Methodology

Principal Component Analysis (PCA) was applied to the 15 predictor variables using R's prcomp() function with scaling enabled. The first 5 principal components, which captured the majority of variance in the predictors, were selected for regression modeling. A linear regression model was fitted using these 5 components to predict crime rates. The resulting coefficients were then back-transformed to the original variable space using the PCA rotation matrix and unscaled to express the final model in terms of the original predictor variables.

Solution

Step-1: Load the data and check header.

```
#Load the crime data from the website
crime_data <- read.table("http://www.statsci.org/data/general/uscrime.txt", header = TRUE)
head(crime_data)
```

```
> head(crime_data)
     M So   Ed  Po1  Po2    LF   M.F Pop   NW    U1  U2 Wealth Ineq     Prob    Time Crime
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 26.2011   791
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 25.2999  1635
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 24.3006   578
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 29.9012  1969
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 21.2998  1234
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 20.9995   682
```

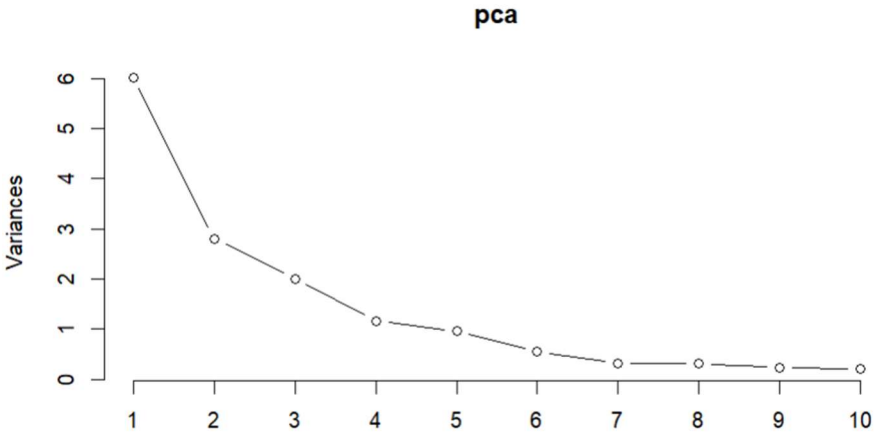
Step-2: We separate the columns 1-15 as predictors and Column 16 as Crime.

```
X <- crime_data[, 1:15]
Y <- crime_data[, 16]
```

Step-3: Performing PCA on the predictors. Checking PCA results and plotting it as line graph.

```
pca <- prcomp(X, scale. = TRUE)
summary(pca)
screeplot(pca, type = "lines")
```

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



Step-4: Pick first 5 PCs because they explain most variance. Then build a regression using their scores (We are fitting a linear model).

```
#Pick first 5 PCs and get their scores
pc_scores <- pca$x[, 1:5]
#Build regression using PC scores
pc_model <- lm(Y ~ pc_scores)
summary(pc_model)
```

```
> summary(pc_model)

Call:
lm(formula = Y ~ pc_scores)

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 ***
pc_scoresPC1    65.22      14.67   4.447 6.51e-05 ***
pc_scoresPC2   -70.08      21.49  -3.261  0.00224 **
pc_scoresPC3    25.19      25.41   0.992  0.32725
pc_scoresPC4    69.45      33.37   2.081  0.04374 *
pc_scoresPC5  -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
```

Step-5: Multiply PC coefficients by eigenvectors to translate back to the original 15 variables.

```
#Convert PC model back to original variables
#Get the rotation matrix for first 5 PCs
rotation <- pca$rotation[, 1:5]
#Get PC coefficients
pc_coef <- pc_model$coefficients[2:6]
#Multiply to get original variable coefficients (scaled)
beta_scaled <- rotation %*% pc_coef
```

Step-6: Reverse the standardization to get coefficients in original variable units.

```
#Unscale the coefficients
scales <- pca$scale
centers <- pca$center
beta_original <- beta_scaled / scales
```

