Richard Bensics

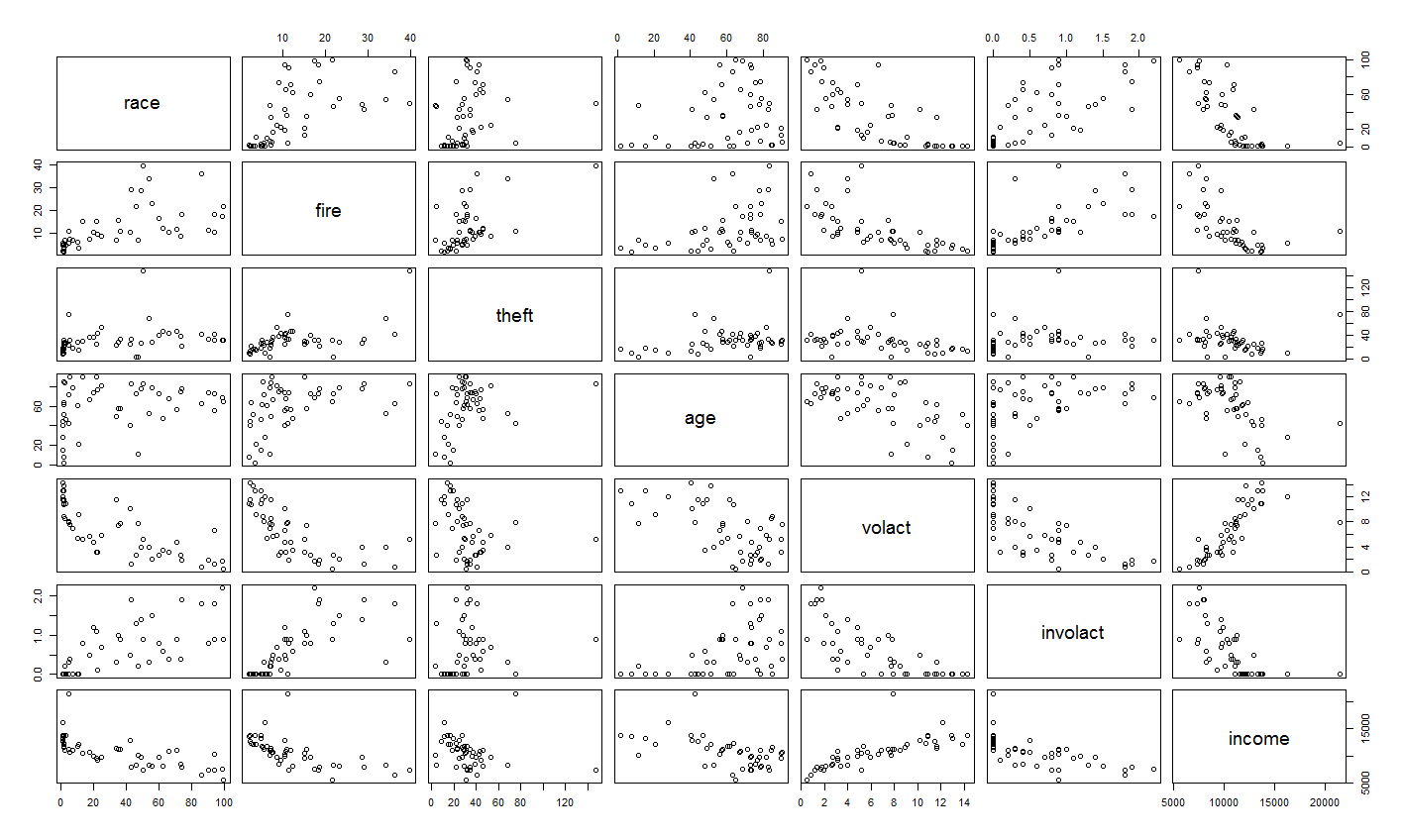
May 8th, 2017

Stat 567, Linear Models and Time Series

Modelling of Chicago Insurance Redlining Data

The purpose of this report is to identify a linear regression model which can predict insurance availability in Chicago zip codes. The data supplied includes demographics such as racial composition, fire frequency, theft frequency, age of housing, insurance availability, and income. Information based on insurance availability was obtained form 1977-1978, and crime data was obtained from 1975.

