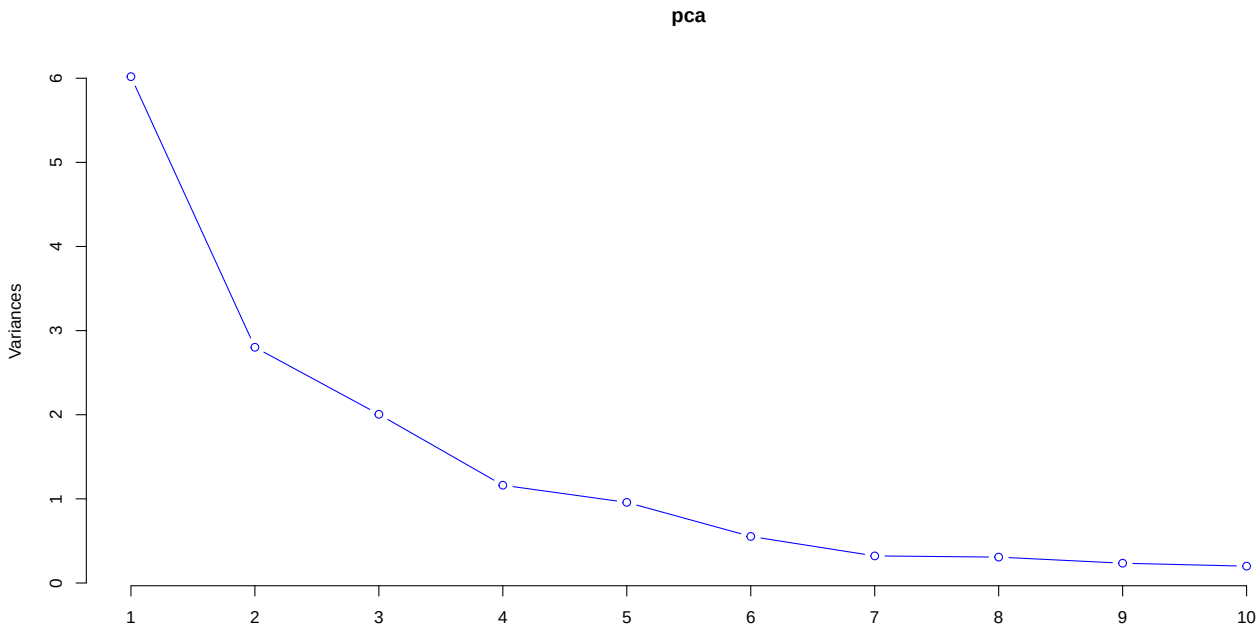
Muh Alif Ahsanul Islam

06/09/2019

Question 9.1

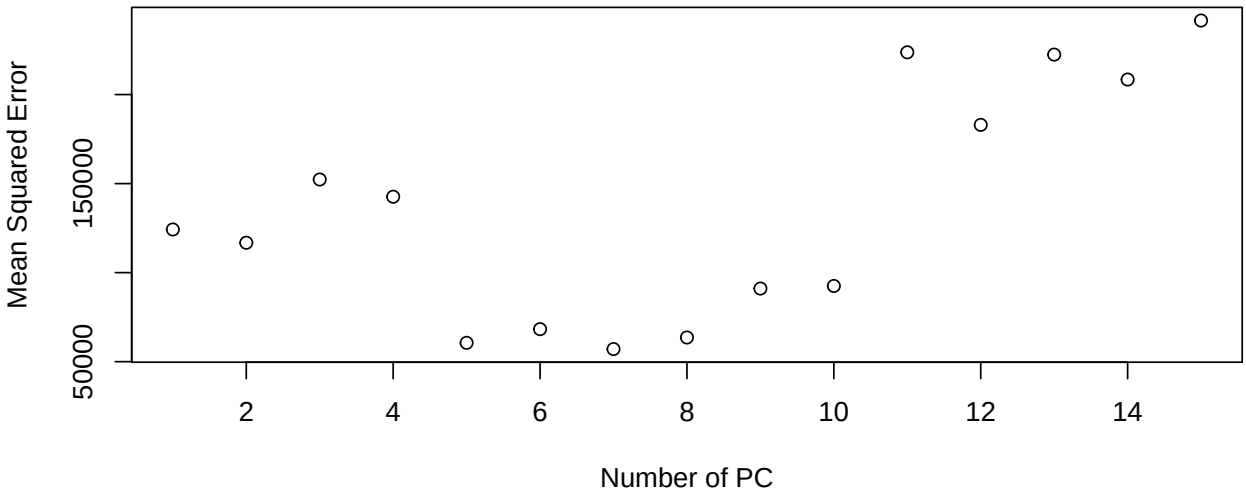
Using the same crime data set uscrime.txt as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your 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)!)

```
rm(list=ls())
crime_df = read.table('uscrime.txt', header=TRUE)
pca = prcomp(x = crime_df[,1:15], scale=TRUE)
screeplot(pca, type='lines', col='blue')
```



