MSDS 6372 Project 1

Infant Mortality Case Study

**Team members:**

Kottegoda Manjula

Messa Araya

Kumar Rajeev

Date:2016-09-13

**Factors of Murder in Cities**

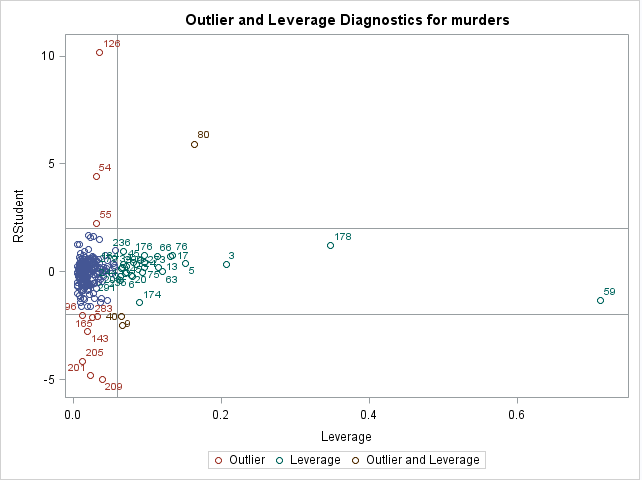
**Introduction**

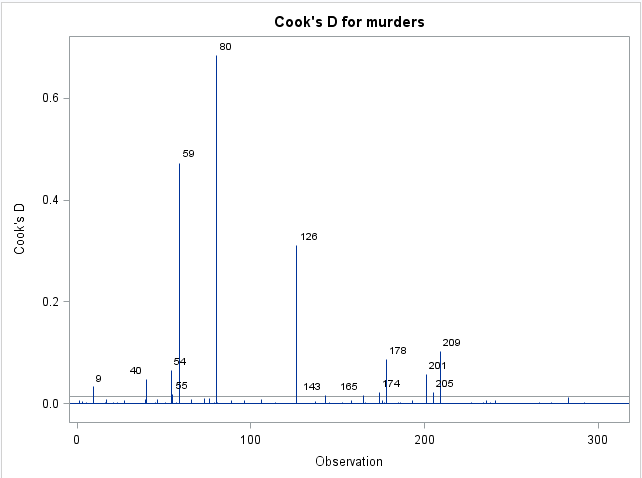
Our group decided to work on a crime data set from the University of California Irvine’s Machine Learning Department. It is the 1990 census data combined with 1995 FBI crime data for each city of interest. The initial data set we procured is over 2200 records but has several records with missing information for each groups of dataset. Our goal of this project is to conduct an analysis of over 120 explanatory variables to determine what factors affect the number of murders in a city using multiple linear regression.

With over a hundred plus explanatory variables for the census such as cultural background, population, level of education, number of houses vacant, median income and immigration population, we see these broken down by groups. We see crime data focusing mostly on law enforcement related information so some of the explanatory variables include number of full time law enforcement officers, number of police cars, law enforcement operating budget, land area of the city, cultural match to the city and the number of law enforcement officers on patrol.

**Descriptive Statistics**

We decided to work with the response variable of Murder and Murder per population. To begin our study we need to determine what factors have an effect on murder in a city. We went through a process of scrubbing through the data and making sure we looked at essential information and complete data for the desired variables we wanted to use. We made sure all the law enforcement data was present and all the explanatory variables had quantifiable value. After exploring different groups of dataset as explanatory variables with complete data we ended up with a dataset of only 302 records from the initial 2215.

We first look at the initial leverage and residual plot of the original (non-transformed) data to determine how much the leverage and outlying points can influence the data.



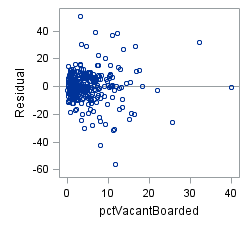
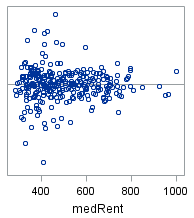
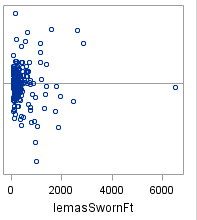
Based on the Outlier and Leverage Diagnostics we excluded three points from the model, observations 80 (Los Angeles, CA), 59 (New York, NY) and 126 (Washington D.C.) as high leverage points or an outlier. These points have significant influence on the linear regression model and will not accurately represent the majority of the data and need to be removed as we continue. (Usually, a reason is needed to exclude some record. You may be able to find the difference between the 3 cities and others).

