Data Cleaning by Influence Diagnostics

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 influence diagnostics to identify possible outliers and high influential points.

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 and Spearman correlation coefficients between the dependent variable Deaths and each independent variables, as well as among the five independent variables.

Table 2. Pearson and Spearman 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

Spearman Correlation Coefficients, N = 50
Prob > |r| under H0: Rho=0

	Deaths	Drivers	People	Mileage	Maxtemp	Fuel
Deaths	1.00000	0.94355 <.0001	0.45213 0.0010	0.58460 <.0001	0.36642 0.0089	0.96710 <.0001
uiDrivers	0.94355 <.0001	1.00000	0.59903 <.0001	0.48386 0.0004	0.26721 0.0607	0.99092 <.0001
People	0.45213 0.0010	0.59903 <.0001	1.00000	-0.17527 0.2234	0.18339 0.2024	0.54929 <.0001

Mileage	0.58460 <.0001	0.48386 0.0004	-0.17527 0.2234	1.00000	-0.08453 0.5595	0.51894 0.0001
Maxtemp	0.36642 0.0089	0.26721 0.0607	0.18339 0.2024	-0.08453 0.5595	1.00000	0.30587 0.0308
Fuel	0.96710 <.0001	0.99092 <.0001	0.54929 <.0001	0.51894 0.0001	0.30587 0.0308	1.00000

According to Table 2, the Pearson correlation shows only the relationship between Deaths and People is not significant, other variables are all significantly positive correlated with Deaths. For the relationship between independent variables, Drivers is significantly positive related to Mileage and Fuel. People is significantly negative correlated with Mileage. And finally, Fuel and Mileage are significantly positive correlated.

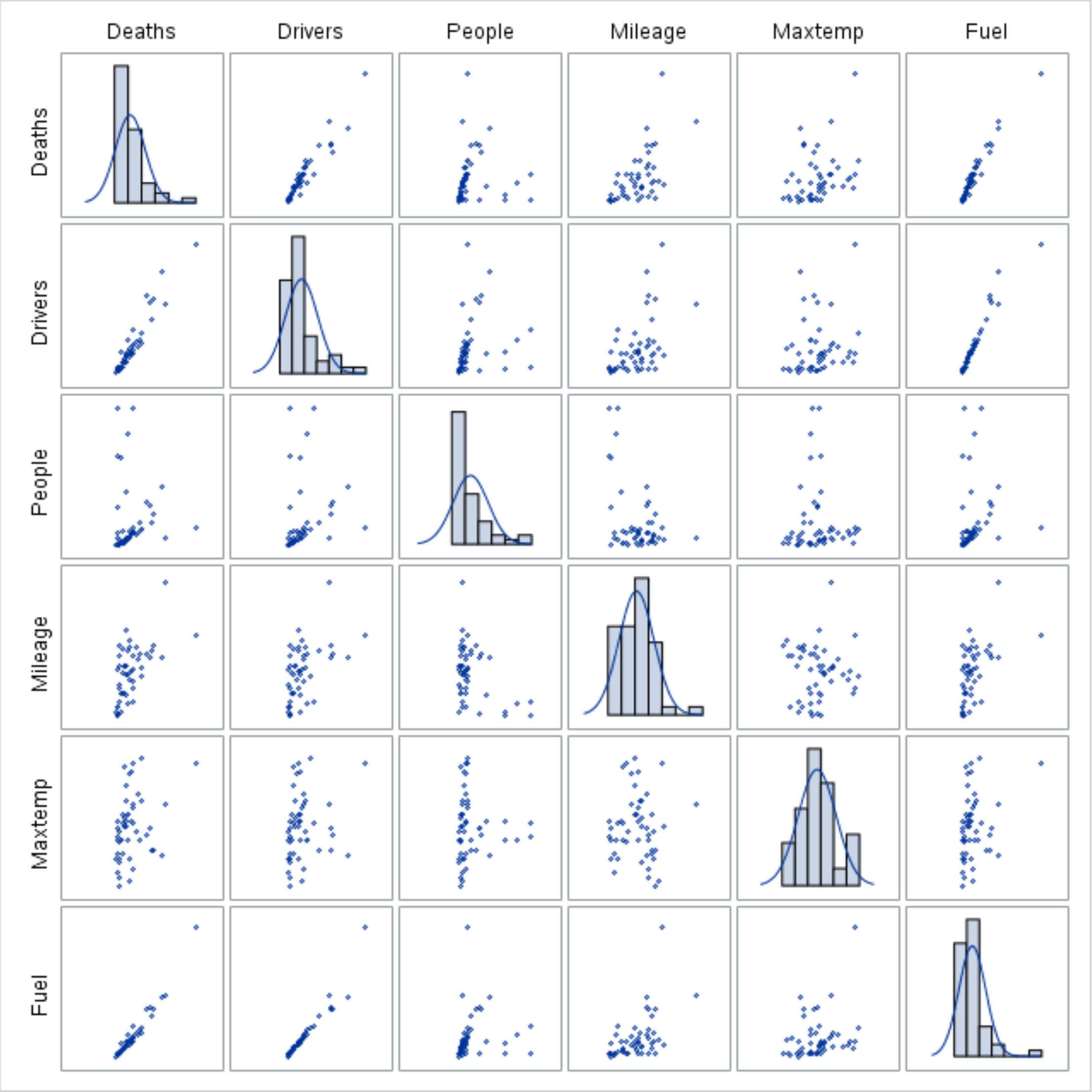
The Spearman correlation on the other hand, shows that all variables are significantly positive correlated with Deaths, all the independent variables are positively correlated except the relationship between People, Mileage and Maxtemp are not significant at α=0.05.

The unequal result from Spearman and Pearson correlation may be due to some extreme data points or outliers in the data set.

