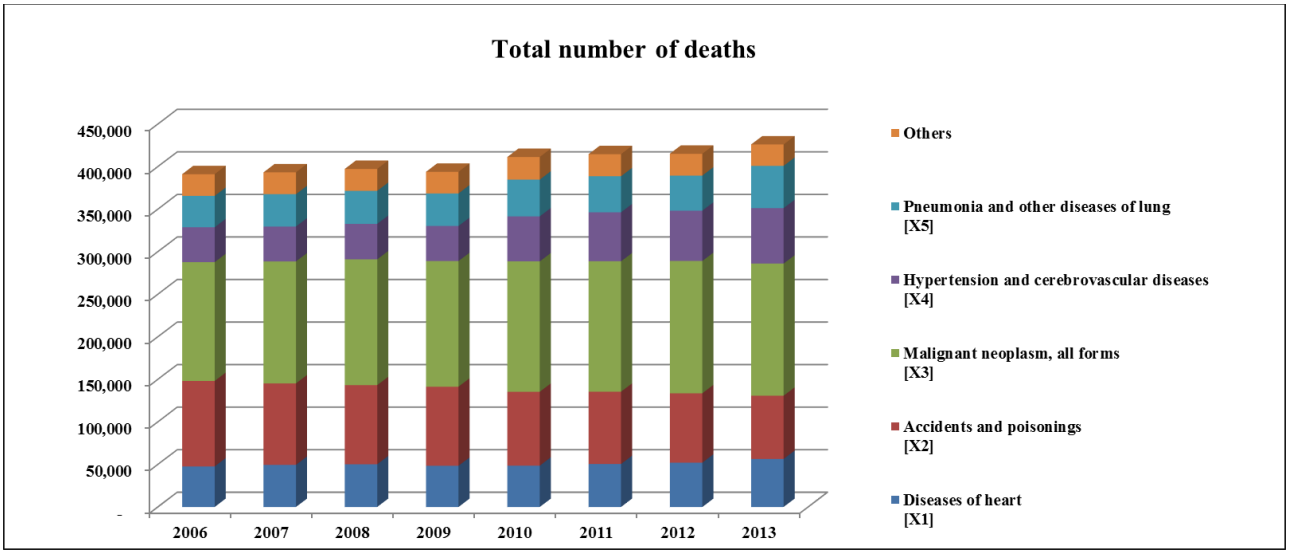
Miss Pattima Ponmatoom
E-Mail: mimkaro@hotmail.com

Regression Analysis Project
Thailand death rate by cause of death

Introduction:

The goal of this study project is to model the Thailand death rate by cause of death from 2006 to 2013. The table below shows statistics of the number of deaths by cause of death. Please note that the number of deaths increased by an average of 1% per year. Therefore, the number of deaths increased throughout the period from year 2006 to 2013, rising a total of 9%.

Year	Total number of death	Leading cause of death					
		Diseases of heart [X1]	Accidents and poisonings [X2]	Malignant neoplasm, all forms [X3]	Hypertension and cerebrovascular diseases [X4]	Pneumonia and other diseases of lung [X5]	Others
2006	391,126	47,677	100,405	139,644	40,996	36,924	25,481
2007	393,254	49,510	95,685	143,374	41,015	38,045	25,624
2008	397,327	50,225	93,007	147,854	41,621	38,808	25,812
2009	393,916	48,393	92,978	147,636	41,211	38,298	25,399
2010	411,331	48,581	86,767	153,345	52,856	43,221	26,561
2011	414,670	50,509	84,980	153,264	57,577	42,364	25,976
2012	415,141	52,092	81,527	155,514	59,117	41,223	25,668
2013	426,065	56,417	74,377	155,268	65,159	49,718	25,126



We use the regression analysis to find the relation between the cause of death and the total number of deaths and then created the regression model for predicting the future number of deaths.

Miss Pattima Ponmatoom
E-Mail: mimkaro@hotmail.com

Data:

The observation data of this project came from the website below;
http://service.nso.go.th/nso/web/statseries/statseries09.html

The variables in this study were assumed follows as:
Y is the total number of deaths
X_1 is the number of death due to heart diseases
X_2 is the number of death due to accidents and poisonings
X_3 is the number of death due to malignant neoplasm, all forms
X_4 is the number of death due to hypertension and cerebrovascular diseases
X_5 is the number of death due to Pneumonia and other diseases of lung

Analysis and Models:

We generate the regression models by using Excel regression add-in function.

Model 1: Perform a regression analysis by using all variables

$Y = a + b_1*X_1 + b_2*X_2 + b_3*X_3 + b_4*X_4 + b_5*X_5$

The table below are the summary output of the model 1 from Excel regression add-in function.

