Chicago Insurance Redlining Final Report

Andrew Siu

May 5, 2018

1 Introduction

Insurance is an important financial instrument in everyday lives, providing a guarantee of compensation for specified loss, damage, illness, or death in return for payment of a premium. Homeowners insurance is specifically important because it can cover damage caused by perils such as fire, windstorms, hail, lightning and vandalism, just to name a few. However, there were allegations made by certain communities in Chicago that insurance companies were redlining their neighborhoods. Redlining is an unethical practice of denying financial services, in this case, homeowners insurance based on race or ethnicity. Victims of redlining are denied insurance solely based on their communities having a large proportion of minorities, or because they are a minority.

In a study of insurance availability in Chicago, the U.S. Commission on Civil Rights attempted to examine charges by several community organizations that insurance companies were redlining their neighborhoods, more specifically, canceling policies or refusing to insure or renew based on ethnicity. To investigate these claims, different types of insurance data was provided. The number of homeowners insurance policies written in Chicago and number of FAIR plan policies were provided by about 70% of insurance companies in the area. FAIR plan policies are available to homeowners who are denied insurance by being deemed as a high risk. Theft and fire rates were provided by the Chicago Police Department and Chicago Fire Department, respectively. The U.S Census Bureau provided data about the race, income, and age of homes of different communities in Chicago. It is important to note that race was reported as percent minority. With the available data provided, the goal is to investigate the main claim that insurance companies are redlining communities in Chicago by using linear regression analysis .

The variables of the data are:

- zipcode: individual zipcode for different areas in Chicago
- race: racial composition in percent minority
- fire: fires per 100 housing units in 1975
- theft: theft per 1000 population in 1975
- age: percent of housing units built before 1939
- vol: new homeowner policies plus renewals minus cancellations and non renewals per 100 housing units
- invol: new FAIR plan policies and renewals per 100 housing units
- income: median family income

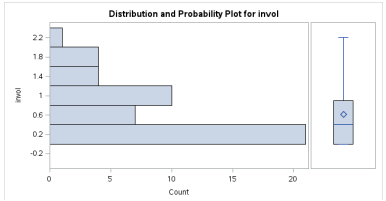
2 Methods

The invol variable was chosen as the response variable because it seems to be the best measure for people that are denied insurance. It is important to keep in mind that this variable is not a perfect measure because not everyone goes into the involuntary market after being denied insurance. Before any analysis, the assumptions of linear regression are defined.

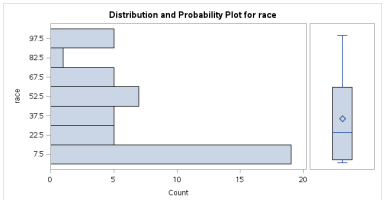
Assumptions for Multiple Linear Regression

- Linear Model is used
- Constant Variance of Residuals
- Residuals are independent
- Errors are normally distributed $N(0, \sigma^2)$
- No multicollinearity
- No outliers: there are a few outliers present, and will be adjusted in future models
- No other variables omitted

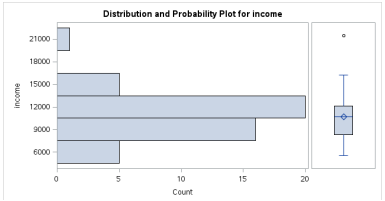
All of the data points were examined, and histograms and box plots were created for each of the variables. It was noticed that there is a wide range in the race variable with some zip codes being almost entirely minority or non-minority. There are several aspects that can be observed from this scatter plot matrix: a negative relationship between the race and income variable; some skewness in the theft and income variables; and the response variable invol has a large number of zeros. This is not good for the assumptions of linear models, but it is still possible to proceed by mitigating as much of the error as possible.