Before performing any technical analysis, it would be useful see a visual representation of the data. Below is a matrix of scatterplots plotting each possible response variable on each possible predictor variable from the CIR dataset. The entry for zip code has been omitted because the numeric value of a zip code is not meaningful as opposed to the measurements obtained from a given zip code. The title within a column of the matrix represents the predictor variable. Each plot in that column will interpret the title in the row of that matrix as the response variable.



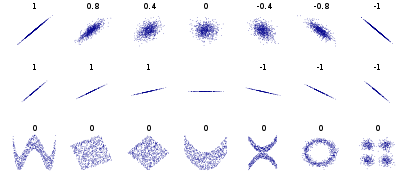
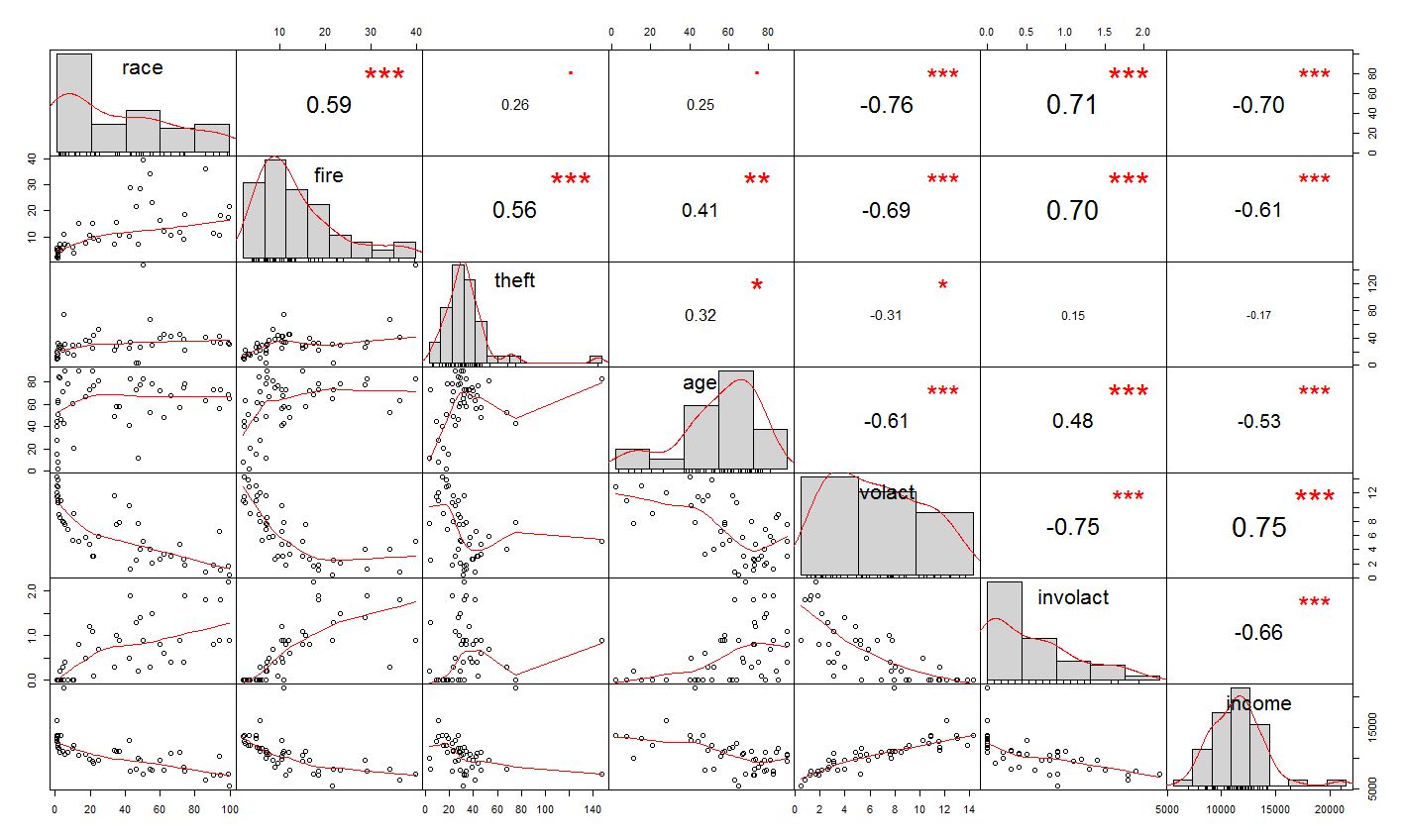
It is desirable to identify any dependencies between these variables. The degree of dependence can be measured by the Pearson Correlation Coefficient *r*. Bivariate data that can be modeled as a linear relationship will have *r*2 ≈ 1. Perhaps it would be easier to identify correlated variables using a correlation matrix. In addition to the scatterplots, the matrix below gives a useful visualization on how each variable is distributed, as well as the correlation coefficient between each variable. Correlation is non-directional, so, for example, Race regressed on Fire results in the same correlation coefficient as Fire regressed on Race.