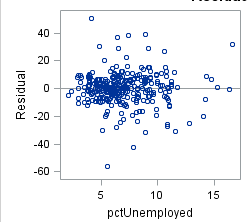
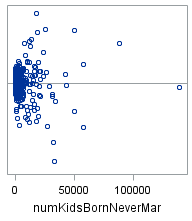
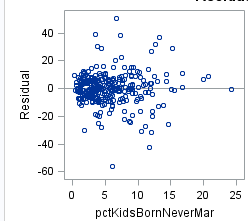
In examining the scatter plots of the data some variables look linear and others have significant variance but may be linear. Below is an example of the first six variables.

|  |
| --- |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

There are significant differences in the distribution of the variables. The data will have to be transformed in order to fit the assumptions of a linear regression model.





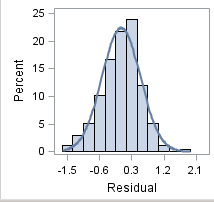
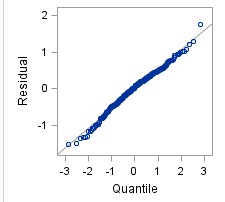
In examining the residual plots the residuals of the variables are not normally distributed. This is another indicator the data will need to be transformed in order to fit the linear regression assumptions.

In looking at the residual plots and the scatter plots the most appropriate options will be the Log() and Logit() transformations. These transformations will reduce the variation and reduce the chance of any outliers. The response variable will need to be transformed as well. The result is a Log() / Log() transformation in which the inference will be on the median vice the mean.

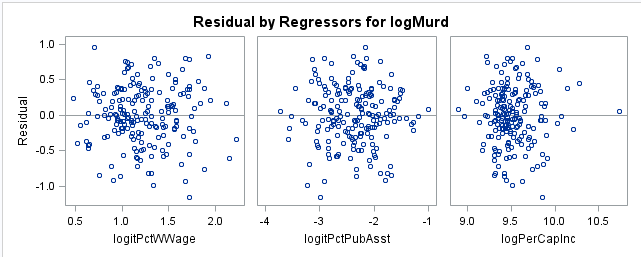
**Analysis**

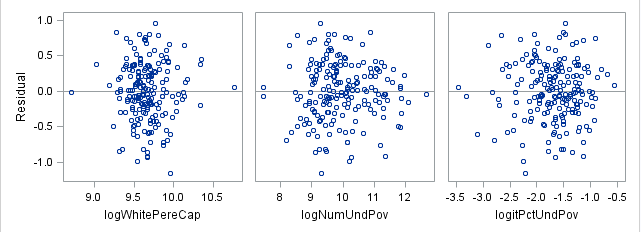
*Assumptions:*

* The residual are independent of each other: This is true as the residuals are independent of each other as indicated by the residual plots.
* The residuals are normally distributed: This is true after we perform the log transformation and we can see the residuals are normally distributed as indicated by the qq plot.

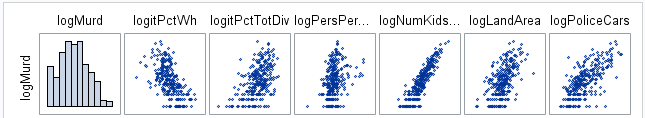


* The residuals have a constant variance: This is true after the log transformation, the variance is within acceptable margin to conclude that the variance is similar. The below are six of the thirty nine explanatory variables. All the plots are similar.





* Observations must be quantified: This is true as each observation is quantifiable by decimal
* Relationship of explanatory and response variables is linear: This is true based on the below scatter plots, the variables are linearly related to the response murder.



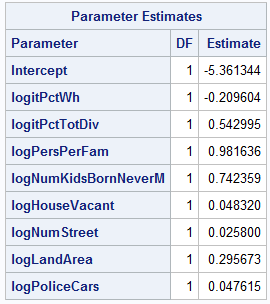
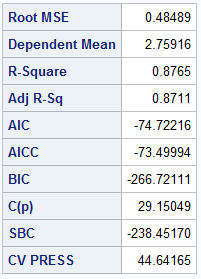
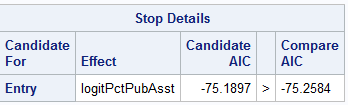
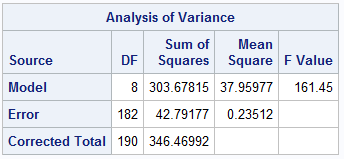
*Build a Model:*

To simplify the model and reduce the number variables analyzed from 125 to 39 we selected one variable from a group of similar variables to reduce correlation redundancy we take out highly correlated variables. For example, for cultural background as a percentage of the community population, we choose, 1 out of 6 to represent cultural percentage of population. Later further analysis can determine if that group of variables is significant.

