Project Report

Name: Yifang Zhang

Email:yzhan377@syr.edu

1. Introduction

1.1 Background

The Department of Motor Vehicles investigated the relationship of the number of driving death to the local population and driving conditions across the 49 States and DC in 1964. The variables in the research included:

(1) One dependent variable:

- Deaths=Y=The number of motor vehicle deaths in 1964.

(2) Five predictor or independent variables:

- Drivers= X_1=The number of drivers in each state / 10^4.
- People= X_2 =The number of people per square mile.
- Mileage=X_3=The total mileage of rural roads / 10^3.
- Maxtemp=X_4=The normal maximum temperature in January.
- Fuel=X^5=The highway fuel consumption in gallons / 10^7.

1.2 Purposes

The purposes of the study are:

(1) Compute descriptive statistics for all variables;

(2) Compute correlations among all variables;

(3) Fit a multiple linear regression model with all independent variables involved full model;

(4) Find a “best” model (reduced model) by using stepwise selection;

(5) Compare the characteristics of the “best” model against the full model.

2. Data and Methods

2.1 Description of the variables

A city was randomly selected and surveyed from each of the 49 States and DC. Data obtained from every city contented one dependent variable Y, and five predictor or independent variables X_1, X_2, X_3, X_4, X_5 as defined above. The descriptive statistics of all variables were listed in Table 1.

Table 1. Descriptive statistics of all variables.

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y	50	926.9400	889.69484	46347.00	43.00	4743.00
X1	50	190.3600	197.41201	9518.00	11.00	952.00
X2	50	135.1040	198.77469	6755.00	0.40	812.00
X3	50	63.1140	38.95398	3156.00	0.00	196.00
X4	50	41.7400	11.75743	2087.00	20.00	67.00
X5	50	140.4640	161.53326	7023.00	6.20	955.00

Table 2 exhibits the correlation coefficients between Y and each X, as well as among the five X variables.

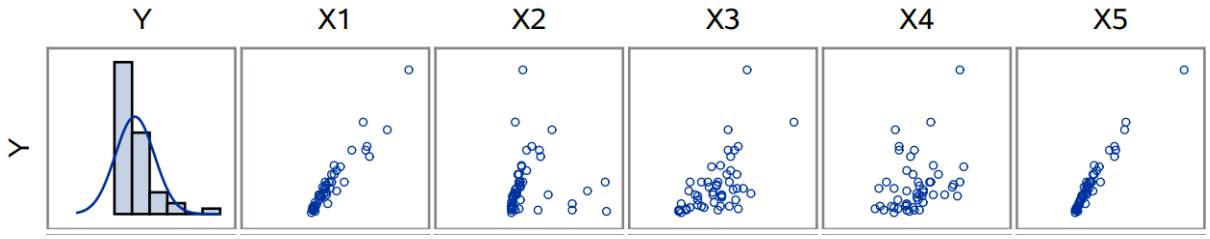
Table 2. Correlations among all variables in the research.

	Y	X1	X2	X3	X4	X5
Y	1.00000	0.95607	0.03947	0.60259	0.32671	0.97448
		<.0001	0.78550	<.0001	0.02060	<.0001
X1	0.95607	1.00000	0.20892	0.49700	0.19445	0.96554
	<.0001		0.14540	0.00020	0.17600	<.0001
X2	0.03947	0.20892	1.00000	-0.41525	-0.03868	0.13115
	0.78550	0.14540		0.00270	0.78970	0.36400
X3	0.60259	0.49700	-0.41525	1.00000	-0.00144	0.51549
	<.0001	0.00020	0.00270		0.99210	0.00010
X4	0.32671	0.19445	-0.03868	-0.00144	1.00000	0.27485
	0.02060	0.17600	0.78970	0.99210		0.05340
X5	0.97448	0.96554	0.13115	0.51549	0.27485	1.00000
	<.0001	<.0001	0.36400	0.00010	0.05340	

Note that Y has significant positive correlations with all five X variables (p-value<0.001 with X_1, X_2, X_3, X_4, and p-value =0.0017<0.05 with X_5). And there were strong positive correlations with X_1 and X_5.

Figure 1 exhibits the linear relationship between Y and five X variables and the histogram with normal distribution. Note that there are a few outliers in the data.

Figure 1. Scatterplot between Y and each X.