Figure 1 Scatterplots with their associated correlation coefficients:



However, one cannot simply rank regression models by *r*2 alone. Adding more predictor variables to a regression equation increases the value of *r*2, but may offer poorer predictive performance due to over-fitting; response variables may overreact to fluctuations in data.

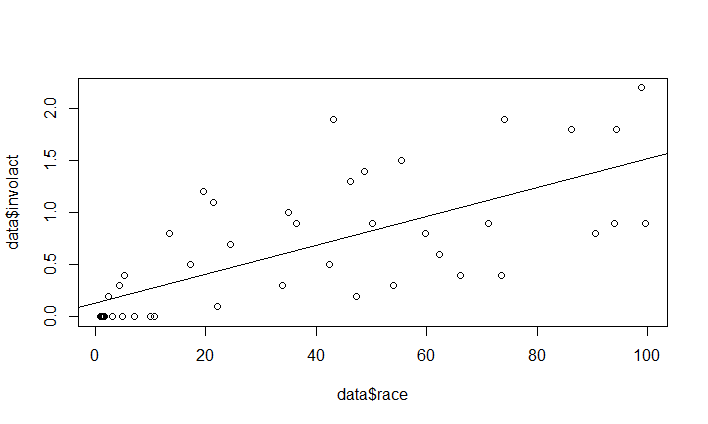
To check for interactions between variables, I have created a simple function[1] which regresses Involact on each possible pair of variable interactions, and stored the significant predictors (determined by p-value) in a vector. The resulting pairs of predictors are as followed:

|  |
| --- |
| > significantInteractions  [1] "race theft" "race age" "race income" "fire age" "fire income"  [6] "theft volact" "age volact" "volact income" |

These significant interactions will be useful for constructing an accurate regression model.

As for interesting features, it is surprising that theft is so poorly correlated with other variables. Even more surprising is that through subsequent tests, Theft becomes a significant variable when Involact is regressed with multiple variables. It is also noticeable that outliers exist in the distributions of Theft and Income, and that most of the distributions are skewed. Consequently, it may be worthwhile to transform the variables, although it is not necessary for predictor variables to be symmetric or normally distributed.

It was suspected by several community organizations that insurance companies were redlining neighborhoods based on ethnicity. We can test the claim that zip codes with high percentage of minorities were being denied insurance by regressing Involact on Race.



|  |
| --- |
| Call:  lm(formula = data$involact ~ data$race)  Residuals:  Min 1Q Median 3Q Max  -0.7496 -0.2479 -0.1487 0.3129 1.1724  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 0.129218 0.096611 1.338 0.188  data$race 0.013882 0.002031 6.836 1.78e-08 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4488 on 45 degrees of freedom  Multiple R-squared: 0.5094, Adjusted R-squared: 0.4985  F-statistic: 46.73 on 1 and 45 DF, p-value: 1.784e-08 |

