Variable Selection:

Forward Selection / Stepwise SAS Macro

**Purpose**:

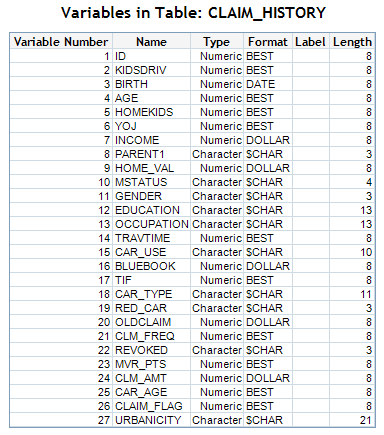
The macro performs an automated forward selection and stepwise variable selection process for PROC GENMOD which does not come with model selection options. Note that the GENMOD procedure in SAS versions prior to 9.4 does not come with model selection options.

**Introduction**:

SAS users of SAS 9.2 and prior versions may face situations where some "powerful" options are only available in certain SAS procedures but not available in others. For example, the model selection options are available in PROC REG, LOGISTIC, PHREG, etc., but not in PROC GENMOD, CATMOD, MIXED, etc. This forward/stepwise selection macro could be used with the procedures GENNMOD, CATMOD, MIXED, GLIMMIX, etc.

**Illustration:**

* Dataset used in this code: **claim\_history.sas7bdat** or **claim\_history.csv**
* The claim\_history.sas7bdat dataset comes from the help library of SAS Enterprise-Miner version 7.1
* Data description: **observations 10302**, **Variables 27**



Interesting target varaibles could be:

**CLAIM\_FLAG** : binary variable 1 if claim occurred and 0 otherwise

**CLM\_AMT**: claim amount including 0

**CLM\_FRQ**: number of claims takes the values 0, 1, 2, etc

**EXPO**: the exposition, I added this variable, takes the value 1 always

* For a **Binomial** (logistic model) the target should be CLAIM\_FLAG
* For a Severity model using **Gamma**, **Inverse Gaussian** or **Log Normal** the target should be CLM\_AMT without the cero values
* For a Frequency model using **Poisson** or **Negative Binomial** and the **Zero Inflated Poisson** and **Zero Inflated** **Negative Binomial**, the target should be CLM\_FRQ
* For the Cost of Claim model (or risk premium) it is possible to use the CLM\_AMT including the cero values

In this example I perform a stepwise variable selection using the variable CLM\_AMT (with ceros) as the target for an Tweedie model.

The next lines contain the two SAS macros for the forward selection/stepwise selection process using a Tweedie error function.

\*-------------------------------------------------------------------------\*;

\*STEPWISE SELECTON MACRO FOR PROC GENMOD;

\*-------------------------------------------------------------------------\*;

\*

The first part of the macro is the construction of all single factor

models and saving their results in a dataset (aggregated), where those

with p-values greater than 0.2 could be easily filtered out

;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* MACRO PARAMETERS \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*Put the name of the modeling dataset:;

%let dataset = GRIDWORK.CLAIM\_HISTORY;

\*Put here the name of the name of the target varible;

%let target = CLM\_AMT;

\*Put here the name of all the potential explanatory variables, the

categorical and the coninous:;

%let explanatory =

GENDER

CAR\_USE

EDUCATION

CAR\_TYPE

MSTATUS

OCCUPATION

RED\_CAR

URBANICITY

REVOKED

PARENT1

AGE

BIRTH

BLUEBOOK

CAR\_AGE

HOMEKIDS

HOME\_VAL

INCOME

KIDSDRIV

MVR\_PTS

TIF

TRAVTIME

YOJ

;

\*Put the name of the character variables with commas:;

%let char=

"GENDER" ,

"CAR\_USE" ,

"EDUCATION" ,

"CAR\_TYPE" ,

"MSTATUS" ,

"OCCUPATION" ,

"RED\_CAR" ,

"URBANICITY" ,

"REVOKED" ,

"PARENT1"

;

\*Put the name of the character variables with commas:;

%let char\_2=

GENDER ,

CAR\_USE ,

EDUCATION ,

CAR\_TYPE ,

MSTATUS ,

OCCUPATION ,

RED\_CAR ,

URBANICITY ,

REVOKED ,

PARENT1

;

\*Put the name of the continous variables. Include the target variable also:;

%let interval =

CLM\_AMT

AGE

BIRTH

BLUEBOOK

CAR\_AGE

HOMEKIDS

HOME\_VAL

INCOME

KIDSDRIV

MVR\_PTS

TIF

TRAVTIME

YOJ

;

\*In order to minimize the number of models that needed to be run, a stepwise

selection model was created, considering susceptible for entry all those

variables with a p-value in a single model less than 0.2, and with a p-value

in the aggregated model of less than 0.25;

%let single\_model=0.2;

%let aggregated\_model=0.05;

\*Set the power Tweedie distribution parameter;

%let power = 1.5;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* END MACRO PARAMETERS \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

/\* Creation of the Aggregated Data Set\*/

**data** explanatory;

set /\*tmp1\*/

&dataset

(keep= &explanatory);

**run**;

**proc** **contents** data = explanatory varnum nodetails noprint

out=explanatory\_names (keep=name);

**run**;

**data** explanatory\_names;

set explanatory\_names;

j = \_n\_;

**run**;

\* Determine the number of observations;

**data** \_NULL\_;

if **0** then

set explanatory\_names nobs=n;

call symputx('nrows',n);

stop;

**run**;

**proc** **datasets** lib=RMTWORK nolist;

delete pva;

**run**;

\*http://support.sas.com/kb/35/591.html;

\*http://www.amadeus.co.uk/sas-training/tips/5/1/64/

using-the-in-operator-with-the-macro-language-in-sas-92.php;

ods select none;

\*https://support.sas.com/documentation/cdl/en/statug/63347/

HTML/default/viewer.htm#statug\_genmod\_sect010.htm;

options minoperator mlogic mprint;

**%macro** ***agrega*** / mindelimiter=',';

%do obs = **1** %to &nrows;

data \_null\_;

set explanatory\_names;

if j = &obs then

call symputx("var", put(name, **30.**));

run;

%if %upcase(&var) in(&char\_2) %then

%do;

ods output Type3=pva&obs (rename=source=parm keep=source ProbChiSq);

proc genmod data=&dataset NAMELEN=**50**;

if \_resp\_ > **0** then

d = **2**\*(\_resp\_\*(\_resp\_\*\*(**1**-&power)-\_mean\_\*\*(**1**-&power))/

(**1**-&power)-(\_resp\_\*\*(**2**-&power)-\_mean\_\*\*(**2**-&power))/

(**2**-&power));

else d = **2**\* \_mean\_\*\*(**2**-&power)/(**2**-&power);

variance var = \_mean\_\*\*&power;

deviance dev = d;

class &var;

model &target = &var / link=log type3 scale=pearson;

\*scwgt expos;

title "&var";

run;

%end;

