Regression Analysis Student Project

Predict transportation accident in Thailand

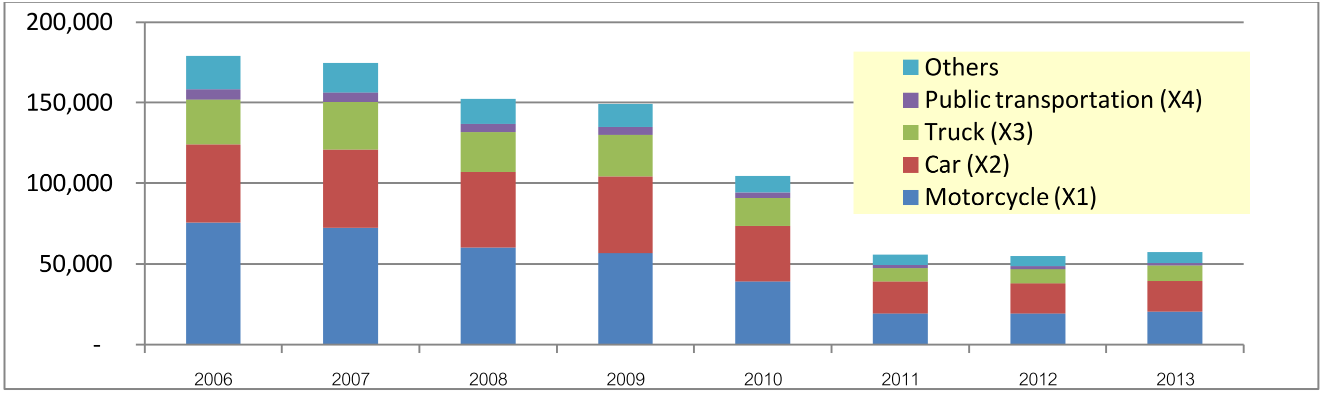
By Areerat Maiam

Introduction

The below table show statistic of transportation accident separated by type of vehicle caused accident from 2006-2013. It shows that no of accident has been reduced in good direction. In 2013, no of accident has reduced by more than half of 2006. Divided by type of vehicles, most of the accidents mainly come from motorcycle, car, truck and public transportation, especially from motorcycle.

Year	TOTAL no of Transportation Accidents	Type of Vehicle				
		Motorcycle (X1)	Car (X2)	Truck (X3)	Public transportation (X4)	Others
2006	178,753	75,752	48,273	27,871	6,531	20,326
2007	174,487	72,373	48,662	29,169	6,168	18,115
2008	152,399	60,248	46,704	24,652	5,074	15,721
2009	149,217	56,658	47,736	25,526	4,815	14,482
2010	104,494	38,815	34,638	17,247	3,401	10,393
2011	55,657	19,311	19,522	8,702	1,833	6,289
2012	55,058	19,122	18,795	8,823	1,847	6,471
2013	57,238	20,239	19,168	9,506	1,726	6,599

This regression analysis takes a look at the relationship between type of vehicles caused accident vs total no of transportation due to accident and create regression model of total no of transportation accident. Graph below show no of accident from each type of vehicle from 2006 - 2013



Data Exploration

The data for this project is obtained from Transport Statistic sub-division, and planning division (web address: http://apps.dlt.go.th/statistics_web/statistics.html)

The available statistic data from this analysis are from 2006-2013 (8 years/observations).
The variables will be assigned as:

1. Y denote the total no of actual transportation accident
2. X_1 represent no of accidents from Motorcycle
3. X_2 represent no of accidents from car
4. X_3 represent no of accidents from truck
5. X_4 represent no of accidents from public transportation