When regressing Involact on Race, the intercept is found to be .129218. It can be interpreted from this that in a zip-code with no minority groups present, then there would be an estimated .13 new FAIR plan policies and renewals per 100 housing units. However, judging by the scatterplot, forcing the regression line through the origin might result in a better fit. The *t*-value, for Race, tests the hypothesis that Race has no predictive power in the linear model for Involact. More specifically, it tests the null hypothesis that β1 = 0 for:

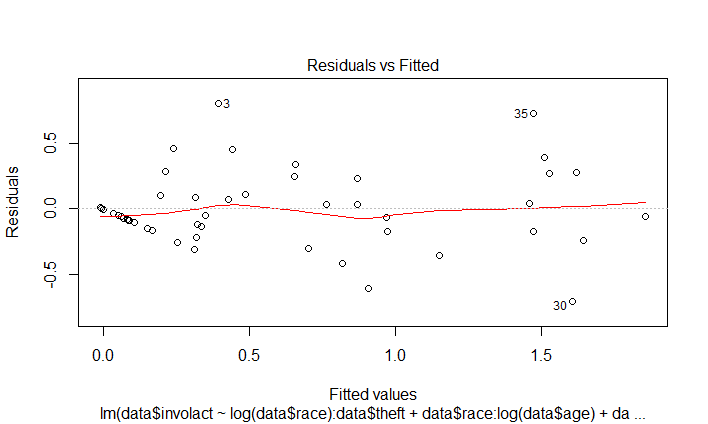
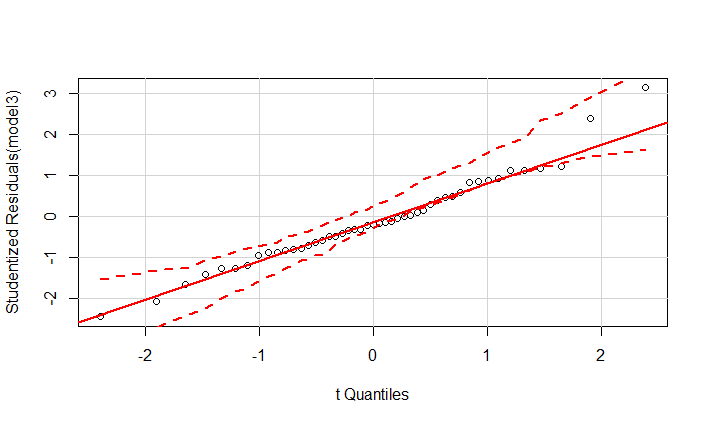
The *p*-value, denoted by Pr(>|t|), is small and falls within the rejection region α = .05. Thus, we can reject the null hypothesis and conclude that there is a significant relationship between Race and Involact. However, it could also be the case that ethnic composition of a zip code could also be related to other factors which may also predict insurance availability. Take the next three possible regression models as examples:

|  |
| --- |
| > summary(model1 <- lm(dataSet$involact ~ dataSet$race + dataSet$fire + dataSet$theft + dataSet$age - 1))  Call:  lm(formula = dataSet$involact ~ dataSet$race + dataSet$fire +  dataSet$theft + dataSet$age - 1)  Residuals:  Min 1Q Median 3Q Max  -0.90775 -0.18548 -0.09804 0.16505 0.83550  Coefficients:  Estimate Std. Error t value Pr(>|t|)  dataSet$race 0.007662 0.001906 4.019 0.000231 \*\*\*  dataSet$fire 0.037965 0.008041 4.721 2.51e-05 \*\*\*  dataSet$theft -0.010567 0.002681 -3.941 0.000293 \*\*\*  dataSet$age 0.004144 0.001598 2.593 0.012958 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.3405 on 43 degrees of freedom  Multiple R-squared: 0.8625, Adjusted R-squared: 0.8497  F-statistic: 67.43 on 4 and 43 DF, p-value: < 2.2e-16  > extractAIC(model1)  [1] 4.00000 -97.46144 |