Multiple collinearity tests were conducted to determine if the remaining 39 variables were measuring similar aspects of the model. Variables such as logPop, lonNumUndPov, logitPctUndPov, logPopDens, logPoliceOperBudg, logpersPerOccupHouse, logNumbUrban, and logPerCapInc were taken out. The remaining variables has a variation inflation factor (VIF) of 28 or less.

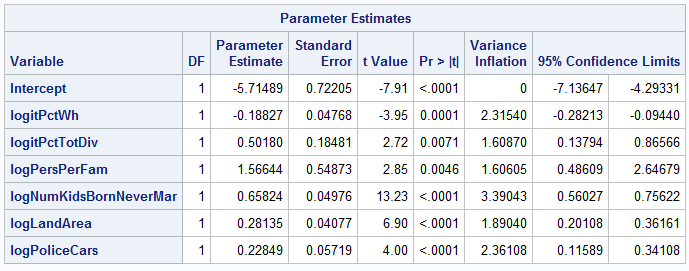
We conducted a variable selection procedure using the LASSO model with the cross validation splitting the data randomly into five groups (4 used to test the variables and 1 to validate) and the stop criteria is AIC. The process selected variables which affect the log of murders in a community (logMurd).

The result of the variable selection process is:

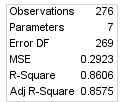
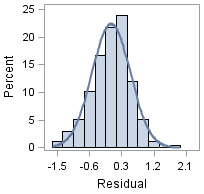
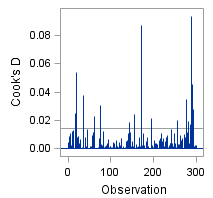
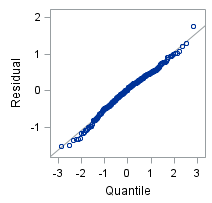
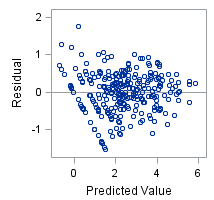
   

*Fit the model:*

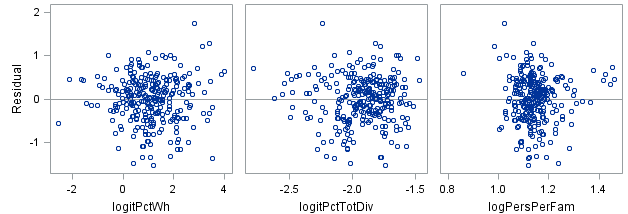
To make sure the variables above fit the model we ran a regression analysis. Per the regression analysis the variable logNumStreet is not significant (p-value = 0.1351). This variable is removed and the final model is as follows:

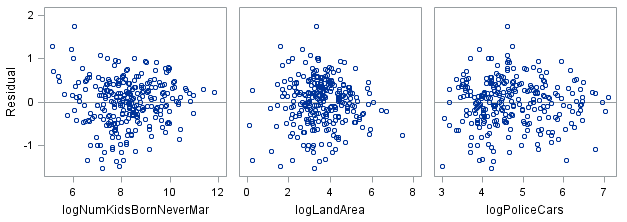


The above information shows the significant variables, the p-values and the 95% confidence intervals of those variables.



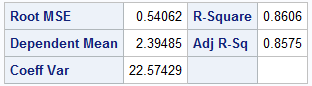
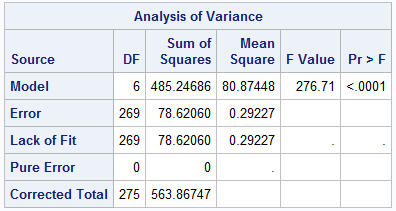
The residual analysis is not the best but no distinct pattern in the residuals is present. The QQ-Plot and the histogram show the residuals are normally distributed. There is no significant leverage points. The Adjusted R2 is 0.8575.





The above residuals do not show a distinct pattern but there is some clustering, especially in logPersPerFam. There does not seem to be a significant reason to reject meeting the assumptions of linear regression.

*Check the Model Fit (lack of fit test):*



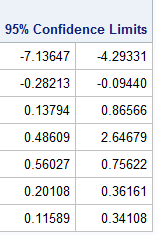
The model is a good fit (p-value < 0.0001) with an Adjusted R2 = 0.8575.