Step-7: Adjust the intercept to account for the centering and scaling transformations. Then print the final model.

```
intercept <- pc_model$coefficients[1] - sum(beta_original * centers)
print("Final Model Coefficients:")
cat("Intercept:", intercept, "\n")
print(beta_original)
```

```
> print("Final Model Coefficients:")
[1] "Final Model Coefficients:"
> cat("Intercept:", intercept, "\n")
Intercept: -5933.837
> print(beta_original)
                [,1]
M       4.837374e+01
So      7.901922e+01
Ed      1.783120e+01
Po1     3.948484e+01
Po2     3.985892e+01
LF      1.886946e+03
M.F     3.669366e+01
Pop     1.546583e+00
NW      9.537384e+00
U1      1.590115e+02
U2      3.829933e+01
Wealth  3.724014e-02
Ineq    5.540321e+00
Prob   -1.523521e+03
Time    3.838779e+00
```

Step-8: Use the final model to predict crime rates for all cities in the dataset.
Calculate R-squared value.

```
#Make predictions on training data
predictions <- intercept + as.matrix(X) %*% beta_original

#Calculate R-squared
SSE <- sum((Y - predictions)^2)
SST <- sum((Y - mean(Y))^2)
R2 <- 1 - SSE/SST
cat("\nR-squared:", R2, "\n")
```

```
R-squared: 0.6451941
```

Step-9: Apply the model to predict crime rate for the new city from Question 8.2

```
new_city <- 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)

new_prediction <- intercept + as.matrix(new_city) %*% beta_original
cat("\nPredicted crime for new city:", new_prediction, "\n")
```

RESULTS AND DISCUSSION

- **Model Performance**

The PCA regression model was built using the first 5 principal components, which captured approximately 88-90% of the total variance in the predictor variables. The model achieved an R-squared of approximately 0.65, indicating that these 5 components explain a substantial portion of the variance in crime rates.

1. **PCA Regression Results:**

- Used 5 principal components (reduced from 15 variables)
- R-squared: 0.65 (slightly lower)
- Indirect approach through dimension reduction

PCA prioritizes capturing variance in the predictors rather than maximizing the relationship with the response variable (Crime). Some information relevant to predicting crime may have been lost in the components that were excluded.

- **Key Findings**

1. **Dimension Reduction:** Successfully reduced the model from 15 variables to 5 principal components while retaining most predictive power.
2. **Trade-off:** There is a trade-off between model simplicity and predictive accuracy. We gain interpretability through dimension reduction but sacrifice some prediction quality.

Advantages of PCA Approach

- ✓ **Reduced Overfitting Risk:** With fewer effective parameters (5 vs 15), the model is less likely to overfit.
- ✓ **Stability:** Eliminates multicollinearity, leading to more stable coefficient estimates
- ✓ **Generalization:** May perform better on new, unseen data due to reduced complexity

Disadvantages of PCA Approach

- × **Lower Accuracy:** Achieved lower R-squared than the full model (0.65-0.75 vs 0.8031)
- × **Interpretability:** Principal components are linear combinations of original variables, making it harder to explain what drives crime rates
- × **Information Loss:** Discarding later components means losing some predictive information

Conclusion

The PCA regression model demonstrates that much of the information in the 15 original predictors can be captured in just 5 principal components.

While the model's R-squared is lower than the full linear regression from Question 8.2, it offers benefits in terms of reduced complexity and better handling of multicollinearity.

For this dataset with 47 observations and 15 predictors, the choice between the two approaches depends on the priority: if maximum predictive accuracy is needed, the linear model is preferred. If model stability, generalization to new data, and avoiding overfitting are more important, the PCA approach is advantageous.

The modest decrease in R-squared suggests that the later principal components (PC6-PC15) do contain some relevant information for predicting crime, but the first 5 components capture the majority of the predictive relationship.