Linear Regression

student

2023-09-27

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:

Variable	Value	Description
M	14.0	percentage of males aged 14–24 in total state population
So	0	indicator variable for a southern state
Ed	10.0	mean years of schooling of the population aged 25 years or over
Po1	12.0	per capita expenditure on police protection in 1960
Po2	15.5	per capita expenditure on police protection in 1959
LF	0.640	labour force participation rate of civilian urban males in the age-group 14-24
M.F	94.0	number of males per 100 females
Pop	150	state population in 1960 in hundred thousands
NW	1.1	percentage of nonwhites in the population
U1	0.120	unemployment rate of urban males 14–24
U2	3.6	unemployment rate of urban males 35–39
Wealth	3200	wealth: median value of transferable assets or family income
Ineq	20.1	income inequality: percentage of families earning below half the median income
Prob	0.04	probability of imprisonment: ratio of number of commitments to number of offenses
Time	39.	average time in months served by offenders in state prisons before their first release
Crime	-	crime rate: number of offenses per 100,000 population in 1960

Show your model (factors(predictors) used and their coefficients), the software output(single predication), and the quality of fit(R squared, p-value). Note that because there are only 47 data points and 15 predictors, you'll probably notice some overfitting.

Summary *I did 3 different models, one on the original data set, one scaled and one logarithmic. The scaled response provide the same result, it indicates that the scaling of the response variable did not significantly alter the regression relationships in this specific case. This suggests that relationships between predictors and the response, and scaling (which is a linear transformation) doesn't necessarily change those fundamental relationships, just the scale on which they're measured. I went ahead and evaluated the log data, however, as you can see, the resulting R-squared values and p-values are close, with a slightly lower p-value with the original and scaled models. The logarithmic transformation is a non-linear transformation which may make the model more robust, or make the residuals more normally distributed. This suggests that the transformation may have changed the model's understanding of the relationships between predictors and the response variable. The log residuals start a bit below 0, rise slightly above 0, and then drop below 0 again, it suggests there might be some non-linearity in the relationship between predictors and response. The log Quantile-Quantile plot points largely follow the straight diagonal line, it suggests that the residuals are approximately normally distributed. The scale location line is waving around 1, this may or may not indicate that the residuals have* *constant variance. Residuals vs. Leverage plot pattern suggests that there are some observations with higher leverage that might also have larger residuals, making them potentially influential. 26, 35, and 29 seem to be possible outlier(26) or other points of interest For the coefficients, given that the response is log-transformed, the interpretation of these coefficients is multiplicative in nature, with regards to the original crime rate:*

Given the results below I lean towards using the original response variable

Positive Coefficients: *as the predictor increases, the crime rate (in log scale) also increases; the crime rate is multiplied by a factor greater than 1.*

Negative Coefficients: *as the predictor increases, the crime rate (in log scale) decreases; the crime rate is multiplied by a factor between 0 and 1.*

R^2 *provides a measure of how well the model's predictions match the actual data. This value of 0.2314751 suggests a low to moderate amount of the variability in the log-transformed crime rate is explained by the selected principal components. Original R^2: 0.2433132 Scaled R^2: 0.2433132*

p-value *The p-value of 0.0043 suggests that the overall regression model is statistically significant at the 5% significance level, there's strong evidence to reject the null hypothesis that all the regression coefficients are zero. This means that the predictors (in this case, the principal components) are collectively useful in predicting the response variable. OG p-value: 0.00317775214807403 Scale p-value: 0.00317775214807403*

Begin Code

Load Libraries

```r
library(readr)
```

Load & View Crime Data

```
##      M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob
## 1 15.1  1  9.1  5.8  5.6 0.510  95.0  33 30.1 0.108 4.1   3940 26.1 0.084602
## 2 14.3  0 11.3 10.3  9.5 0.583 101.2  13 10.2 0.096 3.6   5570 19.4 0.029599
## 3 14.2  1  8.9  4.5  4.4 0.533  96.9  18 21.9 0.094 3.3   3180 25.0 0.083401
## 4 13.6  0 12.1 14.9 14.1 0.577  99.4 157  8.0 0.102 3.9   6730 16.7 0.015801
## 5 14.1  0 12.1 10.9 10.1 0.591  98.5  18  3.0 0.091 2.0   5780 17.4 0.041399
## 6 12.1  0 11.0 11.8 11.5 0.547  96.4  25  4.4 0.084 2.9   6890 12.6 0.034201
##      Time Crime
## 1 26.2011   791
## 2 25.2999  1635
## 3 24.3006   578
## 4 29.9012  1969
## 5 21.2998  1234
## 6 20.9995   682
```

