

10

# STAT 8003, Homework 10

Group # 8

Members: Nooreen Dabbish, Yinghui Lu, Anastasia Vishnyakova

November 21, 2013

**Problem 1.** Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The data set contains the following columns:

Variable	Description
M	percentage of males aged 14-24 in total state population
So	indicator variable for a southern state
Ed	mean years of schooling of the population aged 25 years or over
Po1	per capita expenditure on police protection in 1960
Po2	per capita expenditure on police protection in 1959
LF	labour force participation rate of civilian urban males in the age-group 14-24
M.F	number of males per 100 females
Pop	state population in 1960 in hundred thousands
NW	percentage of nonwhites in the population
U1	unemployment rate of urban males 14-24
U2	unemployment rate of urban males 35-39
Wealth	wealth: median value of transferable assets or family income
Ineq	income inequality: percentage of families earning below half the median income
Prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	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

For the following problems, you should NOT use "lm" function or any existing function which yields the results immediately. Write your own code to solve the problem.

a) Plot the scatter plot matrix between these variables. Describe the pattern you observe. We are interested in studying the adjusted effect of punishment (Time or Prob) on Crime. Is it a good idea to include both variables in the model?

Scatter plot matrix is included below. We observed some correlations between variables from the scatterplot illustration. We observe a rather strong correlation between Ineq and Educ, Ineq and Wealth, Wealth and Educ, Wealth and Po1, Wealth and Po2, Po2 and Po1, Crime and Po1, U2 and U1. Correlation between Po1 and Po2 seemed strongest among all covariates.

We also include a scatterplot of Crime vs. Time and Prob. From the plot we can see more clearly that Prob has a negative correlation with Crime, but the relationship between Crime and Time is not that obvious. This is further confirmed by the correlation coefficient, which are -0.43 (Crime vs. Prob) and 0.15 (Crime vs. Time). There is also a negative correlation between Time and Prob, and the correlation coefficient is -0.44. We believe that it is not a good idea to include both Time and Prob. variables in the model because they are correlated and the collinearity would reduce the quality of the LSE estimate by inflating the parameter variances.

b) Construct a linear model to study the relationship between Crime ( $Y$ ) and Prob, adjusting for the effect of the 13 characteristics variables (M, So, ..., Ineq). Note that we don't want Time to be included in the model. Is the effect of Prob significant on Crime after adjusting for other characteristic variables? What is the p-value? Please use R to show the results.

Denote the Crime rate as  $Y$ . 13 explanatory variables and Prob are included in the multiple linear regression model. Let

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}; X = \begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,p-1} \\ 1 & X_{2,1} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,p-1} \end{pmatrix}; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix}; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

And in this case  $p = 15$ .

Then the linear model is

$$Y = X\beta + \epsilon$$

The model is assumed to satisfy the following assumptions:  $E(\epsilon) = 0$  and  $Var(\epsilon) = \sigma^2 I$ .

First, we work on estimating the parameters in the model:

