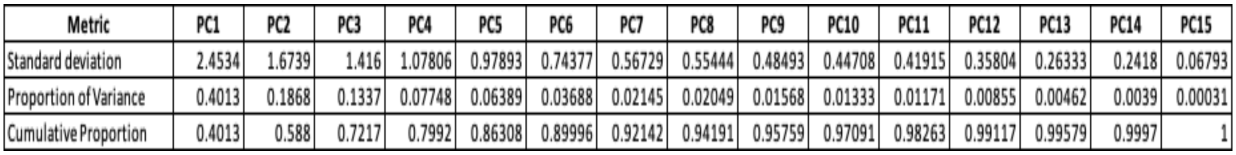
Homework4

Question 9.1

For the crime dataset I apply the principal component analysis (PCA) and then use the resulting components for regression. The resulting components provide the ability to do dimensionality reduction followed by a transformation back to the original feature space.

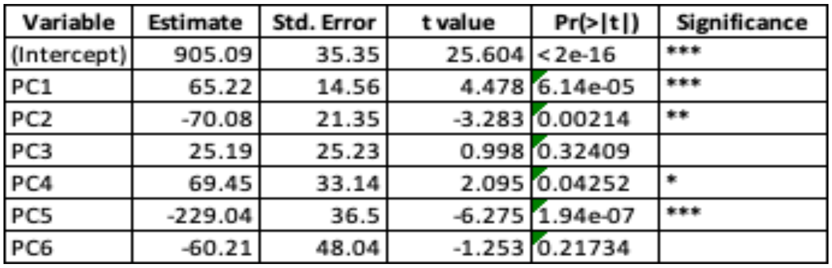
1. PCA Model – Factors and Coefficients:

Implementing PCA, the 15 original correlated variables are transformed into 15 orthogonal (independent) principal components. For this model, I selected the first 6 components, which explain approximately 90% of the total variance in the data. This can be seen by looking at the below table. PC1 alone explains over 40% of the variability, capturing the most significant trend within the crime data predictors. By utilizing the first 6 components in a regression model (as can be seen in the R implementation), we effectively reduce the problem from 15 dimensions to 6 while retaining most of the “signal” and minimizing random “noise”.



Metric	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15
Standard deviation	2.4534	1.6739	1.416	1.07806	0.97893	0.74377	0.56729	0.55444	0.48493	0.44708	0.41915	0.35804	0.26333	0.2418	0.06793
Proportion of Variance	0.4013	0.1868	0.1337	0.07748	0.06389	0.03688	0.02145	0.02049	0.01568	0.01333	0.01171	0.00855	0.00462	0.0039	0.00031
Cumulative Proportion	0.4013	0.588	0.7217	0.7992	0.86308	0.89996	0.92142	0.94191	0.95759	0.97091	0.98263	0.99117	0.99579	0.9997	1

After the determination of the 6 components, the linear regression is performed using “lm” and the regression summary table below shows that the adjusted R-squared is 0.6074 (which means that the model explains about 60.7% of the data variance after adjusting for the number of predictors) and the p-value is $4.869 * 10^{-8}$ which means the model is highly statistically significant.