PCA on Predictors

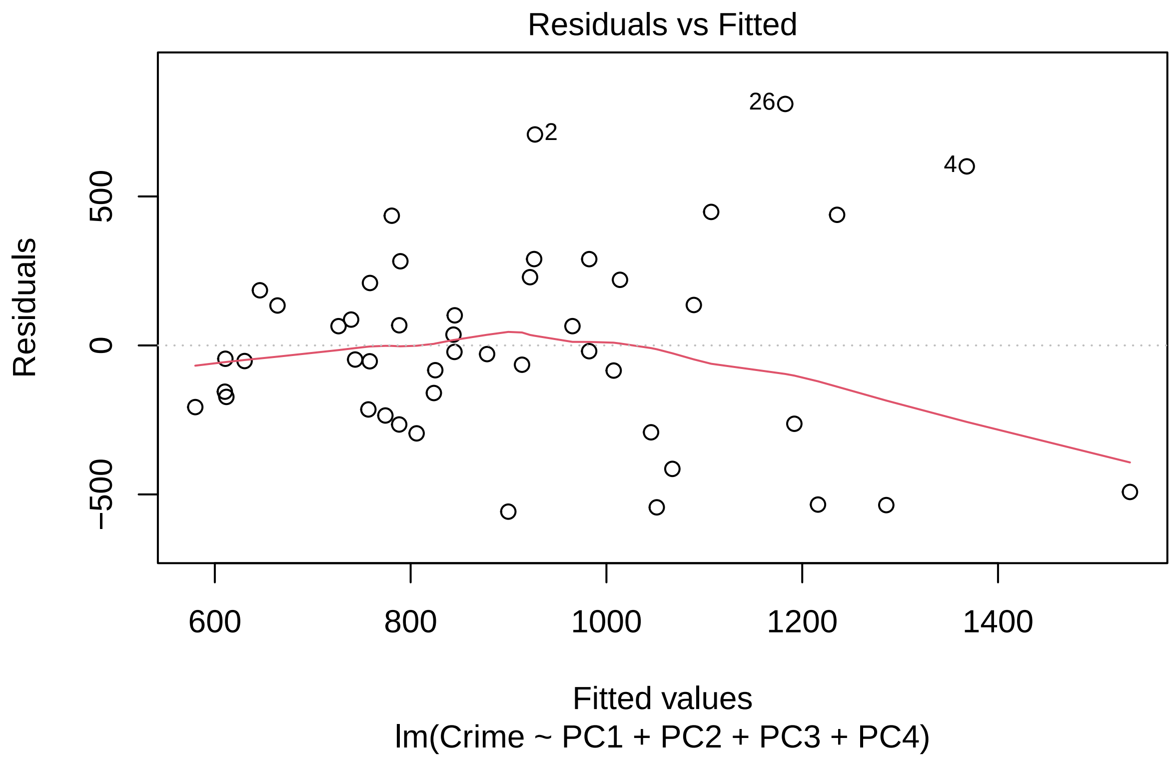
```r
# remove the response variable 'Crime'
predictor_data <- uscrime_data[,-which(names(uscrime_data) == "Crime")]

# perform PCA, scale the predictors
pca_result <- prcomp(predictor_data, scale. = TRUE)

summary(pca_result)
```

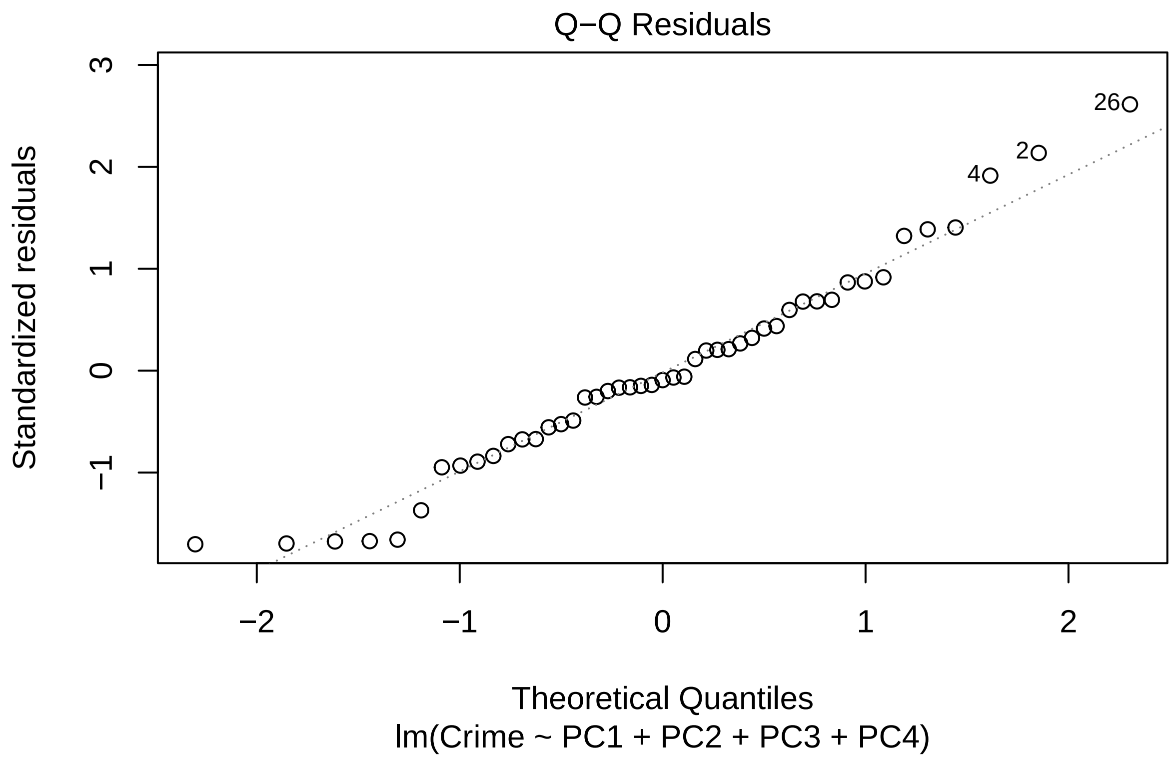
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4534 1.6739 1.4160 1.07806 0.97893 0.74377 0.56729
## Proportion of Variance 0.4013 0.1868 0.1337 0.07748 0.06389 0.03688 0.02145
## Cumulative Proportion  0.4013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