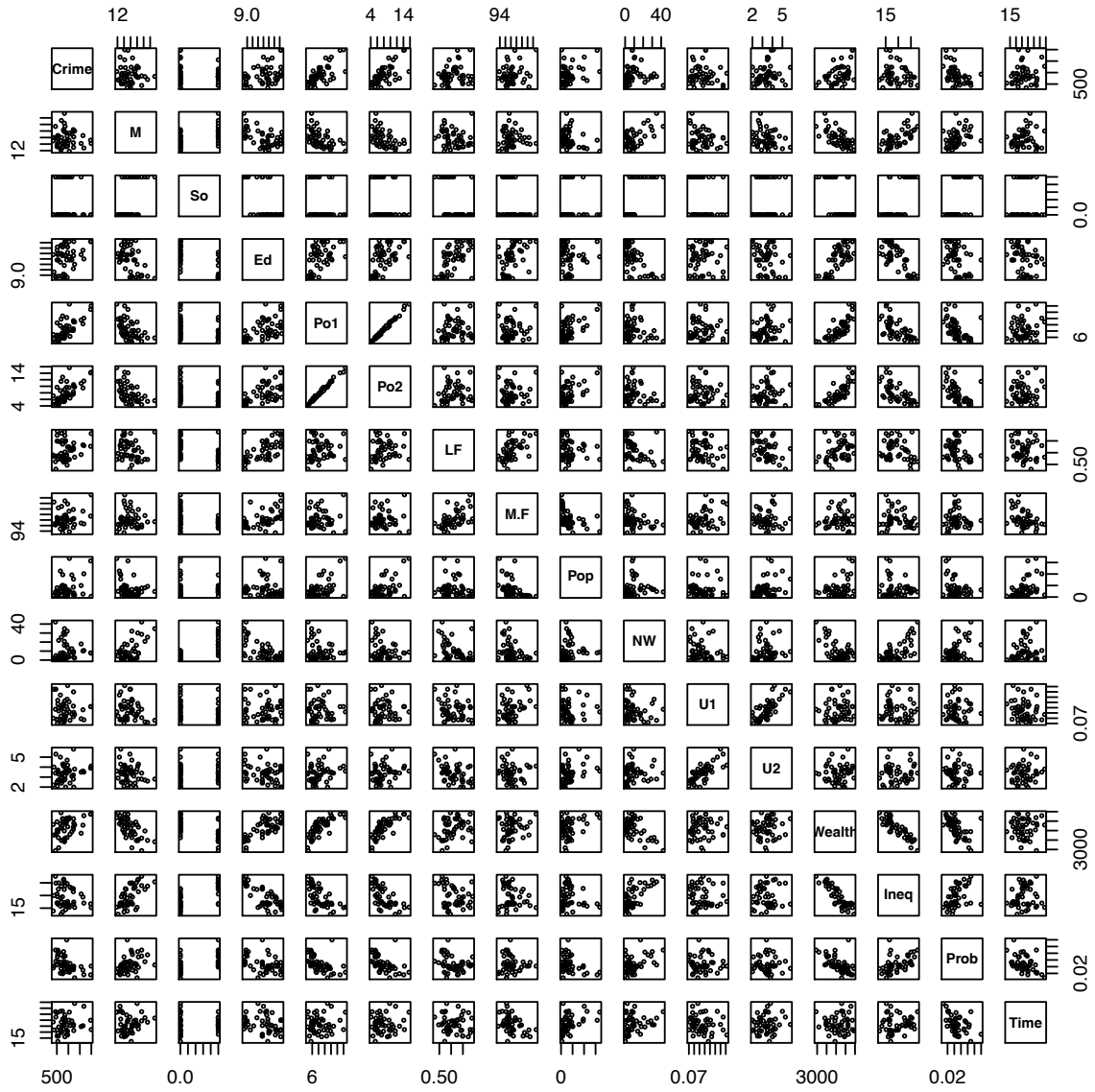


Figure 1: Scatter Plot Matrix of Crime Model

$$\hat{\beta}_{LSE} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}_{LSE}^2 = \frac{Y^T Y - Y^T P_X Y}{n - p}$$

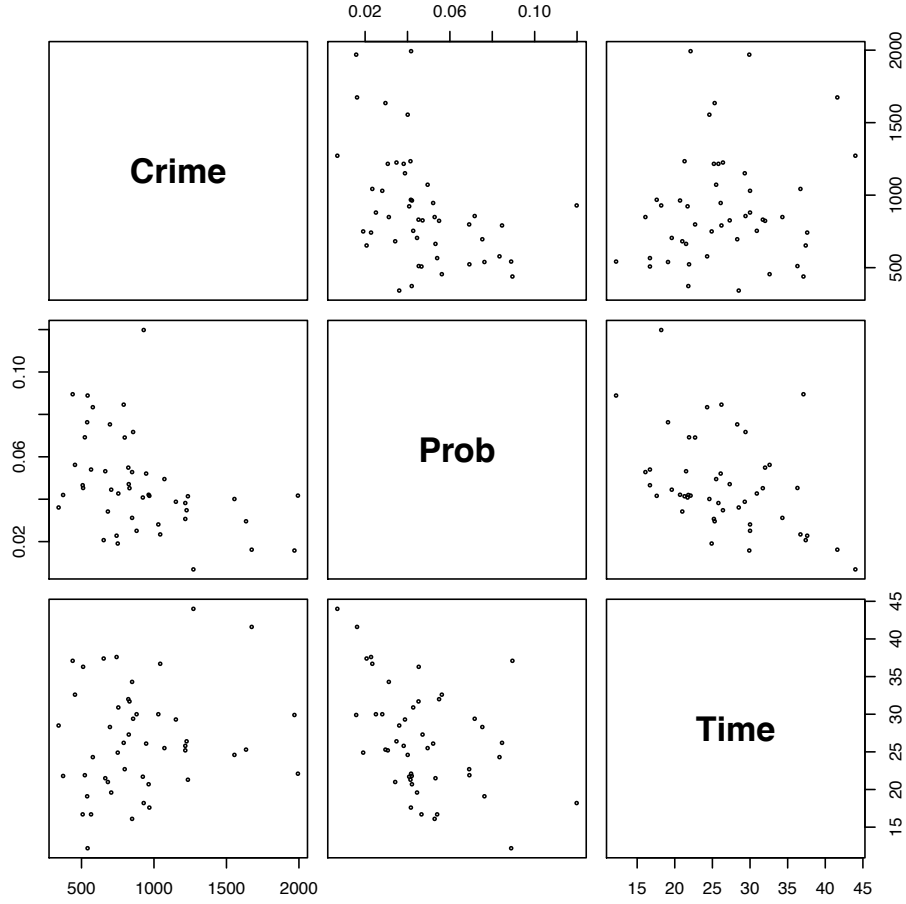


Figure 2: Scatter Plot Matrix Crime vs. Time and vs. Prob.

$$\widehat{Var}(\hat{\beta}_{LSE}) = \hat{\sigma}^2(X^T X)^{-1}$$

Where projection matrix is  $P_X = X(X^T X)^{-1}X^T$

Now we want to test the effect of Prob on Crime adjusting for other characteristic variables.