The final model equation is as follows:

|  |  |  |
| --- | --- | --- |
| Doubling of this variable | Multiplicative Change of the median of murders in a community | Decimal of multiplicative change |
| percentage of white people in the community | 2-0.1827 | 0.88105 |
| percentage of community that is divorced | 20.50180 | 1.41598 |
| mean number of people per family | 21.56644 | 2.96173 |
| number children born to parents who are not married | 20.65824 | 1.57816 |
| land area in square miles | 20.28135 | 1.21533 |
| number of police cars | 20.22849 | 1.17161 |

**Interpretation & Conclusion**

There is significant evidence to suggest, in this sampling of 302 records, the doubling of percentage of white people in a city, the percentage of divorce people in a city, the mean number of people per family, the number of children born to parents who are not married, the land area in square miles and the number of police cars produces multiplicative changes in the median of murders in a city. Below are the 95% confidence intervals of the log of the variables.



The interpretation of these results can be misleading. In our opinion we think further research needs to be conducted to determine what each of these explanatory variables represents in a city. Taken out of context this information can be distorted and manipulated to fit an agenda or narrative but that is not in the scientific purpose of this report.

Do white people make a significant difference in a city to reduce murders? In the 1990’s white people, as a demographic group, may hold more disposable income. There may be merit in researching the amount of disposable income in a city as it relates to crime.

The social factors of divorce percentage, number of people per family and children born to parents who are not married may indicate the lack of mentorship from a respected person (parent or community leader) and an unstable social-economic environment. Further research will need to be conducted to determine what socio-economic factors contribute to the number of murders in a city.

Land area in square miles may indicate number of people and population density of a community. The larger the land area the more people are in the city. As the number of the people increases the crime rate and number of murders may increase. There may be merit in research if larger populations have more murders and higher murder rates.

The number of police cars may indicate similar factors as land area. Police cars may also indicate the operating budget and the need for more police because the crime rate is high thus the number of murders may be high. More law enforcement resources may be needed to react to the crime of a city. Further research needs to be conducted to determine the law enforcement indicators that underlie why the number of police cars would be an indicator for the number of murders in a city.

**Appendix**

SAS Code Explanatory Analysis:

Below is an example of the explanatory analysis looking for outliers and high leverage points. This was done several times in order to determine what was taken out. Then each variable was plotted to check the assumptions.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Begin data exploration for murders

group 02

explanatory variables agePct16t24 - pctWWage;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **sgscatter** data = stats01.zstats02proj01;

matrix murders /\*population houseHoldSize racePctBlack racePctWhite racePctAsian racePctHisp\*/ agePct12t21 agePct12t29 agePct16t24 agePct65Up numbUrban pctUrban /\*medIncome pctWWage pctWFarmSelf pctWInvInc pctWSocSec pctWPubAsst pctWRetire medFamInc perCapInc whitePerCap blackPerCap indianPerCap asianPerCap otherPerCap hispPerCap numUnderPov pctPopUnderPov pctLess9thGrade pctNoHsGrad catPctHsGrad highPctHsGrad medpctHsGrad cnfPopHsGrad pctBSorMore pctUnemployed pctEmploy pctEmplManu pctEmplProfServ pctOccupManu pctOccupMgmtProf malePctDivorce malePctNevMarr femalePctDivorce totalPctDivorce persPerFam pctFam2Par pctKids2Par pctYoungKids2Par pctTeen2Par pctWorkMomYoungKids pctWorkMom numKidsBornNeverMar pctKidsBornNeverMar numImmig pctImmigRecent pctImmigRec5 pctImmigRec8 pctImmigRec10 pctRecentImmig pctRecimmig5 pctRecImmit8 pctRecImmig10 pctSpeakEnglOnly pctNotSpeakEngWell pctLargeHouseGam pctLargeHouseOccup persPerOccupHouse persPerOwnOccHouse persPerRentOccHouse pctPersOwnOccup pctPersDenseHouse pctHouseLess3Br mednumBr HouseVacant pctHouseOccup pctHouseOwnOccup pctVacantBoarded pctVacmore6Mos medYrHouseBuilt pctHouseNoPhone pctWoFullPlumb ownOccLowQuart ownOccMedVal ownOccHiQuart ownOccQRange rentLowQ rentMedian rentHighQ rentQRange medRent medRentPctHouseInc medOwnCostPctInc medOwnCostPctIncNoMtg numInShelters numStreet pctForeignBorn pctBornSameState pctSameHouse85 pctSameCity85 lemasSwornFt lemasSwFtPerPop lemasSwFtFieldOps lemasSwFtFieldPerPop lemasTotalReg lemasTotRegPerPop policeRegPerOff policePerPop recialmatchCommPol pctPoliceWhite pctPoliceBlack pctPoliceHisp pctPoliceAsian

pctPoliceMinor offAssgnDrugUnits numKidsDrugsSeiz policeAveOtWorked landArea popDens pctUsePubTrans policeCars policeOperBudg lemasPctPoliceOnPatr lemasGangUnitDEploy lemasPctOffDrunUn policeBudgePerPop\*/ / diagonal = (histogram);