%else

%do;

ods output Type3=pva&obs (rename=source=parm keep=source ProbChiSq);

proc genmod data=&dataset NAMELEN=**50**;

if \_resp\_ > **0** then

d = **2**\*(\_resp\_\*(\_resp\_\*\*(**1**-&power)-\_mean\_\*\*(**1**-&power))/

(**1**-&power)-(\_resp\_\*\*(**2**-&power)-\_mean\_\*\*(**2**-&power))/

(**2**-&power));

else d = **2**\* \_mean\_\*\*(**2**-&power)/(**2**-&power);

variance var = \_mean\_\*\*&power;

deviance dev = d;

model &target = &var / link=log type3 scale=pearson;

\*scwgt expos;

title "&var";

run;

%end;

data pva&obs;

length parm $**30**;

set pva&obs;

run;

proc append base=pva data=pva&obs force;

run;

%end;

**%mend** agrega;

%***agrega***;

ods select all;

**proc** **sort** data=pva out=aggregate;

by ProbChiSq;

**run**;

**data** elegibles;

set aggregate;

where probchisq le &single\_model;

**run**;

**proc** **sort** data = elegibles;

by probchisq;

**run**;

**data** elegibles;

set elegibles;

rename parm = source;

**run**;

title "Aggregate (all potential explanatory variables)";

**proc** **print** data=aggregate;

**run**;

title "Elegibles (individual model threshold = &single\_model)";

**proc** **print** data=elegibles;

**run**;

/\* STEPWISE PROCESS \*/

%let i = 1;

**proc** **sql** noprint;

select count(source) into: totalfactor from elegibles;

**quit**;

\* Also, the necessary set of macro variables needed to be initialized

as null, as the variables had to exist in order to be used in the rest

of the code;

%let catvar=;

%let invars=;

%let tempvar=;

%let tempcat=;

\*options NOSYNTAXCHECK;/\*Use this oiption before the last macro called

ONE only if you are sure that the failure of the macro is due to a lack

of convergence problem\*/

options mlogic;

**%macro** ***one***;

%do %while (&i le &totalfactor);

data elegibles;

set elegibles;

if \_n\_ = **1** then

do;

call symput('testvar',source);

if upcase(source) in (&char) then

do;

call symput('catvar\_t',source);

end;

else

do;

call symput('catvar\_t','');

end;

delete;

end;

run;

title "Step &i";

title2 "Elegibles (aggregated model threshold = &aggregated\_model)";

ods output parameterestimates = parms type3 = type3;

proc genmod data=&dataset NAMELEN=**50**;

if \_resp\_ > **0** then

d = **2**\*(\_resp\_\*(\_resp\_\*\*(**1**-&power)-\_mean\_\*\*(**1**-&power))/

(**1**-&power)-(\_resp\_\*\*(**2**-&power)-\_mean\_\*\*(**2**-&power))/

(**2**-&power));

else d = **2**\* \_mean\_\*\*(**2**-&power)/(**2**-&power);

variance var = \_mean\_\*\*&power;

deviance dev = d;

class &catvar &catvar\_t;

model &target = &invars &testvar / link=log type3 scale=pearson;

\*scwgt expos;

run;

data stay;

set type3;

If source = "&testvar" and probchisq le &aggregated\_model then

do;

call symput ('invars',"&tempvar"||" "||strip("&testvar"));

call symput ('catvar',"&tempcat"||" "||strip("&catvar\_t"));

end;

else if source = "&testvar" then

delete;

run;

%let tempvar = &invars;

%let tempcat = &catvar;

%let i = %eval(&i+1);

%end;

**%mend** one;

%***one***;

title "Selected Variables";

**proc** **print** data=stay;

**run**;

**proc** **datasets** lib=RMTWORK KILL nolist;

**run**;

/\*

After the process was finished, the resulting model would need to be run one

more time, as two problems could arise.

First, the last factor considered could be not significant, and so, the model

would need to be rerun to have the pvalues

without that variable.

Also, if the last model did not converge, the model would also need to be

rerun to have the values of the next to last model.

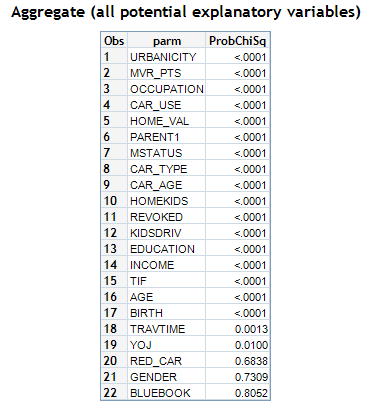
\*/

In order to minimize the number of models that needed to be run, a stepwise selection model was created, considering susceptible for entry all those variables with a p-value in a single model less than 0.2, and with a p-value in the aggregated model of less.

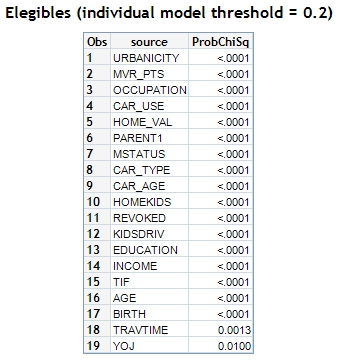
The execution of the above two macros create three outputs:

* A summary table called **Aggregate** with the all potential variables in a single GLM sorted by its p-values
* A summary table named **Eligible** with those selected variables with a threshold p-value less or equal to 0.2 in the single GLM. Sorted by p-value.
* A summary table with the **selected variables**
* The whole variable selection process step by step

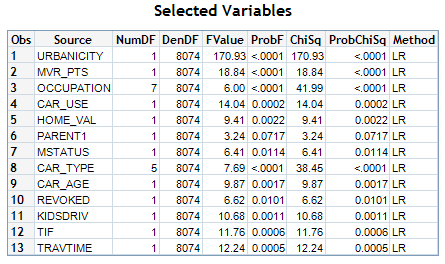
Aggregate table:



Eligible table:

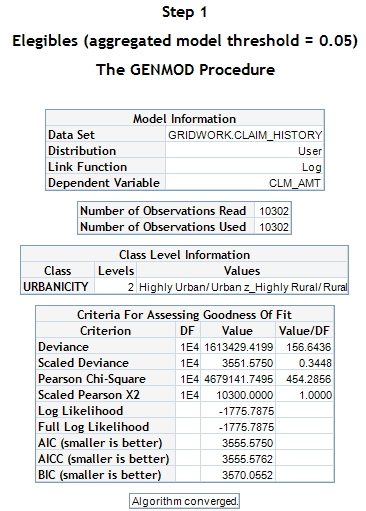


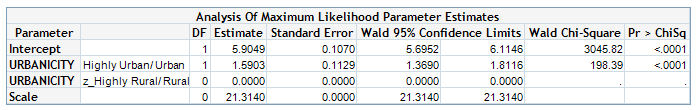
The summary table of the model selection process:



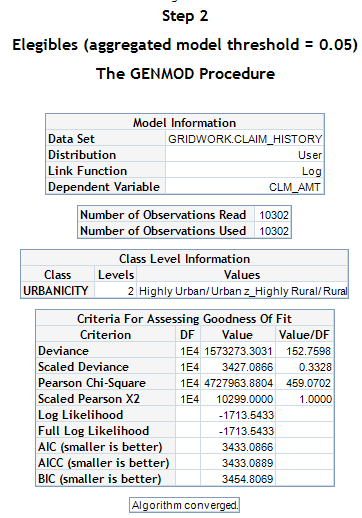
The table shows that only 11 of the 22 original potential variables have been selected