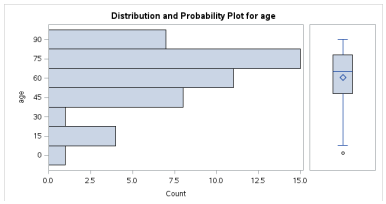
(a) invol



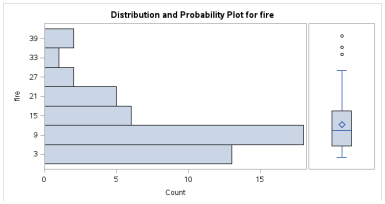
(b) race



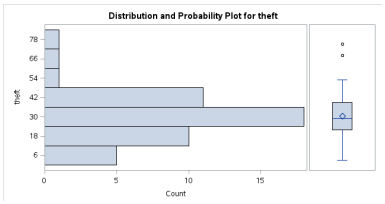
(c) income



(d) age



(e) fire



(f) theft

Figure 1: Histogram and Boxplot for each variable

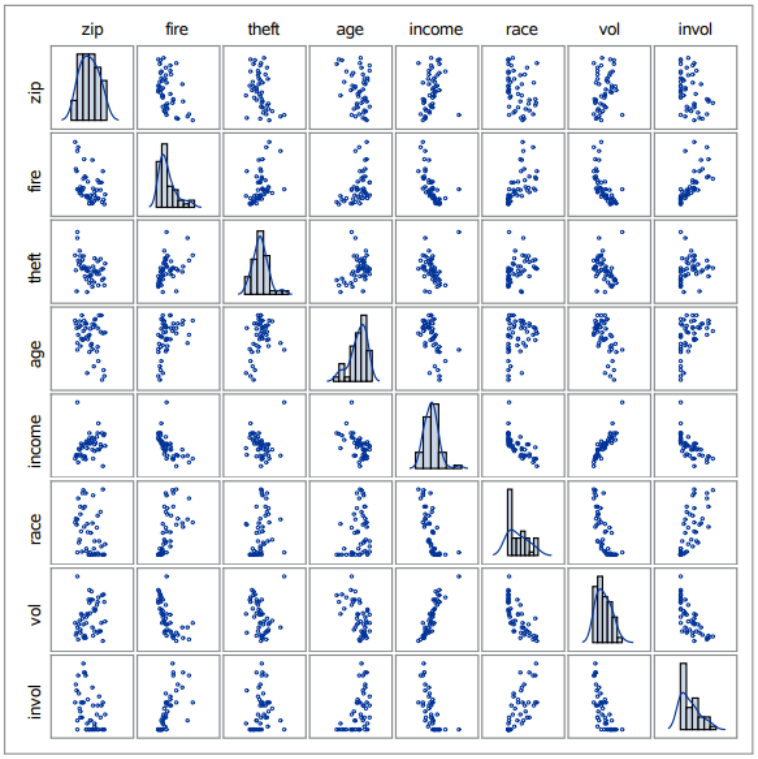
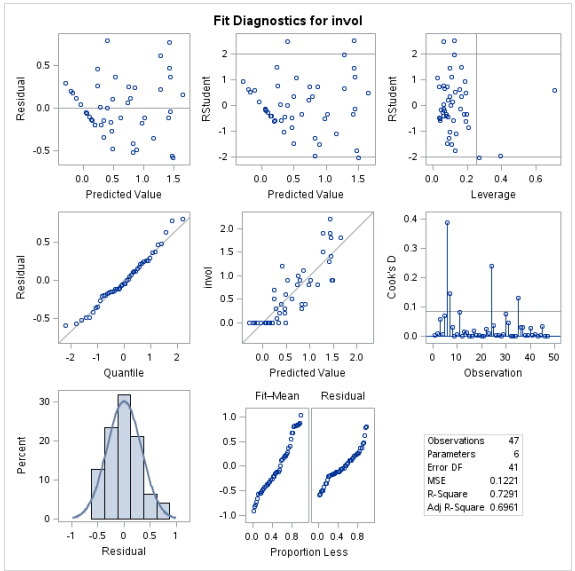


Figure 2: Scatterplot Matrix

The initial model contains all variables except zip and vol. The zip variable was ignored because each data point represents a different zipcode, and zipcodes are an arbitrary number. The variable vol was also left out because it is not needed to estimate the linear regression response variable, invol.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	13.47413	2.69483	22.07	<.0001
Error	41	5.00545	0.12208		
Corrected Total	46	18.47957			

Root MSE	0.34941	R-Square	0.7291
Dependent Mean	0.61489	Adj R-Sq	0.6961
Coeff Var	56.82371		