2.2 Methods

In this research, first least-squares method was applied to fit the following multiple linear regression model, which is called full model, by using Statistical Analysis System (SAS):

$$Y = \beta_0 + \beta_1*X_1 + \beta_2*X_2 + \beta_3*X_3 + \beta_4*X_4 + \beta_5*X_5 + \varepsilon \qquad [1]$$

Where Y, X_1, X_2, X_3, X_4, X_5 are defined as above in the background. β_0, β_1, β_2, β_3, β_4, β_5 are regression coefficients to be estimated, and ε is the model random error. Second, stepwise selection was applied to find the "best" method with SLE = 0.10 and SLS = 0.05.

Finally, the study compared the "best" model with the full model by various criteria including the coefficient estimates, statistical testing at $\alpha = 0.05$, R^2, adjusted R^2, STB, RMSE, AIC, BIC, and PRESS.

3. Results and Discussion

3.1 Full model

Equation [1] was fit the data using least-square method as follows:

$$Y = -292.274 + 1.839*X_1 - 0.251*X_2 + 2.77*X_3 + 8.266*X_4 + 2.728*X_5 \quad [2]$$

The model R^2 was 0.9793 and adjusted R^2 was 0.977, indicating that 97.93% of the total variation in Y can be explained by the five X variables together. And only one of the five slope coefficients (β_2) were not statistically significant at $\alpha = 0.05$ with p-value = 0.0594 > 0.05. Other four slope coefficients (β_1, β_3, β_4, β_5) were statistically significant at $\alpha = 0.05$. Table 3 shows the parameter estimates of the full model. In addition, the PRESS of the full model is 3958594.32, and AIC = 496.061. Table 4 gives the output of the full model by SAS.

Table 3. Estimated regression coefficients for the full model.

		Parameter	Standard			Standardized	95% Confidence Limits	
Variable	DF	Estimate	Error	t Value	Pr > |t|	Estimate		
Intercept	1	-292.27414	95.54989	-3.06	0.0038	0	-484.84229	-99.706
X1	1	1.83859	0.41746	4.4	<.0001	0.40796	0.99726	2.67992
X2	1	-0.2512	0.12981	-1.94	0.0594	-0.05612	-0.51281	0.01041
X3	1	2.77107	0.73859	3.75	0.0005	0.12133	1.28254	4.25959
X4	1	8.26603	1.82174	4.54	<.0001	0.10924	4.59456	11.93751
X5	1	2.7284	0.50686	5.38	<.0001	0.49537	1.70688	3.74991

Table 4. Output of the full model

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	X1	X2	X3
1	MODEL1	PARMS	Y	134.926	3958594.32	-292.274	1.83859	-0.2512	2.77107
X4	X5	Y	_IN_	_P_	_EDF_	_RSQ_	_AIC_	_BIC_	
8.26603	2.7284	-1	5	6	44	0.97935	496.081	499.68	

3.2 Reduced model (the "best" model)

In this study, stepwise selection was chosen to find the "best" model with SLE = 0.10, SLS = 0.05. It produced the "best" model with four X variables as follows:

$$Y = -369.921 + 1.560*X_1 + 2.638*X_3 + 8.695*X_4 + 2.900*X_5 \quad [3]$$

The model R^2 was 0.9776 and adjusted R^2 was 0.956, indicating that 97.76% of the total variation in Y can be explained by the five X variables together. Both slope coefficients were statistically significant at $\alpha = 0.05$. Table 5 shows the parameter estimates of the reduced model.

Table 5. Estimated regression coefficients for the reduced model.

		Parameter	Standard			Standardized		
Variable	DF	Estimate	Error	t Value	Pr > |t|	Estimate	95% Confidence Limits	
Intercept	1	-369.92111	89.32266	-4.14	0.0001	0	-549.82618	-190.01604
X1	1	1.55985	0.40359	3.86	0.0004	0.34611	0.74697	2.37273
X3	1	3.63783	0.60493	6.01	<.0001	0.15928	2.41944	4.85621
X4	1	8.69485	1.86255	4.67	<.0001	0.1149	4.94349	12.44621
X5	1	2.90044	0.514	5.64	<.0001	0.52661	1.8652	3.93569

The "best" model shows that the number of motor vehicle deaths are positively related to the number of drivers in each state, the total mileage of rural roads, the normal maximum temperature in January, and the highway fuel consumption in gallons. In addition, the highway fuel consumption in gallons (X_5) has the strongest positive effects of the number of deaths with the STB = 0.52661 according to Table 5, and the number of drivers in each state (X_1) is the second strongest positive effects.