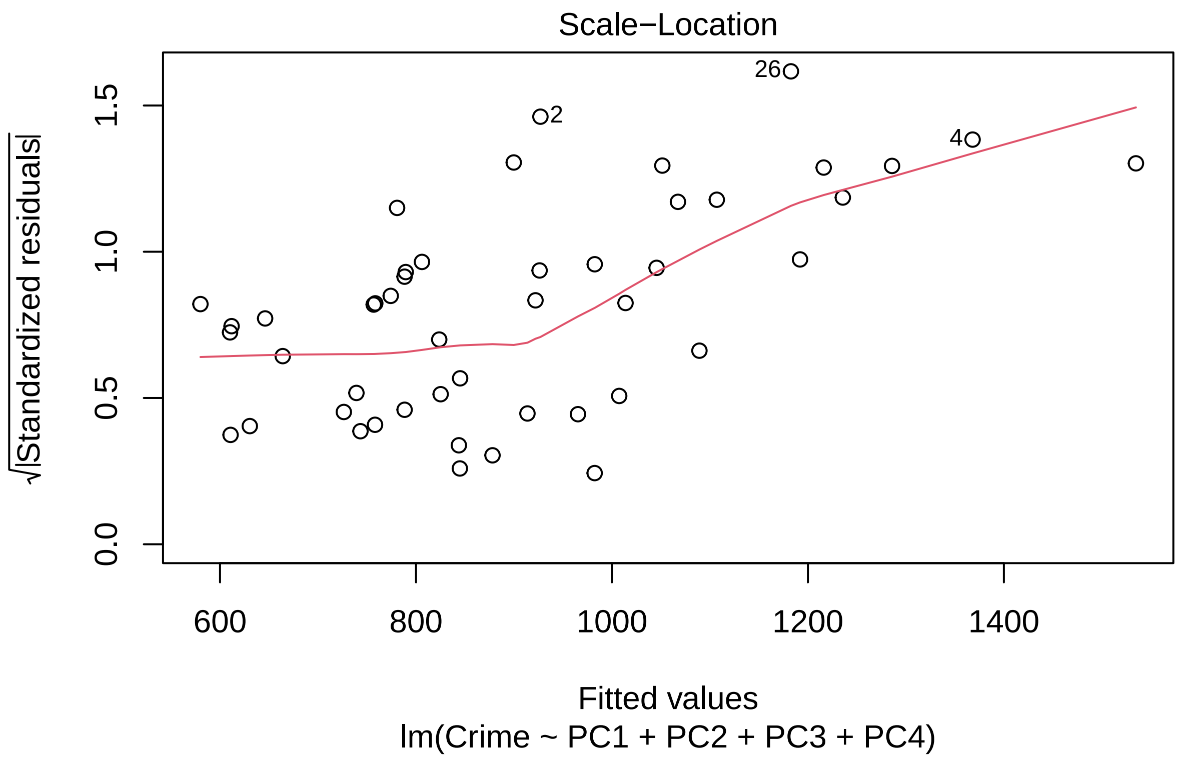
Regression with Original Response using top 4 PCA results that together capture about 80%

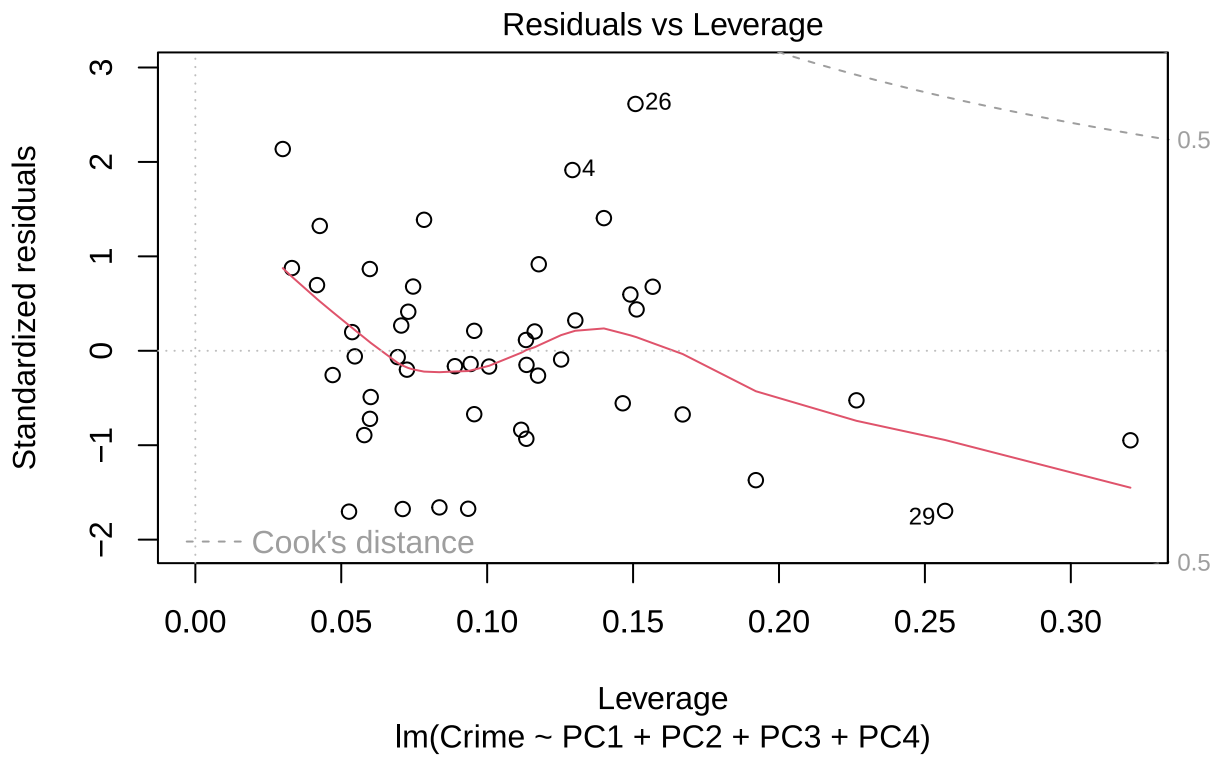
PC1: 40.13% PC1 + PC2: 58.80% (40.13% + 18.68%) PC1 + PC2 + PC3: 72.17% (58.80% + 13.37%) PC1 + PC2 + PC3 + PC4: 79.92% (72.17% + 7.75%)

```
# get scores for the first two PCs
PC_scores <- pca_result$x[,1:4]

# build regression model using the selected PCs
model_pca <- lm(Crime ~ PC1 + PC2 + PC3 + PC4, data = data.frame(Crime = uscrime_data$Crime,
                                                                 PC_scores))

# extract coefficients in terms of the original predictors
original_coeff <- as.matrix(pca_result$rotation[, 1:4]) %*% coef(model_pca)[-1]

# plot and print
plot(model_pca)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Crime ~ PC1 + PC2 + PC3 + PC4)
500
0
−500
600
800
1000
1200
1400
26
2
4


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Crime ~ PC1 + PC2 + PC3 + PC4)
26
2
4
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
26
2
4
600
800
1000
1200
1400
Fitted values
lm(Crime ~ PC1 + PC2 + PC3 + PC4)




```
print(original_coeff)
```

```
##              [,1]
## M      -21.277963
## So      10.223091
## Ed      14.352610
## Po1     63.456426
## Po2     64.557974
## LF     -14.005349
## M.F    -24.437572
## Pop     39.830667
## NW      15.434545
## U1     -27.222281
## U2       1.425902
## Wealth  38.607855
## Ineq   -27.536348
## Prob     3.295707
## Time    -6.612616
```

New Data into PCS space

```
# define the new observation
new_observation <- data.frame(M = 14.0,
                              So = 0,
                              Ed = 10.0,
```

```r
                              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.)