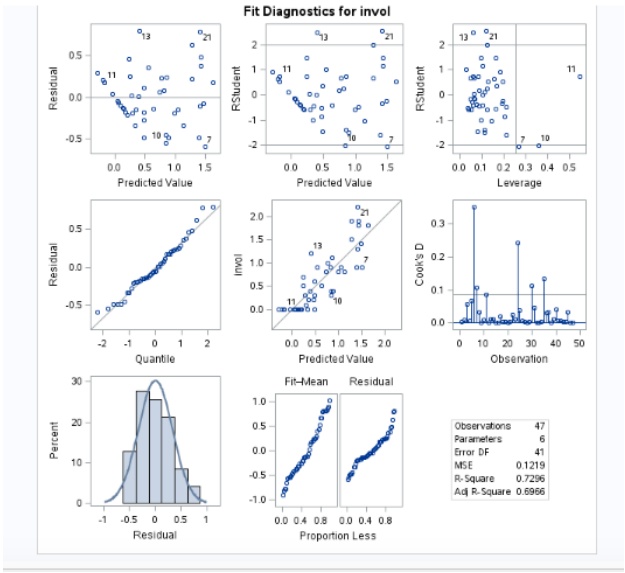
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.74391	0.53700	-1.39	0.1734	0
race	1	0.01131	0.00261	4.34	<.0001	2.71875
fire	1	0.03068	0.00794	3.87	0.0004	2.05447
theft	1	-0.01348	0.00457	-2.95	0.0052	1.66213
age	1	0.00951	0.00309	3.08	0.0037	1.82934
income	1	0.00003928	0.00003583	1.10	0.2793	3.66867

Figure 3: Fit Diagnostic and ANOVA Table for First Model

$$Invol = -0.74391 + 0.01131(race) + 0.03068(fire) - 0.01348(theft) + 0.00951(age) + 0.00003928(income)$$

The initial linear model seems to have constant variance of residuals. A plot of the residuals show they are roughly normal, and as shown in the ANOVA table, all predictors have a vif <10, indicating no multicollinearity. The QQ plot shows that most of the residuals are on the same line. There appears to be a few outliers with high leverage and large Cook's distance. For the current model, it is observed that these outliers-specifically zip codes 7, 10, 11, 13, and 21-have a high Cook's distance and high leverage. For now, these outliers will be noted, and if they reappear in future models, they will be removed from the final model. The residual plot of the full model against income shows decreasing variance. In the scatter plot of income and invol, there is a noticeable skewness because of it's downwards curved shape pattern, which may suggest for a transformation. In addition, the p-value for the income variable is higher than that of the other indicator variables, suggesting that this is not a significant $\alpha = 0.05$ level. A transformation of $log(income)$ was tested to account for the skewness. This made sense because income is best looked at in a multiplicative scale, rather than an additive scale. For example, $100 is worth a lot more to a low income family than a high income family.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	13.48206	2.69641	22.12	<.0001
Error	41	4.99751	0.12189		
Corrected Total	46	18.47957			

Root MSE	0.34913	R-Square	0.7296
Dependent Mean	0.61489	Adj R-Sq	0.6966
Coeff Var	56.77865		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-4.97883	4.26097	-1.17	0.2494	0
race	1	0.01166	0.00279	4.18	0.0001	3.11178
fire	1	0.03190	0.00832	3.83	0.0004	2.26222
theft	1	-0.01322	0.00443	-2.98	0.0048	1.56562
age	1	0.00942	0.00301	3.13	0.0033	1.74723
adj_income	1	0.50021	0.44403	1.13	0.2665	4.58011

Figure 4: Fit Diagnostic and ANOVA Table for Model with log(income)

$$Invol = -4.97883 + 0.01166(race) + 0.03190(fire) - 0.01322(theft) + 0.00942(age) + 0.50021(log(income))$$

Observing the $log(income)$ model, it was noticed that the same zip codes (7, 10, 11, 13, and 21) are still present as outliers and high leverage points in this model. There is also a slightly higher R^2 value than the initial model. The new model seems to follow all but one assumption: The residual plot of the full model against $log(income)$ still shows decreasing error variance. Again, the p-value of the income variable is higher than $\alpha = 0.05$, which means this variable is not significant to the model. In order to validate this, a selection procedure was ran:

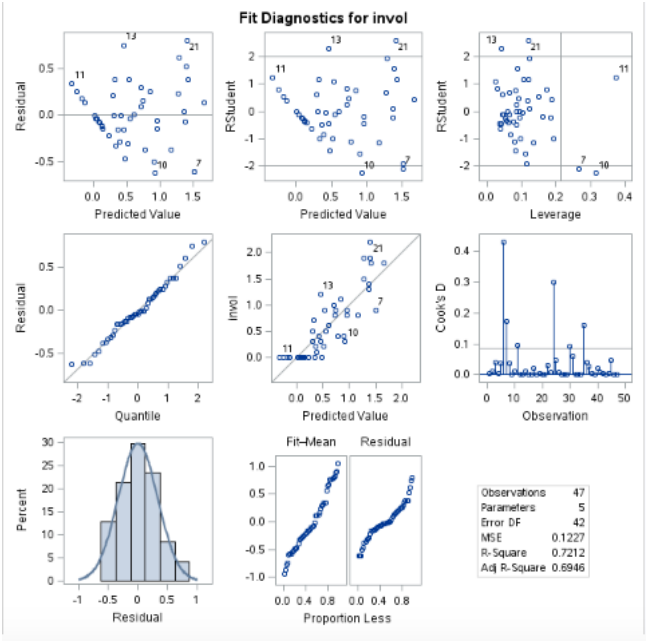
Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	Variables in Model
3	0.5987	0.5707	21.8381	-78.7885	-71.38790	fire age adj_income
3	0.5961	0.5680	22.2295	-78.4870	-71.08645	fire theft adj_income
3	0.5713	0.5414	25.9981	-75.6798	-68.27918	theft race adj_income
3	0.5701	0.5401	26.1802	-75.5483	-68.14774	fire theft age
3	0.5133	0.4793	34.7951	-69.7139	-62.31328	theft age adj_income
4	0.7212	0.6946	5.2690	-93.9041	-84.65333	fire theft race age
4	0.6709	0.6395	12.8956	-86.1078	-76.85709	fire race age adj_income
4	0.6651	0.6333	13.7668	-85.2942	-76.04344	fire theft race adj_income
4	0.6327	0.5977	18.6900	-80.9440	-71.69327	theft race age adj_income

Figure 5: Model Selection Procedure

By running the selection procedure, it verified that there are no interactions between the variables. In addition, the optimal model was chosen by finding the lowest AIC and SBC(BIC), and C(P) closest to the number of predictors. The optimal model has values of C(p)=4.0832, AIC=-103.7456, SBC=-94.71226 and results in a model with the following indicator variables: fire, theft, race, and age. This also validated the earlier suggestion that the income variable is not significant enough to include in the model. This will be the basis for the final model that will be utilized to prove the initial hypothesis.

$$Invol = -.18224 + 0.00944(race) + 0.02779(fire) - 0.01105(theft) + 0.00762(age)$$

In this new model (Figure 6), it can be seen that the p-values for all of the indicator variables are below α=0.05. However, the outliers were observed again, and major outliers (zipcodes 7, 10, 11, 13, and 21) are still present. Among these outliers, zipcodes 7 and 10 seem to have the highest Cook's distance and high leverage. Remedial measures were taken by removing the two of these, and the resulting model and diagnostics plot are shown in Figure 7.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	13.32738	3.33184	27.16	<.0001
Error	42	5.15219	0.12267		
Corrected Total	46	18.47957			