Our null hypothesis is  $H_0: \beta_{14} = 0$  vs.  $H_a: \beta_{14} \neq 0$ . We transform the hypothesis in this way:  $H_0: q^T \beta = 0$  vs.  $H_a: q^T \beta \neq 0$ . In this case  $q = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)^T$ .

We know that

$$\hat{\beta}_{LSE} \sim N(\beta, \sigma^2(X^T X)^{-1})$$

then

$$q^T \hat{\beta}_{LSE} \sim N(q^T \beta, \sigma^2 q^T (X^T X)^{-1} q)$$

Since  $\sigma^2$  is unknown, we use  $\hat{\sigma}^2$  to replace it, then we have

$$\frac{q^T \hat{\beta}_{LSE} - q^T \beta}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \sim T_{n-p}$$

Under  $H_0$ ,

$$\frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \sim T_{n-p}$$

The GLR test leads to rejecting  $H_0$  when

$$\left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right|$$

is large. Therefore, to control type I error at level 5%, we reject  $H_0$  if

$$\left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right| > t_{n-p}^{-1}(1 - 2.5\%)$$

Calculating it using R, we got

$$\left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right| = 2.28 > t_{n-p}^{-1}(1 - 2.5\%) = 2.04$$

According to this result, we tend to reject the null hypothesis and say that the effect of Prob might be significant on crime after adjusting for other characteristic variables. Then

$$p\text{-value} = 2P \left( T_{n-p} > \left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right| \right)$$

Calculate it using R, the result is

$$p\text{-value} = 0.029 < 0.05$$

c) Perform the same analysis to study the adjusted effect of Time on Crime, adjusting for 13 characteristic variables. Compare to model b), which variable do you think has a stronger adjusted effect on Crime? Why?

Here we performed a similar analysis as in part b testing the effect of Time on Crime adjusting for other characteristic variables. We use a similar model as part b), except changing the last variable of 'Prob' into 'Time'. Then, calculating it using R, we got

$$\left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right| = 0.829 < t_{n-p}^{-1}(1 - 2.5\%) = 2.04$$

$$p\text{-value} = 2P \left( T_{n-p} > \left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right| \right) = 0.413 > 0.05$$

According to the above result, we tend not to reject the null hypothesis, that is to say adjusted effect of time on crime may not be significant, adjusting for the 13 variables.

Compare to model b), we see that Prob has a significant effect on crime, but Time does not, adjusting for the 13 characteristic variables. In other word, Prob might have a significant effect on crime rate on and beyond the other 13 variables, but within the context of the other 13 variables, time might not has much to do with the crime rate. So we think that Prob has a stronger adjusted effect on Crime, comparing to Time according to the above two models.

d) If the purpose is to only study the adjusted effect of Time on Crime, is it a serious problem that we include 13 characteristic variables which might be correlated in the model? If the purpose is to study the effect of all these covariates (Time and 13 characteristic variables) on Crime, is it a serious problem that we include in total 14 variables in the model? Use the data to justify your answer.

From the correlation coefficient matrix we can see that Time does correlated with some of the other variables (See Appendix 1). So if the researcher's main interest is to study the relationship between Time and Crime, and we only construct a simple model to study the relationship of Time and Crime, it might give us a bias result, since we did not take into account the other factors that might affect the their relationship. It would be better to adjust for other covariates, to see whether Time has a significant effect on and beyond the other factors. In that case, whether the other 13 variables are highly correlated or not may not be a serious problem.