# transform the new observation into the PCA space
new_observation_PC <- as.matrix(scale(new_observation,
                                center = pca_result$center,
                                scale = pca_result$scale)) %*% pca_result$rotation[,1:4]

# predict using the model
predicted_value <- predict(model_pca, newdata = data.frame(PC1 = new_observation_PC[,1],
                                                           PC2 = new_observation_PC[,2],
                                                           PC3 = new_observation_PC[,3],
                                                           PC4 = new_observation_PC[,4]))

print(predicted_value)
```

```
##      PC1
## 1112.678
```

Original: R^2

```r
summary(model_pca)$adj.r.squared
```

```
## [1] 0.2433132
```

Original: p-value

```r
# F-statistic
f_stat <- summary(model_pca)$fstatistic[1]

# degrees of freedom
df1 <- summary(model_pca)$fstatistic[2]
df2 <- summary(model_pca)$fstatistic[3]

# calculate p-value
p_value <- 1 - pf(f_stat, df1, df2)

# print the results
print(paste("F-statistic:", f_stat))
```

```
## [1] "F-statistic: 4.69783386512055"
```

```r
print(paste("p-value:", p_value))
```

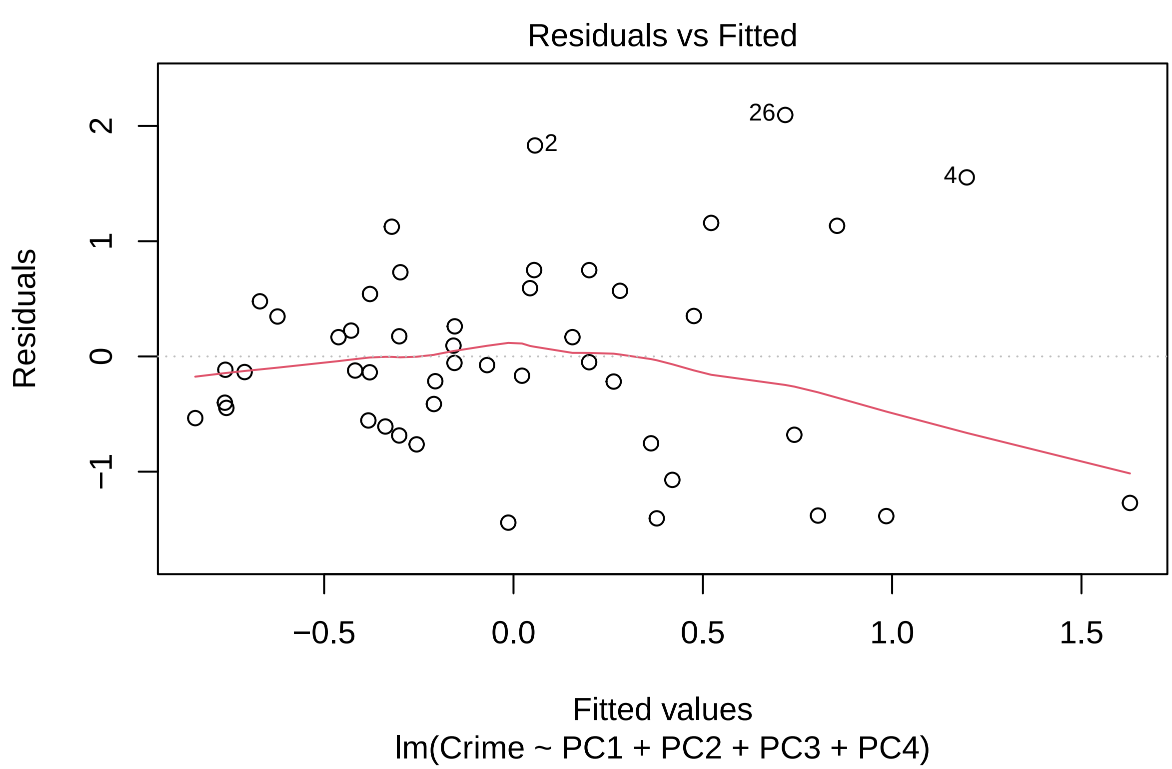
```
## [1] "p-value: 0.00317775214807403"
```

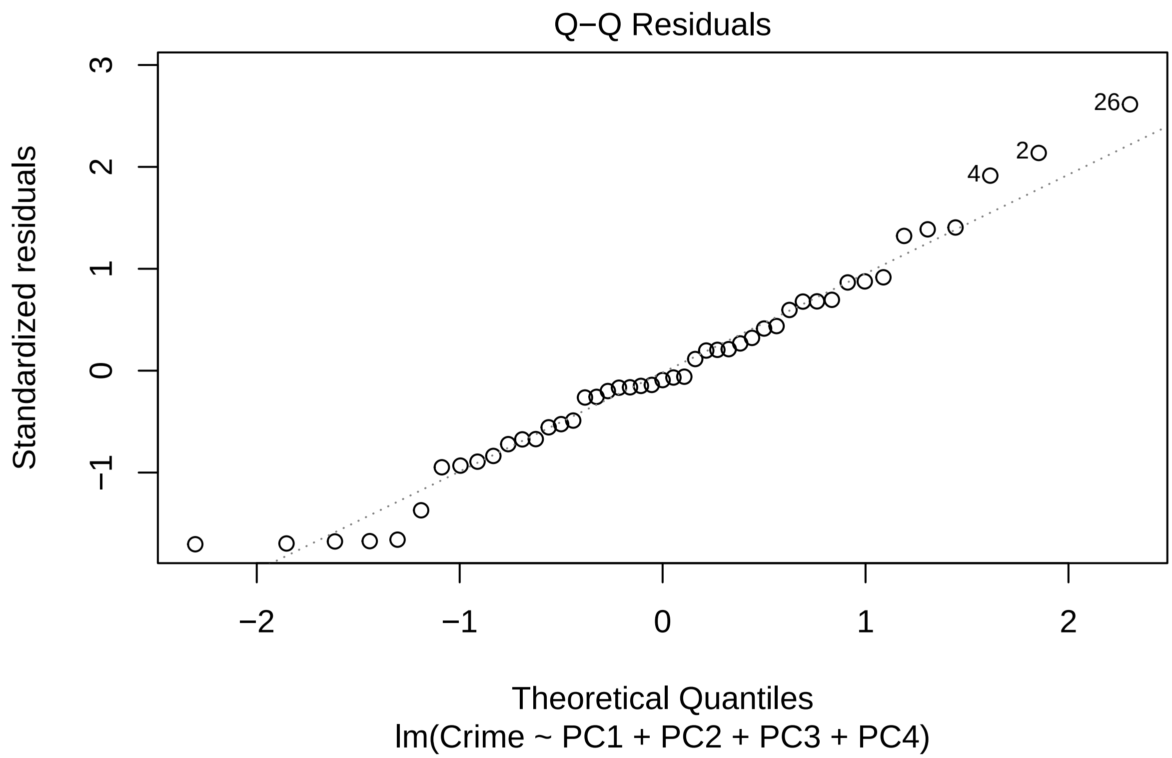
Response Transformation: Scaling

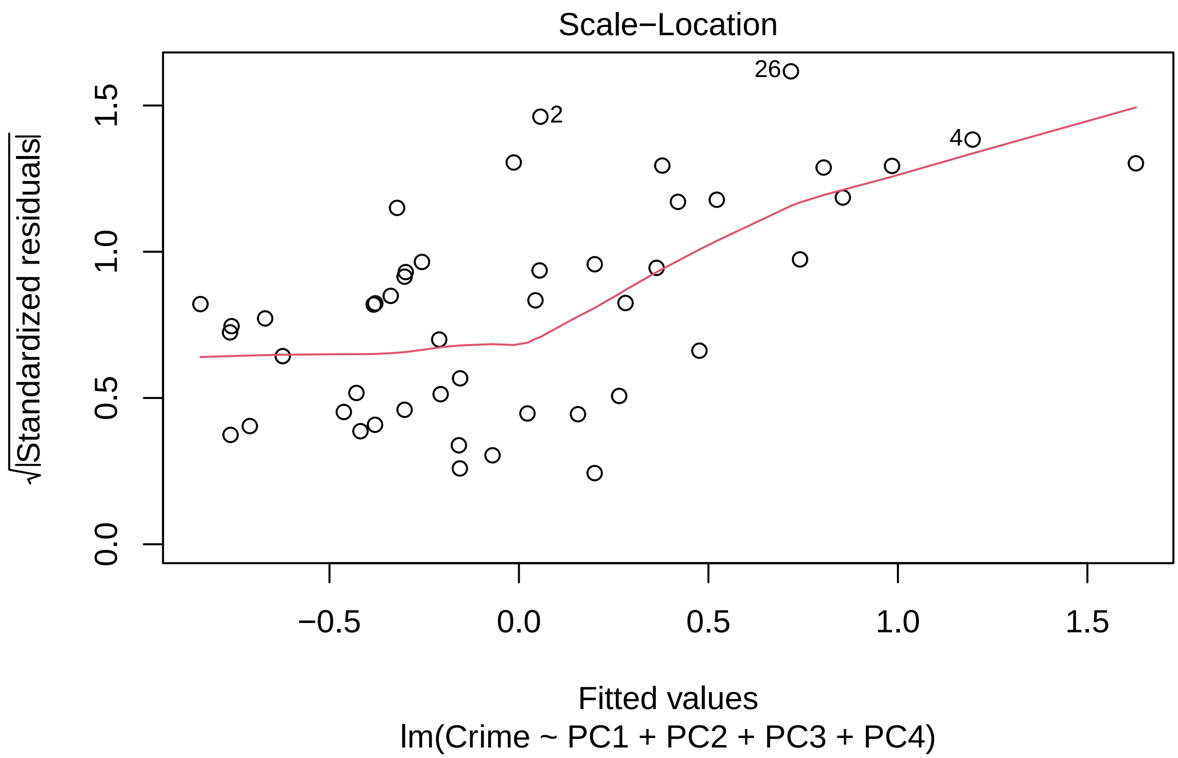
```r
uscrime_data <- read.csv("C:\\Users\\Public\\Documents\\gatech\\crime.csv")

