Variable Selection:
Backwards Elimination SAS Macro

Purpose:

The macro performs an automated backward elimination 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 backwards selection macro could be used with the procedures GENNMOD, CATMOD, MIXED, GLIMMIX, etc.

Illustration:

The following SAS statements simulate 5000 observations, which are based on an underlying Tweedie generalized linear model (GLM) that exploits its connection with the compound Poisson distribution. A natural logarithm link function is assumed for modeling the response variable (yTweedie), and there are five categorical variables (C1–C5), each of which has four numerical levels and two continuous variables (D1 and D2). By design, two of the categorical variables, C3 and C4, and one of the two continuous variables, D2, have no effect on the response. The dispersion parameter is set to 0.5, and the power parameter is set to 1.5.

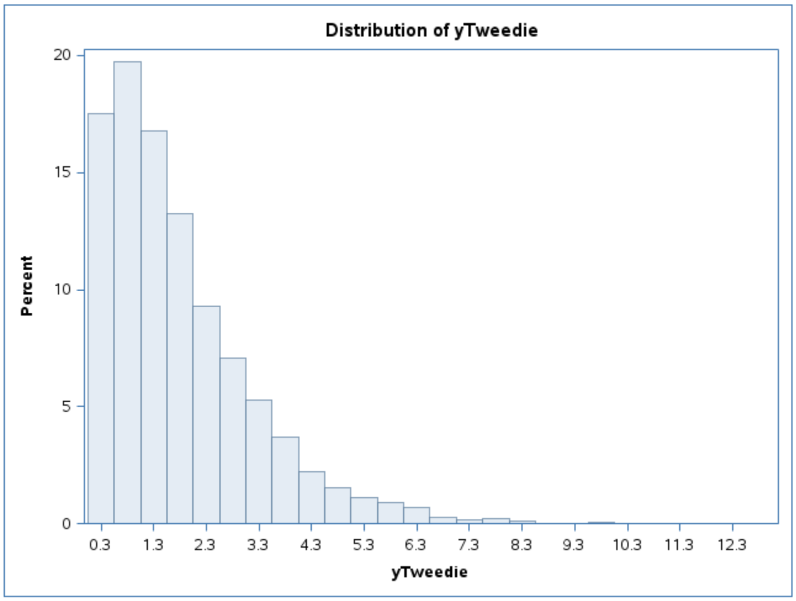
```
%let nObs = 5000;
%let nClass = 5;
%let nLevs = 4;
%let seed = 1234;

data tmp1;
   array c{&nClass};

   keep c1-c&nClass yTweedie d1 d2;

   /* Tweedie parms */
   phi=0.5;
   p=1.5;

   do i=1 to &nObs;

      do j=1 to &nClass;
         c{j} = int(ranuni(1)*&nLevs);
      end;

      d1 = ranuni(&seed);
      d2 = ranuni(&seed);

      xBeta = 0.5*((c2<2) - 2*(c1=1) + 0.5*c&nClass + 0.05*d1);
      mu = exp(xBeta);

      /* Poisson distributions parms */
      lambda = mu**(2-p)/(phi*(2-p));
      /* Gamma distribution parms */
      alpha = (2-p)/(p-1);
      gamma = phi*(p-1)*(mu**(p-1));

      rpoi = ranpoi(&seed,lambda);
      if rpoi=0 then yTweedie=0;
      else do;
         yTweedie=0;
         do j=1 to rpoi;
         yTweedie = yTweedie + rangam(&seed,alpha);
         end;
         yTweedie = yTweedie * gamma;
      end;
      output;
   end;
run;
```

The following code generates a basic explanatory data analysis for the dependent and independent variables:

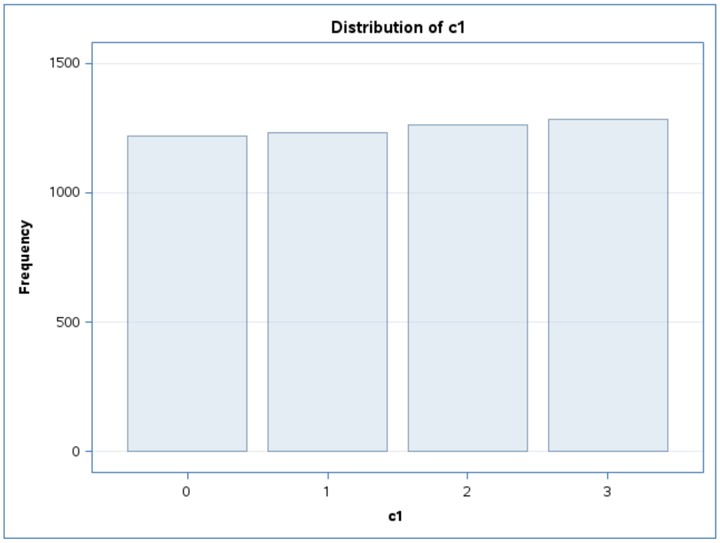
```
/* EDA */

%let var_char = yTweedie c1 c2 c3 c4 c5 d1 d2;

%put &var_char;


data var_char;
      set tmp1
      (keep= &var_char);
run;

proc contents data = var_char varnum nodetails noprint
      out=var_char_names (keep=name);
run;

data var_char_names;
      set var_char_names;
      j = _n_;
run;

* Determine the number of observations;
data _NULL_;
      if 0 then set var_char_names nobs=n;
      call symputx('nrows',n);
      stop;
run;

%put &nrows;

%macro do_eda_uni;
%do obs = 1 %to &nrows;

data _null_;
      set var_char_names;
      if j = &obs  then call symputx("var", put(name, 10.));
run;
     %if (%upcase(&var)=YTWEEDIE) or (%upcase(&var)=D1) or (%upcase(&var)=D2)   %then %do;

      ods graphics on;
            proc means data=tmp1 fw=12 printalltypes chartype
```