|  |
| --- |
| > summary(model2 <- lm(dataSet$involact ~ log(dataSet$race):dataSet$age + sqrt(dataSet$theft):dataSet$logIncome + sqrt(dataSet$fire) - 1))  Call:  lm(formula = dataSet$involact ~ log(dataSet$race):dataSet$age +  sqrt(dataSet$theft):dataSet$logIncome + sqrt(dataSet$fire) -  1)  Residuals:  Min 1Q Median 3Q Max  -0.65869 -0.17733 0.03131 0.13610 0.95979  Coefficients:  Estimate Std. Error t value Pr(>|t|)  sqrt(dataSet$fire) 0.2190319 0.0585522 3.741 0.000527 \*\*\*  log(dataSet$race):dataSet$age 0.0031513 0.0006312 4.992 9.89e-06 \*\*\*  sqrt(dataSet$theft):dataSet$logIncome -0.0132027 0.0027384 -4.821 1.74e-05 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.3184 on 44 degrees of freedom  Multiple R-squared: 0.8769, Adjusted R-squared: 0.8685  F-statistic: 104.5 on 3 and 44 DF, p-value: < 2.2e-16  > extractAIC(model2)  [1] 3.0000 -104.6716 |

|  |
| --- |
| > summary(model3 <- lm(data$involact ~ data$race:data$theft+ data$race:log(data$age) + data$fire:data$age - 1))  Call:  lm(formula = data$involact ~ data$race:data$theft + data$race:log(data$age) +  data$fire:data$age - 1)  Residuals:  Min 1Q Median 3Q Max  -0.63699 -0.22062 -0.05725 0.12687 0.78225  Coefficients:  Estimate Std. Error t value Pr(>|t|)  data$race:data$theft -2.377e-04 4.908e-05 -4.842 1.62e-05 \*\*\*  data$race:log(data$age) 3.710e-03 4.822e-04 7.693 1.12e-09 \*\*\*  data$fire:data$age 5.183e-04 7.025e-05 7.378 3.21e-09 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.2906 on 44 degrees of freedom  Multiple R-squared: 0.8975, Adjusted R-squared: 0.8905  F-statistic: 128.4 on 3 and 44 DF, p-value: < 2.2e-16  > extractAIC(model3)  [1] 3.000 -113.276 |

Each of these three models were chosen such that each predictor term was significant. Note that for each model, the intercept such that the model is forced through the origin. This decision was made because the response variable contains several zeros. The first model was constructed by first regressing the full model, determining which predictors, as well as the intercept, were significant, and then constructing a sub-model of those significant predictors. The second equation was constructed by randomly checking for interactions between variables and applying transformations to them to increase Adjusted R-Squared. The last model was constructed using the pairs of significant interactions found earlier in which the pairs were composed of both significant variables, then applying transformations to increase Adjusted R-Squared. Notice that both the Adjusted R-Squared increases for each model and the value for AIC decreases for each model. From this, it is fair to say that of these three models, the third is the best fit.

A common diagnostic is the analysis of residuals, the estimated and observed values of the dependent variable. Neither independent nor dependent variables need to be normally distributed to fit a valid regression equation, but an assumption for regression analysis is that the residuals (the difference between observed and estimated values) are normally distributed. If the model assumptions are correct, then the residual scatter plot would show a random pattern. Similarly, a Q-Q plot can also be used to determine if the residuals fit a normal distribution, in which the residuals would fall approximately along the reference line.

The residual plots for the third model indicate that the residuals are approximately normally distributed, which indicates that it is a good fit for the response Involact.