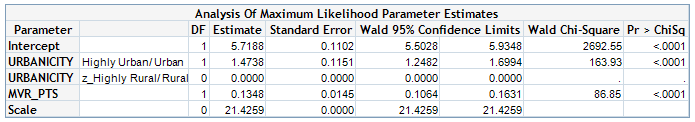
The whole variable selection process step by step:

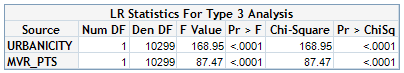


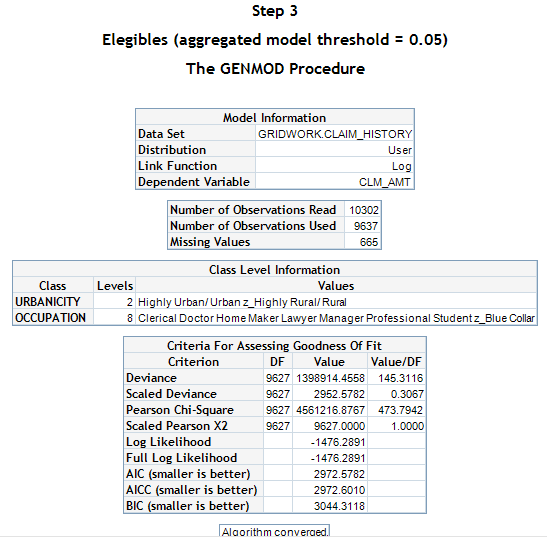


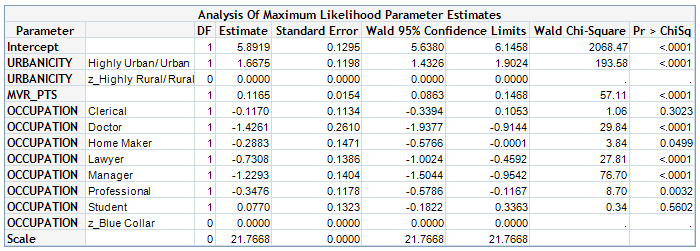


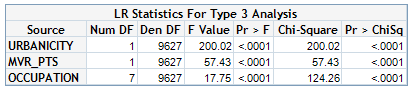


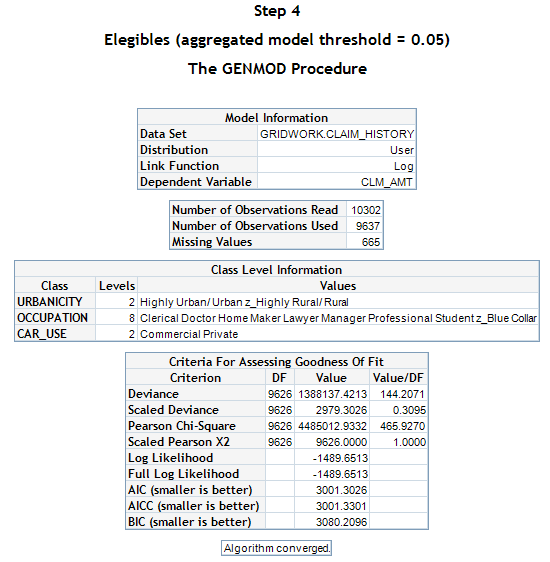


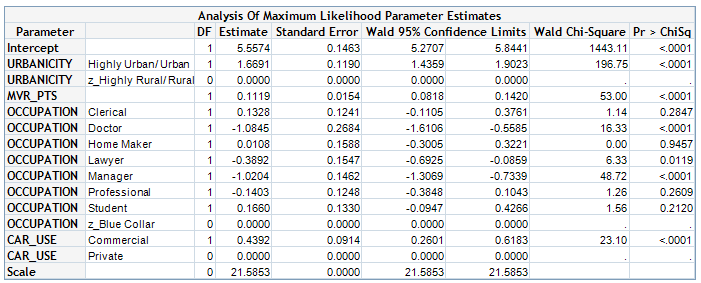


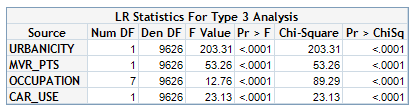


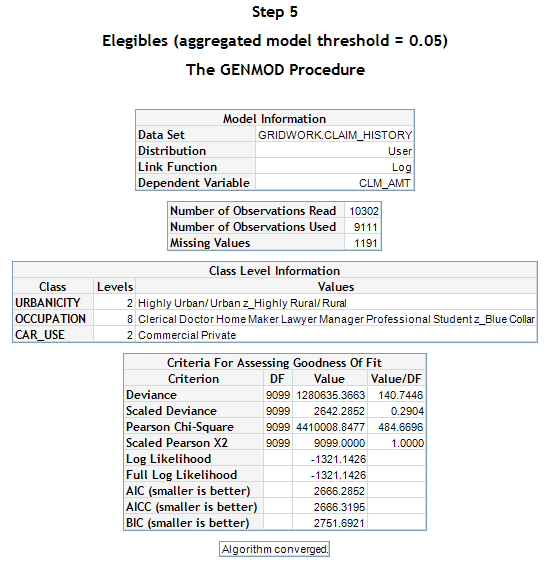


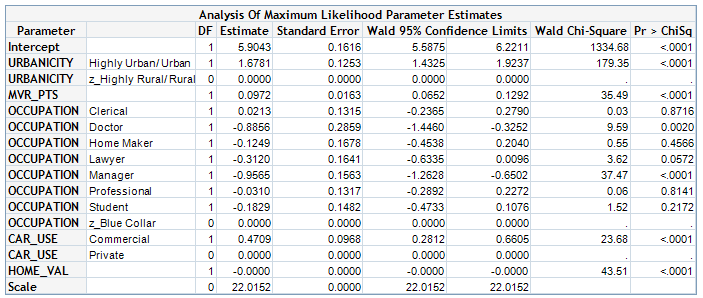


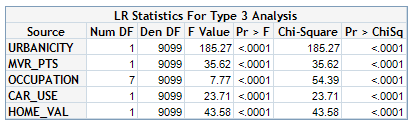


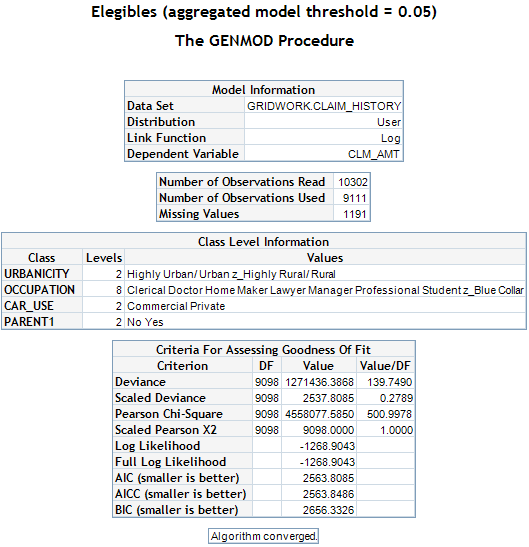


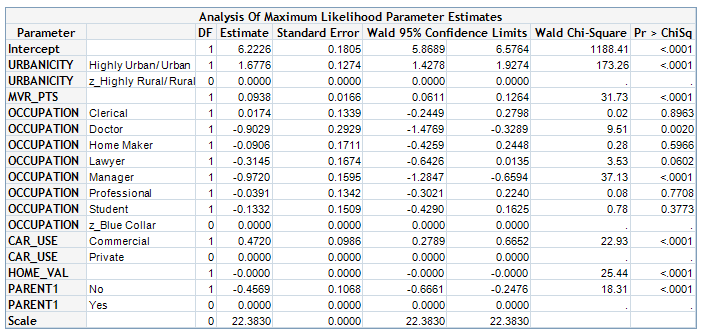


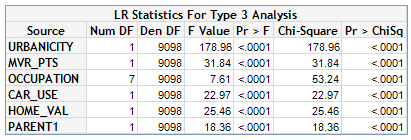


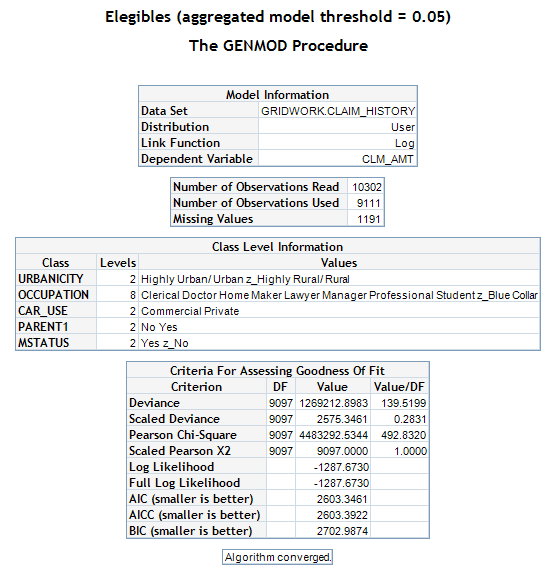


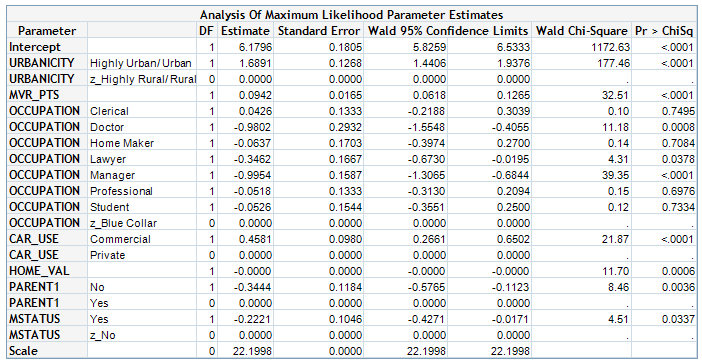


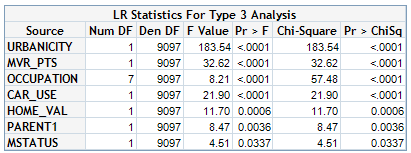