But if the researcher's aim is to study the effect of each individual variable, the collinearity among these 14 variables would be a serious problem. As there exists high collinearity among these variables (ie. Po1 vs Po2), the standard errors of the correspondingly LSE of coefficients would be very unstable, usually very large. Then it may not give valid results about any individual variable, or about which variables are redundant with respect to others. We calculate it in R, and listed the standard

error of each LSE below. We can see that many of the standard errors are even bigger than the LSE themselves (ie. So, Po2, LF, M.F, Pop, NW, Time)

	LSE	SE
Intercept	-7212.7437856	1606.0622748
M	94.9946763	43.8339698
So	-49.5555210	155.1898240
Ed	183.7237633	65.4164605
Po1	141.2917050	108.9389835
Po2	-41.8161134	119.2723821
LF	-372.5932631	1542.7641307
M.F	19.6162752	21.4300541
Pop	-0.7388713	1.3594933
NW	-0.1040350	6.4932578
U1	-5503.8047531	4435.7730361
U2	169.6183482	86.7969303
Wealth	0.1185483	0.1087291
Ineq	77.8200200	23.6874697
Time	5.1678449	6.2336979

Also, we observed that some of the coefficients have signs not as anticipated and high errors. For example, from examining the scatter plot matrix, we expected a positive correlation between Po2 and Crime, however we observed a negative sign on the coefficient. In another example, we expected a positive slope for Pop but got a negative sign on the coefficient.

But multicollinearity might not reduce the reliability of the model as a whole. We also calculate the overall fit of the whole model in R. It turns out that,  $F - statistic = 7.832$  on 14 and 32df,  $p\text{-value} = 8.216e - 07$ . That is though there are colliearity among the variables, we can treat them as an entire bundle to predict the outcome variable.

Hence, we say that the collearity among the covariates might not reduce reliability of the model as a whole, but it does affects calculations of the effects regarding each individual variables.

e) Now suppose we take the first four principal components of  $X = (X_1; \dots; X_{13})$  (13 characteristic variables) to be the variables included in the model, together with Prob. Please write out the model. What is the LSE of the coefficient for Prob? What is its variance?

First of all, we will try to find the first four principal components of  $X = (X_1, \dots, X_{13})$ .  $X_1$  through  $X_{13}$  denote M, So, Ed, Po1, Po2, LF, M.F, Pop, NW, U1, U2, Wealth and

Ineq, respectively. All of the above values are centered  
According to singular value decomposition,

$$X = UDV^T$$

Here, U and V are  $n \times p$  and  $p \times p$  orthogonal matrices. The eigenvalue decomposition is

$$X^T X = VD^2V^T$$

The columns of  $V$  are eigenvectors of  $X^T X$ . Then the first four principal component of  $X$  is

$$\text{pc1: } z_1 = Xv_1$$

$$\text{pc2: } z_2 = Xv_2$$

$$\text{pc3: } z_3 = Xv_3$$

$$\text{pc4: } z_4 = Xv_4$$

We then include the first four component of  $X$ , together with  $Prob$  (denoted as  $P$ , centered) to construct another linear model,

$$Y = X\beta + \epsilon$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} z_{1.1} & z_{1.2} & z_{1.3} & z_{1.4} & p_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ z_{n.1} & z_{n.2} & z_{n.3} & z_{n.4} & p_n \end{pmatrix}; \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}; \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Then,

$$\hat{\beta}_{LSE} = (X^T X)^{-1} X^T Y$$

$$\begin{aligned} \widehat{Var}(\hat{\beta}_{LSE}) &= \hat{\sigma}^2 (X^T X)^{-1} \\ &= \frac{Y^T Y - Y^T P_X Y}{n - 5} (X^T X)^{-1} \end{aligned}$$

Calculating it in R, we have

$$\hat{\beta}_5 = -0.28$$



$$\widehat{Var}(\hat{\beta}_5) = 0.012$$

f) Let's get back to Model b). Now suppose the investigator is not only interested in the adjusted effect of Prob, but also the effect of other variables. The investigator noticed that the variables Po1 and Po2 are highly correlated. It is reasonable to include  $(Po1 + Po2)/2$  in the model to replace these two variables. How to test whether such replacement is 0? Formulate this problem and use hypothesis testing to solve it.

In model b), coefficients for Po1 and Po2 are denoted as  $\beta_4$  and  $\beta_5$ , respectively. To test whether the replacement is fine, we first construct a hypothesis as follows

$$H_0 : \beta_4 = \beta_5$$

$$H_1 : \beta_4 \neq \beta_5$$

It can be transformed to

$$H_0 : q^T \beta = 0 \quad \text{vs.} \quad H_1 : q^T \beta \neq 0$$

where  $q = (0 \cdots 1 - 1 \cdots 0)^T$  and  $\beta = (\beta_0 \cdots \beta_{14})^T$

From what has been calculated previously, we know that

$$\hat{\beta}_{LSE} \sim N(\beta, \sigma^2(X^T X)^{-1})$$

then

$$q^T \hat{\beta}_{LSE} \sim N(q^T \beta, \sigma^2 q^T (X^T X)^{-1} q)$$

Since  $\sigma^2$  is unknown, we use  $\hat{\sigma}^2$  to replace it, then we have

$$\frac{q^T \hat{\beta}_{LSE} - q^T \beta}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \sim T_{n-15}$$

Under  $H_0$ ,

$$\frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \sim T_{n-15}$$

Then

$$p\text{-value} = 2P \left( T_{n-15} > \left| \frac{q^T \hat{\beta}_{LSE}}{\hat{\sigma} \sqrt{q^T (X^T X)^{-1} q}} \right| \right)$$

Calculate it using R, the result is

$$p\text{-value} = 0.207 > 0.05$$

According to the result, we tend not to reject  $H_0$ , that is according to the hypothesis testing above, we say that it might be reasonable to include  $(Po1 + Po2)/2$  in the model to replace these two variables.