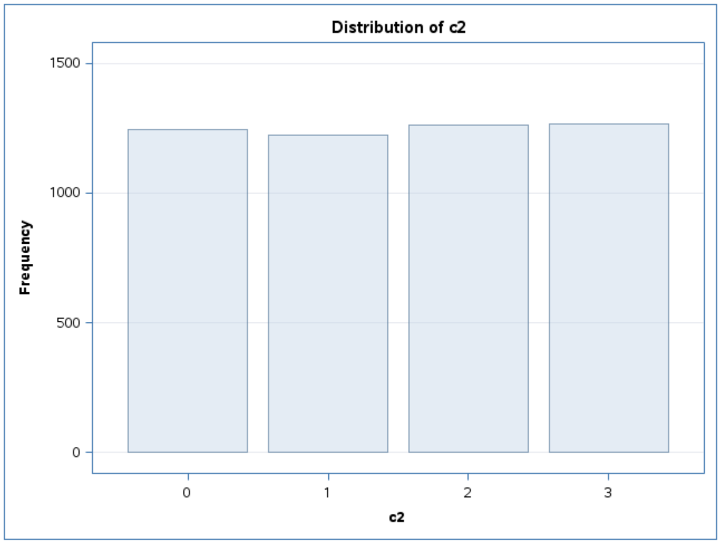
```
                  qmethod=os maxdec=2

                  mean
                  min
                  max
                  mode
                  range
                  n
                  nmiss
                  p1
                  p5
                  median
                  p95
                  p99   ;
                  var &var;
            run;

            title "histograms";
            proc univariate data=tmp1     noprint;
                  var &var;
                  histogram ;
            run;
      ods graphics off;
      %end;

      %else %do;
      ods graphics on;
            proc freq data=tmp1
            order=internal;
            tables &var /  scores=table plots(only)=freq;
            run;
      ods graphics off;
      %end;

%end;
%mend do_eda_uni;

%do_eda_uni;
```

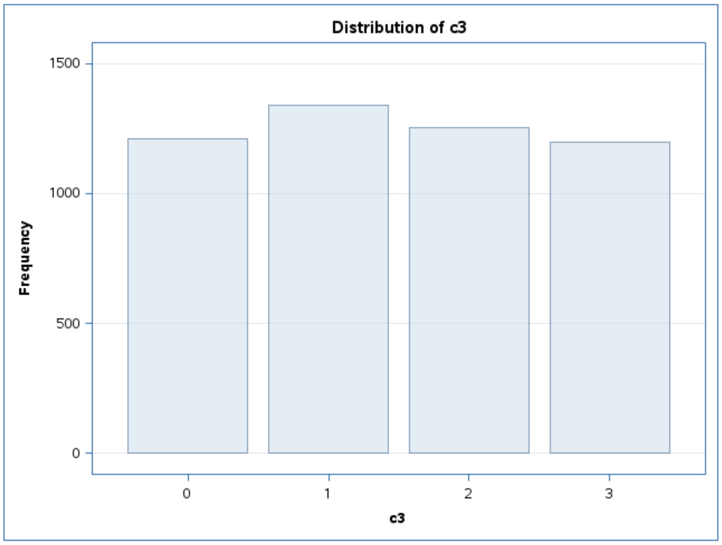
The histogram for the yTweedie dependent variable:

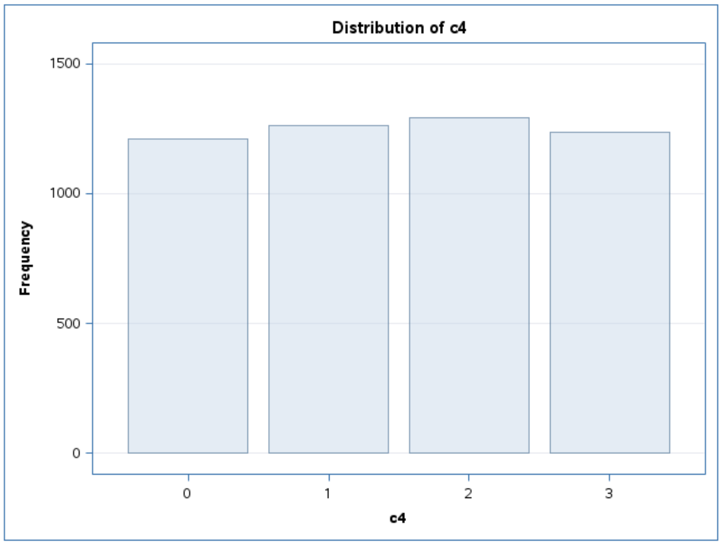


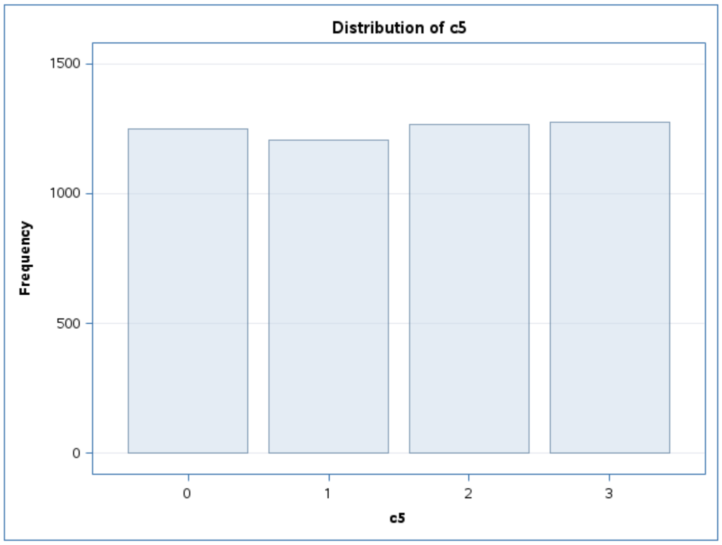
The independent character variable c1-c5:



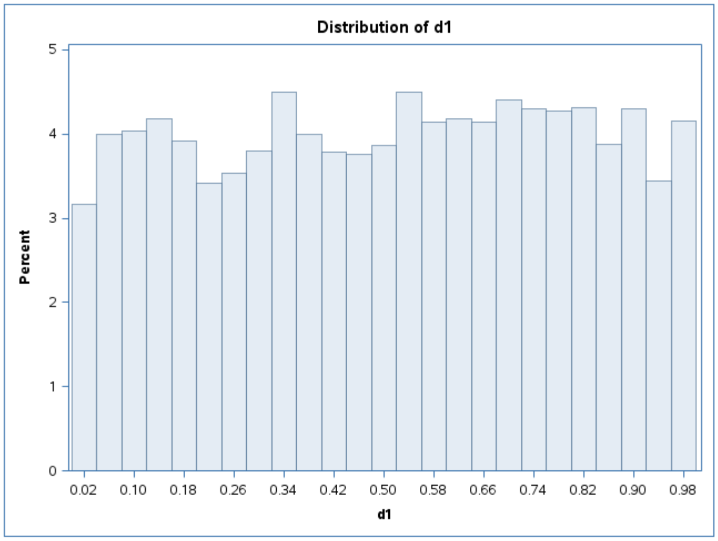


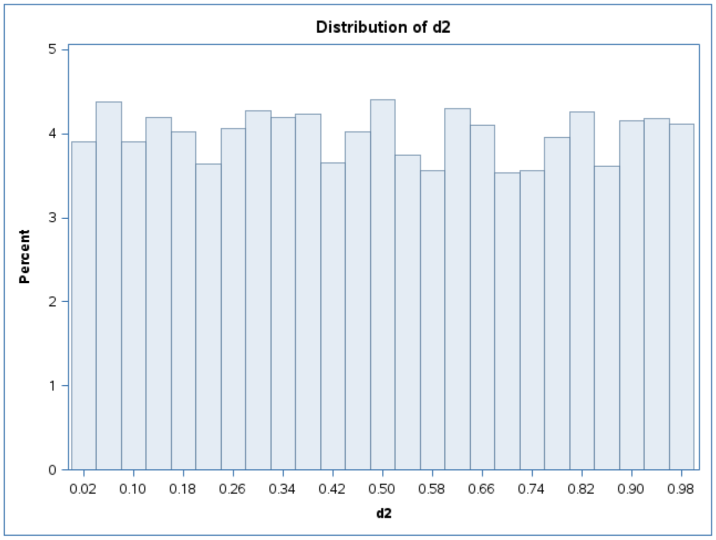






The histogram for the two independent continuous variables d1 and d2:





The next lines contain the two SAS macros for the backwards elimination selection process using a Tweedie error function.

The first macro %MdStmt is a stand-alone macro. The main macro, %MdSelect, consists of multiple calls to the macro %MdStmt.

```
/* Variable Selection Macro: Backwards elimination */
%let p=1.5;
options mlogic;
%macro MdStmt(
          resvar = /*response variable */
         ,expvar = /*list of explanatory variables, separated by ' ' */
         ,clsvar = /*classification variables in the CLASS statement separated by ' ' */
         ,p =
         );

        ods output Type3=pval(rename=source=parm);
            proc genmod data=tmp1 NAMELEN=50;
                  if _resp_ > 0 then
                  d = 2*(_resp_*(_resp_**(1-&p)-_mean_**(1-&p))/
                  (1-&p)-(_resp_**(2-&p)-_mean_**(2-&p))/(2-&p));
                  else d = 2* _mean_**(2-&p)/(2-&p);
                  variance var = _mean_**&p;
                  deviance dev = d;
                  class &clsvar;
                  model &resvar =  &expvar /link=log type3 scale=pearson;
                  *scwgt expos;
            title "&resvar = &expvar";
            run;
            ods output close;
 %mend MdStmt;
```

There are five macro parameters in the macro %MdSelect: &VAR, &INTVAR, &CATVAR, &SLSTAY and &POWER:

- &VAR is the response variable which will be passed into &RESVAR when calling the macro %MdStmt;
- &INTVAR includes all the potential explanatory variables which will be passed into &EXPVAR in %MdStmt only forthe first call;

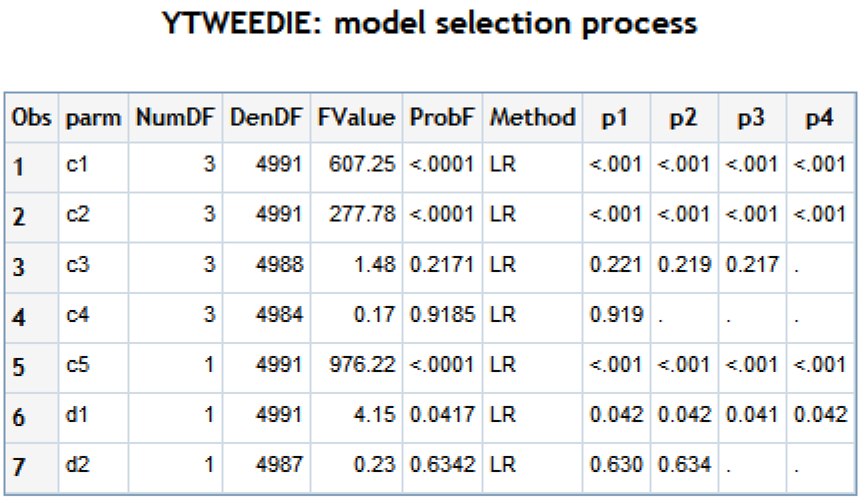
- &CATVAR contains all the categorical explanatory variables which will be passed into %CLSVAR in %MdStmt;
- &SLSTAY is the criteria for removing variable;
- and &POWER is the power parameter of the Tweedie distribution

```
%macro MdSelect(
         var= /*response variable */
         ,intvar= /*initial explanatory variables for full model */
         ,catvar= /*categorical explanatory variables */
         ,slstay= /*criterion for removing variable */
         ,power=
         );
      %let var=%upcase(&var);
      %let intvar=%upcase(&intvar);
      %let catvar=%upcase(&catvar);
      %let power =&power;
%*-------------------------------------------------------------------------*;
%* Create empty dataset "step" with only one column "parm". It will be *;
%* merged with "pval" from PROC GENMOD by "parm" *;
%*-------------------------------------------------------------------------*;
 proc sql;
      create table step_&var (parm char(9));
 quit;
%*------------------------------------------------------------------------------*;
%* %do %until performs multivariate backward model selection: *;
%* In each iteration: *;
%* 1. Run the logistic regression model *;
%* 2. Update the dataset "step_&var" *;
%* 3. Create &pmax as the maximum p-value, and &varlist as the list of *;
%* variables without the one with the max p-value *;
%* 4. Check whether the max p-value <= &SLSTAY *;
%* 5. If NO, then eliminate the variable with max p-value, repeat step 1 to 4.*;
%* If YES, the loop stops *;
%*------------------------------------------------------------------------------*;
 %let i=1;
 %do %until (&pmax<=&slstay);

      %if &i = 1 %then
            %MdStmt(resvar=&var ,expvar=&intvar, clsvar=&catvar, p=&power); %*initial
model;
      %else %do;
            %MdStmt(resvar=&var ,expvar=&varlist, clsvar=&catvar, p=&power); %*reduced
model;
      %end;
      proc sort data=step_&var; by parm;
      proc sort data=pval; by parm;
      data step_&var;
            merge step_&var pval;
            by parm;
            p&i=put(ProbChiSq, pvalue6.3);
            drop ProbChiSq ChiSq DF;
      run;
      proc sql noprint;
            select max(ProbChiSq) into :pmax
            from pval;
            select distinct parm into :varlist separated by ' '
            from pval
            having ProbChiSq^=max(ProbChiSq);
      quit;

      %let i=%eval(&i+1);
 %end;
 proc print data=step_&var;
      title "&var: model selection process";
 run;
%mend MdSelect;

%MdSelect(var=yTweedie, intvar=c1 c2 c3 c4 c5 d1 d2, catvar=c1 c2 c3 c4 c5, slstay=0.05,
power=1.5);
```