SUMMARY OUTPUT	Model 1
Regression Statistics	
Multiple R	0.999936292
R Square	0.999872587
Adjusted R Square	0.999554055
Standard Error	275.4933351
Observations	8

ANOVA

	df	SS	MS	F	Significance F
Regression	5	1191198038	238239607.7	3139.003299	0.000318501
Residual	2	151793.1553	75896.57767		
Total	7	1191349832			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-53704.6401	45450.65477	-1.181603222	0.358825973	-249263.0239	141853.7437	-249263.0239	141853.7437
Diseases of heart[X1]	1.073028776	0.171177344	6.268521019	0.024516924	0.336512112	1.809545441	0.336512112	1.809545441
Accidents and poisonings[X2]	1.317215685	0.179492895	7.338539411	0.018066977	0.54492009	2.08951128	0.54492009	2.08951128
Malignant neoplasm, all forms[X3]	1.260122077	0.123380871	10.21326944	0.009451037	0.729257037	1.790987117	0.729257037	1.790987117
Hypertension and cerebrovascular diseases[X4]	1.050266582	0.042439777	24.74722189	0.001628865	0.867662959	1.232870205	0.867662959	1.232870205
Pneumonia and other diseases of lung[X5]	1.15070882	0.085533725	13.45327605	0.005479774	0.782686903	1.518730738	0.782686903	1.518730738

The regression model is;

$Y = -53704.64 + 1.07303 *X_1 + 1.31722 *X_2 + 1.26012*X_3 + 1.05027*X_4 + 1.15071*X_5$

Model 2: Perform a regression analysis by excluding the variable of hypertension and cerebrovascular diseases [X4] due to we obtain the lowest coefficient in X_4

$Y = a + b_1*X_1 + b_2*X_2 + b_3*X_3 + b_5*X_5$

The table below are the summary output of the model 2 from Excel regression add-in function.

SUMMARY OUTPUT	Model 2
Regression Statistics	
Multiple R	0.980233244
R Square	0.960857212
Adjusted R Square	0.908666828
Standard Error	3942.61986
Observations	8

ANOVA

	df	SS	MS	F	Significance F
Regression	4	1144717077	286179269.4	18.41061771	0.018905852
Residual	3	46632754.07	15544251.36		
Total	7	1191349832			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	631350.1993	515888.5801	1.223811156	0.308373222	-1010437.506	2273137.905	-1010437.506	2273137.905
Diseases of heart[X1]	-1.243752811	2.050908804	-0.606439842	0.587036014	-7.770659958	5.283154336	-7.770659958	5.283154336
Accidents and poisonings[X2]	-1.710611779	1.879510697	-0.910136762	0.429853142	-7.692053652	4.270830094	-7.692053652	4.270830094
Malignant neoplasm, all forms[X3]	-0.277113932	1.525612573	-0.181641091	0.867444093	-5.132294028	4.578066165	-5.132294028	4.578066165
Pneumonia and other diseases of lung[X5]	0.728025947	1.199430441	0.60697638	0.586723101	-3.089097028	4.545148923	-3.089097028	4.545148923

The regression model is;

$Y = 631350.20 - 1.24375*X_1 - 1.71061*X_2 - 0.27711*X_3 + 0.72803*X_5$

Model 3: Perform a regression analysis by excluding the variable of diseases of heart [X1] due to the low coefficient in X_1

$Y = a + b2*X2 + b3*X3 + b4*X4 + b5*X5$

The table below are the summary output of the model 3 from Excel regression add-in function.

Miss Pattima Ponmatoom
E-Mail: mimkaro@hotmail.com

SUMMARY OUTPUT **Model 3**

Regression Statistics	
Multiple R	0.998683777
R Square	0.997369286
Adjusted R Square	0.993861668
Standard Error	1022.105715
Observations	8

ANOVA

	df	SS	MS	F	Significance F
Regression	4	1188215731	297053932.8	284.3437411	0.000336794
Residual	3	3134100.28	1044700.093		
Total	7	1191349832			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	210626.205	62921.21962	3.347459033	0.044142063	10382.80209	410869.6079	10382.80209	410869.6079
Accidents and poisonings[X2]	0.295283226	0.278632481	1.059758806	0.367022806	-0.591449685	1.182016136	-0.591449685	1.182016136
Malignant neoplasm, all forms[X3]	0.574596747	0.211938003	2.711154854	0.073093112	-0.099884569	1.249078062	-0.099884569	1.249078062
Hypertension and cerebrovascular diseases[X4]	0.904770754	0.13182084	6.863639752	0.006332521	0.485258011	1.324283498	0.485258011	1.324283498
Pneumonia and other diseases of lung[X5]	0.911704545	0.284065586	3.209486081	0.048975462	0.00768107	1.815728019	0.00768107	1.815728019

