By Xiaofeng Qian

Regression analysis of heart disease (Cardiovascular Health)

Introduction

The purpose of this project is to build a regression equation to explain the variation in heart disease mortality rate among the US states. A few factors which may affect the heart disease rate are taken into consideration and tested in my model. These factors include: Obesity, Smoking, Physical activities, hypertension and poverty. All the calculations were performed in the attached excel file (RA_Xiaofeng Qian_spring2010_Heart disease.xls).

Model Construction

1. Build the original linear regression equation.

My original equation is displayed as follow:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

where,

Y is the heart disease death rate /100000 by state (data obtained from http://www.statemaster.com/).

X_1 is the obesity rate by state (data obtained from http://www.statemaster.com/).

X_2 is the percentage of smokers by state (data obtained from http://www.statemaster.com/).

X_3 is the No Leisure-Time Physical Activity rate by state (data obtained from http://www.cdc.gov/).

X_4 is the hypertension rate by state (data obtained from http://healthyamericans.org/).

X_5 is the percentage below poverty by state (data obtained from http://www.statemaster.com/).

The original model with all the variables were calculated in the attached Excel Spreadsheet and shown in Table 1. From Table 1, I found the regression model has strong statistics on R square (0.766042), adjust R square (0.740047) and *F* statistics (Significance of *F* =3.8E-13). However, I noticed a few problems in the preliminary results. First of all, the variable of the obesity rate has a very poor *t* statistics (0.743493) and a high *P*-Value (0.461047) which is a big surprising to me. Because the obesity rate is thought to be very close to the occurrence of the heart disease from medical research and experience. Secondary, the standard errors for most of the independent variables are

too high. All these indicate there must be multicollinearity among the independent variables.

Table 1. Original regression Statistics summary

Regression Statistics	
Multiple R	0.875238
R Square	0.766042
Adjusted R square	0.740047
Standard Error	19.22071
Observations	51

ANOVA

	df	*SS*	*MS*	*F*	*F Significance*
Regression	5	54433.54	10886.71	29.47	3.80E-13
Residual	45	16624.61	369.43		
Total	50	71058.15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-51.66	26.54	-1.95	0.0579	-105.12	1.79877
Obesity Rate	133.10	179.02	0.74	0.461	-227.46	493.652
Per of Smokers	-135.76	111.63	-1.22	0.230	-360.58	89.07133
Physical inactivity	303.96	131.69	2.31	0.026	38.717	569.2072
Hypertension	716.10	165.77	4.32	8.5E-05	382.22	1049.98
Poverty	112.58	103.19	1.10	0.281	-95.26	320.4193

By Xiaofeng Qian

So I calculated the sample autocorrelations between each pair of variables which are shown in Table 2. From Table 2, I found there are high correlations among the variables of the obesity rate, physical inactivity and hypertension rate (71%, 71%, 76%).

Table 2. Correlations between pairs of variables

	H.D. Death Rate	Obesity Rate	Per of Smokers	Physical Inactivity	Hyper-tension Rate	Per below Poverty
Heart Disease Death Rate/10^5(Y)	1	68%	49%	77%	84%	59%
Obesity Rate	68%	1	59%	71%	71%	51%
Percentage of Smokers	49%	59%	1	54%	66%	29%
No Leisure-Time Physical Activity	77%	71%	54%	1	76%	52%
Hypertension Rate (% Adults)	84%	71%	66%	76%	1	58%
Percent below Poverty	59%	51%	29%	52%	58%	1

Then I ran the regression models of the heart disease death rate over each variable (the obesity rate, the physical inactivity and the hypertension rate) separately (Details seen in the attached excel spreadsheets). The comparison and summary of these statistics over the original multiple independent variable regression model (MVR) were displayed in Table 3. In comparison with the original multivariable model, all these three separated regression models have much stronger *F* statistics, *t* statistics and lower standard errors. It further confirms the existence of the multicollinearity in the original model. Although both the obesity rate and physical inactivity are good explanation of the variable of the heart disease death rate, I would remove them from my original model. Since the hypertension rate has a high correlation with these two variables (71%, 76%) and the highest *F* statistics, *t* statistics and lowest standard error among these three variables, it would be a good reprehensive for the other two variables to explain the variable of the heart disease rate. A likely and intuitively explanation of this result is: Hypertension may directly cost heart disease. Most of fat or physical inactivity people may be easier to get a hypertension which leads to the heart disease.