The execution of the above two macros create two outputs:

- A summary table of the model selection process
- The whole model selection process step by step

The summary table of the model selection process:



YTWEEDIE: model selection process

Obs	parm	NumDF	DenDF	FValue	ProbF	Method	p1	p2	p3	p4
1	c1	3	4991	607.25	<.0001	LR	<.001	<.001	<.001	<.001
2	c2	3	4991	277.78	<.0001	LR	<.001	<.001	<.001	<.001
3	c3	3	4988	1.48	0.2171	LR	0.221	0.219	0.217	.
4	c4	3	4984	0.17	0.9185	LR	0.919	.	.	.
5	c5	1	4991	976.22	<.0001	LR	<.001	<.001	<.001	<.001
6	d1	1	4991	4.15	0.0417	LR	0.042	0.042	0.041	0.042
7	d2	1	4987	0.23	0.6342	LR	0.630	0.634	.	.

The table shows that the variable C4 is eliminated in the second step of the process. The variable D2 is eliminated in the third step. And the variable C3 is eliminated in the fourth step. After the fourth step the algorithm arrive at final main effects model.

The whole variable selection process step by step:

YTWEEDIE = C1 C2 C3 C4 C5 D1 D2

The GENMOD Procedure

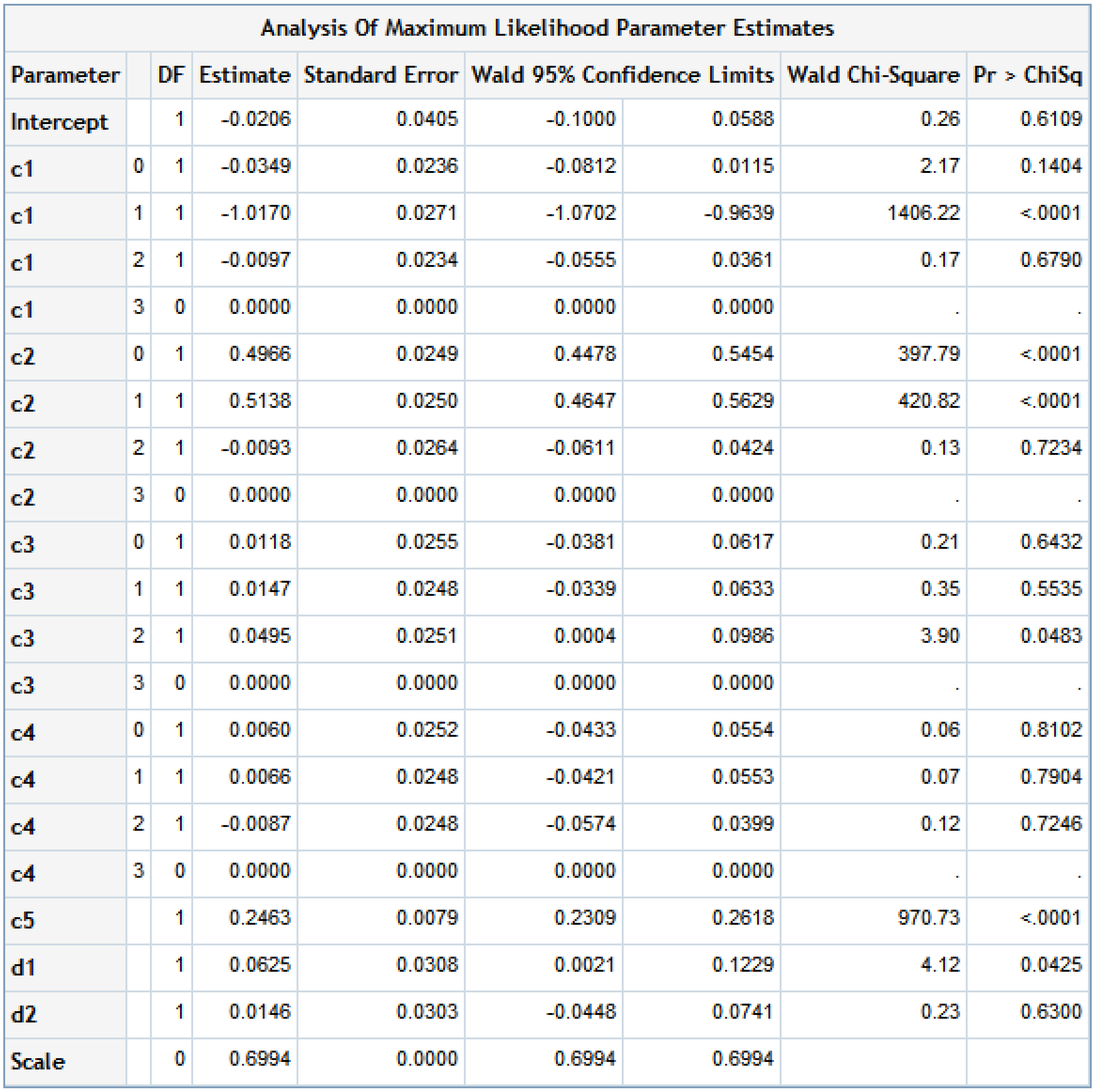
Model Information	
Data Set	WORK.TMP1
Distribution	User
Link Function	Log
Dependent Variable	yTweedie

Number of Observations Read	5000
Number of Observations Used	5000

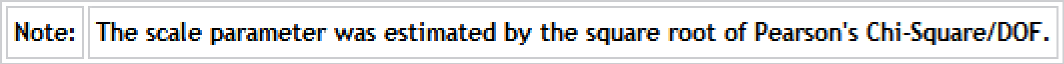
Class Level Information		
Class	Levels	Values
c1	4	0 1 2 3
c2	4	0 1 2 3
c3	4	0 1 2 3
c4	4	0 1 2 3