# scaling the response
uscrime_data$Crime_scaled <- scale(uscrime_data$Crime)
```

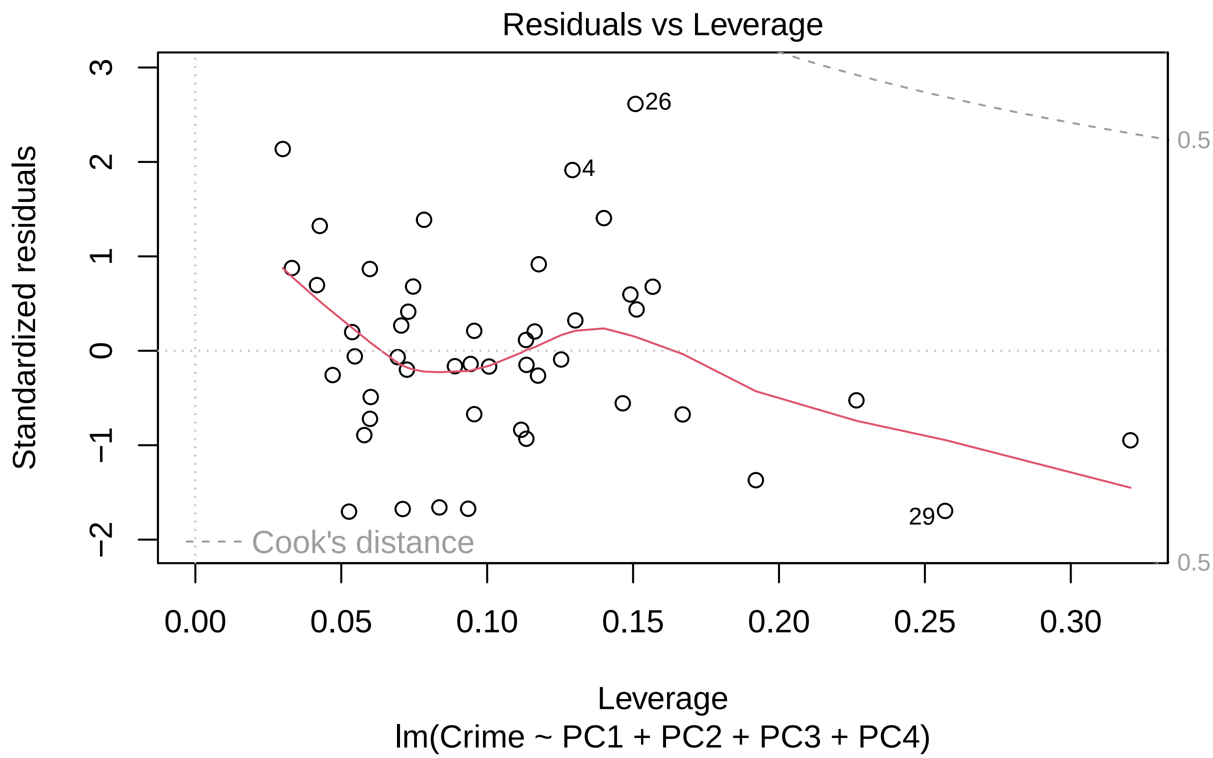
Regression with Scaled Response