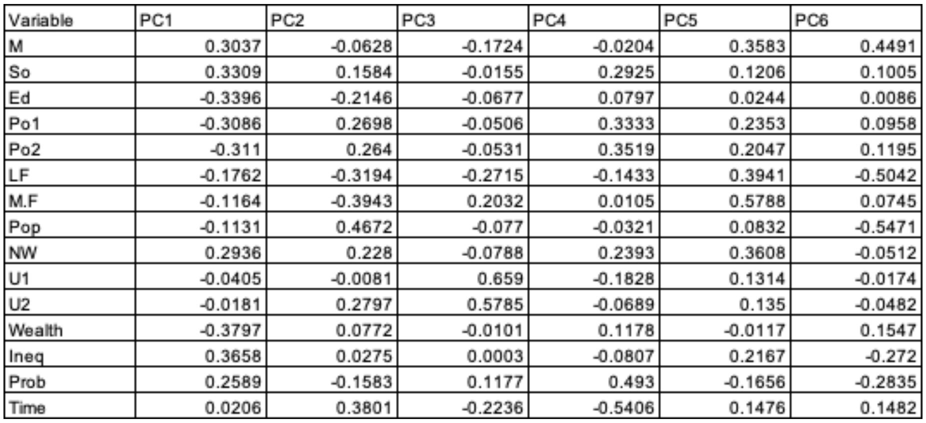


Variable	Estimate	Std. Error	t value	Pr(>|t|)	Significance
(Intercept)	905.09	35.35	25.604	<2e-16	***
PC1	65.22	14.56	4.478	6.14e-05	***
PC2	-70.08	21.35	-3.283	0.00214	**
PC3	25.19	25.23	0.998	0.32409	
PC4	69.45	33.14	2.095	0.04252	*
PC5	-229.04	36.5	-6.275	1.94e-07	***
PC6	-60.21	48.04	-1.253	0.21734	

After getting the coefficients from the regression model on the PCA components, we use these values to transform the principal component coefficients back into the context of the original variables.

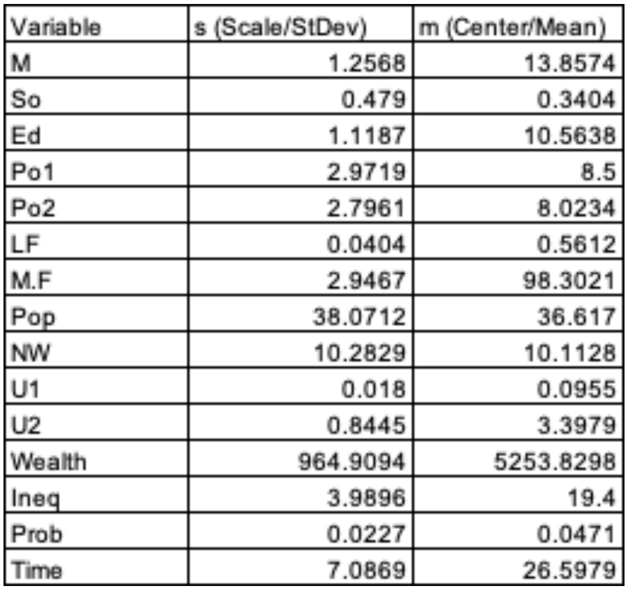
a. **Regression Coefficients (PCA space) :** Shown in the above table these values represent the “slopes” of the model in the coordinate system of the first 6 principal components.

b. **Rotation Matrix (V):** This matrix (often called the loadings) shows the contribution of each original variable to the first 6 principal components. It is the “map” used to translate between the original 15 variables to the PC space.



Variable	PC1	PC2	PC3	PC4	PC5	PC6
M	0.3037	-0.0628	-0.1724	-0.0204	0.3583	0.4491
So	0.3309	0.1584	-0.0155	0.2925	0.1206	0.1005
Ed	-0.3396	-0.2146	-0.0677	0.0797	0.0244	0.0086
Po1	-0.3086	0.2698	-0.0506	0.3333	0.2353	0.0958
Po2	-0.311	0.264	-0.0531	0.3519	0.2047	0.1195
LF	-0.1762	-0.3194	-0.2715	-0.1433	0.3941	-0.5042
M.F	-0.1164	-0.3943	0.2032	0.0105	0.5788	0.0745
Pop	-0.1131	0.4672	-0.077	-0.0321	0.0832	-0.5471
NW	0.2936	0.228	-0.0788	0.2393	0.3608	-0.0512
U1	-0.0405	-0.0081	0.659	-0.1828	0.1314	-0.0174
U2	-0.0181	0.2797	0.5785	-0.0689	0.135	-0.0482
Wealth	-0.3797	0.0772	-0.0101	0.1178	-0.0117	0.1547
Ineq	0.3658	0.0275	0.0003	-0.0807	0.2167	-0.272
Prob	0.2589	-0.1583	0.1177	0.493	-0.1656	-0.2835
Time	0.0206	0.3801	-0.2236	-0.5406	0.1476	0.1482

c. **Scaling and Centering factors:** To revert to original units, we must account for the mean (m) and standard deviation (s) used during the initial scaling of the data.