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4984	2731.4307	0.5480
Scaled Deviance	4984	5584.1513	1.1204
Pearson Chi-Square	4984	2437.8728	0.4891
Scaled Pearson X2	4984	4984.0000	1.0000
Log Likelihood		-2792.0756	
Full Log Likelihood		-2792.0756	
AIC (smaller is better)		5616.1513	
AICC (smaller is better)		5616.2605	
BIC (smaller is better)		5720.4264	

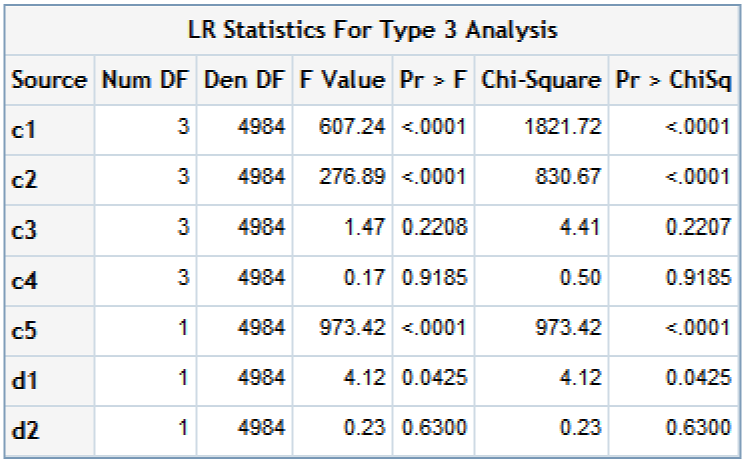
Algorithm converged.



Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0206	0.0405	-0.1000	0.0588	0.26	0.6109
c1	0	1	-0.0349	0.0236	-0.0812	0.0115	2.17	0.1404
c1	1	1	-1.0170	0.0271	-1.0702	-0.9639	1406.22	<.0001
c1	2	1	-0.0097	0.0234	-0.0555	0.0361	0.17	0.6790
c1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c2	0	1	0.4966	0.0249	0.4478	0.5454	397.79	<.0001
c2	1	1	0.5138	0.0250	0.4647	0.5629	420.82	<.0001
c2	2	1	-0.0093	0.0264	-0.0611	0.0424	0.13	0.7234
c2	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c3	0	1	0.0118	0.0255	-0.0381	0.0617	0.21	0.6432
c3	1	1	0.0147	0.0248	-0.0339	0.0633	0.35	0.5535
c3	2	1	0.0495	0.0251	0.0004	0.0986	3.90	0.0483
c3	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c4	0	1	0.0060	0.0252	-0.0433	0.0554	0.06	0.8102
c4	1	1	0.0066	0.0248	-0.0421	0.0553	0.07	0.7904
c4	2	1	-0.0087	0.0248	-0.0574	0.0399	0.12	0.7246
c4	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c5		1	0.2463	0.0079	0.2309	0.2618	970.73	<.0001
d1		1	0.0625	0.0308	0.0021	0.1229	4.12	0.0425
d2		1	0.0146	0.0303	-0.0448	0.0741	0.23	0.6300
Scale		0	0.6994	0.0000	0.6994	0.6994		



Note:	The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.



LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
c1	3	4984	607.24	<.0001	1821.72	<.0001
c2	3	4984	276.89	<.0001	830.67	<.0001
c3	3	4984	1.47	0.2208	4.41	0.2207
c4	3	4984	0.17	0.9185	0.50	0.9185
c5	1	4984	973.42	<.0001	973.42	<.0001
d1	1	4984	4.12	0.0425	4.12	0.0425
d2	1	4984	0.23	0.6300	0.23	0.6300

YTWEEDIE = c1 c2 c3 c5 d1 d2

The GENMOD Procedure

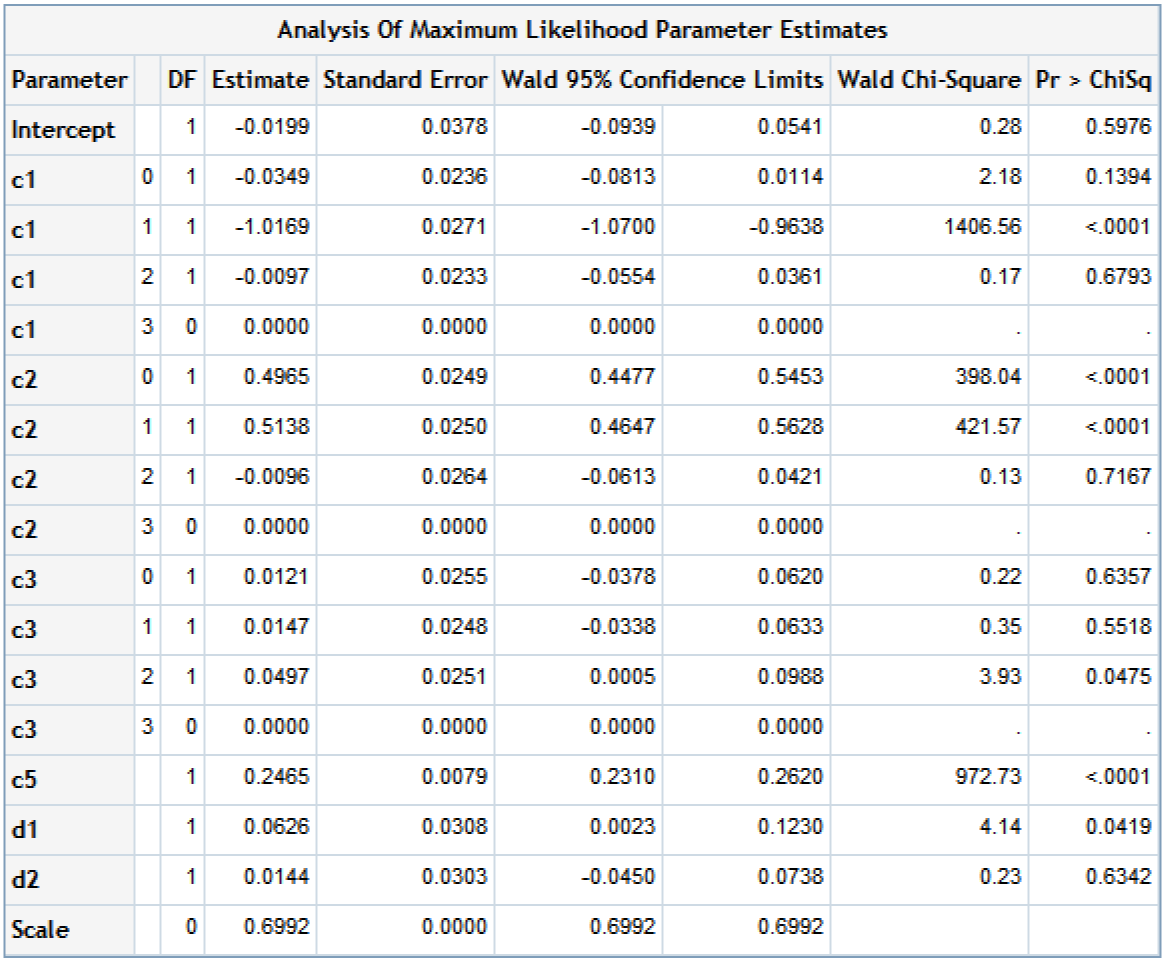
Model Information	
Data Set	WORK.TMP1
Distribution	User
Link Function	Log
Dependent Variable	yTweedie