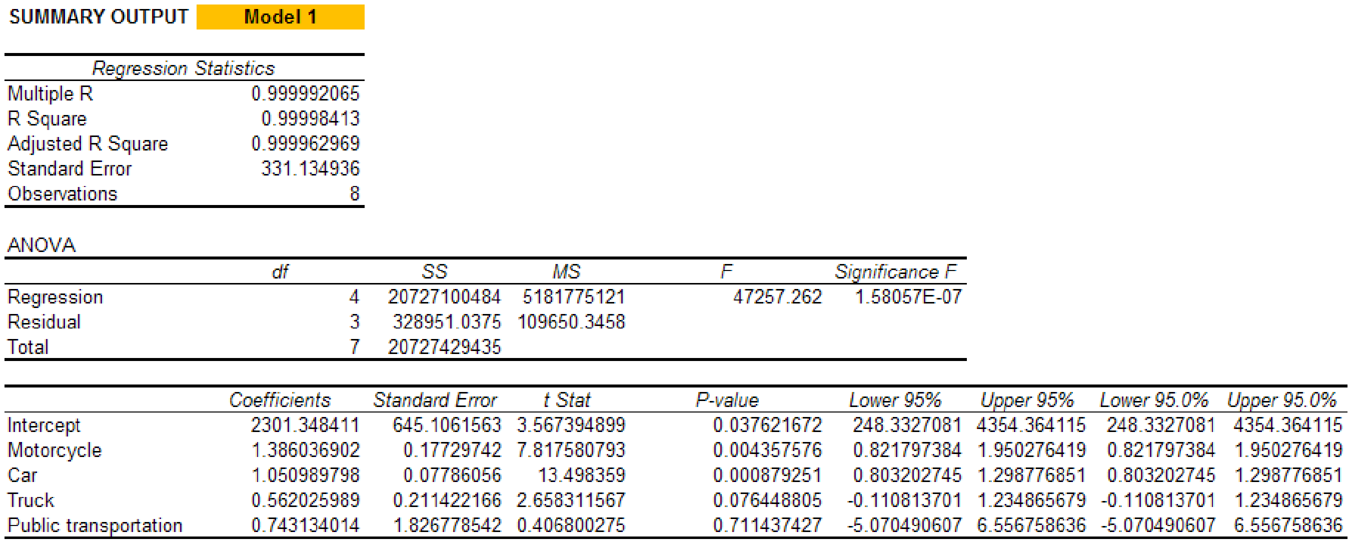
Analysis & Model

Model 1: use all 4 variables to create regression model

All data analysis will be done using Microsoft Excel's regression add-in. We will first take a look at the first scenario, where all variables are used in the following model:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Using Excel's regression add-in, the following summary output is obtained:



SUMMARY OUTPUT Model 1

Regression Statistics	
Multiple R	0.999992065
R Square	0.99998413
Adjusted R Square	0.999962969
Standard Error	331.134936
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	20727100484	5181775121	47257.262	1.58057E-07
Residual	3	328951.0375	109650.3458		
Total	7	20727429435			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2301.348411	645.1061563	3.567394899	0.037621672	248.3327081	4354.364115	248.3327081	4354.364115
Motorcycle	1.386036902	0.17729742	7.817580793	0.004357576	0.821797384	1.950276419	0.821797384	1.950276419
Car	1.050989798	0.07786056	13.498359	0.000879251	0.803202745	1.298776851	0.803202745	1.298776851
Truck	0.562025989	0.211422166	2.658311567	0.076448805	-0.110813701	1.234865679	-0.110813701	1.234865679
Public transportation	0.743134014	1.826778542	0.406800275	0.711437427	-5.070490607	6.556758636	-5.070490607	6.556758636

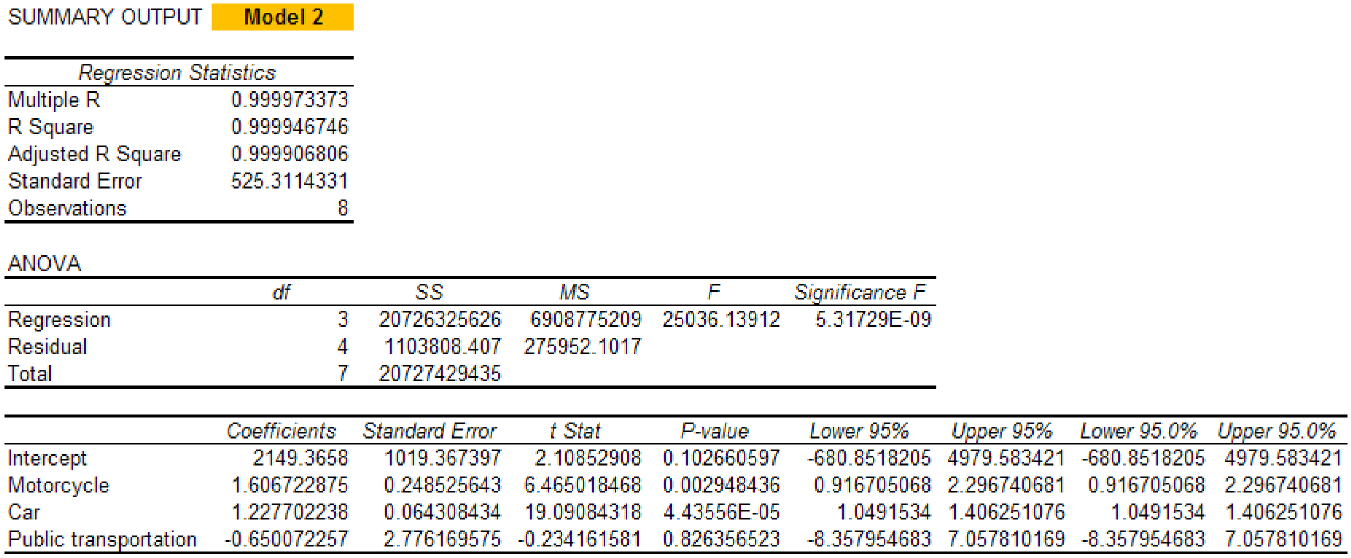
Y = 2301.34841149647 + 1.38603690163428X1 + 1.05098979763321X2 + 0.562025988643144X3 + 0.743134014221157X4+ ε

With above model, R square, adjusted R square are very close to 1 which implies that the model is best fit linear regression correlated with these variables. However, it uses 4 variables to create regression model. Therefore, we will minimize no of variables from 4 to be 3 in next model (model 2).