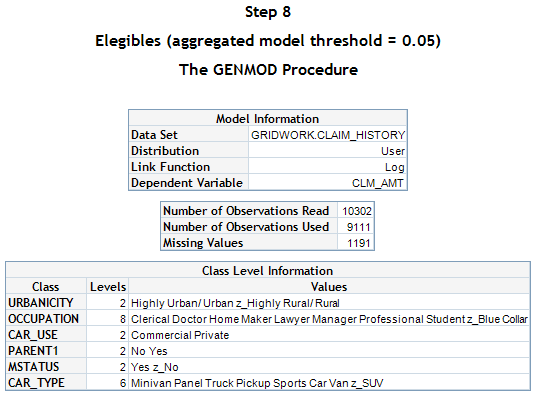


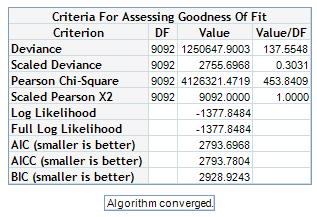


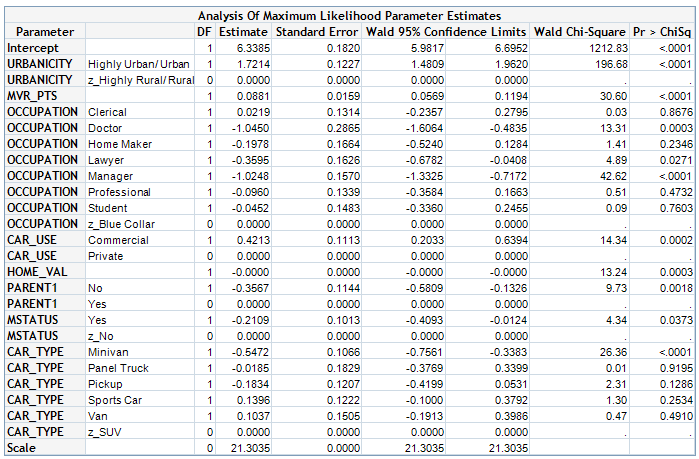


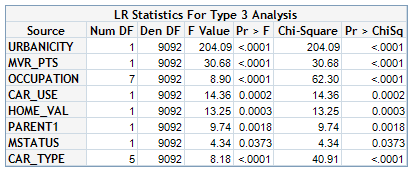


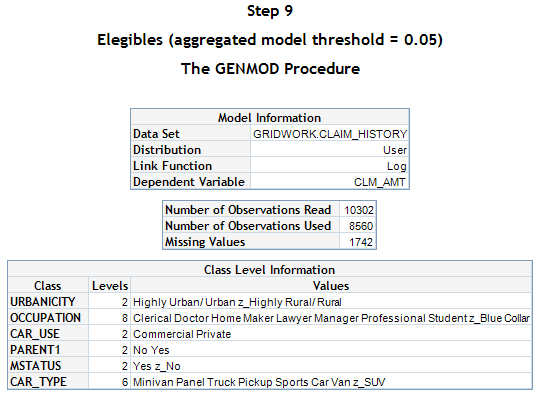


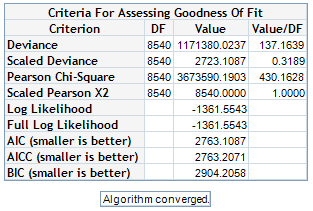


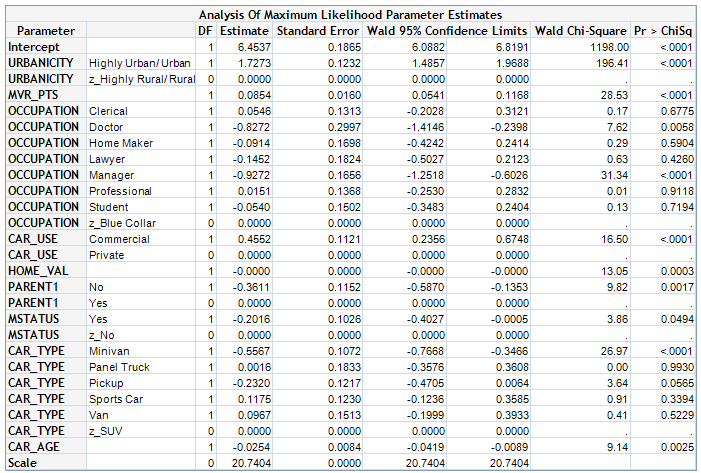


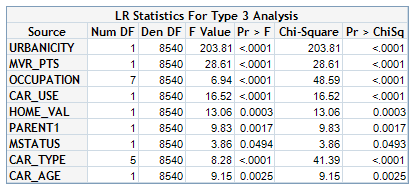


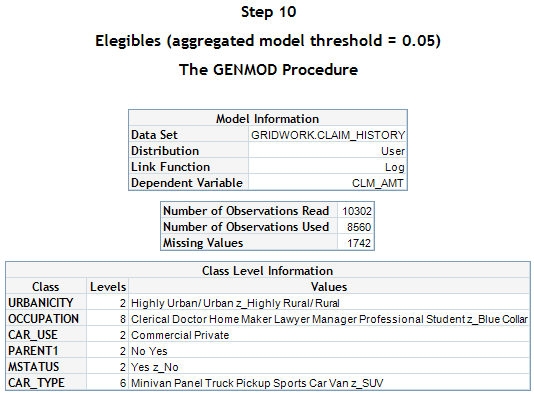


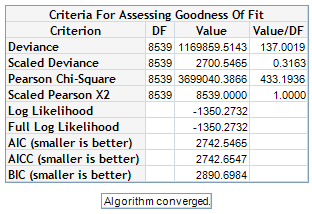


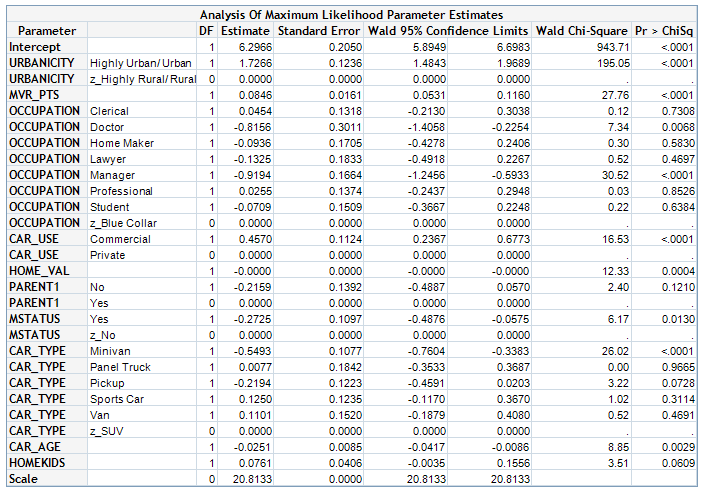


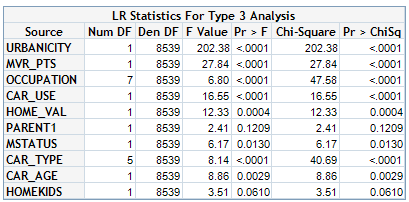


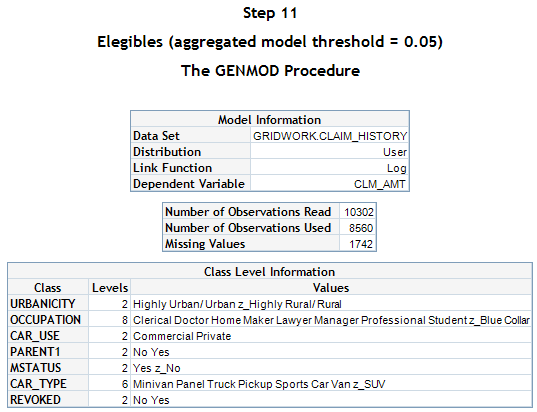


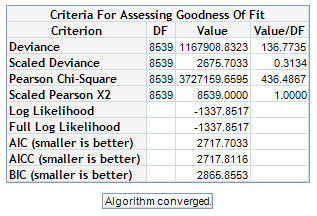


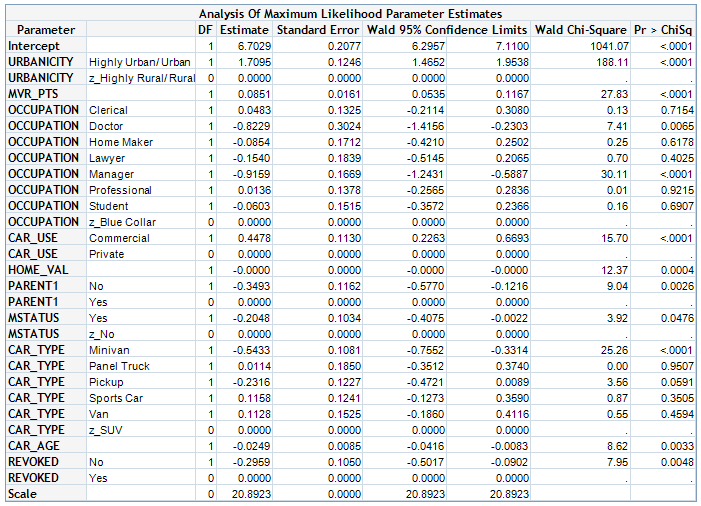


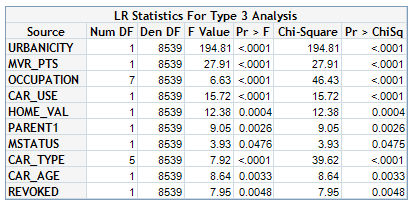


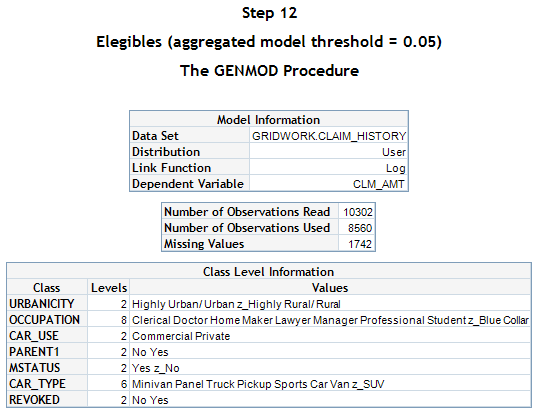


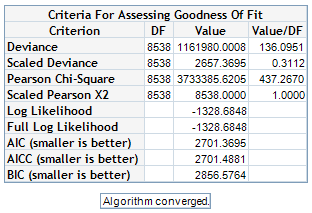


