Homework6

Pablo Diaz

2/19/2019

Contents

Question 9.1

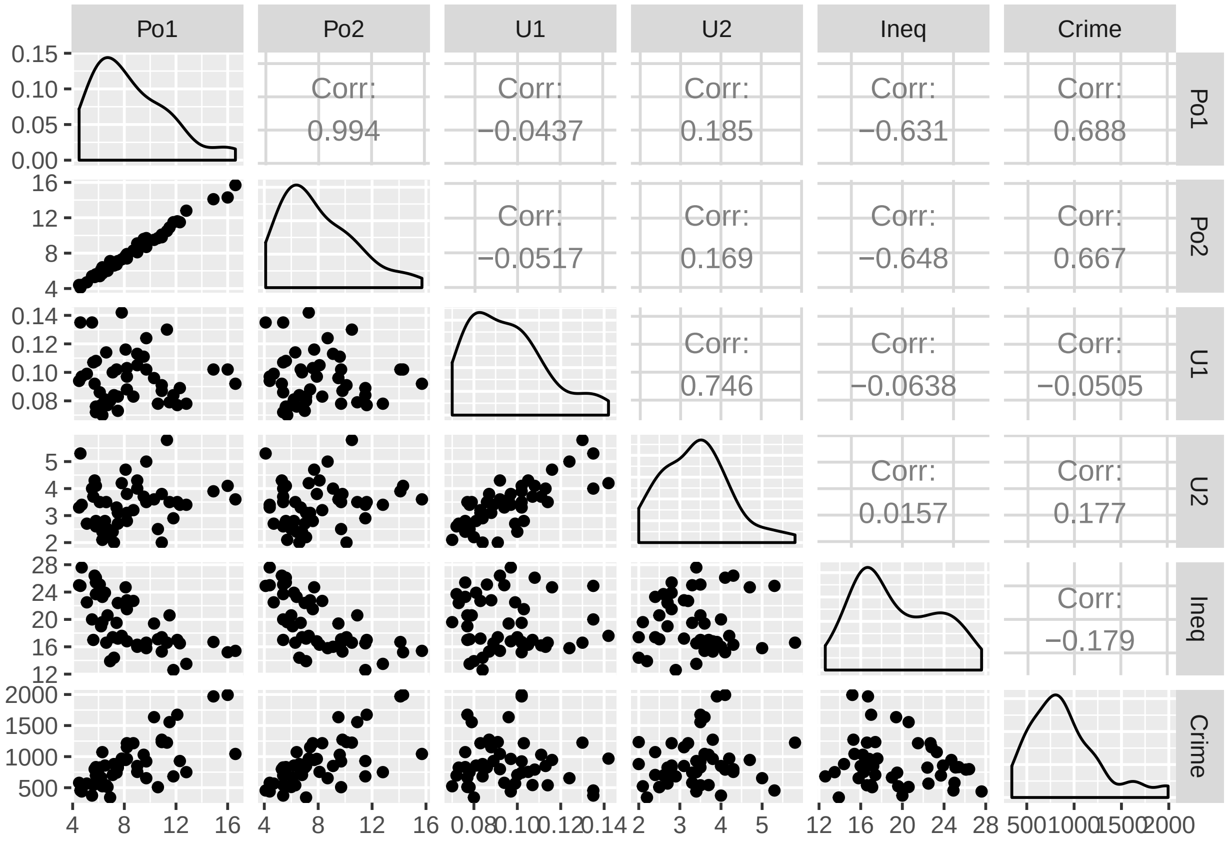
Using the same crime data set uscrime.txt as in Question 8.2, apply Principal Component Analysis and then create a regression model using the first few principal components. Specify your new model in terms of the original variables (not the principal components), and compare its quality to that of your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

```
set.seed(42)
# import data
uscrime <- read.delim("~/Documents/R/GeorgiaTech/DataPreparation/uscrime.txt")
#The TA suggested this graph

library(GGally)
```

```
## Loading required package: ggplot2
```

```
# TA suggsted this predictos
ggpairs(uscrime, columns = c("Po1", "Po2", "U1", "U2", "Ineq", "Crime"))
```