First I will choose how many principal components to use by doing cross validation on linear model and choose number of principal component to use.

```
plot(x=cv_res$n_pca_comp, y=cv_res$mean_squared_error,
     xlab='Number of PC', ylab='Mean Squared Error')
```



Answer to Question 7.2

From the summary of the linear regression model, we understand that the coefficient of year is 0.072 with p value 0.828. The p value is higher than our standard 0.05 so we regard there are no strong evidence for positive trend. This means we do not have evidence that summer end is getting later.

Question 8.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a linear regression model would be appropriate. List some (up to 5) predictors that you might use.

Answer

Analytics can be used to help increase productivity of your banana farm. Using linear regresion model, we can know what type of stimulation we need to give to the plant to make it produce more bananas. Target variable (y): number of bananas produced
Predictor:
1. Temperature
2. Humidity
3. Fertilizer
4. Water
5. Precipitation

Question 8.2

Using crime data, use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the test data below. Show your model (factors used and their coefficients), the software output, and the quality of fit.

```
rm(list=ls())
raw_data = read.table('uscrime.txt', header=TRUE)
test_data = data.frame(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)
model = lm(Crime~., data=raw_data)
summary(model)
```

```
##
## Call:
## lm(formula = Crime ~ ., data = raw_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -395.7  -98.1   -6.7  113.0  512.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.98e+03   1.63e+03   -3.68  0.00089 ***
## M            8.78e+01   4.17e+01    2.11  0.04344 *
```

```
## So          -3.80e+00   1.49e+02   -0.03  0.97977
## Ed           1.88e+02   6.21e+01    3.03  0.00486 **
## Po1          1.93e+02   1.06e+02    1.82  0.07889 .
## Po2         -1.09e+02   1.17e+02   -0.93  0.35883
## LF          -6.64e+02   1.47e+03   -0.45  0.65465
## M.F          1.74e+01   2.04e+01    0.86  0.39900
## Pop         -7.33e-01   1.29e+00   -0.57  0.57385
## NW           4.20e+00   6.48e+00    0.65  0.52128
## U1          -5.83e+03   4.21e+03   -1.38  0.17624
## U2           1.68e+02   8.23e+01    2.04  0.05016 .
## Wealth       9.62e-02   1.04e-01    0.93  0.36075
## Ineq         7.07e+01   2.27e+01    3.11  0.00398 **
## Prob        -4.86e+03   2.27e+03   -2.14  0.04063 *
## Time        -3.48e+00   7.17e+00   -0.49  0.63071
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 209 on 31 degrees of freedom
## Multiple R-squared:  0.803,  Adjusted R-squared:  0.708
## F-statistic: 8.43 on 15 and 31 DF,  p-value: 3.54e-07
```

Based on model summary above, we can see the significant predictors are (based on p value): Ed, Ineq, M, Prob, U2, Po1. The R-squared is 0.8031 and Adjusted R-squared 0.7078.

Redo modelling with only siginificant predictors, we will call this new model 'significant model' and old model 'old model'.

```
model_significant = lm(Crime ~ Ed + Ineq + M + Prob + U2 + Po1, data=raw_data)
summary(model_significant)
```

```
##
## Call:
## lm(formula = Crime ~ Ed + Ineq + M + Prob + U2 + Po1, data = raw_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -470.7  -78.4  -19.7  133.1  556.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5040.5      899.8   -5.60  1.7e-06 ***
## Ed             196.5       44.8    4.39  8.1e-05 ***
## Ineq            67.7       13.9    4.85  1.9e-05 ***
## M              105.0       33.3    3.15   0.0031 **
## Prob         -3801.8     1528.1   -2.49   0.0171 *
## U2              89.4       40.9    2.18   0.0348 *
## Po1            115.0       13.8    8.36  2.6e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 201 on 40 degrees of freedom
## Multiple R-squared:  0.766,  Adjusted R-squared:  0.731
## F-statistic: 21.8 on 6 and 40 DF,  p-value: 3.42e-11
```

```
predict(model_significant, test_data)
```

```
##    1
## 1304
```

The R-squared for this model is is 0.7659 and Adjusted R-squared 0.7307. The significant model has lowest R-squared but higher adjusted R-squared.
Source click here:
The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance.

This means the significant model is better than old model. And we will use it to predict the test data.

Answer to Question 7.2

The predicted crime is:

```
predict(model_significant, test_data)
```

```
##    1
## 1304
```