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

looking for significant outliers & leverage

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **univariate** data = stats01.zstats02proj01;

var numbUrban pctUrban;

histogram;

**run**;

**proc** **reg** data = stats01.zstats02proj01 corr plots(label) = (rstudentleverage cooksd);

model murders = agePct16t24 agePct65Up numbUrban pctUrban medIncome pctWWage;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of exploratory analysis

group 02

significant outliers of the population are:

obs 80 -> Los Angeles City; outlier and leverage

obs 59 -> New York City; leverage

obs 126 -> Washington DC; significant outlier

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

SAS Code Variable Transformation:

Most of the variables, only two did not need to be transformed, used a log() or logit() transformation. This transformation meet the assumptions of the linear regression assumtions.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This is the data that was transformed;

There were only several variables that were not transformed;

Most were transformed using a log() or logit() transformation;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** stats01.zstats02proj01\_otlv\_trans;

set stats01.zstats02proj01\_otlv;

/\* response variable \*/

logMurder = log(murders);

/\* explanatory variables \*/

logPop = log(population);

logHouseHoldSize = log(houseHoldSize);

logitPctBl = log(racePctBlack / (**100** - racePctBlack));

logitPctWh = log(racePctWhite / (**100** - racePctWhite));

logitPctAs = log(racePctAsian / (**100** - racePctAsian));

logitPctHsp = log(racePctHisp / (**100** - racePctHisp));

logitPct12t21 = log(agePct12t21 / (**100** - agePct12t21));

logitPct12t29 = log(agePct12t29 / (**100** - agePct12t29));

logitPct16t24 = log(agePct16t24 / (**100** - agePct16t24));

logitPct65Up = log(agePct65Up / (**100** - agePct65Up));

lognumbUrban = log(numbUrban);

logPctUrb = log(pctUrban);

logMedInc = log(medIncome);

logitPctWWage = log(pctWWage / (**100** - pctWWage));

logitPctWFarmSelf = log(pctWFarmSelf / (**100** - pctWFarmSelf));

logitPctWInvInc = log(pctWInvInc / (**100** - pctWInvInc));

logitPctWSocSec = log(pctWSocSec / (**100** - pctWSocSec));

logitPctPubAsst = log(pctWPubAsst / (**100** - pctWPubAsst));

logitPctWRetire = log(pctWRetire / (**100** - pctWRetire));

logMedFamInc = log(medFamInc);

logPerCapInc = log(perCapInc);

logWhitePereCap = log(whitePerCap);

logBlackPerCap = log(blackPerCap);

logIndPerCap = log(indianPerCap);

logMurd = log(murders);

logAsianPerCap = log(asianPerCap);

logOtherPerCap = log(otherPerCap);

logHispPerCap = log(hispPerCap);

logNumUndPov = log(numUnderPov);

logitPctUndPov = log(pctPopUnderPov / (**100** - pctPopUnderPov));

logitPctLess9Grade = log(pctLess9thGrade / (**100** - pctLess9thGrade));

logitNoHsGrad = log(pctNoHsGrad / (**100** - pctNoHsGrad));

logCnfPopHsGrad = log(cnfPopHsGrad);

logitPctBSorMore = log(pctBSorMore / (**100** - pctBSorMore));

logitPctUnempl = log(pctUnemployed / (**100** - pctUnemployed));

logitPctEmpl = log(pctEmploy / (**100** - pctEmploy));

logitPctEmplMan = log(pctEmplManu / (**100** - pctEmplManu));

logitPctEmplProfServ = log(pctEmplProfServ / (**100** - pctEmplProfServ));

logitPctEmplOccupMan = log(pctOccupManu / (**100** - pctOccupManu));

logitPctEmplOccupProf = log(pctOccupMgmtProf / (**100** - pctOccupMgmtProf));

logitPctMaleDiv = log(malePctDivorce / (**100** - malePctDivorce));

logitPctMaleNevMarr = log(malePctNevMarr / (**100** - malePctNevMarr));

logitPctFemaleDiv = log(femalePctDivorce / (**100** - femalePctDivorce));

logitPctTotDiv = log(totalPctDivorce / (**100** - totalPctDivorce));

logPersPerFam = log(persPerFam);

