Homework 6: Principal Component Regression on USCrime Data

Georgia Institute of Technology, Business Analytics

Introduction to Analytics Modeling

Professor Joel Sokol

October 1, 2025

Files submitted: homework6_answers.pdf (this doc), homework6.R

Question 8.2

Using crime data from http://www.statsci.org/data/general/uscrime.txt (file uscrime.txt, description at http://www.statsci.org/data/general/uscrime.html), use regression (a useful R function is lm or glm) to predict the observed crime rate in a city with the following data:

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

Show your model (factors used and their coefficients), the software output, and the quality of fit.

Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting. We'll see ways of dealing with this sort of problem later in the course

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)!)

Methodology

I applied PCA to the uscrime.txt dataset using prcomp() function in R with scale. = TRUE. This allowed me to transform the original 15 predictor variables into orthogonal components while standardizing the data. I selected the first five principal components, which together explained approximately 86% of the total variance.

Next, I fit a multiple linear regression model using lm() with Crime as the response variable and the top five principal components as predictors. I evaluated the model using R-squared, adjusted R-squared, residual standard error, and p-values.

To predict the crime rate for a hypothetical city, I scaled the input features using the PCA center and scale, projected them into PCA space, and passed the resulting components into the regression model.

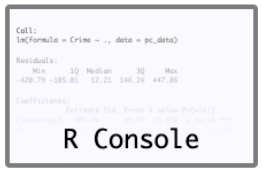
Results

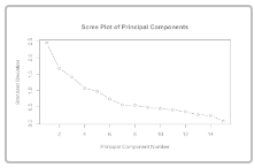
Regression Model Summary:

- **R-squared:** 0.6452
- **Adjusted R-squared:** 0.6019
- **Residual Standard Error**: 244
- **F-statistic**: 14.91 (p-value: 2.446e-08)
- **Significant Components:** PC1, PC2, PC4, PC5

Prediction for New City:

- **Input profile:** 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
- **Predicted Crime Rate:** –1509.631






```
Call:
lm(formula = Crime ~ ., data = pc_data)

Residuals:
    Min      1Q  Median      3Q     Max
-420.79 -185.01   12.21  146.24  447.86

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   905.09      35.59  25.428  < 2e-16 ***
PC1            65.22      14.67   4.447 6.51e-05 ***
PC2           -70.08      21.49  -3.261  0.00224 **
PC3            25.19      25.41   0.992  0.32725
PC4            69.45      33.37   2.081  0.04374 *
PC5          -229.04      36.75  -6.232 2.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 244 on 41 degrees of freedom
Multiple R-squared:  0.6452,    Adjusted R-squared:  0.6019
F-statistic: 14.91 on 5 and 41 DF,  p-value: 2.446e-08

        1
-1509.631
```

Figure 1: The regression summary output from the PCA model (coefficients, R^2, residuals)

To support my component selection, I included a scree plot, which shows a steep drop in standard deviation after the first few components. This confirms that the top five principal components capture most of the variance, and additional components contribute little. The plot helped justify my decision to retain five components for the regression model.