We can interpret the replacement in another way. Since Po1 and Po2 are highly correlated, if we include both Po1 and Po2 simultaneously into the model, then  $X'X$  would be close to a singular matrix, which can hardly be inverted. In that case, the estimate of coefficients for Po1 ( $\hat{\beta}_4$ ) and Po2 ( $\hat{\beta}_5$ ) would be very unstable. To confirm this, we calculated  $\widehat{Var}(\hat{\beta}_4)$  and  $\widehat{Var}(\hat{\beta}_5)$  using R, and it turns out that  $\widehat{Var}(\hat{\beta}_4) = 9990$  and  $\widehat{Var}(\hat{\beta}_5) = 11799$ . After the replacement,  $\widehat{Var}(\hat{\beta}_4)$  ( $\beta_4$  denote the coefficient for the new covariate  $'(Po1+Po2)/2'$ ) is 605, which could reflect from another facet that the replacement might be more reasonable.

Then we construct another hypothesis testing to test the effect before and after the replacement.

**Before the replacement:**

hypothesis testing 1:

$$H_0 : \beta_4 = 0 \quad \text{vs.} \quad \beta_4 \neq 0$$

then

$$p\text{-value} = 2P \left( T_{n-15} > \left| \frac{\hat{\beta}_4}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{44}}} \right| \right)$$

Using R to calculate the result is

$$p\text{-value} = 0.086 > 0.05$$

hypothesis testing 2:

$$H_0 : \beta_5 = 0 \quad \text{vs.} \quad \beta_5 \neq 0$$

then

$$p\text{-value} = 2P \left( T_{n-15} > \left| \frac{\hat{\beta}_5}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{55}}} \right| \right)$$

Using R to calculate the result is

$$p\text{-value} = 0.417 > 0.05$$

From the result we can see that before the replacement, neither Po1 nor Po2 has a significant effect on crime rate.

**After the replacement:** After the replacement, the linear model would be changed accordingly. The number of covariate included in the model would be reduced to 13, with  $X_4 = (Po1 + Po2)/2$ . Denote the coefficient of  $X_4$  as  $\beta_4$ , then our hypothesis testing would be

$$H_0 : \beta_4 = 0 \quad \text{vs.} \quad \beta_4 \neq 0$$

then

$$p\text{-value} = 2P \left( T_{n-14} > \left| \frac{\hat{\beta}_4}{\hat{\sigma} \sqrt{(X^T X)^{-1}_{44}}} \right| \right)$$

Using R to calculate the result is

$$p\text{-value} \approx 0 < 0.05$$

Hence, after the replacement,  $(Po1 + Po2)/2$  seems to have a significant effect on the crime rate.

Combining the above analysis, we say that such replacement may be fine.

g) How to test whether the effect of the first five characteristic variables (M, So, . . . , Po2) are all zero? What's your test result?

For the linear model in b),

$$Y = X\beta + \epsilon$$

LSE of  $\beta$  and  $\sigma^2$ .

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^2 = \frac{Y^T Y - Y^T P_X Y}{n - p}$$

To test whether the effect of the first five characteristic variables (M, So, . . . , Po2) are all zero, we construct a hypothesis testing as below.

$$H_0 : K^T \beta = 0$$

$$H_1 : K^T \beta \neq 0$$

We will reject  $H_0$  if

$$F = \frac{(K^T \hat{\beta} - m)^T (K^T (X^T X)^{-1} K)^{-1} (K^T \hat{\beta} - m)}{s \hat{\sigma}^2}$$

is large. Where

$$K^T = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \end{pmatrix}_{s \times p}$$

$$m = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{s \times 1}$$

$\text{rank}(X) = p$ ,  $\text{rank}(K) = s$ , and in this case we have  $p = 15$ ,  $s = 5$ .  
Plug in the data, and solve it using R, we got

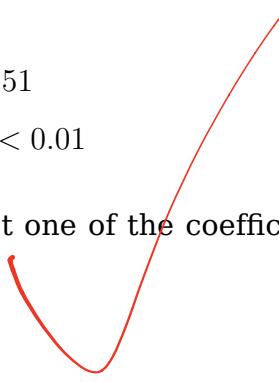
$$F = 6.51$$

Under  $H_0$ ,  $F$  follows  $F$  distribution  $F_{s, n-p}$ , then

$$F = 6.51 > F_{s, n-p}^{-1}(0.95) = 2.51$$

$$p\text{-value} = P(F_{s, n-p} > 6.51) \approx 0 < 0.01$$

Hence, we tend to reject  $H_0$  and conclude that at least one of the coefficients is not zero.



