Model Selection and Variable Reduction

Introduction

The project invests the relationship of the the number of motor vehicle death to their driving conditions across the United States, the data were collected from 49 States and DC in the 1964. The variables studied are as below:

Dependent variable:
(Y) Deaths = the number of motor vehicle deaths in 1964.
Independent variabes:
(X1) Drivers = the number of drivers in each state x 10-4
(X2) People = the number of people per square mile.
(X3) Mileage = the total mileage of rural roads x 10-3
(X4) Maxtemp = the normal maximum temperature in January.
(X5) Fuel = the highway fuel consumption in gallons x 10-7

The purposes of this project are (1) find the correlation relationship between each variable, (2) fit a multiple linear regression model with all independent variables involved (full model), (3) Conduct multicollinearity diagnostics and identify the problems of multicollinearity among the independent variables. (4) Remedies for multicollinearity by using Ridge Regression and Principal Component Regression.

Data and Methods

Variables data listed above were collected from 49 States and 1 DC, Table 1 gives a brief summary of the variables.

Table 1. Describe Statistics of All Variables.

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Deaths(Y)	50	926.94000	889.69484	46347	43.00000	4743
Drivers(X1)	50	190.36000	197.41201	9518	11.00000	952.00000
People(X2)	50	135.10400	198.77469	6755	0.40000	812.00000
Mileage(X3)	50	63.11400	38.95398	3156	0	196.00000
Maxtemp(X4)	50	41.74000	11.75743	2087	20.00000	67.00000
Fuel(X5)	50	140.46400	161.53326	7023	6.20000	955.00000

Table 2 presents the Pearson correlation coefficients between the dependent variable Deaths and each independent variables, as well as among the five independent variables.

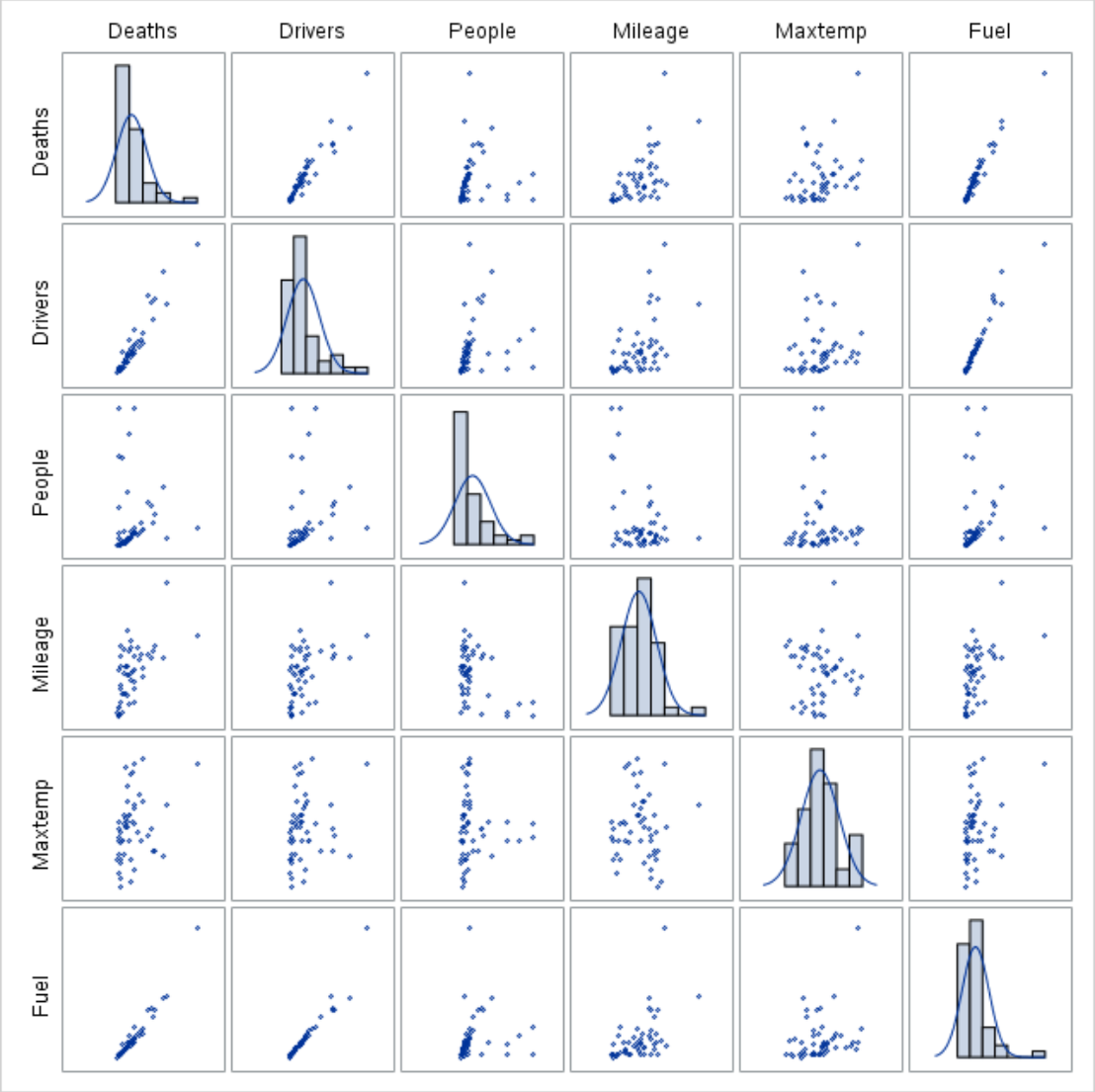
Table 2. Pearson Correlations Among All Variables.