logitPctFam2Par = log(pctFam2Par / (**100** - pctFam2Par));

logitPctKids2Par = log(pctKids2Par / (**100** - pctKids2Par));

logitPctYoungKids2Par = log(pctYoungKids2Par / (**100** - pctYoungKids2Par));

logitPctTeen2Par = log(pctTeen2Par / (**100** - pctTeen2Par));

logitPctWorkMomYoungKids = log(pctWorkMomYoungKids / (**100** - pctWorkMomYoungKids));

logitPctWorkMom = log(pctWorkMom / (**100** - pctWorkMom));

logNumKidsBornNeverMar = log(numKidsBornNeverMar);

logitPctKidsBornNeverMar = log(pctKidsBornNeverMar / (**100** - pctKidsBornNeverMar));

logNumImmig = log(numImmig);

logPctImmigRecent = log(pctImmigRecent / (**100** - pctImmigRecent));

logPctImmigRec5 = log(pctImmigRec5 / (**100** - pctImmigRec5));

logPctImmigRec8 = log(pctImmigRec8 / (**100** - pctImmigRec8));

logPctImmigRec10 = log(pctImmigRec10 / (**100** - pctImmigRec10));

logitPctRecentImmig = log(pctRecentImmig / (**100** - pctRecentImmig));

logitPctRecimmig5 = log(pctRecimmig5 / (**100** - pctRecimmig5));

logitPctRecImmit8 = log(pctRecImmit8 / (**100** - pctRecImmit8));

logitPctRecImmig10 = log(pctRecImmig10 / (**100** - pctRecImmig10));

logitPctSpeakEnglOnly = log(pctSpeakEnglOnly / (**100** - pctSpeakEnglOnly));

logitPctNotSpeakEngWell = log(pctNotSpeakEngWell / (**100** - pctNotSpeakEngWell));

logitPctLargeHouseGam = log(pctLargeHouseGam / (**100** - pctLargeHouseGam));

logitPctLargeHouseOccup = log(pctLargeHouseOccup / (**100** - pctLargeHouseOccup));

logPersPerOccupHouse = log(persPerOccupHouse);

logPersPerOwnOccHouse = log(persPerOwnOccHouse);

logPersPerRentOccHouse = log(persPerRentOccHouse);

logitPctPersOwnOccup = log(pctPersOwnOccup / (**100** - pctPersOwnOccup));

logitPctPersDenseHouse = log(pctPersDenseHouse / (**100** - pctPersDenseHouse));

logitPctHouseLess3Br = log(pctHouseLess3Br / (**100** - pctHouseLess3Br));

logHouseVacant = log(HouseVacant);

logitPctHouseOccup = log(pctHouseOccup / (**100** - pctHouseOccup));

logitPctHouseOwnOccup = log(pctHouseOwnOccup / (**100** - pctHouseOwnOccup));

logitPctVacantBoarded = log(pctVacantBoarded / (**100** - pctVacantBoarded));

logitPctVacmore6Mos = log(pctVacmore6Mos / (**100** - pctVacmore6Mos));

logMedYrHouseBuilt = log(medYrHouseBuilt);

logitPctHouseNoPhone = log(pctHouseNoPhone / (**100** - pctHouseNoPhone));

logitPctWoFullPlumb = log(pctWoFullPlumb / (**100** - pctWoFullPlumb));

logOwnOccLowQuart = log(ownOccLowQuart);

logOwnOccMedVal = log(ownOccMedVal);

logOwnOccHiQuart = log(ownOccHiQuart);

logOwnOccQRange = log(ownOccQRange);

logRentLowQ = log(rentLowQ);

logRentMedian = log(rentMedian);

logRentHighQ = log(rentHighQ);

logRentQRange = log(rentQRange);

logMedRent = log(medRent);

logitMedRentPctHouseInc = log(medRentPctHouseInc / (**100** - medRentPctHouseInc));

logitMedOwnCostPctInc = log(medOwnCostPctInc / (**100** - medOwnCostPctInc));

logitMedOwnCostPctIncNoMtg = log(medOwnCostPctIncNoMtg / (**100** - medOwnCostPctIncNoMtg));

logNumInShelters = log(numInShelters);

logNumStreet = log(numStreet);

logitPctForeignBorn = log(pctForeignBorn / (**100** - pctForeignBorn));

logitPctBornSameState = log(pctBornSameState / (**100** - pctBornSameState));

logitPctSameHouse85 = log(pctSameHouse85 / (**100** - pctSameHouse85));

logitPctSameCity85 = log(pctSameCity85 / (**100** - pctSameCity85));

