Regression Analysis Winter 2012

Zhuo Jin

Reference Number: 39797167

Factors that affect the crime in small cities

Introduction

This project aims to determine the relationship of crime in small cities with a number of explanatory variables using regression analysis. The data comes from "Life In America's Small Cities", By G.S. Thomas.

Y = Total overall reported crime rate per 1 million residents

X_1= Reported violent crime rate per 100,000 residents

X_2 = Annual police funding in $/resident

X_3 = % of people 25 years+ with 4 yrs. of high school

X_4 = % of 16 to 19 year-olds not in high school and not high school graduates

X_5 = % of 18 to 24 year-olds in college

X_6 = % of people 25 years+ with at least 4 years of college

Objectives of the Study:

1. To identify factors those significantly affect crime in small cities

2. To find a regression model that best estimates the crime in small cities

Methodology

My methodology was simple, use the data analysis add-in for Microsoft excel and find the minimal amount of variables necessary to affect crime in small cities.

I mainly assume the variables listed from X1-X6 all have significant effect to choose the necessary variables. After using the 6 variables to describe the health condition, I came up with the following statistics:

Regression Statistics	
Multiple R	0.783045776
R Square	0.613160687
Adjusted R Square	0.559183109
Standard Error	195.1578302
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	6	2595877.037	432646.2	11.35954	1.42427E-07
Residual	43	1637722.883	38086.58		
Total	49	4233599.92			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	100.3936116	370.6931749	0.270827	0.78782
X_1	0.332336477	0.059617416	5.574486	1.52E-06
X_2	3.998173898	2.682482516	1.490475	0.143399
X_3	1.857912471	5.240872552	0.354504	0.724694
X_4	7.838860632	7.759872204	1.010179	0.31806
X_5	2.558769325	3.426951869	0.74666	0.459332
X_6	-3.231161942	10.71537117	-0.30154	0.764453

$Y = 100.39 + 0.33X_1 + 4.00X_2 + 1.86X_3 + 7.84X_4 + 2.56X_5 - 3.23X_6$

From the P-values of X_3 and X_6, we can see that 0.724694 and 0.764453 are fairly high. Based on the 95% confidence level, it is sufficient to indicate that % of people 25 years+ with 4 yrs. of high school and % of people 25 years+ with at least 4 years of college are not necessary factors to affect the crime condition in small cities. It is shown that the high education level is not a significant factor to affect the crime condition. Knowing this, I remove the two variables X_3 and X_6 to continue doing the regression analysis.

After using the 4 variables to predict crime condition I came up with the following statistics:

Regression Statistics	
Multiple R	0.782291275
R Square	0.611979639
Adjusted R Square	0.57748894
Standard Error	191.0626989
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	2590876.949	647719.2	17.74332	8.29075E-09
Residual	45	1642722.971	36504.95		
Total	49	4233599.92			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	211.2947736	164.744804	1.282558	0.206214
X_1	0.3268637	0.056050074	5.831637	5.55E-07
X_2	4.070755261	2.356701108	1.727311	0.090971
X_4	6.454484263	5.984638329	1.078509	0.286554
X_5	1.744066077	2.383084628	0.731852	0.468053

$Y = 211.29 + 0.33X_1 + 4.07X_2 + 6.45X_3 + 1.74X_5$

Considering the 4 variables, the regression results are still not satisfied even though it has some improvement. The F statistics is 17.74332, which is better than 11.35945 and shows some advantage in this model. In addition, R is 0.782291275 and does not change much comparing to 0.783045776 in the previous model.

However, the P-value of X_5 and X_4 are 0.468053 and 0.286554, which are still too high to accept the necessity of the variables of % of 16 to 19 year-olds not in high school and not high school graduates and % of 18 to 24 year-olds in college. So I further removed these two variables to continue the regression analysis.

After using the2 variables to predict crime condition I came up with the following statistics:

Regression Statistics	
Multiple R	0.775776452
R Square	0.601829103
Adjusted R Square	0.584885661
Standard Error	189.382888
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2547903.642	1273952	35.519884	3.99607E-10
Residual	47	1685696.278	35865.88		
Total	49	4233599.92			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	350.8864621	79.25300386	4.427422	5.653E-05
X_1	0.335467291	0.054795407	6.122179	1.758E-07
X_2	4.246962457	2.274778918	1.866978	0.0681491

$Y = 350.89 + 0.036X_1 + 4.25X_5$

We see from the analysis result that F statistics is improved to 35.519884 and R is still 0.775776452, which shows that this model is better fit the observations. Moreover, the P-value of X_1 is perfect to indicate the necessity of the variable. While the variable X_2 (Annual police funding in $/resident) does not show the significance based on the 95% confidence level. I will try to remove the variable X_2 to continue test the feasibility of the model.

Thus the variable only contains X_1 (Reported violent crime rate per 100,000 residents). After using this solely variable to predict crime condition I came up with the following statistics:

Regression Statistics	
Multiple R	0.756505129
R Square	0.57230001
Adjusted R Square	0.563389593
Standard Error	194.2244537
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	2422889.276	2422889	64.2282	2.09562E-10
Residual	48	1810710.644	37723.14		
Total	49	4233599.92			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	479.1448702	40.52692949	11.82288	7.99E-16
X_1	0.387573647	0.048360565	8.01425	2.1E-10

$Y = 479.14 + 0.39\ X_1$

The one-variable model produced splendid results. Compared to the previous models this one had a fairly high R-square value 0.756505129. The F-statistics is much improved to 64.2282. The P-value of the variable X_1 is sufficient small to demonstrate the necessity of Reported violent crime rate per 100,000 residents. This model also has the least number of variables. It is so far the best model to predict the crime condition in small cities.

Moreover, I continue to remove the variable and leave the rest 5 variables. The regression results is shown below:

Regression Statistics	
Multiple R	0.57758352
R Square	0.333602722
Adjusted R Square	0.257875759
Standard Error	253.2183509
Observations	50

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	1412340.459	282468.1	4.405336	0.002443987
Residual	44	2821259.461	64119.53		
Total	49	4233599.92			

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	489.648597	472.3659239	1.036587	0.305592
X2	10.98067026	3.077783886	3.56772	0.000884
X3	6.088529386	6.543685388	-0.93044	0.357219
X4	5.480304202	10.05349938	0.545114	0.588428
X5	0.377044314	4.417396035	0.085354	0.932367
X6	5.500471224	13.75390717	0.399921	0.69115

From the results we see that R=0.57758352 is much lower comparing to the previous 4 models with X_1. In addition, the F statistics 4.405336is fairly small to accept the omnibus null hypothesis. Unfortunately, this model is not suitable to describe the relationship of crime condition with explanatory variables.

Conclusion

The one-variable model, using reported violent crime rate per 100,000 residents, is the best predictors of total overall reported crime rate per 1 million residents. Intuitively thinking, this result is rather reasonable. The total overall reported crime rate per 1 million residents should have a strong linear relationship to the reported violent crime rate of a smaller sample set. Furthermore, although the 6 and 4 variable models had higher R-square values, the superfluous variables made them less efficient than our one-variable model which yielded similar results. At last, the variable X_1 show its necessity comparing with the other 5 variables.

Our final model for predicting delinquency rates is as follows:

The one-variable model, using reported violent crime rate per 100,000 residents

$Y = 479.14 + 0.39\ X_1$