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0						
	Deaths	**Drivers**	**People**	**Mileage**	**Maxtemp**	**Fuel**
Deaths	1.00000	0.95607 <.0001	0.03947 0.7855	0.60259 <.0001	0.32671 0.0206	0.97448 <.0001
Drivers	0.95607 <.0001	1.00000	0.20892 0.1454	0.49700 0.0002	0.19445 0.1760	0.96554 <.0001
People	0.03947 0.7855	0.20892 0.1454	1.00000	-0.41525 0.0027	-0.03868 0.7897	0.13115 0.3640
Mileage	0.60259 <.0001	0.49700 0.0002	-0.41525 0.0027	1.00000	-0.00144 0.9921	0.51549 0.0001
Maxtemp	0.32671 0.0206	0.19445 0.1760	-0.03868 0.7897	-0.00144 0.9921	1.00000	0.27485 0.0534
Fuel	0.97448 <.0001	0.96554 <.0001	0.13115 0.3640	0.51549 0.0001	0.27485 0.0534	1.00000

According to Table 2, only the relationship between Deaths and People is not significant, other variables are all significantly positive correlated with Deaths. From the multicollinearity point of view, Drivers is strongly positive correlated to Fuel, Fuel and Mileage are significantly highly positive correlated.

Figure 1 represents the linear relationship between Deaths and each independent variable and each predictor. According to the plot, there are positive linear relationships between Deaths and Drivers, Deaths and People, Deaths and Fuel, Fuel and Drivers, Fuel and Mileage. Thus, there might be multicollinearity according to the correlation but they do not always indicate the actual nature or the extent of the multicollinearity. Further analysis is needed.

Figure 1. Matrix Scatterplot of Deaths and Five Predictor Variables



OLS was applied to fit the following multiple linear regression model (full model):

Deaths = β0 + β1*Drivers + β2* People + β3*Mileage + β4* Maxtemp + β5*Fuel + ε

β0 – β5 are regression coefficients to be estimated, and ε is the model random error.

Remedies like PCR and Ridge Regression are applied if there is multicollinearity effect.

Results and Discussion

1. Full model

Put the estimated coefficients back to the model and we obtained the estimated full model:

Deaths = -292.27414 + 1.83859*Drivers – 0.2612* People + 2.77107*Mileage + 8.26603* Maxtemp + 2.7284*Fuel

Table 3. R-Square of the Full Model

R-Square	0.9793
Adj R-Sq	0.9770

Table 4. Estimated Regression Coefficients for the Full Model.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-292.27414	95.54989	-3.06	0.0038	0
Drivers	1	1.83859	0.41746	4.40	<.0001	18.27998
People	1	-0.25120	0.12981	-1.94	0.0594	1.79193
Mileage	1	2.77107	0.73859	3.75	0.0005	2.22798
Maxtemp	1	8.26603	1.82174	4.54	<.0001	1.23481
Fuel	1	2.72840	0.50686	5.38	<.0001	18.04295

The R-square is 97.93% and the adjusted R-square is 97.7% (Table 3), and the coefficient of People is not significant at $\alpha = 0.05$ (Table 4.), and is negative, it is opposite with what we may expect, since usually the death will increase as the people increases. Thus, there might be multicollinearity among the predictors.