In R, the leaps() function will search for the best subsets of predictors using the specified criterion. The following output is a logical matrix which lists the possible sub-models, with the first column indicating the number of predictors in each row, and TRUE indicating that the variable is included in the sub-model and FALSE indicating that the variable is not included. This output would be the same regardless of the criterion chosen:

|  |
| --- |
| > leaps(subset(dataSet, select=c(-zip,-involact)), dataSet$involact, names=names(subset(dataSet, select=c(-zip,-involact))))[1]  $which  race fire theft age volact logIncome  1 FALSE FALSE FALSE FALSE TRUE FALSE  1 TRUE FALSE FALSE FALSE FALSE FALSE  1 FALSE FALSE FALSE FALSE FALSE TRUE  1 FALSE TRUE FALSE FALSE FALSE FALSE  1 FALSE FALSE FALSE TRUE FALSE FALSE  1 FALSE FALSE TRUE FALSE FALSE FALSE  2 TRUE TRUE FALSE FALSE FALSE FALSE  2 FALSE TRUE FALSE FALSE TRUE FALSE  2 TRUE FALSE FALSE FALSE TRUE FALSE  2 TRUE FALSE FALSE TRUE FALSE FALSE  2 FALSE TRUE FALSE FALSE FALSE TRUE  2 FALSE FALSE FALSE FALSE TRUE TRUE  2 FALSE TRUE TRUE FALSE FALSE FALSE  2 TRUE FALSE FALSE FALSE FALSE TRUE  2 FALSE FALSE TRUE FALSE TRUE FALSE  2 FALSE FALSE FALSE TRUE TRUE FALSE  3 TRUE TRUE TRUE FALSE FALSE FALSE  3 FALSE TRUE TRUE FALSE TRUE FALSE  3 TRUE TRUE FALSE TRUE FALSE FALSE  3 TRUE TRUE FALSE FALSE TRUE FALSE  3 TRUE TRUE FALSE FALSE FALSE TRUE  3 FALSE TRUE TRUE FALSE FALSE TRUE  3 FALSE TRUE TRUE TRUE FALSE FALSE  3 FALSE TRUE FALSE FALSE TRUE TRUE  3 FALSE TRUE FALSE TRUE TRUE FALSE  3 TRUE FALSE FALSE TRUE TRUE FALSE  4 TRUE TRUE TRUE TRUE FALSE FALSE  4 TRUE TRUE TRUE FALSE TRUE FALSE  4 TRUE TRUE TRUE FALSE FALSE TRUE  4 FALSE TRUE TRUE TRUE TRUE FALSE  4 FALSE TRUE TRUE FALSE TRUE TRUE  4 TRUE TRUE FALSE TRUE TRUE FALSE  4 TRUE TRUE FALSE TRUE FALSE TRUE  4 FALSE TRUE TRUE TRUE FALSE TRUE  4 TRUE TRUE FALSE FALSE TRUE TRUE  4 TRUE FALSE TRUE TRUE TRUE FALSE  5 TRUE TRUE TRUE TRUE FALSE TRUE  5 TRUE TRUE TRUE TRUE TRUE FALSE  5 TRUE TRUE TRUE FALSE TRUE TRUE  5 FALSE TRUE TRUE TRUE TRUE TRUE  5 TRUE TRUE FALSE TRUE TRUE TRUE  5 TRUE FALSE TRUE TRUE TRUE TRUE  6 TRUE TRUE TRUE TRUE TRUE TRUE |

Using adjusted coefficient of multiple determination gives the following result:

|  |
| --- |
| > subsets <- leaps(subset(dataSet, select=c(-zip,-involact)), dataSet$involact, names=names(subset(dataSet, select=c(-zip,-involact))),method="adjr2")  > subsets[4]  $adjr2  [1] 0.5473059680 0.4985435890 0.4845861643 0.4830263212 0.2091252577 0.0006647144 0.6134546099  [8] 0.6088751541 0.5903207094 0.5854906358 0.5693190100 0.5677708704 0.5594955201 0.5500467191  [15] 0.5449880707 0.5379363527 0.6718178973 0.6666566401 0.6476771705 0.6377317972 0.6155145500  [22] 0.6140022031 0.6106633586 0.6098604905 0.6008260184 0.5995329883 0.7231141763 0.6924449273  [29] 0.6668625864 0.6662476636 0.6611510702 0.6445012405 0.6395471332 0.6314430958 0.6300378448  [36] 0.6065647411 0.7214343796 0.7169408422 0.6853492944 0.6597892447 0.6358306554 0.5994128481  [43] 0.7159709599  > subsets[[1]][27,]  race fire theft age volact logIncome  TRUE TRUE TRUE TRUE FALSE FALSE  > subsets[[4]][27]  [1] 0.7231142 |

Since *r*2 can only increase with more predictors, adjusted *r*2 take the number of predictors into consideration. It is desirable to have adjusted *r*2 as close to 1 as possible. The highest value shown in the table is .7231142, corresponding to the 27th model which consists of Race, Fire, Theft, and Age.