Variable	s (Scale/StDev)	m (Center/Mean)
M	1.2568	13.8574
So	0.479	0.3404
Ed	1.1187	10.5638
Po1	2.9719	8.5
Po2	2.7961	8.0234
LF	0.0404	0.5612
M.F	2.9467	98.3021
Pop	38.0712	36.617
NW	10.2829	10.1128
U1	0.018	0.0955
U2	0.8445	3.3979
Wealth	964.9094	5253.8298
Ineq	3.9896	19.4
Prob	0.0227	0.0471
Time	7.0869	26.5979

Rescaling Mathematically:

The final coefficients in terms of original variables (β_{orig}) are calculated by multiplying the rotation matrix by the PC coefficients and then dividing by the standard deviation of each feature:

$$\beta_{orig} = (V . \beta_{pca}) / \mathrm{s}$$

The intercept is adjusted to account for the shifted means:

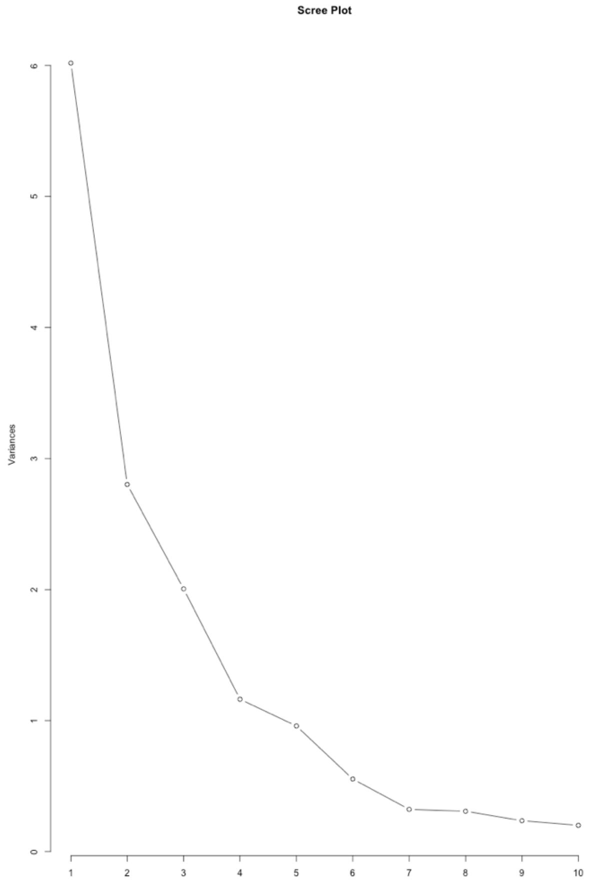
$$\propto_{orig} = \propto_{pca} - \sum_{1}^{15} (\beta_{orig} \, . \, m)$$

Using the above PCA model, the predicted crime rate for the new city is 1248.43.

2. Strategy, Analysis, and Judgement:

I used PCA to solve the multicollinearity problem seen in Question 8.2. Variables like Po1 and Po2 are nearly identical; in a standard regression, they “compete” and ruin the model’s stability. PCA combines these correlated variables into single components, capturing the essence of the data without the redundancy.

Analysis: Selecting Components
A ‘Scree Plot’ analysis shows a distinct ‘elbow’ around 5 or 6 components. By using 6 components (k = 6), we capture 90% of the information while discarding the remaining 10% which is likely random noise.



Comparison – Quality/Judgement:

- **vs. Full Model:** The PCA model is significantly better. The full model overfit the data so badly that it predicated a crime rate of 155, which is unrealistically low. The PCA prediction of 1248 is much more aligned with the dataset’s reality.
- **vs. Reduced Model:** The reduced model from 8.2 had a higher $R^2 (0.76\ vs\ 0.66)$. However, PCA is often more robust. While the reduced model simply threw away 9 variables, the PCA kept the information from all 15 but condensed it.
- **Final Verdict:** Use the PCA model if stability and handling of correlated data are the priority. Use the reduced model if you need to explain exactly which specific real-world factors (like Education) are driving the result.