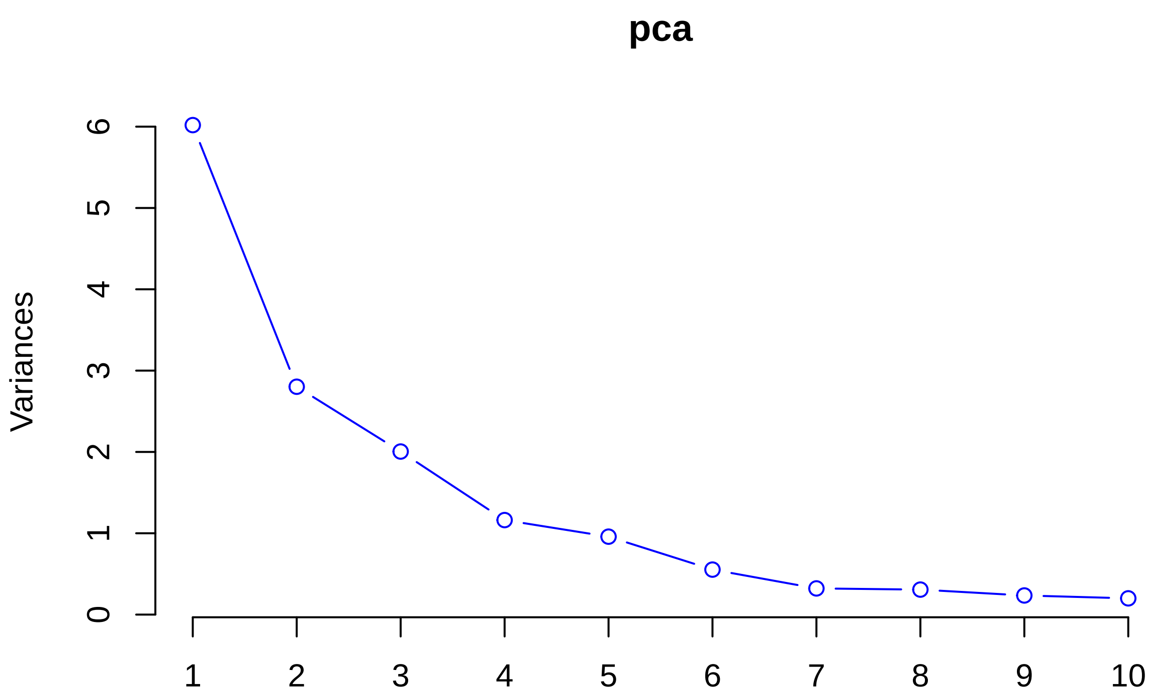
```
# here we can see that Po1 and Po2 are highly correlated between them and there seems to be some correl

# apply the PCA model
pca <- prcomp(uscrime[,1:15], scale = TRUE)
summary(pca)
```

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

```
# the summary of PCA is for each predictor of the original data set
# in the summary we can say that more than 85% of the variance can be explanied with  the first 5 princ

# this graph helps to choose how many PCA to use.
# the prcomp, will show the best ones at first.
screeplot(pca, type="lines",col="blue")
```



```
# according to the grapg 5 components should be enough

#obtain the 5 principal components from result matrix
principal_components <- pca$x[,1:5]
# now create a new matrix whith components and crime response
matrix_crime <- cbind(principal_components, uscrime[,15])
model <- lm(V6~., data = as.data.frame(matrix_crime))
summary(model)
```

```
##
## Call:
## lm(formula = V6 ~ ., data = as.data.frame(matrix_crime))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0798 -1.3405 -0.4174  1.4309  6.8161
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  26.5979     0.3975  66.912  < 2e-16 ***
## PC1          -0.1462     0.1638  -0.893   0.3773
## PC2          -2.6941     0.2400 -11.223 4.46e-14 ***
## PC3           1.5844     0.2838   5.583 1.69e-06 ***
## PC4          -3.8311     0.3727 -10.279 6.50e-13 ***
## PC5          -1.0464     0.4105  -2.549   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.725 on 41 degrees of freedom
## Multiple R-squared:  0.8682, Adjusted R-squared:  0.8521
## F-statistic: 54.02 on 5 and 41 DF,  p-value: < 2.2e-16
```

```r
# I don't quite understand how to unscale the data
transformed_coeff <- pca$rotation[,1:5] %*% model$coefficients[2:6]
# This homework is taking me too much time
# I will leave this homework unfinished, so I can prepare for the first quizz
```

Result

The goal of Principal Component Analysis is a tool to extract the features of data that is used that to reduce a large set of variables to a small set that still contains most of the information in the large set. Basically, PCA removes correlation of predictos and ranks the coordinate dimensions according to the variability.

The steps of this homework is to run the PCA model (the ggpairs can show an which predictors are the most correlated), extract the relevant components that describe most of the data (this can be done using screeplot and the pca shows the components in order), use the scaled rotation matrix from the result and uscale it to build a linear regression model. Calculate the city's crime using the data from last homework and measure its R^2 and adjusted R^2 to compare the quality of the model (other factors can be used to see the quality of the model).

To scale the data this is used $a'_j = (a_j - mu_j)/sigma_j$ To unscale the data this should be used $a_j = a'_j * sigma_j + mu_j$ where j = 1, …, 15

The result should almost or the same as the past homework. The idea is to see that the same result can be achived using less predictors. I don't have the time to finish the math matrix part. Cross validation can also be applied to measure a more realistic quality of the lm and also according to the TA the binary column of the data set should be removed because pca doesn't work well with binary data.

The follwing way could be used to remove the binary data `pca <- prcomp(cbind(uscrime[,1],data[3:15]),scale=TRUE)` and to then build the lm with all the columns `crime_data_with_all_columns <- cbind(uscrime[,2],pca2$x[,1:5],uscri`