Number of Observations Read	5000
Number of Observations Used	5000

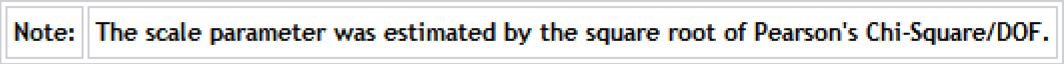
Class Level Information		
Class	Levels	Values
c1	4	0 1 2 3
c2	4	0 1 2 3
c3	4	0 1 2 3
c4	4	0 1 2 3

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4987	2731.6761	0.5478
Scaled Deviance	4987	5586.8460	1.1203
Pearson Chi-Square	4987	2438.3827	0.4889
Scaled Pearson X2	4987	4987.0000	1.0000
Log Likelihood		-2793.4230	
Full Log Likelihood		-2793.4230	
AIC (smaller is better)		5612.8460	
AICC (smaller is better)		5612.9190	
BIC (smaller is better)		5697.5695	

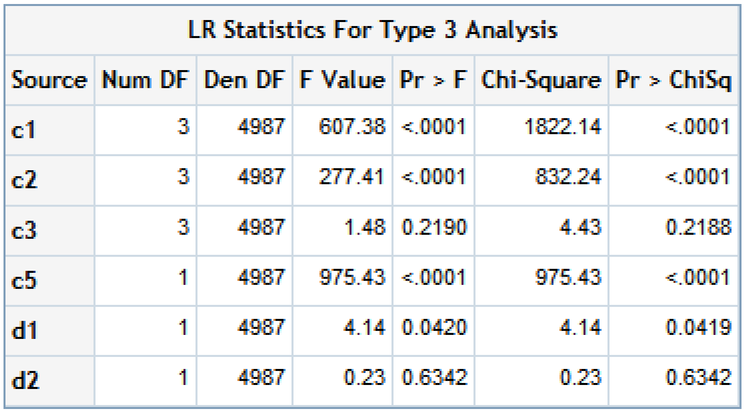
Algorithm converged.



Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0199	0.0378	-0.0939	0.0541	0.28	0.5976
c1	0	1	-0.0349	0.0236	-0.0813	0.0114	2.18	0.1394
c1	1	1	-1.0169	0.0271	-1.0700	-0.9638	1406.56	<.0001
c1	2	1	-0.0097	0.0233	-0.0554	0.0361	0.17	0.6793
c1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c2	0	1	0.4965	0.0249	0.4477	0.5453	398.04	<.0001
c2	1	1	0.5138	0.0250	0.4647	0.5628	421.57	<.0001
c2	2	1	-0.0096	0.0264	-0.0613	0.0421	0.13	0.7167
c2	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c3	0	1	0.0121	0.0255	-0.0378	0.0620	0.22	0.6357
c3	1	1	0.0147	0.0248	-0.0338	0.0633	0.35	0.5518
c3	2	1	0.0497	0.0251	0.0005	0.0988	3.93	0.0475
c3	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c5		1	0.2465	0.0079	0.2310	0.2620	972.73	<.0001
d1		1	0.0626	0.0308	0.0023	0.1230	4.14	0.0419
d2		1	0.0144	0.0303	-0.0450	0.0738	0.23	0.6342
Scale		0	0.6992	0.0000	0.6992	0.6992		



Note:	The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.



LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
c1	3	4987	607.38	<.0001	1822.14	<.0001
c2	3	4987	277.41	<.0001	832.24	<.0001
c3	3	4987	1.48	0.2190	4.43	0.2188
c5	1	4987	975.43	<.0001	975.43	<.0001
d1	1	4987	4.14	0.0420	4.14	0.0419
d2	1	4987	0.23	0.6342	0.23	0.6342