# Appendix

## Appendix 1-Correlation coefficient matrix

	M	So	Ed	Po1	Po2	LF
M	1.00000000	0.58435534	-0.53023964	-0.50573690	-0.51317336	-0.1609488
So	0.58435534	1.00000000	-0.70274132	-0.37263633	-0.37616753	-0.5054695
Ed	-0.53023964	-0.70274132	1.00000000	0.48295213	0.49940958	0.5611780
Po1	-0.50573690	-0.37263633	0.48295213	1.00000000	0.99358648	0.1214932
Po2	-0.51317336	-0.37616753	0.49940958	0.99358648	1.00000000	0.1063496
LF	-0.16094882	-0.50546948	0.56117795	0.12149320	0.10634960	1.00000000
M.F	-0.02867993	-0.31473291	0.43691492	0.03376027	0.02284250	0.5135588
Pop	-0.28063762	-0.04991832	-0.01722740	0.52628358	0.51378940	-0.1236722
NW	0.59319826	0.76710262	-0.66488190	-0.21370878	-0.21876821	-0.3412144
U1	-0.22438060	-0.17241931	0.01810345	-0.04369761	-0.05171199	-0.2293997
U2	-0.24484339	0.07169289	-0.21568155	0.18509304	0.16922422	-0.4207625
Wealth	-0.67005506	-0.63694543	0.73599704	0.78722528	0.79426205	0.2946323
Ineq	0.63921138	0.73718106	-0.76865789	-0.63050025	-0.64815183	-0.2698865
Prob	0.36111641	0.53086199	-0.38992286	-0.47324704	-0.47302729	-0.2500861
Time	0.11451072	0.06681283	-0.25397355	0.10335774	0.07562665	-0.1236404
Crime	-0.08947240	-0.09063696	0.32283487	0.68760446	0.66671414	0.1888663
	M.F	Pop	NW	U1	U2	
M	-0.02867993	-0.28063762	0.59319826	-0.224380599	-0.24484339	
So	-0.31473291	-0.04991832	0.76710262	-0.172419305	0.07169289	
Ed	0.43691492	-0.01722740	-0.66488190	0.018103454	-0.21568155	
Po1	0.03376027	0.52628358	-0.21370878	-0.043697608	0.18509304	
Po2	0.02284250	0.51378940	-0.21876821	-0.051711989	0.16922422	
LF	0.51355879	-0.12367222	-0.34121444	-0.229399684	-0.42076249	
M.F	1.00000000	-0.41062750	-0.32730454	0.351891900	-0.01869169	
Pop	-0.41062750	1.00000000	0.09515301	-0.038119948	0.27042159	
NW	-0.32730454	0.09515301	1.00000000	-0.156450020	0.08090829	
U1	0.35189190	-0.03811995	-0.15645002	1.000000000	0.74592482	
U2	-0.01869169	0.27042159	0.08090829	0.745924815	1.00000000	
Wealth	0.17960864	0.30826271	-0.59010707	0.044857202	0.09207166	
Ineq	-0.16708869	-0.12629357	0.67731286	-0.063832178	0.01567818	
Prob	-0.05085826	-0.34728906	0.42805915	-0.007469032	-0.06159247	
Time	-0.42769738	0.46421046	0.23039841	-0.169852838	0.10135833	
Crime	0.21391426	0.33747406	0.03259884	-0.050477918	0.17732065	
	Wealth	Ineq	Prob	Time	Crime	
M	-0.6700550558	0.63921138	0.361116408	0.1145107190	-0.08947240	

So	-0.6369454328	0.73718106	0.530861993	0.0668128312	-0.09063696
Ed	0.7359970363	-0.76865789	-0.389922862	-0.2539735471	0.32283487
Po1	0.7872252807	-0.63050025	-0.473247036	0.1033577449	0.68760446
Po2	0.7942620503	-0.64815183	-0.473027293	0.0756266536	0.66671414
LF	0.2946323090	-0.26988646	-0.250086098	-0.1236404364	0.18886635
M.F	0.1796086363	-0.16708869	-0.050858258	-0.4276973791	0.21391426
Pop	0.3082627091	-0.12629357	-0.347289063	0.4642104596	0.33747406
NW	-0.5901070652	0.67731286	0.428059153	0.2303984071	0.03259884
U1	0.0448572017	-0.06383218	-0.007469032	-0.1698528383	-0.05047792
U2	0.0920716601	0.01567818	-0.061592474	0.1013583270	0.17732065
Wealth	1.0000000000	-0.88399728	-0.555334708	0.0006485587	0.44131995
Ineq	-0.8839972758	1.00000000	0.465321920	0.1018228182	-0.17902373
Prob	-0.5553347075	0.46532192	1.0000000000	-0.4362462614	-0.42742219
Time	0.0006485587	0.10182282	-0.436246261	1.0000000000	0.14986606
Crime	0.4413199490	-0.17902373	-0.427422188	0.1498660617	1.00000000