Using Mallows’ Cp criterion gives the following results:

|  |
| --- |
| > subsets <- leaps(subset(dataSet, select=c(-zip,-involact)), dataSet$involact, names=names(subset(dataSet, select=c(-zip,-involact))),method="Cp")  > subsets[4]  $Cp  [1] 28.722354 36.447998 38.659335 38.906468 82.301847 115.329190 18.881191 19.590611  [9] 22.464950 23.213195 25.718402 25.958230 27.240195 28.703944 29.487598 30.580006  [17] 10.684463 11.465841 14.339200 15.844859 19.208394 19.437353 19.942831 20.064379  [25] 21.432135 21.627890 3.943717 8.478846 12.261763 12.352693 13.106338 15.568385  [33] 16.300960 17.499322 17.707119 21.178139 5.211348 5.859996 10.420281 14.109911  [41] 17.568368 22.825331 7.000000  > subsets[[1]][27,]  race fire theft age volact logIncome  TRUE TRUE TRUE TRUE FALSE FALSE  > subsets[[4]][27]  [1] 3.943717 |

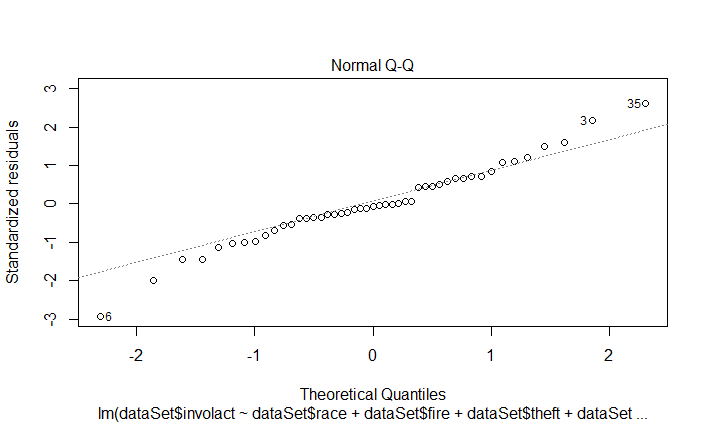
A small value of Cp means that the model is relatively precise. In the output table for Cp, the lowest value is 3.943717, corresponding to the 27th model which consists of Race, Fire, Theft, and Age.

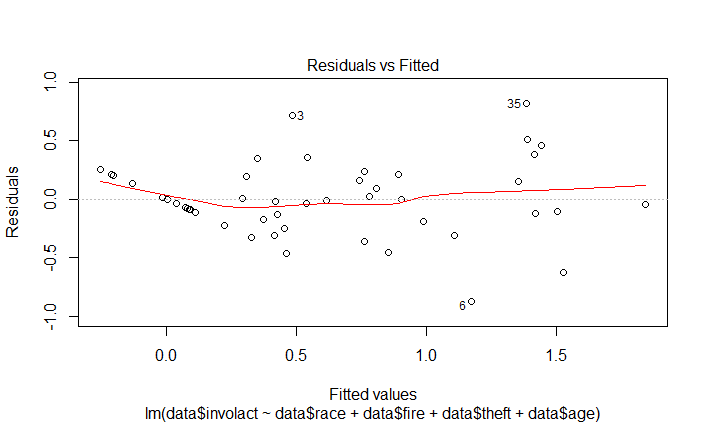
If the Involact column is repositioned to be the last column in the data frame, then the bestglm() function can be used to find the best subset generalized linear model using information criterion.