```r
# get scores for the first two PCs
PC_scores <- pca_result$x[,1:4]

# build regression model using the selected PCs
model_pca <- lm(Crime ~ PC1 + PC2 + PC3 + PC4,
                data = data.frame(Crime = uscrime_data$Crime_scaled,
                                  PC_scores))

# extract coefficients in terms of the original predictors
original_coeff <- as.matrix(pca_result$rotation[, 1:4]) %*% coef(model_pca)[-1]

plot(model_pca)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Crime ~ PC1 + PC2 + PC3 + PC4)
26
2
4
3
2
1
0
−1
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.5
26
2
4
Fitted values
lm(Crime ~ PC1 + PC2 + PC3 + PC4)




```r
print(original_coeff)
```

```
##                [,1]
## M      -0.055015551
## So      0.026432464
## Ed      0.037109603
## Po1     0.164070698
## Po2     0.166918823
## LF     -0.036211737
## M.F    -0.063184924
## Pop     0.102984769
## NW      0.039907016
## U1     -0.070384971
## U2      0.003686762
## Wealth  0.099823110
## Ineq   -0.071197011
## Prob    0.008521265
## Time   -0.017097346
```

New Data: scaled response

```r
# transform the new observation into the PCA space
new_observation_PC <- as.matrix(scale(new_observation,
                                center = pca_result$center,
                                scale = pca_result$scale)) %*% pca_result$rotation[,1:4]
```

```r
# predict using the model
predicted_value <- predict(model_pca, newdata = data.frame(PC1 = new_observation_PC[,1],
                                                            PC2 = new_observation_PC[,2],
                                                            PC3 = new_observation_PC[,3],
                                                            PC4 = new_observation_PC[,4]))

original_scale_prediction <- (predicted_value * sd(uscrime_data$Crime)) + mean(uscrime_data$Crime)

print(original_scale_prediction)
```

```
##      PC1
## 1112.678
```

Scaling: R^2

```r
summary(model_pca)$adj.r.squared
```

```
## [1] 0.2433132
```

Scaling: p-value

```r
# extract F-statistic
f_stat <- summary(model_pca)$fstatistic[1]

# degrees of freedom
df1 <- summary(model_pca)$fstatistic[2]
df2 <- summary(model_pca)$fstatistic[3]

# calculate p-value
p_value <- 1 - pf(f_stat, df1, df2)

# results
print(paste("F-statistic:", f_stat))
```

```
## [1] "F-statistic: 4.69783386512055"
```

```r
print(paste("p-value:", p_value))
```

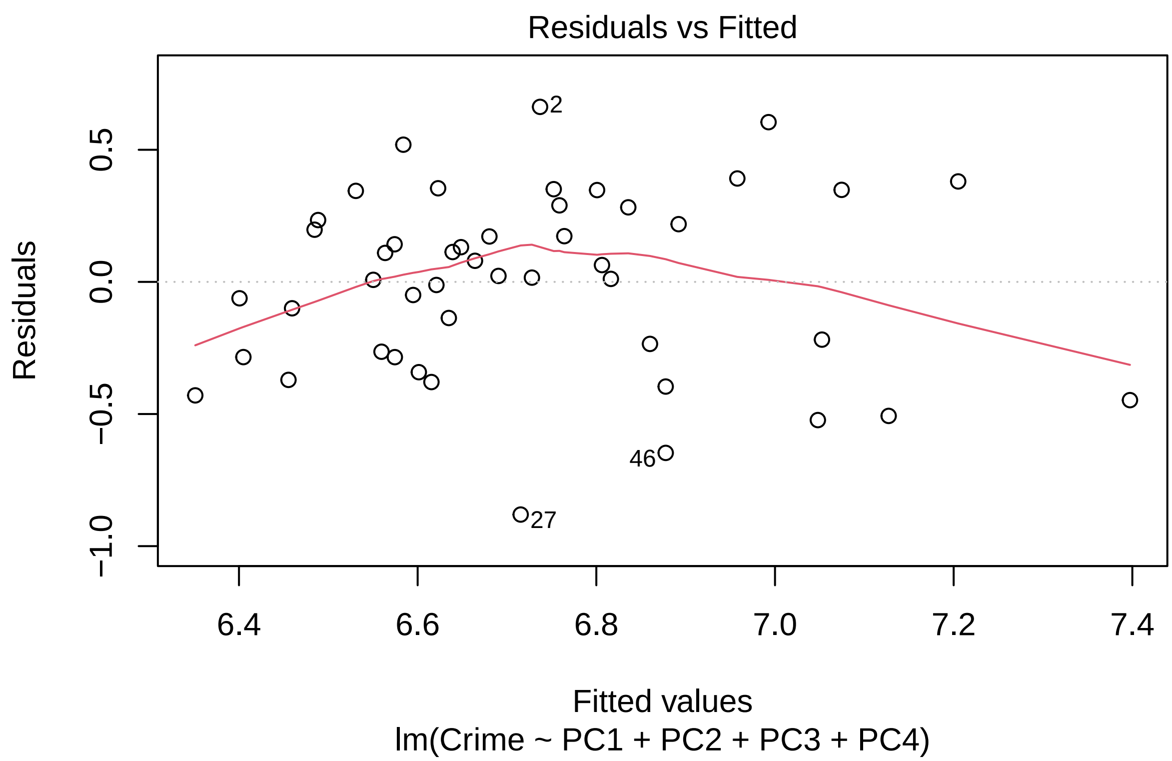
```
## [1] "p-value: 0.00317775214807403"
```