YTWEEDIE = c1 c2 c3 c5 d1

The GENMOD Procedure

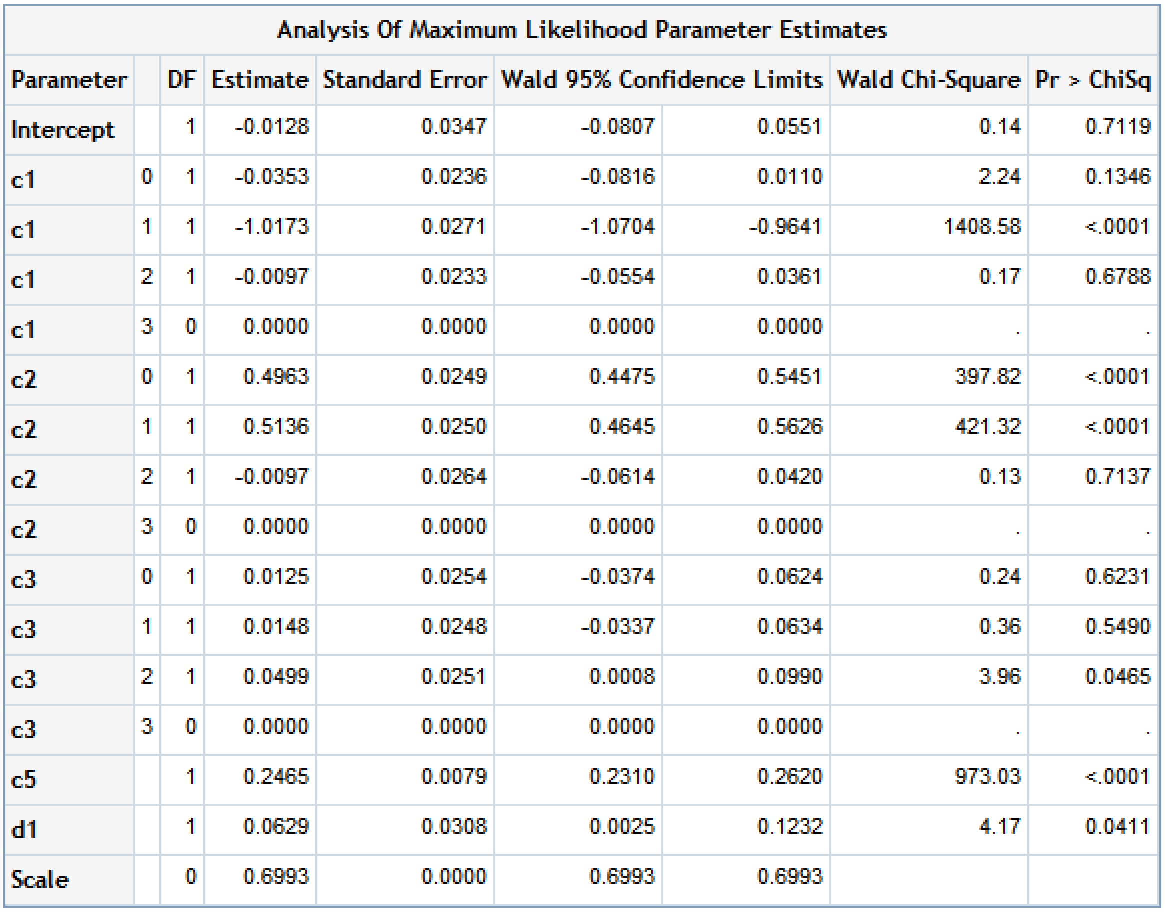
Model Information	
Data Set	WORK.TMP1
Distribution	User
Link Function	Log
Dependent Variable	yTweedie

Number of Observations Read	5000
Number of Observations Used	5000

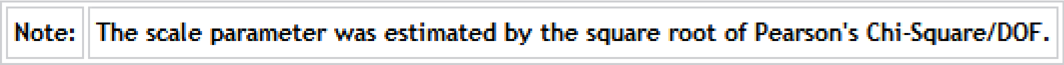
Class Level Information		
Class	Levels	Values
c1	4	0 1 2 3
c2	4	0 1 2 3
c3	4	0 1 2 3
c4	4	0 1 2 3

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4988	2731.7869	0.5477
Scaled Deviance	4988	5586.6502	1.1200
Pearson Chi-Square	4988	2439.0560	0.4890
Scaled Pearson X2	4988	4988.0000	1.0000
Log Likelihood		-2793.3251	
Full Log Likelihood		-2793.3251	
AIC (smaller is better)		5610.6502	
AICC (smaller is better)		5610.7128	
BIC (smaller is better)		5688.8565	

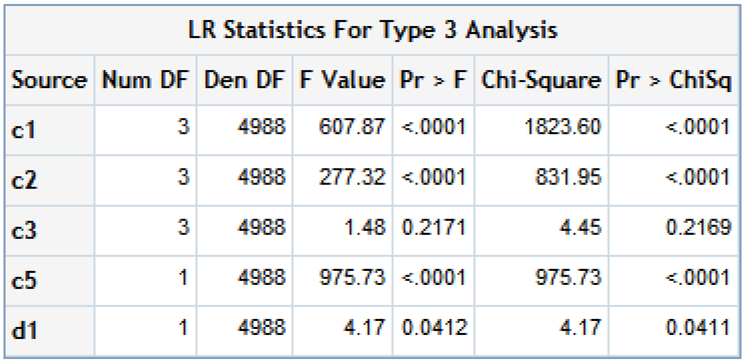
Algorithm converged.



Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.0128	0.0347	-0.0807	0.0551	0.14	0.7119
c1	0	1	-0.0353	0.0236	-0.0816	0.0110	2.24	0.1346
c1	1	1	-1.0173	0.0271	-1.0704	-0.9641	1408.58	<.0001
c1	2	1	-0.0097	0.0233	-0.0554	0.0361	0.17	0.6788
c1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c2	0	1	0.4963	0.0249	0.4475	0.5451	397.82	<.0001
c2	1	1	0.5136	0.0250	0.4645	0.5626	421.32	<.0001
c2	2	1	-0.0097	0.0264	-0.0614	0.0420	0.13	0.7137
c2	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c3	0	1	0.0125	0.0254	-0.0374	0.0624	0.24	0.6231
c3	1	1	0.0148	0.0248	-0.0337	0.0634	0.36	0.5490
c3	2	1	0.0499	0.0251	0.0008	0.0990	3.96	0.0465
c3	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c5		1	0.2465	0.0079	0.2310	0.2620	973.03	<.0001
d1		1	0.0629	0.0308	0.0025	0.1232	4.17	0.0411
Scale		0	0.6993	0.0000	0.6993	0.6993		



Note: The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.



LR Statistics For Type 3 Analysis						
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
c1	3	4988	607.87	<.0001	1823.60	<.0001
c2	3	4988	277.32	<.0001	831.95	<.0001
c3	3	4988	1.48	0.2171	4.45	0.2169
c5	1	4988	975.73	<.0001	975.73	<.0001
d1	1	4988	4.17	0.0412	4.17	0.0411