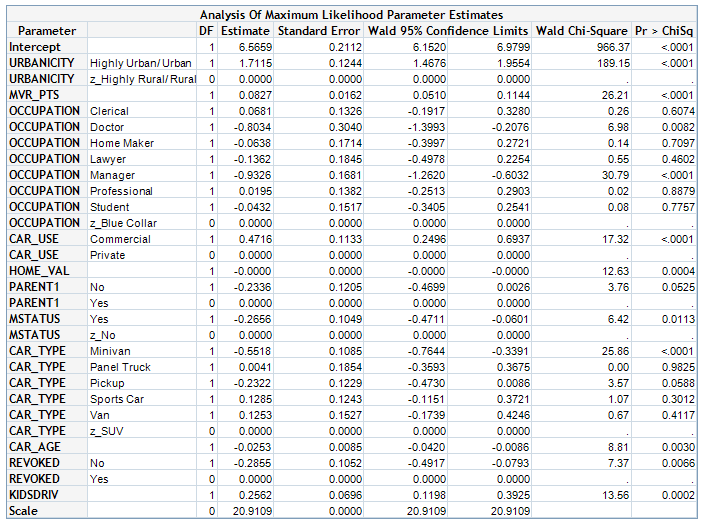


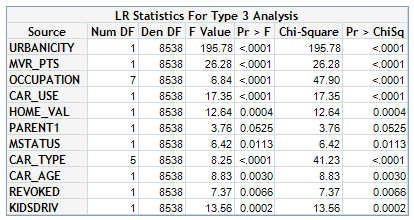


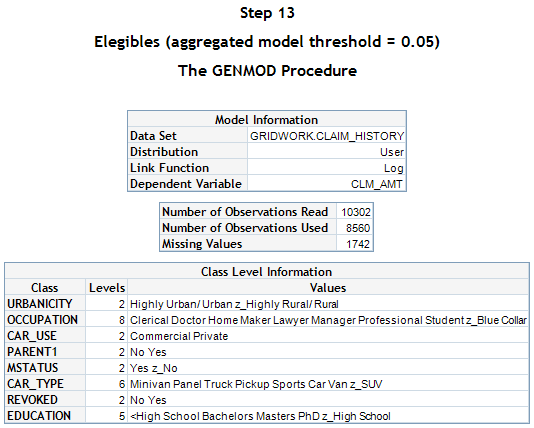


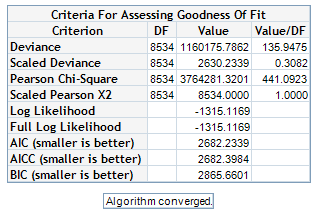


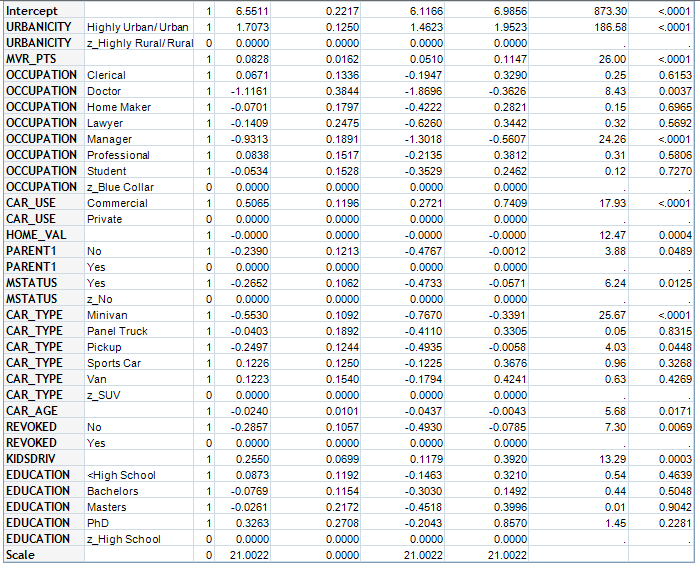


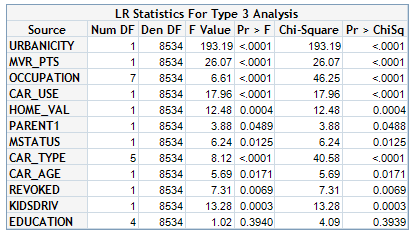


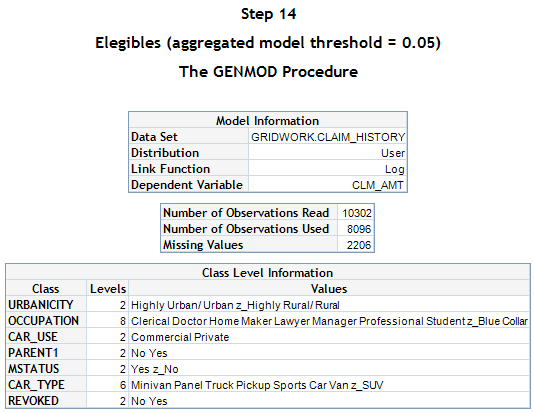


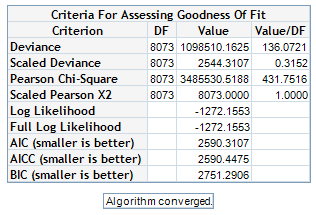


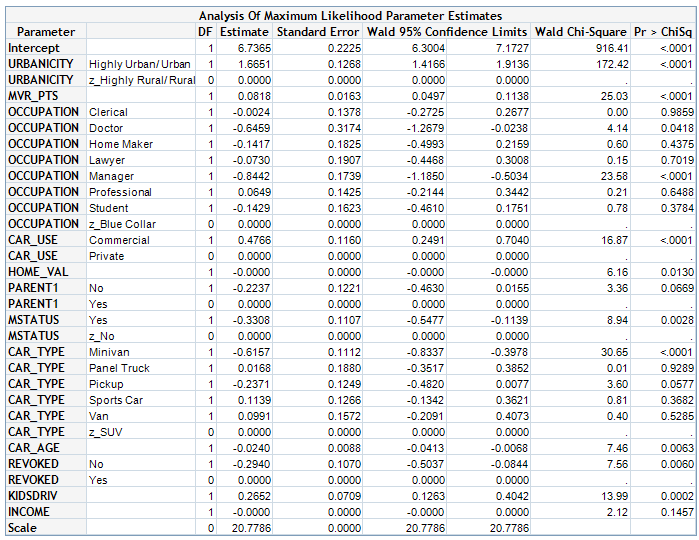


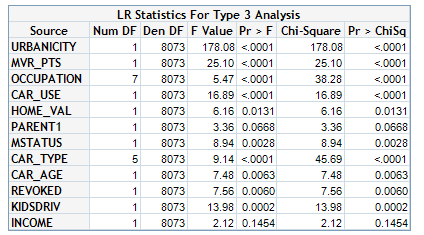


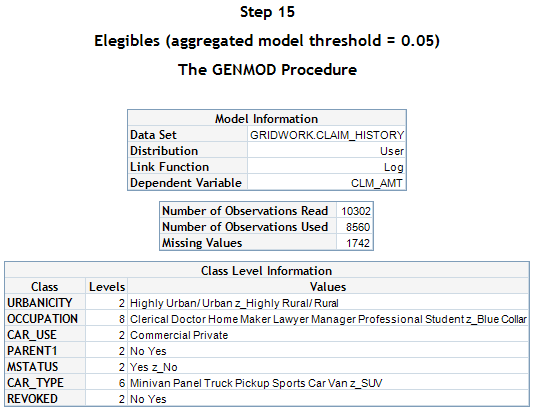


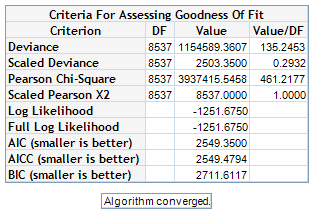


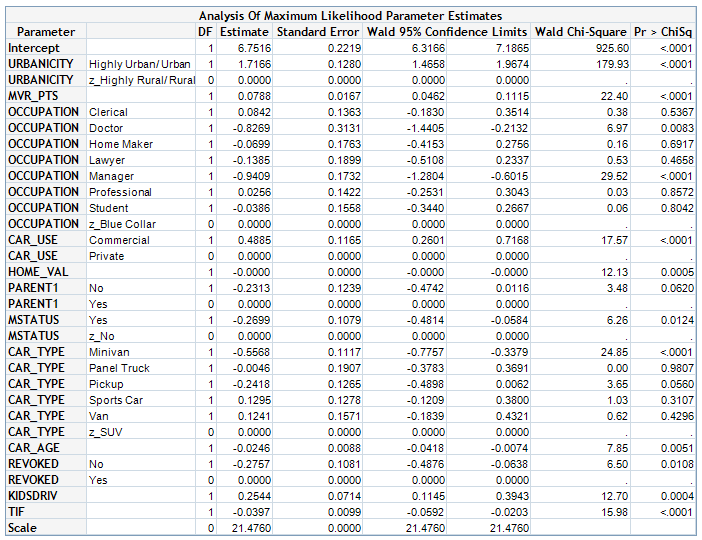


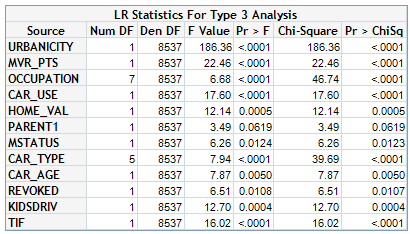


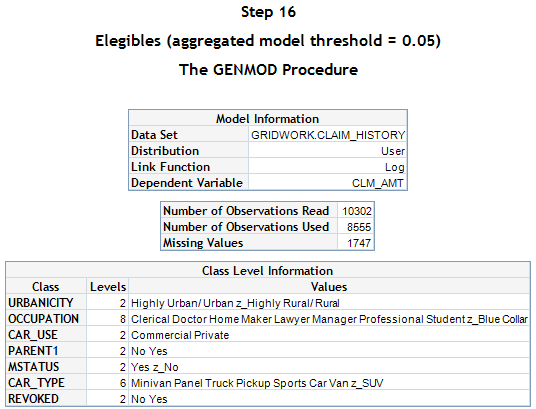


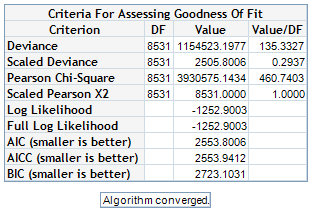


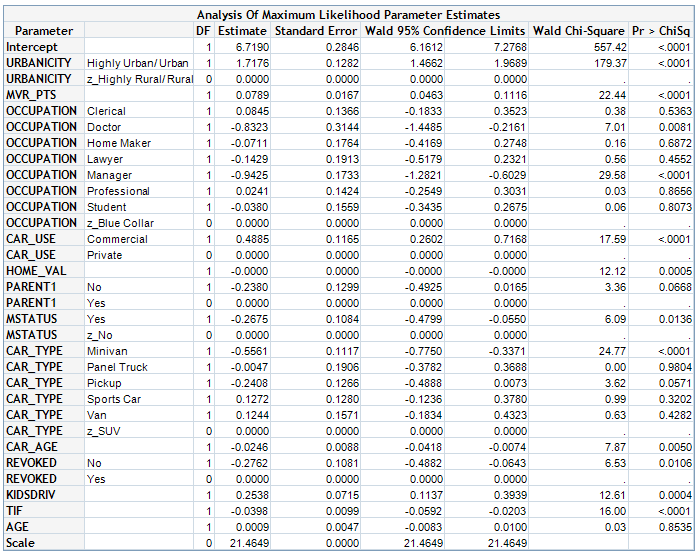


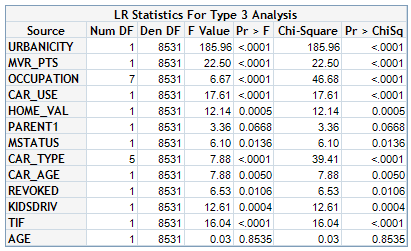