|  |
| --- |
| > bestglm(dataSet, IC = "AIC")$Subsets  (Intercept) race fire theft age volact logIncome logLikelihood AIC  0 TRUE FALSE FALSE FALSE FALSE FALSE FALSE 21.93682 -43.87363  1 TRUE FALSE FALSE FALSE FALSE TRUE FALSE 41.07798 -80.15597  2 TRUE TRUE TRUE FALSE FALSE FALSE FALSE 45.31832 -86.63665  3 TRUE TRUE TRUE TRUE FALSE FALSE FALSE 49.70507 -93.41015  4\* TRUE TRUE TRUE TRUE TRUE FALSE FALSE 54.25218 -100.50436  5 TRUE TRUE TRUE TRUE TRUE FALSE TRUE 54.67633 -99.35267  6 TRUE TRUE TRUE TRUE TRUE TRUE TRUE 54.80017 -97.60035 |

A smaller AIC indicates a better fit. The results show that the best subset includes Race, Fire, Theft, and Age when using AIC.

It is apparent that each criterion has selected the same sub-set of predictors, however, this will not be the case in all scenarios. This is due to how different criteria are penalized directly to the number of predictors used in the model.





Based on these residual plots, it is a fair assumption that a model containing Race, Fire, Theft, and Age can be used to predict Involact.

The best subset of predictors for Involact can also be found using backwards stepwise regression using the step() function in R.

|  |
| --- |
| > step(lm(dataSet$involact~., dataSet),data,direction="backward")  Start: AIC=-95.6  dataSet$involact ~ race + fire + theft + age + volact + logIncome  Df Sum of Sq RSS AIC  - volact 1 0.02412 4.5882 -97.353  - logIncome 1 0.09813 4.6622 -96.601  <none> 4.5641 -95.600  - age 1 0.61847 5.1826 -91.628  - race 1 1.03947 5.6036 -87.957  - theft 1 1.43409 5.9982 -84.758  - fire 1 2.03392 6.5980 -80.279  Step: AIC=-97.35  dataSet$involact ~ race + fire + theft + age + logIncome  Df Sum of Sq RSS AIC  - logIncome 1 0.08357 4.6718 -98.504  <none> 4.5882 -97.353  - age 1 1.03268 5.6209 -89.812  - theft 1 1.49356 6.0818 -86.108  - race 1 1.63030 6.2185 -85.063  - fire 1 2.31330 6.9015 -80.165  Step: AIC=-98.5  dataSet$involact ~ race + fire + theft + age  Df Sum of Sq RSS AIC  <none> 4.6718 -98.504  - age 1 0.99734 5.6691 -91.410  - theft 1 1.41436 6.0862 -88.074  - race 1 2.05375 6.7256 -83.379  - fire 1 2.38365 7.0554 -81.128  Call:  lm(formula = dataSet$involact ~ race + fire + theft + age, data = dataSet)  Coefficients:  (Intercept) race fire theft age  -0.243118 0.008104 0.036646 -0.009592 0.007210 |

The results of backwards step- regression show that the best model is one that includes the variables Age, Theft, Race, and Fire. This is the same sub-set of variables obtained previously.

These functions, although useful, do not take variable interactions into account. Because of this, I recommend my proposed model:

Each variable, and each variable interaction used in this regression model is significant. Out of all the subsets of variables mentioned in this report, it has the highest Adjusted R-Squared of 0.7949 and the lowest AIC value of -113.5013. In conclusion, race is a component in determining insurance availability in Chicago, but not the determining factor. Considering the interactions between theft, age of houses, and frequency of fires along with racial composition creates a more powerful model for predicting insurance availability

Additional R Code

[1]: (Function to determine significant interactions)

|  |
| --- |
| x = 0  significantInteractions <- NULL  for(i in 2:8) #1st column is zip code  {  if( i != 7 && (i+1)<=8) #7th column is involact  {  for(j in (i+1):8) #Dont interact variables with themselves  {  if(j != 7)  {  if(coef(summary(lm(data$involact~data[[i]]:data[[j]], data)))[2,4] < .05)  {  significantInteractions[x] <-paste(names(data[i]),names(data[j]))  x = x + 1  }  }  }  }  } |