logLemasSwornFt = log(lemasSwornFt);

logLemasSwFtPerPop = log(lemasSwFtPerPop);

logLemasSwFtFieldOps = log(lemasSwFtFieldOps);

logLemasSwFtFieldPerPop = log(lemasSwFtFieldPerPop);

logLemasTotalReg = log(lemasTotalReg);

logLemasSwFtPerPop = log(lemasSwFtPerPop);

logPoliceRegPerOff = log(policeRegPerOff);

logPolicePerPop = log(policePerPop);

logitRecialmatchCommPol = log(recialmatchCommPol / (**100** - recialmatchCommPol));

logitPctPoliceWhite = log(pctPoliceWhite / (**100** - pctPoliceWhite));

logitPctPoliceBlack = log(pctPoliceBlack / (**100** - pctPoliceBlack));

logitPctPoliceHisp = log(pctPoliceHisp / (**100** - pctPoliceHisp));

logitPctPoliceAsian = log(pctPoliceAsian / (**100** - pctPoliceAsian));

logitPctPoliceMinor = log(pctPoliceMinor / (**100** - pctPoliceMinor));

logOffAssgnDrugUnits = log(offAssgnDrugUnits);

logNumKidsDrugsSeiz = log(numKidsDrugsSeiz);

logPoliceAveOtWorked = log(policeAveOtWorked);

logLandArea = log(landArea);

logPopDens = log(popDens);

logitPctUsePubTrans = log(pctUsePubTrans / (**100** - pctUsePubTrans));

logPoliceCars = log(policeCars);

logPoliceOperBudg = log(policeOperBudg);

logitLemasPctPoliceOnPatr = log(lemasPctPoliceOnPatr / (**100** - lemasPctPoliceOnPatr));

logitLemasPctOffDrunUn = log(lemasPctOffDrunUn / (**100** - lemasPctOffDrunUn));

logPoliceBudgePerPop = log(policeBudgePerPop);

**run**;

SAS Code Variable Selection:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This is the analysis portion of the test;

- Model name: TransAllGlmselectAfterColinCheck

- linear regression model (included the transformed variables and the few that are not transformed

- variable selection

- cross validation

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **glmselect** data = stats01.zstats02proj01\_otlv\_trans;

title "Stats 02 Project 01; GLMSELECT; Model: TransAllGlmselectAfterColinCheck";

/\* model \*/

model logMurd = /\*logPop\*/ logHouseHoldSize /\*logitPctBl\*/ logitPctWh /\*logitPctAs logitPctHsp logitPct12t21 logitPct12t29\*/ logitPct16t24 /\*logitPct65Up lognumbUrban logPctUrb\*/ logMedInc logitPctWWage /\*logitPctWFarmSelf logitPctWInvInc logitPctWSocSec\*/ logitPctPubAsst /\*logitPctWRetire logMedFamInc logPerCapInc\*/ logWhitePereCap /\*logBlackPerCap logIndPerCap logAsianPerCap logOtherPerCap logHispPerCap logNumUndPov logitPctUndPov logitPctLess9Grade\*/ logitNoHsGrad /\*logCnfPopHsGrad\*/ logitPctBSorMore /\*logitPctUnempl\*/ /\*logitPctEmpl logitPctEmplMan logitPctEmplProfServ logitPctEmplOccupMan logitPctEmplOccupProf logitPctMaleDiv\*/ logitPctMaleNevMarr /\*logitPctFemaleDiv\*/ logitPctTotDiv logPersPerFam logitPctFam2Par

/\*logitPctKids2Par logitPctYoungKids2Par logitPctTeen2Par logitPctWorkMomYoungKids logitPctWorkMom\*/ logNumKidsBornNeverMar /\*logitPctKidsBornNeverMar logNumImmig logPctImmigRecent logPctImmigRec5

logPctImmigRec8 logPctImmigRec10 logitPctRecentImmig logitPctRecimmig5 logitPctRecImmit8 logitPctRecImmig10\*/ logitPctSpeakEnglOnly /\*logitPctNotSpeakEngWell logitPctLargeHouseGam logitPctLargeHouseOccup

logPersPerOccupHouse logPersPerOwnOccHouse logPersPerRentOccHouse logitPctPersOwnOccup logitPctPersDenseHouse logitPctHouseLess3Br mednumBr\*/ logHouseVacant /\*logitPctHouseOccup logitPctHouseOwnOccup\*/ logitPctVacantBoarded logitPctVacmore6Mos /\*logMedYrHouseBuilt logitPctHouseNoPhone logitPctWoFullPlumb\*/ logOwnOccLowQuart /\*logOwnOccMedVal logOwnOccHiQuart logOwnOccQRange logRentLowQ\*/ logRentMedian /\*logRentHighQ logRentQRange logMedRent logitMedRentPctHouseInc logitMedOwnCostPctInc logitMedOwnCostPctIncNoMtg\*/ logNumInShelters logNumStreet /\*logitPctForeignBorn logitPctBornSameState logitPctSameHouse85 logitPctSameCity85\*/ logLemasSwornFt /\*logLemasSwFtPerPop logLemasSwFtFieldOps logLemasSwFtFieldPerPop logLemasTotalReg logPoliceRegPerOff logPolicePerPop\*/ logitRecialmatchCommPol