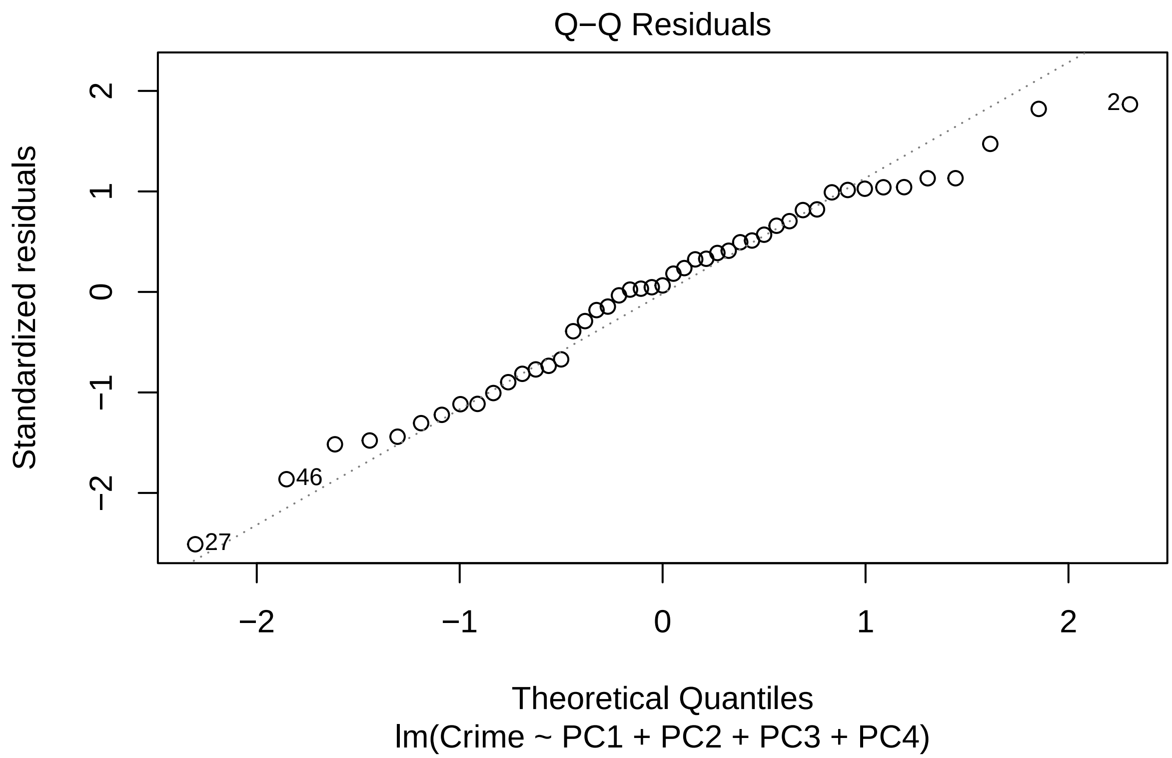
Log

```r
# crime data
uscrime_data <- read.csv("C:\\Users\\Public\\Documents\\gatech\\crime.csv")
# log-transforming the response
uscrime_data$Crime_log <- log(uscrime_data$Crime)
```

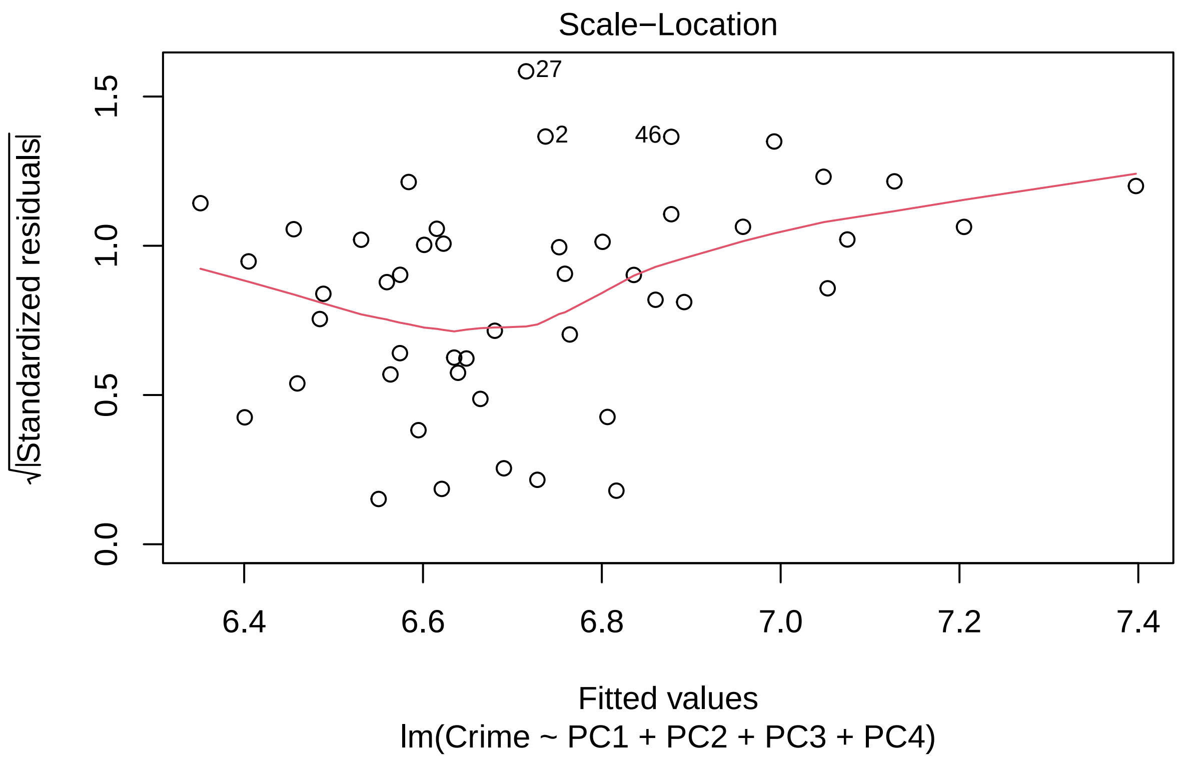
Regression with Log-Transformed Response

```r
# get scores for the first two PCs
PC_scores <- pca_result$x[,1:4]

# build regression model using the selected PCs
model_pca <- lm(Crime ~ PC1 + PC2 + PC3 + PC4,
                data = data.frame(Crime = uscrime_data$Crime_log,
                                  PC_scores))

# extract coefficients in terms of the original predictors
original_coeff <- as.matrix(pca_result$rotation[, 1:4]) %*% coef(model_pca)[-1]

plot(model_pca)
```