## Appendix 2-R code

```

a)
#create a scatter plot matrix
dev.new
pdf(file ="scatter_matrix.pdf", width=6.5,heigh=6.5)
pairs(Crime ~ M + So + Ed + Po1 + Po2 + LF + M.F + Pop + NW + U1 + U2 + Wealth +
      Ineq + Prob + Time, data = uscrime, cex = 0.4, font.labels=2)
dev.off()
#create a scatter plot with time and probability
dev.new
pdf(file ="scatter_time_prob.pdf", width=6.5,heigh=6.5)
pairs(Crime ~ Prob + Time, data = uscrime, cex = 0.4, font.labels=2)
dev.off()
#calculate the correlation coefficient
cor(uscrime)

b)

> #Work with Prob first
> Y <- matrix(uscrime$Crime, n,1)
> p <- ncol(X)
> n <- nrow(uscrime)
> X <- data.matrix(uscrime)

```

```

> X <- cbind(rep(1,n), X[,-c(match(c("Time", "Crime"), names(uscrime)))]))
>
> #regression
> b <- solve(t(X)%*%X)%*%t(X)%*%Y
> sigma_sq <- sum((Y-X%*%b)^2)/(n-p)
> sigma_sq
[1] 42664.06
> b.cov <- sigma_sq*solve(t(X)%*%X)
> b.sd <- sqrt(diag(b.cov))
>
> ## Test the effect of prob controlling for other parameters
> #find position of Prob in the design matrix
> match("Prob", colnames(X))
[1] 15
>
> #T-test
> alpha <- 0.05
> q <- matrix(c(rep(0,14), 1), 15, 1)
> t_stat <- (t(q)%*%b)/(sqrt(sigma_sq*t(q)%*%solve(t(X)%*%X)%*%q))
> t_stat
      [,1]
[1,] -2.28429
> qt((1-alpha/2),n-p)
[1] 2.036933
> p_value <- 2*(pt(abs(t_stat), df= n-p, lower.tail = FALSE))
> p_value
      [,1]
[1,] 0.02913165

```

c)

```

> #Part c: Perform the same analysis to study the adjusted
effect of Time on Crime
> n <- nrow(uscrime)
> Y <- matrix(uscrime$Crime, n,1)
> p <- ncol(X)
> X <- data.matrix(uscrime)
> X <- cbind(rep(1,n), X[,-c(match(c("Prob", "Crime"), names(uscrime)))]))
> b <- solve(t(X)%*%X)%*%t(X)%*%Y
> sigma_sq <- sum((Y-X%*%b)^2)/(n-p)

```

```

> b.cov <- sigma_sq*solve(t(X)%*%X)
> b.sd <- sqrt(diag(b.cov))
>
> ## Test the effect of time controlling for other parameters
> #find position of Time in the design matrix
> match("Time", colnames(X))
[1] 15
> #T-test
> alpha <- 0.05
> q <- matrix(c(rep(0,14), 1), 1, 15, 1)
> t_stat <- (t(q)%*%b)/(sqrt(sigma_sq*t(q)%*%solve(t(X)%*%X)%*%q))
> t_stat
      [,1]
[1,] 0.8290175
> p_value <- 2*(pt(abs(t_stat), df= n-p, lower.tail = FALSE))
> p_value
      [,1]
[1,] 0.4132358

```

d)

```

> n <- nrow(uscrime)
> Y <- matrix(uscrime$Crime, n,1)
> X <- data.matrix(uscrime)
> X <- cbind(rep(1,n), X[, -c(match(c("Prob", "Crime"), names(uscrime)))]))
> p <- ncol(X)
> b <- solve(t(X)%*%X)%*%t(X)%*%Y
> sigma_sq <- sum((Y-X)%*%b)^2/(n-p)
> var<- sigma_sq*solve(t(X)%*%X)
> se <- qt(0.975,n-p)*sqrt(diag(var))
> se <- sqrt(diag(var))
> summary <- cbind(b, se)
> summary

```

e)

```

#Part e: select 4 principal components
> n <- nrow(uscrime)
> uscrime_s <- scale(uscrime,scale=TRUE, center = TRUE )
> Y <- uscrime_s[,match("Crime", colnames(uscrime_s))]