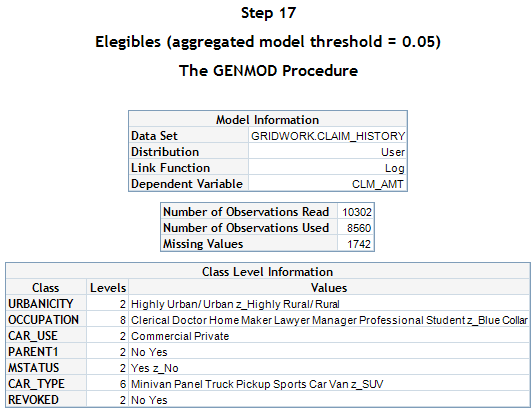


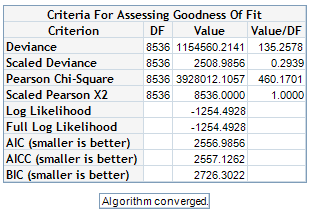


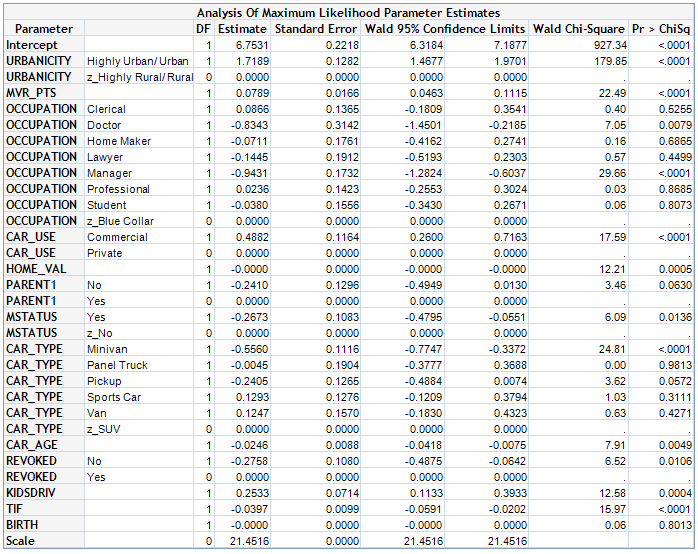


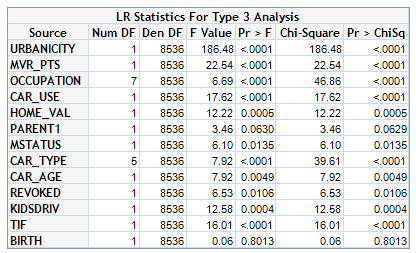


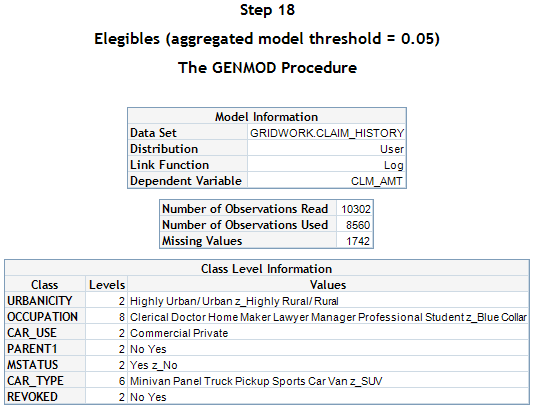


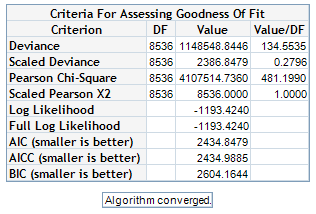


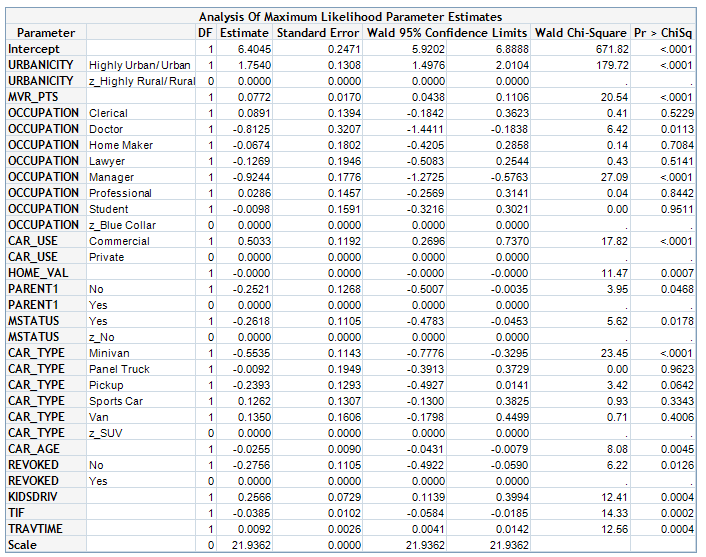


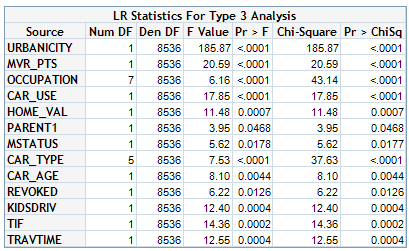


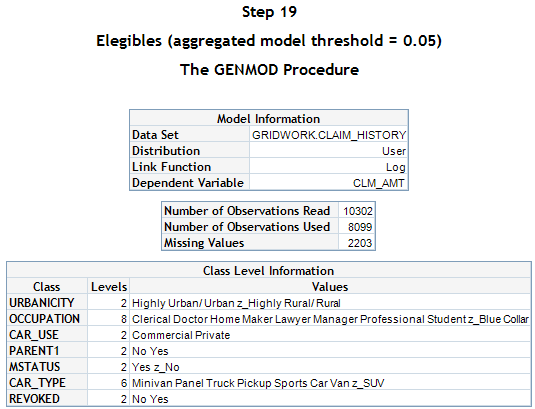


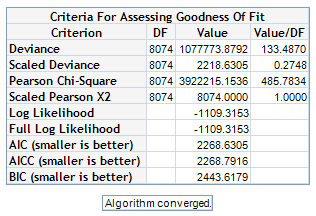


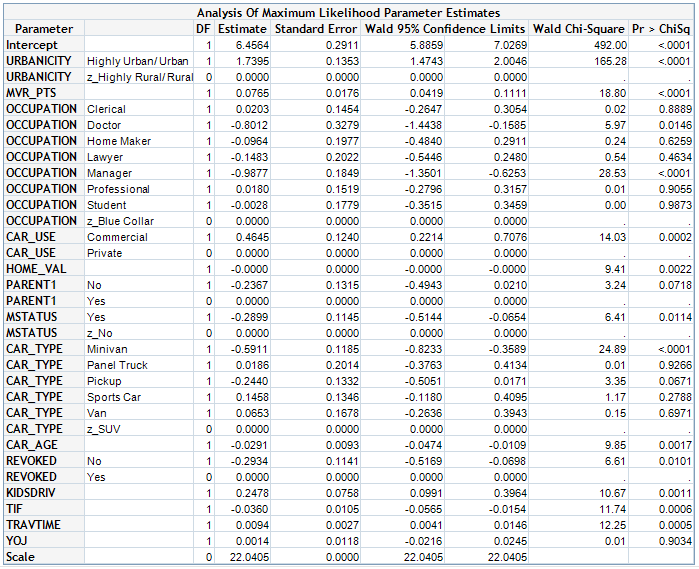


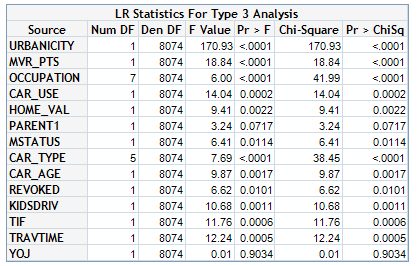


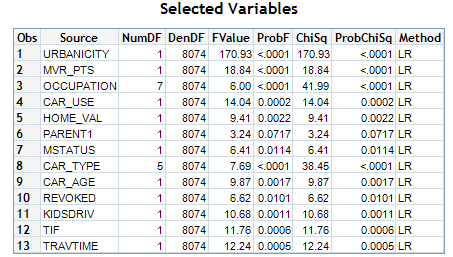












**Conclusion:**

The above lines show how the variable selection algorithm selects only 11 of the 22 original potential variables.

The SAS macros:

* Performs a stepwise selection variable based mainly on a forward selection process