2. Multicollinearity Diagnostics

(1). VIF

According to Table 4 above, the last column represents variance inflation factor(VIF), as VIF exceeds 10, we can claim that at least some concerns on multicollinearity in the data. Since both Drivers and Fuel have VIF larger than 10, there might be multicollinearity in the Driver and Fuel.

Table 5. Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation					
			Intercept	Drivers	People	Mileage	Maxtemp	Fuel
1	4.49443	1.0000	0.00168	0.00115	0.00719	0.00482	0.00245	0.00122
2	0.75051	2.4471	0.00024	0.00067804	0.39837	0.02345	0.00010742	0.00180
3	0.59763	2.7423	0.01479	0.01151	0.01638	0.00947	0.01785	0.01458
4	0.11670	6.20592	0.00134	0.00085653	0.25549	0.57192	0.15599	0.01484
5	0.02674	12.96532	0.73953	0.12492	0.30090	0.38340	0.54569	0.03544
6	0.01400	17.91525	0.24242	0.86088	0.02166	0.00695	0.27792	0.93211

(2). Conditional Number (Index).

An exclusive large conditional number or index is evidence that the regression coefficients are unstable, when the index exceeds 30, we can claim that there might be multicollinearity. Table 5 shows that the largest conditional index is 18, we cannot tell if there is any multicollinearity effect in the data.

(3). Variance Proportions

The proportion tells the percentage of the variance of the parameter estimate coefficient is associated with each eigenvalue. A high proportion of variance of an independent variable coefficient reveals a strong association with the eigenvalue, if an eigenvalue is small enough and some independent variables show high proportions of variation with respect to the eigenvalue, we may conclude that these independent variables have significant linear-dependency.

Table 5 shows that the eigenvalue five is fairly small (0.014) and the variance proportion for predictor Driver and Fuel are large(> 0.8), thus, Driver and Fuel might be collinear.

3. Remedies for Multicollinearity

(1). Ridge Regression

there are three criteria to find the best ridge penalty k.

• VIF close to 1.

• Estimated coefficients should be “stable”

• look for only “modest” change in R-square.

Table 6 shows that when k equals 0.10, the coefficients tend to be stable (β1(Drivers) = 1.83, β2(People) = -0.148, β3(Mileage) = 3.38, β4(Maxtemp) = 8.74, β5(Fuel) = 2.39),

At k = 0.10, RMSE changed about 5.4% compared to the RMSE at k = 0.

Table 7 gives the VIF of the predictors at different ks. When k = 0.11, all VIFs close to 1.

Table 6. Ridge Regression Analysis for Death (Partition)

Obs	_TYPE_	_RIDGE_	_RMSE_	Intercept	Drivers	People	Mileage	Maxtemp	Fuel
3	RIDGE	0.00	134.926	-292.27	1.8385	-0.251	2.7710	8.26603	2.728
19	RIDGE	0.08	139.911	-315.86	1.8490	-0.165	3.2935	8.70680	2.433
21	RIDGE	0.09	141.036	-316.24	1.8408	-0.156	3.3381	8.72651	2.413
23	RIDGE	0.10	142.242	-316.30	1.8323	-0.148	3.3794	8.74232	2.394
25	RIDGE	0.11	143.523	-316.06	1.8237	-0.141	3.4174	8.75468	2.376