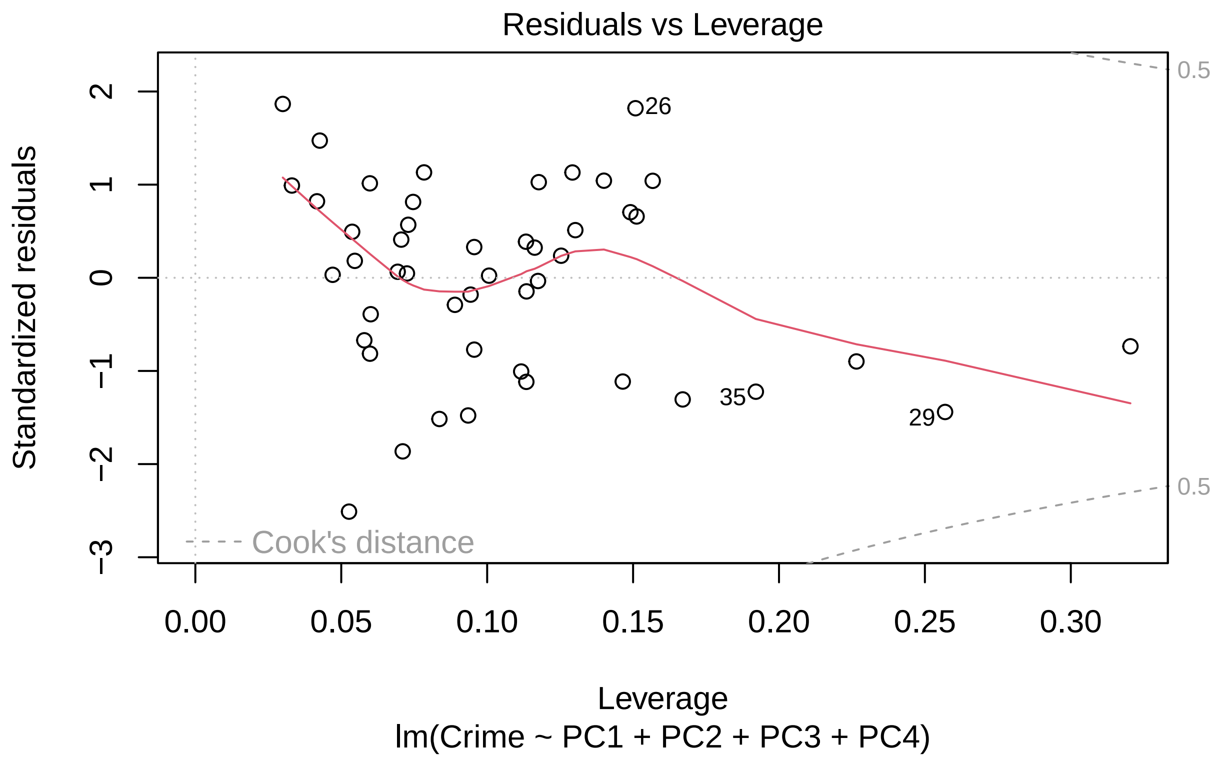




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Crime ~ PC1 + PC2 + PC3 + PC4)
27
46
2
−2
−1
0
1
2



Scale−Location
27
2
46
1.5
1.0
0.5
0.0
√|Standardized residuals|
6.4
6.6
6.8
7.0
7.2
7.4
Fitted values
lm(Crime ~ PC1 + PC2 + PC3 + PC4)




```r
print(original_coeff)
```

```
##                [,1]
## M      -0.019819845
## So      0.015464890
## Ed      0.013084055
## Po1     0.068681152
## Po2     0.069937915
## LF     -0.016320770
## M.F    -0.029523955
## Pop     0.043682383
## NW      0.021128738
## U1     -0.034619214
## U2     -0.001736897
## Wealth  0.039550711
## Ineq   -0.027197648
## Prob    0.006702271
## Time   -0.007102780
```

Transform

```r
# transform the new observation into the PCA space
new_observation_PC <- as.matrix(scale(new_observation,
                                      center = pca_result$center,
                                      scale = pca_result$scale)) %*% pca_result$rotation[,1:4]
```

```r
# predict using the model
predicted_value <- predict(model_pca, newdata = data.frame(PC1 = new_observation_PC[,1],
                                                            PC2 = new_observation_PC[,2],
                                                            PC3 = new_observation_PC[,3],
                                                            PC4 = new_observation_PC[,4]))

original_scale_prediction <- exp(predicted_value)

print(original_scale_prediction)
```

```
##      PC1
## 1037.789
```

R-Squared

```r
summary(model_pca)$adj.r.squared
```

```
## [1] 0.2314751
```

P-value

```r
# extract F-statistic
f_stat <- summary(model_pca)$fstatistic[1]

# degrees of freedom
df1 <- summary(model_pca)$fstatistic[2]
df2 <- summary(model_pca)$fstatistic[3]

# p-value
p_value <- 1 - pf(f_stat, df1, df2)

# results
print(paste("F-statistic:", f_stat))
```

```
## [1] "F-statistic: 4.46373082519864"
```

```r
print(paste("p-value:", p_value))
```

```
## [1] "p-value: 0.00426909291234767"
```