Model 2: Take out the truck variable from model 1 due to its coefficients, 0.562026 is less than coefficient of other variables

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$$

Again, using Excel's regression add-in, we obtain the following output:



SUMMARY OUTPUT Model 2

Regression Statistics	
Multiple R	0.999973373
R Square	0.999946746
Adjusted R Square	0.999906806
Standard Error	525.3114331
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	20726325626	6908775209	25036.13912	5.31729E-09
Residual	4	1103808.407	275952.1017		
Total	7	20727429435			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2149.3658	1019.367397	2.10852908	0.102660597	-680.8518205	4979.583421	-680.8518205	4979.583421
Motorcycle	1.606722875	0.248525643	6.465018468	0.002948436	0.916705068	2.296740681	0.916705068	2.296740681
Car	1.227702238	0.064308434	19.09084318	4.43556E-05	1.0491534	1.406251076	1.0491534	1.406251076
Public transportation	-0.650072257	2.776169575	-0.234161581	0.826356523	-8.357954683	7.057810169	-8.357954683	7.057810169

Y = 2149.36580015109 + 1.60672287455829X1 + 1.22770223803468X2 + -0.650072257323921X4 + ε

Again, for model 2, R square, adjusted R square are still very close to 1 which implies that the model is best fit linear regression correlated with these variables.

Results

Both models are also best fit linear regression and can be best model to use. Look at their correlation matrix as below. All variables are almost perfectly correlation with other variables.

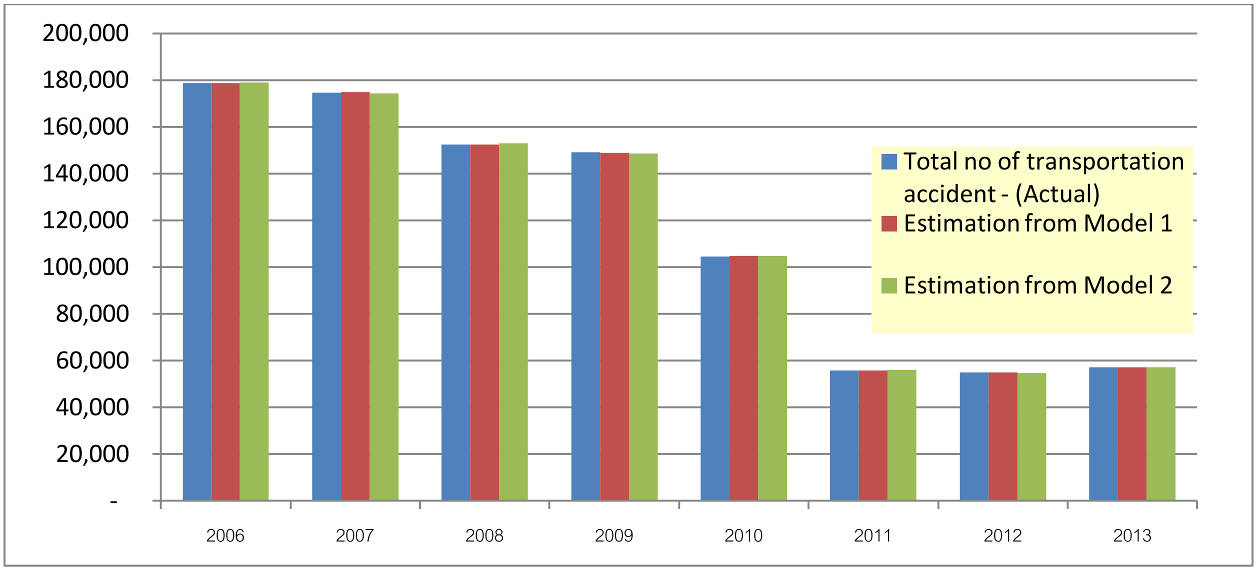
Correlation

	Motorcycle	*Car*	*Truck*	*Public transportation*	*Others*
Motorcycle	1				
Car	0.970106494	1			
Truck	0.988949501	0.991214108	1		
Public transportation	0.999194555	0.96498747	0.985096266	1	
Others	0.996418827	0.952441102	0.974473338	0.997040561	1

In order to compare estimation value from model 1& 2 vs actual value, we have taken into account regression function, then summarize and plot graph as shown below.

Comparison of actual transportation accident vs estimation from model 1 and 2

Year	Total no of transportation accident - (Actual)	Estimation from Model 1	Estimation from Model 2
2006	178,753	178,548.48	178,881.09
2007	174,487	174,733.65	174,165.52
2008	152,399	152,518.45	152,991.34
2009	149,217	148,925.94	148,658.57
2010	104,494	104,725.22	104,828.57
2011	55,657	55,837.44	55,952.41
2012	55,058	54,889.82	54,747.10
2013	57,238	57,123.99	57,078.40



Conclusion

From result above, the estimation from either model 1 and model 2 are also able to be best fit linear regression. If all information are available, we prefer to use regression from model 1.