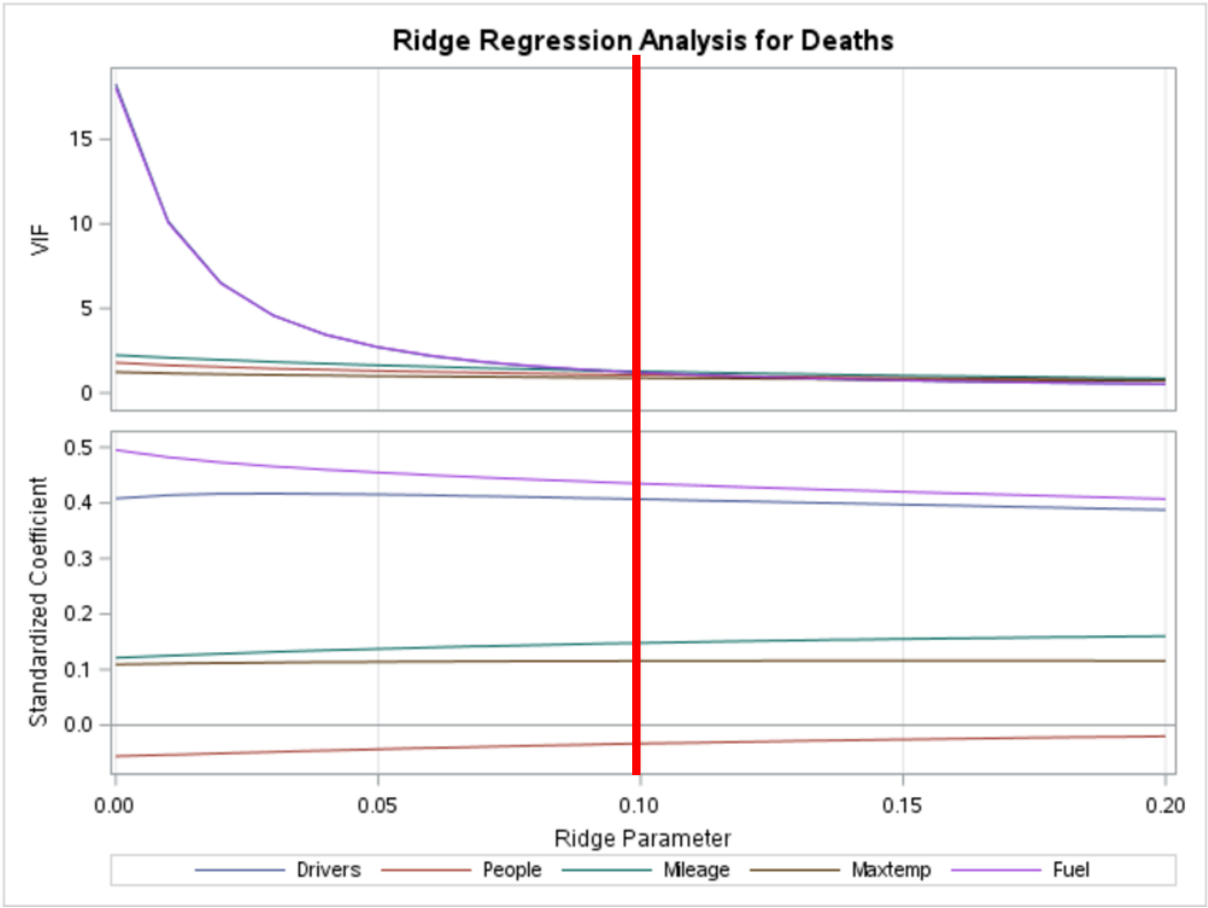
Table 7. VIF for Ridge Regression Analysis for Death (Partition)

Obs	_RIDGE_	Intercept	Drivers	People	Mileage	Maxtemp	Fuel

10	0.04	.	3.4343	1.36908	1.73174	1.03043	3.4411
12	0.05	.	2.6919	1.30299	1.63703	1.00066	2.7070
14	0.06	.	2.1839	1.24375	1.55084	0.97355	2.2037
16	0.07	.	1.8203	1.19007	1.47215	0.94845	1.8425
18	0.08	.	1.5505	1.14107	1.40006	0.92496	1.5737
20	0.09	.	1.3442	1.09608	1.33384	0.90282	1.3677
22	0.10	.	1.1827	1.05458	1.27283	0.88184	1.2059
24	0.11	.	1.0535	1.01615	1.21650	0.86189	1.0762
26	0.12	.	0.9484	0.98044	1.16434	0.84286	0.9703

Figure 2 summarized the information in Table 6 and Table 7, when k = 0.10 or 0.11, VIFs get closer to 1 and all the coefficients become stable afterwards.

Figure 2. Ridge Regression Analysis for Deaths.



Thus, the best k = 0.10 or 0.11, corresponding coefficients are shown is table 8.

Table 8. Ridge Regression Coefficients (k=0.10).

Intercept	Drivers	People	Mileage	Maxtemp	Fuel
-316.30	1.8323	-0.148	3.3794	8.74232	2.394

(2). Principal Component Regression

Table 9 shows that the first three principal components explained more than 94% variance in the Death variable, all three eigenvalues are larger or round to one.

Table 9. Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.40794193	1.07416745	0.4816	0.4816
2	1.33377448	0.36392366	0.2668	0.7483
3	0.96985082	0.70955558	0.1940	0.9423
4	0.26029524	0.23215772	0.0521	0.9944
5	0.02813753		0.0056	1.0000

Table 10. OLS RMSE Estimates

Obs	_TYPE_	_RMSE_
1	PARMS	134.96

Table 11. Principal Component Analysis with IPC.