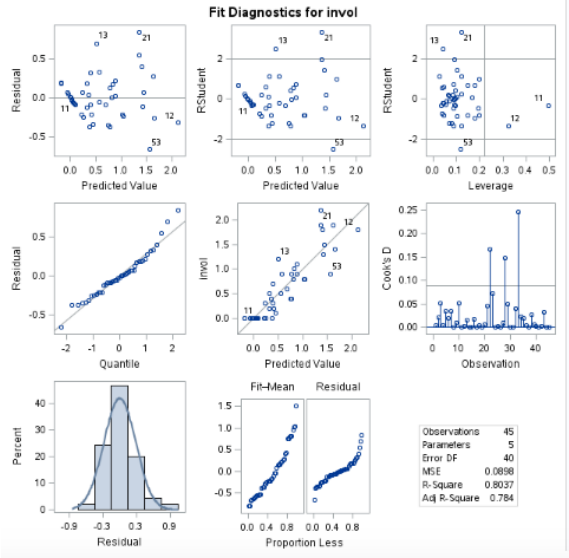
Root MSE	0.35024	R-Square	0.7212
Dependent Mean	0.61489	Adj R-Sq	0.6946
Coeff Var	56.96020		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.18224	0.16136	-1.13	0.2651
race	1	0.00944	0.00198	4.77	<.0001
fire	1	0.02779	0.00750	3.70	0.0006
theft	1	-0.01105	0.00400	-2.76	0.0085
age	1	0.00762	0.00256	2.97	0.0049

Figure 6: Fit Diagnostic and ANOVA Table for Final Model without removing outliers

Removing the two zipcodes significantly changes the modelâĂŹs p-values for fire and theft. This results in theft and age becoming insignificant to the model with p-values above α=0.05. Fire still remains relevant, however this outcome is acceptable because of the assumption that fire rates are important no matter what.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	14.70613	3.67653	40.93	<.0001
Error	40	3.59298	0.08982		
Corrected Total	44	18.29911			

Root MSE	0.29971	R-Square	0.8037
Dependent Mean	0.61556	Adj R-Sq	0.7840
Coeff Var	48.68895		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-0.26787	0.13967	-1.92	0.0623	-0.55015	0.01441
race	1	0.00649	0.00184	3.53	0.0011	0.00278	0.01020
fire	1	0.04906	0.00823	5.96	<.0001	0.03243	0.06568
theft	1	-0.00581	0.00373	-1.56	0.1271	-0.01334	0.00173
age	1	0.00469	0.00233	2.01	0.0514	-0.00002910	0.00940

Figure 7: Fit Diagnostic and ANOVA Table for Final Model with outliers (zip codes 7 and 10) removed

Final model:

$$Invol = -.26787 + 0.00649(race) + 0.04906(fire) - 0.00581(theft) + 0.00469(age)$$

In the new model table (Figure 7), it can be seen that there are many more outliers with high CookâĂŹs Distance values and high leverage values. Analyzing the table, the p-value of race is 0.0126, with two zip codes having race values higher than 98.9. In addition, zip codes 11, 12, and 13 have high CookâĂŹs Distance and leverage. The zip codes 21 and 53 will be removed for their high race values, 11 for theft, and 12 and 13 for high theft, fire and race values. This new model is not shown, however it is observed that the p-values for theft and age skyrocket to p-values over 0.3, while the race p-value hardly moves, increases to 0.0140, which is an increase of only 0.0014. From this, it can be concluded that the high leverage zip codes for theft and age must be removed, while the two race zip codes can be removed from the model and will have minimal impact on the p-value of race significantly. Thus it shows that removing high value race points does not heavily affect the p-value of race.

3 Results

Based on the claim in the introduction, the initial hypothesis was that the insurance companies are redlining higher ethnic dense areas, forcing homeowners in that area to purchase FAIR plan policies. Lower ethnic dense areas will have lower FAIR plan policy rates. Or in terms of hypothesis testing:
$H_0 : \beta_{race} = 0 \; and \; H_a : \beta_{race} \neq 0$. If $\beta \neq 0$, where $\alpha = 0.05$, then it shows that the insurance companies are denying people based on their race or ethnicity. As shown in Figure 7, the p-value for β_{race} is 0.0011. This means that at $\alpha = 0.05$, the null hypothesis is rejected and it is concluded that β_{race} is significant from the model. Further, the 95% confidence interval for β_{race} is (0.00278, 0.01020). As this interval does not include 0, this verifies the rejection of the null hypothesis.