```



```

> X <- uscrime_s[,-c(14,15,16)]
> eig <- eigen(t(X)%*%X)
> eig$values
[1] 259.9810928 115.9303637 87.6721127
44.8813008 29.1785952 16.2078459
12.0080627 10.8970121 8.7930985
6.1335336 3.2391747 2.8372938
[13] 0.2405135
> V <- eig$vectors
> Z <- X%*%V
> X <- cbind(Z[,1:4],uscrime_s[,14])
> p <- ncol(X)
> b <- solve(t(X)%*%X)%*%t(X)%*%Y
> sigma_sq <- sum((Y-X%*%b)^2)/(n-p)
> var<- sigma_sq*solve(t(X)%*%X)
> b[5]
[1] -0.2771373
> var[5,5]
[1] 0.01185295

```

f)

```

#test replacement of Po variables
#Look at ffect of Po1 and Po2
#Model as in b without replacement
Y <- matrix(uscrime$Crime, n,1)
n <- nrow(uscrime)
X <- data.matrix(uscrime)
X <- cbind(rep(1,n), X[,-c(match(c("Time", "Crime"), names(uscrime)))]])
p <- ncol(X)
#Linear regression
b <- solve(t(X)%*%X)%*%t(X)%*%Y
sigma_sq <- sum((Y-X%*%b)^2)/(n-p)
#T-test for b4 and b5 both equal zero.
alpha <- 0.05
q <- matrix(c(rep(0,4), 1,-1, rep(0,9)), 15, 1)
t_stat <- (t(q)%*%b)/(sqrt(sigma_sq*t(q)%*%solve(t(X)%*%X)%*%q))
t_stat
qt((1-alpha/2),n-p)
p_value <- 2*(pt(abs(t_stat), df= n-p, lower.tail = FALSE))

```

```

p_value
#Test b4 and b5 separately
#test for b4
q <- matrix(c(rep(0,4), 1, rep(0,10)), 15, 1)
t_stat <- (t(q)%*%b)/(sqrt(sigma_sq*t(q)%*%solve(t(X)%*%X)%*%q))
t_stat
qt((1-alpha/2),n-p)
p_value <- 2*(pt(abs(t_stat), df= n-p, lower.tail = FALSE))
p_value
#test for b5
q <- matrix(c(rep(0,5), 1, rep(0,9)), 15, 1)
t_stat <- (t(q)%*%b)/(sqrt(sigma_sq*t(q)%*%solve(t(X)%*%X)%*%q))
t_stat
qt((1-alpha/2),n-p)
p_value <- 2*(pt(abs(t_stat), df= n-p, lower.tail = FALSE))
p_value
#After the replacement
X <- data.matrix(uscrime)
Po <- (uscrime$Po1+uscrime$Po2)/2
X <- cbind(rep(1,n), X[, -c(match(c("Time", "Crime", "Po1", "Po2"), names(uscrime)))])
p <- ncol(X)
#Linear regression
b <- solve(t(X)%*%X)%*%t(X)%*%Y
sigma_sq <- sum((Y-X%*%b)^2)/(n-p)
#T-test for b4 and b5 both equal zero.
alpha <- 0.05
q <- matrix(c(rep(0,13),1), 14, 1)
t_stat <- (t(q)%*%b)/(sqrt(sigma_sq*t(q)%*%solve(t(X)%*%X)%*%q))
t_stat
qt((1-alpha/2),n-p)
p_value <- 2*(pt(abs(t_stat), df= n-p, lower.tail = FALSE))
p_value

g)
> #test whether the effect of the first 5 characteristic variables (M, So,... ,Po2)
> n <- nrow(uscrime)
> Y <- matrix(uscrime$Crime, n,1)
> p <- ncol(X)
> p1 <- p-1
> X <- data.matrix(uscrime)

```

```

> X <- cbind(rep(1,n), X[, -c(match(c("Time", "Crime"), names(uscrime))))])
> #regression
> b <- solve(t(X)%*%X)%*%t(X)%*%Y
> sigma_sq <- sum((Y-X%*%b)^2)/(n-p)
> sigma_sq
[1] 42664.06
> b.cov <- sigma_sq*solve(t(X)%*%X)
> b.sd <- sqrt(diag(b.cov))
> #Test first five vars
> s <- 5
> I <- diag(s)
> K <- t(cbind(rep(0,s), I, matrix(rep(0,45), s, p-s-1)))
> Kb.mid <- t(K)%*%solve(t(X)%*%X)%*%K
> F.num <- t(t(K)%*%b)%*%solve(Kb.mid)%*%(t(K)%*%b)
> F.den <- s*sigma_sq
> F <- F.num/F.den
> F
      [,1]
[1,] 6.512417
> qf(0.95,s,n-p)
[1] 2.512255

```