YTWEEDIE = c1 c2 c5 d1

The GENMOD Procedure

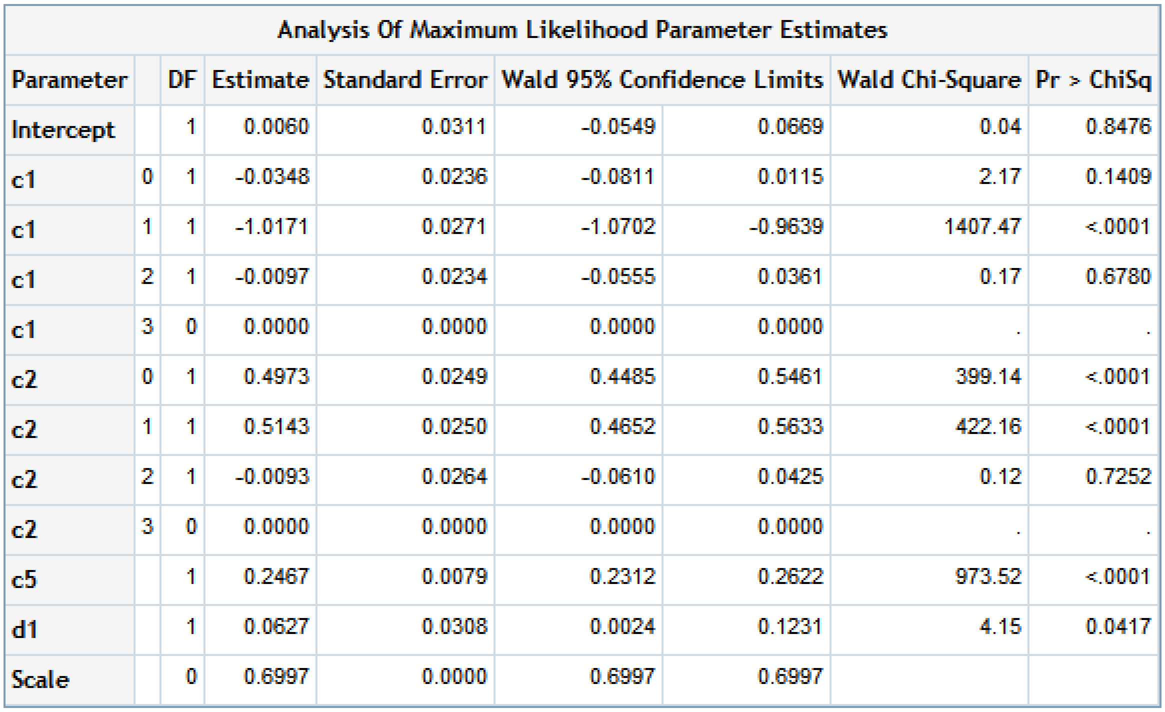
Model Information	
Data Set	WORK.TMP1
Distribution	User
Link Function	Log
Dependent Variable	yTweedie

Number of Observations Read	5000
Number of Observations Used	5000

Class Level Information		
Class	Levels	Values
c1	4	0 1 2 3
c2	4	0 1 2 3
c3	4	0 1 2 3
c4	4	0 1 2 3

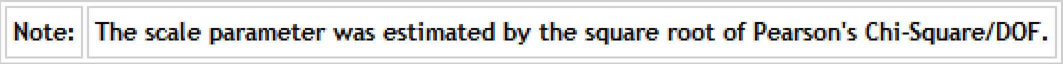
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4991	2733.9621	0.5478
Scaled Deviance	4991	5584.3541	1.1189
Pearson Chi-Square	4991	2443.4705	0.4896
Scaled Pearson X2	4991	4991.0000	1.0000
Log Likelihood		-2792.1770	
Full Log Likelihood		-2792.1770	
AIC (smaller is better)		5602.3541	
AICC (smaller is better)		5602.3901	
BIC (smaller is better)		5661.0088	

Algorithm converged.

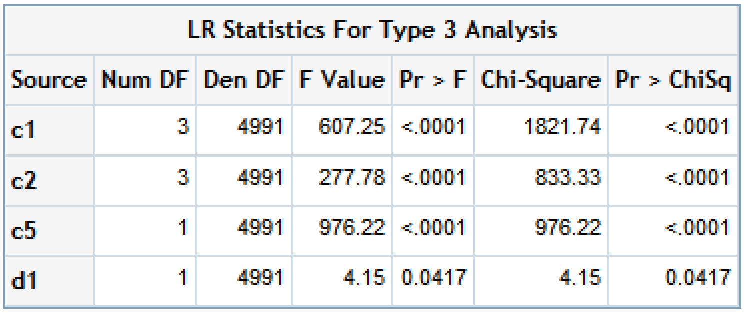


Analysis Of Maximum Likelihood Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	0.0060	0.0311	-0.0549	0.0669	0.04	0.8476
c1	0	1	-0.0348	0.0236	-0.0811	0.0115	2.17	0.1409
c1	1	1	-1.0171	0.0271	-1.0702	-0.9639	1407.47	<.0001
c1	2	1	-0.0097	0.0234	-0.0555	0.0361	0.17	0.6780
c1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c2	0	1	0.4973	0.0249	0.4485	0.5461	399.14	<.0001
c2	1	1	0.5143	0.0250	0.4652	0.5633	422.16	<.0001
c2	2	1	-0.0093	0.0264	-0.0610	0.0425	0.12	0.7252
c2	3	0	0.0000	0.0000	0.0000	0.0000	.	.
c5		1	0.2467	0.0079	0.2312	0.2622	973.52	<.0001
d1		1	0.0627	0.0308	0.0024	0.1231	4.15	0.0417
Scale		0	0.6997	0.0000	0.6997	0.6997		



Note:	The scale parameter was estimated by the square root of Pearson's Chi-Square/DOF.



LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
c1	3	4991	607.25	<.0001	1821.74	<.0001
c2	3	4991	277.78	<.0001	833.33	<.0001
c5	1	4991	976.22	<.0001	976.22	<.0001
d1	1	4991	4.15	0.0417	4.15	0.0417

Conclusion:

The above lines shows how the variable selection algorithm eliminates those variables (C3, C4 and D2) no associated with the dependent variable yTweedie - remember that the illustrative dataset was arterially created with this aim. Therefore, the macro works accurately.

The SAS macros %MdStmt and %MdSelect:

- Performs a backwards elimination variable selection process
- The last step in the elimination process shows the selected model and a summary table of the elimination process
- The macro needs around 15 minutes to get results with a dataset of one million observations and around 13 variables
- The elimination 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
- This macro could be useful as a template to create Forward and Stepwise variable selection processes
- One drawback of the backwards elimination process is that if the full model with all potential main factors does not converge the macro does not work. That is one of the reasons because a forward option is interesting
- 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:

A detailed explanation of the algorithm and the code appears here:

Using Macro and ODS to Overcome Limitations of SAS® Procedures Jing Su and Wei (Lisa) Lin, Merck & Co, Inc., North Wales, PA

The dataset for the example comes from here:

http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_genmod_examples12.htm

I made some changes in order to get coherence results.