In addition, the PRESS of the reduced model is 2872592.18 and AIC = 498.165. Table 6 gives the output of the reduced model.

Table 6. Output of the reduced model.

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	X1	X3
1	MODEL1	PARMS	Y	138.98	2872592.18	-369.921	1.55985	3.63783
X4	X5	Y	_IN_	_P_	_EDF_	_RSQ_	_AIC_	_BIC_
8.69485	2.90044	-1	4	5	45	0.97759	498.165	501.252

3.3 Model Comparison

The characteristics of the full model and the "best" model were list in Table 7. The result shows that the "best" model (equation [3]) with X_1, X_3, X_4, and X_5 have a slightly lower model R2 than the full model (equation [2]) with all five X variables. Also, its AIC and BIC are lower than those of the full model. But the difference between two model's AIC is equal to 2, which means that those two models are essentially indistinguishable.

Most importantly, the PRESS of the reduced model is much smaller than that of the full model, indicating that the "best" model will perform better in predicting Y than the full model.

Table 7. Comparison between the full model and the reduced model.

Model	Independent	R^2	Adj R^2	RMSE	PRESS	AIC	BIC
Full Model	X1 X2 X3 X4 X5	0.97935	0.9770	134.926	3958594.32	496.081	499.680
Reduced Model	X1 X3 X4 X5	0.97759	0.9756	138.980	2872592.18	498.165	501.252

4. Summary

In this research, multiple linear regression models were applied to figure out the relationships between the number of driving death to five potential elements that might affect the amount of death. Stepwise selection was applied to find the "best" model, and shows that four of the five elements (X_1, X_3, X_4, and X_5) can explain about 97.56% of the total variation in the number of driving death.

The analysis indicated that the “best” model might have a better prediction capacity than the full model since the “best” model has a much lower PRESS than that of the full model.

The “best” model suggests that among those four variables, the highway fuel consumption in gallons (X_5) has the strongest positive effects of the number of driving deaths, and the number of drivers in each state (X_1) was the second strongest positive effects. And the only non-significance variable is the number of people per square mile.

5. SAS programs

5.1 Descriptive statistics and correlation

```
OPTIONS NODATE LS=76 PS=45 PAGENO=1 NOLABEL;
PROC IMPORT
DATAFILE='/home/yifangz02120/sasuser.v94/MotorDeath-
RENAMED.xls'
OUT=MOTORDEATH
DBMS=XLS
REPLACE;
RUN;
DATA ALL;
SET MOTORDEATH;
RUN;
*=== DESCRIPTIVE STATISTICS AND CORRELATION===;
PROC CORR DATA=ALL;
VAR Y X1 X2 X3 X4 X5;
TITLE 'MOTORDEATH-Descriptive Statistics and Correlation';
```

```
RUN;
PROC SGSCATTER DATA=ALL;
MATRIX Y X1 X2 X3 X4 X5 / DIAGONAL= (HISTOGRAM NORMAL);
RUN;
```

5.2 Full model and reduced model

```
*===FULL REGRESSION MODEL WITH 5 VARIABLES===;
TITLE 'FULL REGRESSION MODEL WITH 5 VARIABLES';
PROC REG DATA=ALL OUTEST=OUT1;
MODEL Y = X1 X2 X3 X4 X5 / STB CLB CLM AIC BIC PRESS
RSQUARE RMSE;
PROC PRINT DATA=OUT1;
RUN;
*===STEPWISE SELECTION===;
TITLE 'STEPWISE SELECTION OF VARIABLES';
PROC REG DATA=ALL;
MODEL Y = X1 X2 X3 X4 X5 / SELECTION=STEPWISE SLE=0.1
SLS=0.05 SS2;
RUN;
```

5.3 Comparison

```
*===FINAL REGRESSION MODEL WITH 5 VARIABLES===;
PROC REG DATA=ALL OUTEST=OUT1;
MODEL Y = X1 X2 X3 X4 X5 / STB CLB CLM AIC BIC RSQUARE
RMSE PRESS;
PROC PRINT DATA=OUT1;
```

```
RUN;
*===FINAL REGRESSION MODEL WITH 4 VARIABLES===;
PROC REG DATA=ALL OUTEST=OUT2;
MODEL Y = X1 X3 X4 X5 / STB CLB CLM AIC BIC RSQUARE RMSE
PRESS;
PROC PRINT DATA=OUT2;
RUN;
```