/\*logitPctPoliceWhite logitPctPoliceBlack logitPctPoliceHisp logitPctPoliceAsian logitPctPoliceMinor logOffAssgnDrugUnits\*/ logNumKidsDrugsSeiz logPoliceAveOtWorked logLandArea /\*logPopDens\*/ logitPctUsePubTrans logPoliceCars /\*logPoliceOperBudg logitLemasPctPoliceOnPatr lemasGangUnitDEploy\*/ logitLemasPctOffDrunUn

/\*logPoliceBudgePerPop\*/ /

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

OPTIONS

- variable selection criteria = LASSO

- variable stop criteria = AIC

- break up data = cross validation, method random in 5 parts

- additional statistics:

- adjusted R^2 -> linear correlation

- cp -> Mallows C(p)

- bic -> sawa bayesian information criterion

- sbc -> schwarz bayesian information criterion

- sl -> significance level of the F-Stat for entering and departing effects

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

selection = lasso (choose = cv stop = aic) cvmethod = random(**5**) stats = (adjrsq cp bic sbc sl);

**run**;

title ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of the model TransAllGlmselectAfterColinCheck:

- seven variables show as selected to the AIC criterial using cross validation

- logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logNumStreet logLandArea logPoliceCars

- the next step is to run the model in proc reg to look at the vif's to check the collinearity of the remaining variables

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

SAS Code Model Validation:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Collinearity check of model TransAllGlmselectAfterColinCheck:

- will use proc reg to check the model fit and the vif's of the seven variables

- variables: logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logNumStreet logLandArea logPoliceCars

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **reg** data = stats01.zstats02proj01\_otlv\_trans;

title "Stats 02 Project 01; REG; Model: TransAllGlmselectAfterColinCheck";

model logMurd = logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logNumStreet logLandArea logPoliceCars /

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

OPTIONS

- lack of fit test

- variance inflation factors to check for collinearity different variables are measuring the same thing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

lackfit vif ;

**run**;

title ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of collinearity check of model TransAllGlmselectAfterColinCheck:

Lack of Fit Test:

- pass; looks like a good fit

Variable (VIF) P-value

- logitPctWh 2.37945 < 0.0001

- logitPctTotDiv 1.59913 0.002

- logPersPerFam 1.66595 0.0022

- logNumKidsBornNeverMar 3.31191 < 0.0001

- logNumStreet 1.35758 0.1351

- logLandArea 2.04826 < 0.0001

- logPoliceCars 2.86895 0.0375

Stats:

- the adjusted R^2 = 0.8658

Interesting findings:

- logNumStreet does not appear significant in the model

- re-run proc reg without logNumStreet in model

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Collinearity check of model TransAllGlmselectAfterColinCheck:

- will use proc reg to check the model fit and the vif's of the six variables

- took out logNumStreet because the analysis showed it was significantly greater than p-value of 0.05

- variables: logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logLandArea logPoliceCars

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **reg** data = stats01.zstats02proj01\_otlv\_trans;

title "Stats 02 Project 01; REG; Model: TransAllGlmselectAfterColinCheck";

model logMurd = logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logLandArea logPoliceCars /

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

OPTIONS

- lack of fit test

- variance inflation factors to check for collinearity different variables are measuring the same thing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

lackfit vif ;

**run**;

title ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of collinearity check of model TransAllGlmselectAfterColinCheck:

Lack of Fit Test:

- pass; looks like a good fit

Variable (VIF) P-value

- logitPctWh 2.31540 0.0001

- logitPctTotDiv 1.60870 0.0071

- logPersPerFam 1.60605 0.0046

- logNumKidsBornNeverMar 3.39043 < 0.0001

- logLandArea 1.89040 < 0.0001

- logPoliceCars 2.36108 < 0.0001

Stats:

- the adjusted R^2 = 0.8575

Interesting findings:

- analysis complete

- the next step is the write-up and conclusions

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/