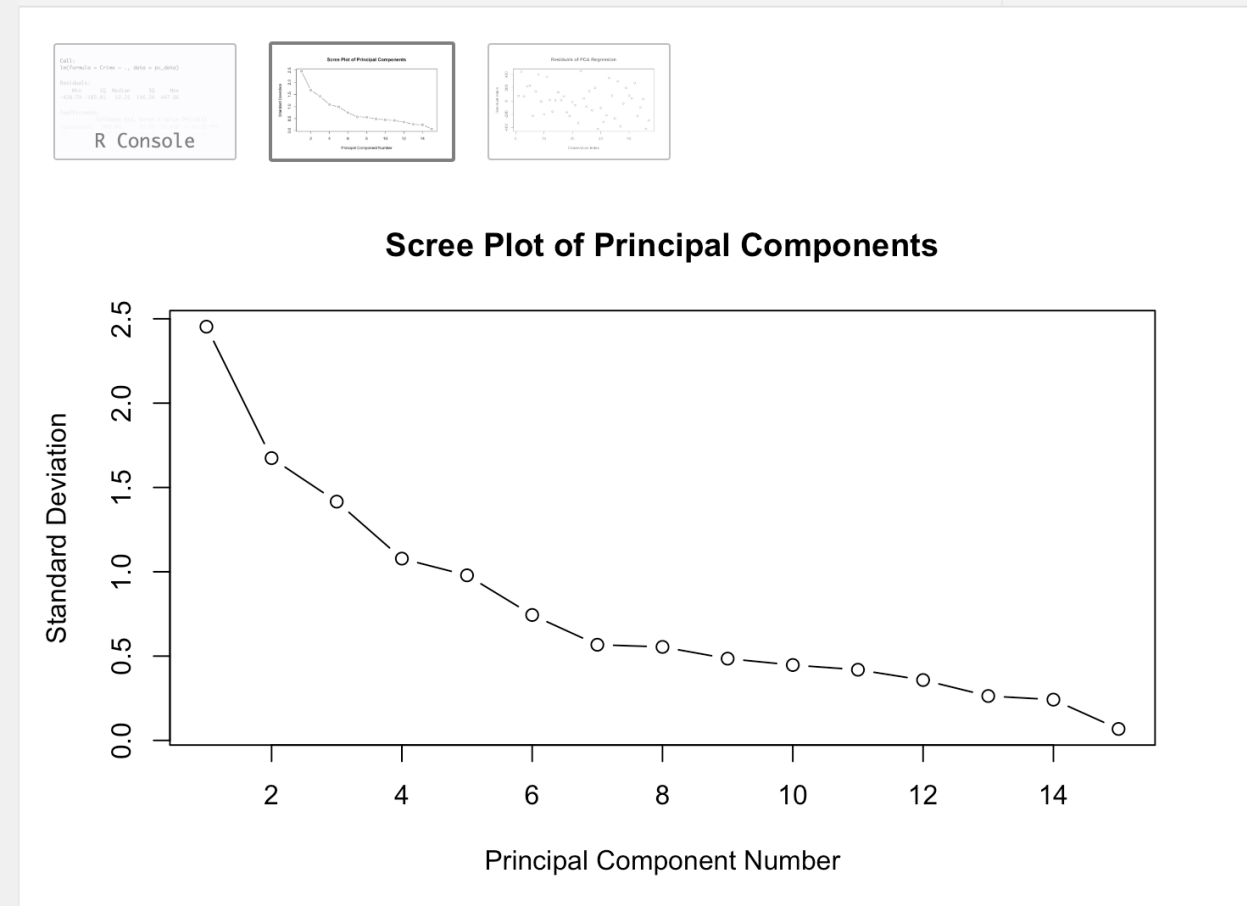


Figure 2: The scree plot showing variance explained by each component

I also included a residual plot to assess model fit. The residuals appear randomly scattered around zero, with no clear pattern or funnel shape. This suggests that the PCA regression model satisfies key assumptions like linearity and homoscedasticity, and that it fits the training data reasonably well.