The regression model is;

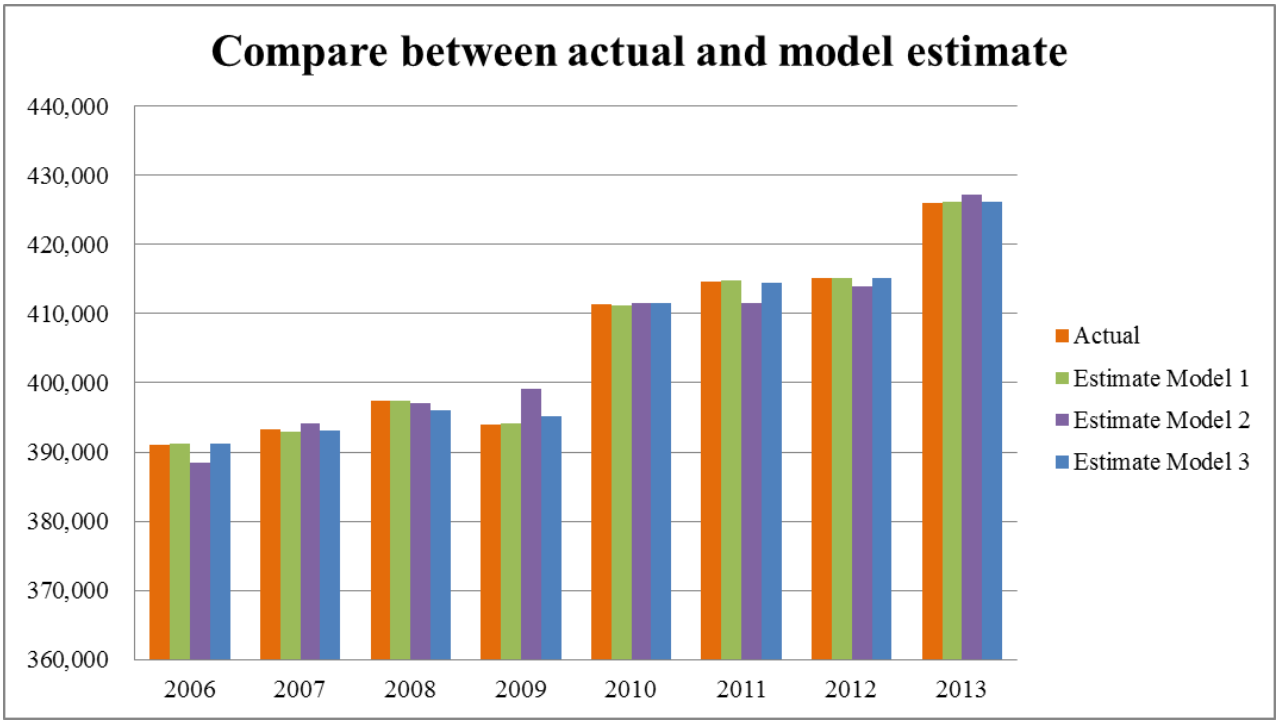
$$\mathbf{Y = 210626.21 + 0.29528*X_2 + 0.57460*X_3 + 0.90477*X_4 + 0.91170*X_5}$$

Result:

The R^2 statistics in regression models are really closed to 1, this means that these models are fit to the observed data.
The comparison between actual and model results are shown in the table below.

Year	Actual	Estimate Model 1	Estimate Model 2	Estimate Model 3
2006	391,126	391,222	388,482	391,268
2007	393,254	392,984	394,058	393,058
2008	397,327	397,383	397,065	396,085
2009	393,916	394,088	399,082	395,116
2010	411,331	411,196	411,475	411,586
2011	414,670	414,781	411,533	414,502
2012	415,141	415,071	414,016	415,128
2013	426,065	426,105	427,120	426,086

Miss Pattima Ponmatoom
E-Mail: mimkaro@hotmail.com



Conclusion:

The R2 statistics results in model 1, model 2 and model 3 are really closed to 1, this means that these models are fit to the observed data. However, the model 3 is model results in lower of the p-value. Therefore we conclude that model 3 is a better fit for the observed data.

Best fit model:

$$\mathbf{Y = 210626.21 + 0.29528 * X_2 + 0.57460 * X_3 + 0.90477 * X_4 + 0.91170 * X_5}$$