Obs	TYPE	PCOMIT	_RMSE_	Intercept	Drivers	People	Mileage	Maxtemp	Fuel
3	IPC	1	133.65	-301.2	2.002	-0.265	2.751	8.4762	2.5304
5	IPC	2	158.60	-590.9	1.777	0.1840	5.750	11.552	2.2039
7	IPC	3	157.71	-547.8	1.788	0.2024	5.826	10.282	2.207
9	IPC	4	163.00	-530.3	1.750	0.0382	6.372	9.8521	2.174

Table 12. Principal Component Analysis with IPCVIF.

Obs	_TYPE_	_PCOMIT_	_RMSE_	Drivers	People	Mileage	Maxtemp	Fuel
2	IPCVIF	1	.	0.4738	1.6533	2.2182	1.13040	0.5621
4	IPCVIF	2	.	0.2003	0.5455	0.3250	0.94893	0.1763

6	IPCVIF	3	.	0.1810	0.4907	0.2892	0.02935	0.1746
8	IPCVIF	4	.	0.1560	0.00008	0.0805	0.01753	0.16115

Compare table 10, 11 and 12, we find out that when PC = 3, the RMSE is 157.71, which is 16.8% larger than the original OLS RMSE, and all predictors have VIF smaller than 1.1. Thus, we choose to include three PCs.

Table 13. PCs Predictors' Coefficients Estimation.

Eigenvectors					
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**
Drivers	0.612979	0.182544	-.136838	-.266812	-.707828
People	0.013489	0.808975	-.230475	0.536990	0.062450
Mileage	0.440265	-.527653	-.186230	0.701992	0.016580
Maxtemp	0.205440	0.125546	0.944383	0.217335	-.054204
Fuel	0.622926	0.134377	-.040191	-.316899	0.701332

Table 14. Three PCs Parameter Estimation

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	926.94000	22.43002	41.33	<.0001	0
Prin1	1	563.84286	14.60137	38.62	<.0001	1.00000
Prin2	1	40.33047	19.61893	2.06	0.0455	1.00000
Prin3	1	15.80773	23.00722	0.69	0.4955	1.00000

Now we conduct the regression with three PCs. From table 14, none of the PCs is multicollinearity, the R-square of the new PC regression is 97.02%.

After combining the result from table 13 and 14, the PC regression is

$\widehat{Death}$ ***= 926.94 + 563.84286*Prin1 + 40.33047*Prin2 + 15.80773*Prin3***

While:

$\widehat{Prin1}$ ***= 0.612979*Z1 + 0.013489*Z2 + 0.440265*Z3 + 0.20544*Z4 + 0.622926*Z5***

$\widehat{Prin2}$ ***= 0.182544*Z1 + 0.808975*Z2 - 0.527653*Z3 + 0.125546*Z4 + 0.134377*Z5***

$\widehat{Prin3}$ ***= -0.136838*Z1 - 0.230475*Z2 - 0.18623*Z3 + 0.944383*Z4 - 0.040191*Z5***

Where Z1-Z5 are standardized Drivers, People, Mileage, Maxtemp and Fuel, respectively.

According to the coefficients of the PCs, we may say that PC1 represent the Driver and Fuel, PC2 represent the relationship between People and the Mileage, PC3 focus on the temperature.

Comparison and Summary

After remedying for the multicollinearity, VIFs are significantly decreased to around 1, some of the ridge regression's coefficients are closer to zero, PCR is composed by three components and there is no multicollinearity effect among these PCs.

Table 15. Comparison between OLS, Ridge Regression and PCR.

	OLS		Ridge Regression			PCR		
Variable	**Parameter Estimate**	**Variance Inflation**	**Parameter Estimate**	**Variance Inflation**		**Variable**	**Parameter Estimate**	**Variance Inflation**
Intercept	-292.274	0	-316.305	.		**Intercept**	926.94	0
Drivers	1.83859	18.2799	1.83233	1.1827		**Prin1**	563.8428	1
People	-0.2512	1.79193	-0.14885	1.05458		**Prin2**	40.33047	1
Mileage	2.77107	2.22798	3.3794	1.27283		**Prin3**		1
Maxtemp	8.26603	1.23481	8.74232	0.88184				
Fuel	2.7284	18.0429	2.39465	1.2059				

To sum up, both Ridge Regression and PCR are trying to simplify the model. Ridge Regression try to release some bias to get a smaller MSE and push some of the coefficients to zero, this approach is efficient in testing data prediction. PCR reduce the dimensions of the regressions and try to form a simpler equation. PCR is easier to interpret when the predictors in the PC are highly correlated and can represent a specific feature of the dependent variable.