* The penultimate step in the selection process shows the selected model and after the last step a summary table of the selection process is showed too
* The macro needs around 50 minutes to get results with a dataset of one million observations and around 13 variables
* The selection criteria is based on the p-values of the type 3 analysis
* With small changes the macro is useful in a context with a GENMOD procedure under Gamma, Inverse Gaussian, Log-Normal, Binomial, Gaussian, Poisson, Negative Binomial, Zero Inflated Poisson and Zero inflated Negative Binomial error functions
* In case of one or several model doesn’t converge use the NODMSSYNCHK option in the SAS code to avoid any stop in the execution
* The specification of the model is the same that the Tweedie macro used in the NAR project
* This macro only admits main factors. So, it is not possible to include interactions in the model statement of the GENMODE procedure. To include interactions it is needed create a new variable with the interaction

**References:**

[PharmaSUG 2012 - SP09 Automated forward selection for Generalized Linear Models with Categorical and Numerical Variables using PROC GENMOD Manuel Sandoval, Pharmanet-i3, Mexico City, Mexico](http://www.pharmasug.org/proceedings/2012/SP/PharmaSUG-2012-SP09.pdf)

[Paper 327-2013 1 An Overview of Syntax Check Mode and Why it is Important Thomas E. Billings, Union Bank, San Francisco, California](http://support.sas.com/resources/papers/proceedings13/327-2013.pdf)