Figure 1 illustrates the linear relationship between Deaths and each independent variable and the linear relationship between each predictor. According to the plot, there are positive linear relationships between Deaths and Drivers, Deaths and People, Deaths and Fuel, Fuel and Drivers. Thus, each variable appears some relationship between each other and we are not able to delete any variables at this stage.

From figure 1, the scatterplot between Death and People, most data points are near each other, but there are some extreme points in both x-axis and y-axis. for People and Mileage scatterplot, extreme points exists in the y-asis, the same situation happens in People vs. Maxtemp and Maxtemp vs. Mileage. These extreme points may be outliers or highly influential points in model fitting.

Figure 1. Matrix Scatterplot of Deaths and Five Predictor Variables



Least-squares method was applied to fit the following multiple linear regression 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.

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.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-292.27414	95.54989	-3.06	0.0038	0
Drivers	1	1.83859	0.41746	4.40	<.0001	0.40796
People	1	-0.25120	0.12981	-1.94	0.0594	-0.05612
Mileage	1	2.77107	0.73859	3.75	0.0005	0.12133
Maxtemp	1	8.26603	1.82174	4.54	<.0001	0.10924
Fuel	1	2.72840	0.50686	5.38	<.0001	0.49537

The R-square is 97.93% and the adjusted R-square is 97.7% (Table 3), which indicants that 98% of the dependent variable variation can be explained by these five predictors in the model. However, the coefficient of People is not significant at α = 0.05 (Table 4.), other reduced models may be constructed to fit the data. Or we can check the data, if there are many influential points or outliers that impacts the model fitting.

2. Influence Diagnostics

It is common that influential points can influence the coefficient and have great impact on the regression fitting. Other outliers in the data may not be as influential as the influential points, but the can still vary the model little and put extra difficulties in the data analysis. Now we conduct influence diagnostics to detect the influential points and outliers in the dataset.

Table 5 below gives the criteria for influence statistics

Table 5. Criteria for Influence Statistics

n=50 p=6	
Influence Statistics	Observation I May be Influential IF
R-student	>3.53($> t_{\alpha/2}$, df=n-p-1=43)
hii	> 2p/n = 0.24
DFFITS	>$2\sqrt{p/n}$ = 0.693
DFBETAS	>$2/\sqrt{n}$ = 0.283
Cook's D	>4/n = 0.08
COVARATIO	<1-3p/n = 0.64 >1+3p/n = 1.36

By the criteria from table 5, we can pick out the data that are outliers or influential. Table 6 give a summary of the data.

Table 6. Summary of Influence Diagnostics for the Data

Obs	hii	R-student	DFFITS	Cook’s D	COVARATIO	DFBETAS					
						Intercept	Drivers	People	Mileage	Maxtemp	Fuel
5	*	*	*	*		*	*	*	*	*	*
16			*			*			*		
18										*	
22			*		*	*				*	
30	*				*			*			
32	*				*		*				*
33					*					*	
38			*	*	*		*				*
39	*		*	*	*		*	*			
43	*		*	*		*	*	*	*		
45						*					

(1). hii
point 30, 32, 39 and 43 may be considered to have unusual set of values for the predictor variables, which are high leverage points. They have extreme value in X-axis, but they give no information of the dependent variable. Thus, these observations may have exerted undue influence on at least one regression coefficient as well as performance criteria.

(2). R-student
A large R-student value indicates that the observations may have unusually large error in model fitting. Point 5 has R-student value larger than 3.53, indicates its’ large error in model fitting and may be considered as outlier, we have no information on the point’s influence nor the influenced statistics.

(3). DFFITS
DFFITS for the point shows its’ influence on model fitting, large value indicates the point is very influential in neighborhood of the X-axis space.
The DFFITS of point 5, 16, 22, 38, 39, and 43 are larger than 0.693, thus, they are considered high influential points.
Usually, points with large R-student or large hii will have large DFFITS, but for point 16, 22, and 38, neither R-student nor hii is large, this indicates these point’s influences are not due to the leverage on X-axis nor the error in Y-axis.
Point 30 has high leverage but small DFFITS, this in return indicates this points has near zero error, that is extreme small R-student.

(4). Cook’s D
Cook’s D measures the influence of the point on the set of regression coefficients. Large Cook’s D indicates the point exert undue influence on the set of coefficients. Further diagnose DFBETAS determines which coefficients are affected.

Point 5, 38, 39 and 43 has large Cook's D value, thus, they have heavy influence on at least one regression coefficient.

(5). COVRATIO
Point 22, 30, 32, 33, 38 and 39 have COVRATIO value out of the range [0.64, 1.36], they measure the change in the determinant of the covariance matrix of the coefficient estimates, these points provide improvement in the regression equation. The achieve better performance on the model fitting if included.

(6). DFBETAS
DFBETAS are the scaled measures of the change in each coefficient estimate $\widehat{\beta\iota}$, a large absolute value indicates the point has sizable impact on the i-th regression coefficient. Table 6 gives the summary of DFBETAS, however, as Cook's D, DFFITS and hii can detect the influence of the point on the at least one regression coefficients, none of the points 18, 33 and 45 have large value in Cook's D, DFFITS nor hii, thus, we choose to ignore the influence of these three points.

Summary

Point 5, 16, 22, 30, 32, 38, 39 and 43 can be considered as influential points, where points 30, 32, 39 and 43 have high leverage (extreme value on X-axis), point 5 have extreme large error (residual) on Y-axis and can be viewed as outlier, all off these points have impacts on model fitting, 5, 38, 39 and 43 impacts the coefficient estimation, Point 22, 30, 32, 33, 38 and 39 improve the covariance matrix of the coefficient estimates and improvement in the regression equation.