4 Conclusion

Through linear regression analysis, it was found that there is enough evidence of redlining within the Chicago community for the years 1977 and 1978. Using $\alpha = 0.05$, it was found that ethnic density is a significant factor in determining the proportion of people with new FAIR plan policies. It was also recognized that the p-value for fire is also less than $\alpha = 0.05$, which means that fire rate is also a significant factor for the involuntary market activity in homeowner's insurance policies.

There are many different errors that could have been unaccounted for. The major area of error coming from the fact that the variable invol does not include the rate for everyone that was denied voluntary insurance. This

is due to the problem that some people who are denied voluntary insurance, do not sign up for the FAIR plan policies. It also does not factor in the few irrational people that voluntarily signed up for FAIR plan policy even if they were approved for voluntary insurance plans. However, this event of happening is so small and slim that it would not have made a noticeable difference. Another factor to consider is the fact that all of the data used in this analysis comes from one source, the Illinois Department of Insurance, which accounts for more than 70%, but not 100%. Other smaller factors that were not considered are natural factors, such as snow storms and floods that can also contribute to the denial of voluntary insurance. This proves the alternate hypothesis that race is a significant factor, and can safely reject the null hypothesis which states that race is insignificant to the response variable invol.

5 SAS Code

```
ods graphics on;
option ls=100 nodate; /* sets line width at 100 characters */
title 'Chicago Redlining Insurance'; /* gives the output a title */


data Project525;
infile '~/Homework Data/525Project.txt';


input zip fire theft age income race vol invol;
run;
proc print data=Project525;
run;


/*Stem-and-leaf plots and box plots for each predictor*/
proc univariate data=Project525 plots;
var fire theft age income race invol;
run;


proc sgscatter data=Project525;
  matrix invol fire theft age income race /diagonal=(histogram kernel);
run;

/* the code below calculates the correlation matrix for all variables*/
proc corr data=Project525;
var invol fire theft age income race;
run;

/* This is the initial model, with all variables as predictors excluding zip and vol*/
proc reg data=Project525;
model invol=race fire theft age income/vif;
plot r.*p. r.*nqq.; /*r.*fire r.*theft r.*age r.*income;*/
run;
```

```
/*shows exact data points with high leverage or outliers, and large Cook's distance*/
proc reg data=Project525 plots(label)=(cooksd rstudentbyleverage);
model invol=race fire theft age income;
id zip;
run;


/*Make new data set, with log income.*/
data NewInsurance;
        set Project525;
        adj_income = log(income);
run;


proc reg data=NewInsurance;
model invol=race fire theft age adj_income/vif;
plot r.*p. r.*nqq. r.*fire r.*theft r.*age r.*adj_income;
run;


proc reg data=NewInsurance plots(label)=(CooksD RStudentByLeverage);
model invol=race fire theft age adj_income;
id zip;
run;


/*This gives different models*/
proc reg data=NewInsurance outest=out;
model invol = fire theft race age adj_income / selection = rsquare adjrsq cp aic sbc press;
plot rsq.*np.;
plot adjrsq.*np.;
plot cp.*np. / chocking=red  cmallows=blue;
plot aic.*np.;
plot sbc.*np.;
run;


/*This is the final model, based on model selection*/
proc reg data=NewInsurance plots(label)=(CooksD RStudentByLeverage);
model invol=race fire theft age;
id zip;
run;


/*Removed outliers*/
data FinalInsurance;
        set NewInsurance;
        if zip = 7  then delete;
        if zip = 10 then delete;
run;
```

```
proc reg data=FinalInsurance plots(label)=(cooksd rstudentbyleverage);
model invol=race fire theft age/p clb clm cli alpha=0.05;
id zip;
run;


/*   A few outliers based on new model */
/*   See if removing them has any influence on the model */
data OutlierSet;
        set NewInsurance;
        if zip = 21 then delete;
        if zip = 53 then delete;
run;


proc reg data=OutlierSet plots(label)=(cooksd rstudentbyleverage);
model invol=race fire theft age/p clb clm cli alpha=0.05;
id zip;
run;


/*proc glmselect data=NewInsurance plot=CriterionPanel;*/
/*      model invol = zip fire theft age race*/
/*                  / selection=stepwise(select=SL slentry=.01 slstay=.15) stats=all;*/
/*run;
```