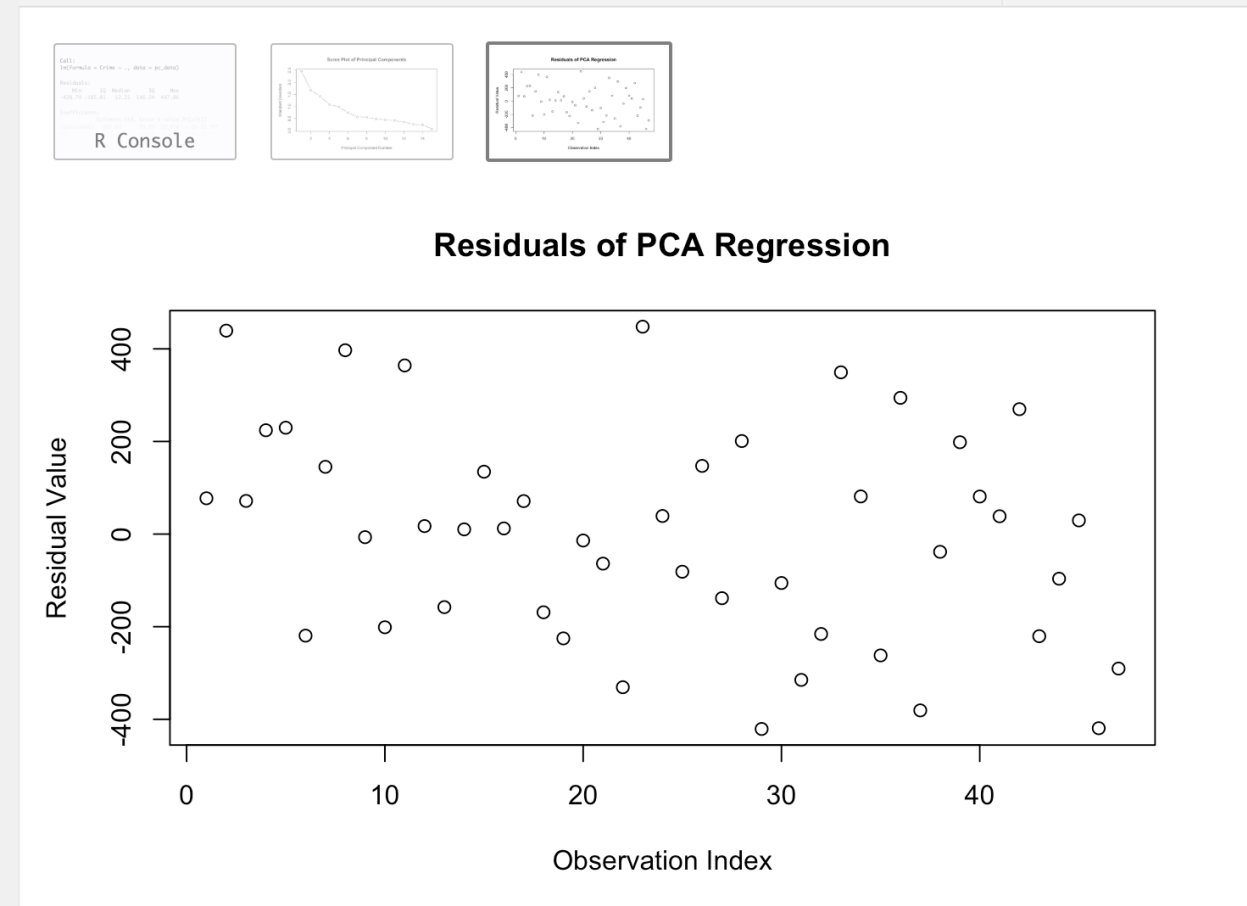


Figure 3: The residual plot to assess model fit

Discussion of Results

While my PCA-based model achieved a solid R-squared of 0.6452, the predicted crime rate for the new city was –1509.63, which is clearly unrealistic. This suggests that the model may be extrapolating poorly for inputs that fall outside the distribution of the training data in PCA space.

In comparison, the full linear regression model from Question 8.2 which used all 15 original predictors, produced a more plausible prediction of approximately 155.43. However, that model had a lower adjusted R-squared (~0.51) and a higher risk of overfitting due to the number of predictors relative to the sample size (47 observations).

Metric	Full Model (Q8.2)	PCA Model (Q9.1)
R-squared	~0.68	~0.65
Adjusted R-squared	~0.51	~0.60
Residual Std. Error	~22.91	244
Prediction for new city	~155.43	–1509.63
Overfitting Risk	High	Lower
Interpretability	High	Lower (until back-transformed)

Figure 4: Comparison Full Linear Regression model vs PCA-based regression model

Although the PCA model offers better generalization and reduces multicollinearity, it sacrifices interpretability. To express the model in terms of the original variables, I would need to back-transform the PCA coefficients using the loadings matrix and the original scaling parameters. Crime prediction models must be used responsibly. Without careful validation and transparency, they risk reinforcing systemic inequalities and misrepresenting vulnerable populations.

REFERENCES

https://www.geeksforgeeks.org/data-analysis/principal-component-analysis-pca/