By Xiaofeng Qian

Table 3. Summary and comparison of Regressions

	Obesity Rate		**Physical Inactivity**		**Hypertension Rate**	
Regression Models	Single Variable	MVR	Single Variable	MVR	Single Variable	MVR
F statistics	42.4	36.1	73.0	36.1	116.4	36.1
t statistics	6.5	0.74	8.5	2.3	10.8	4.3
P-value	3.84E-08	0.46	2.86E-11	0.026	1.52E-14	8.5E-05
Standard errors/Coeff.	15%	135%	12%	43%	10%	23%

Table 4. Summary of the statistics for the regression model after removal of obesity factor and physical inactivity factor

Regression Statistics	
Multiple R	0.849057
R Square	0.720898
Adjusted R square	0.709268
Standard Error	20.32676
Observations	51

ANOVA

	df	*SS*	*MS*	*F*	*F Significance*
Regression	5	51445.9	17148.6	41.1	3.5E-13
Residual	45	19612.2	417.3		
Total	50	71058.2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-39.77	26.447	-1.504	0.139	-92.96	13.42
Per of Smokers	-82.99	114.22	-0.727	0.471	-312.78	146.80
Hypertension	975.99	146.89	6.644	2.83E-08	680.49	1271.50
Poverty	168.98	106.65	1.584	0.120	-45.58	383.54

By Xiaofeng Qian

The regression of the heart disease rate by state over the independent variables without obesity rate and physical inactivity was calculated and summarized in Table 4. Both R square and the adjusted R square are slightly decreased from the original model: R square is reduced from 0.766 to 0.724; the adjusted R square is reduced from 0.740 to 0.706. But the F statistics is improved from original 29.5 to 41.1. More importantly, the *t* statistics is much more significant for the variables of hypertension and poverty rate now. The *t* statistics for hypertension is raised from 4.32 to 6.64 and the *t* statistics for the poverty rate is raised from 1.09 to 1.58. However, the variable of percentage of smokers in the model is still not good. It has a high standard error/coefficient ratio 137% and p-value 0.47109. Additionally, the percentage of smokers by state has a high correlation with hypertension rate by state (66%) and low correlation with heart disease death rate (49%). So I removed this variable from my model in the next step.

So our final regression model of the heart disease rate is built with only two independent variables: hypertension rate and poverty rate. The statistics data is shown in table 5. The R square slightly decreases from 0.724 in previous model to 0.721 but the adjust R square increases from 0.706 to 0.709. Also *F* statistics is improved from 41.1 to 62.0 in this model. The t statistics is greatly enhanced as well. So in conclusion, I believed the final model is the best fit model with all these variables available. The equation will be:

$$Y_i = -40.5 + 909.3 \times X_{1i} + 180.1 \times X_{2i}$$

where,

Y is the heart disease death rate /100000 by state,

X_1 is the hypertension rate by state,

X_2 is the percentage below poverty by state.

The figure 1 shows the regression model predicts the Y variable pretty well.

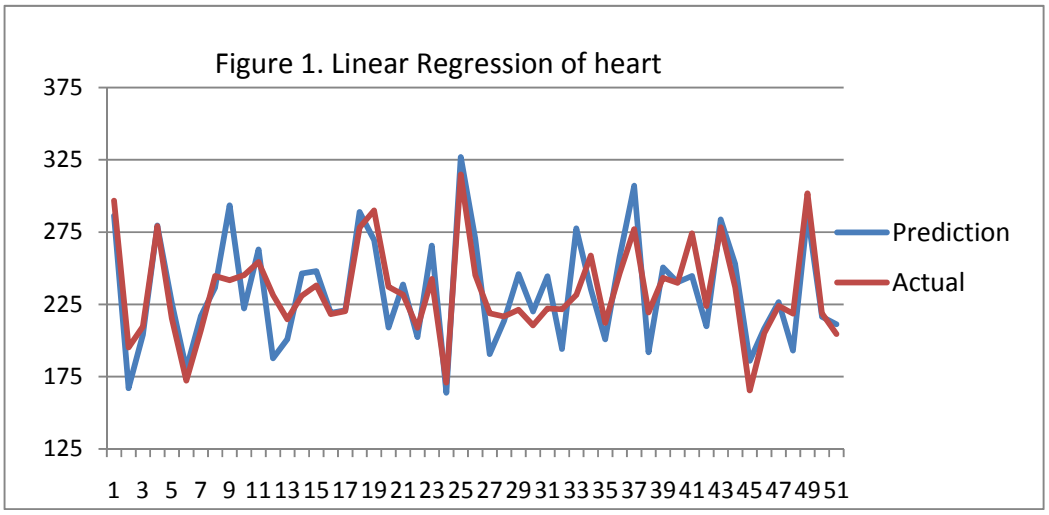

Table 4. Summary of statistics for the regression model after removal of obesity factor and physical inactivity factor

Regression Statistics	
Multiple R	0.849057
R Square	0.720898
Adjusted R square	0.709268
Standard Error	20.32676
Observations	51

ANOVA

	df	*SS*	*MS*	*F*	*F Significance*
Regression	2	51225.66	25612.83	61.99	5.0E-14
Residual	48	19832.49	413.177		
Total	50	71058.15			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-40.51	26.29	-1.54	0.130	-93.37	12.35
Hypertension	909.31	114.12	7.97	2.46E-10	679.85	1138.77
Poverty	180.10	105.03	1.71	0.093	-31.08	391.27