Paper 247-26

Analysis of Count Data Using the SAS® System

Alex Pedan, Vasca Inc., Tewksbury, MA

ABSTRACT

Count data is increasingly common in clinical research (Gardner, Mulvey and Shaw (1995); Glynn and Buring (1996)). Examples include the number of adverse events occurring during a follow up period, the number of hospitalizations, the number of seizures in epileptics, etc. It is straightforward to analyze the count data by using PROCEDURE GENMOD of SAS/STAT, but as it is going to be shown below, correctly specifying the statistical model is of the utmost importance in getting proper inferences.

In this paper, we review the statistical methodology and present a SAS macro that creates convenient output for the presentation of count data.

INTRODUCTION

Recurrent events are frequent outcomes in longitudinal clinical and epidemiological studies. One natural and clinically interpretable measure of occurrence is the event rate, defined as the number of events divided by the total person-years of experience. The challenge in analyzing the event rates arises because some individuals are more prone to recurrences than others. To illustrate the problem, we will consider a data set from a randomized clinical trial which was conducted to evaluate a novel transcutaneous approach to hemodialysis vascular access offered by the totally implantable Lifesite® Hemodialysis Access System (Vasca, Inc.,Tewksbury, MA). 36 patients were treated with the LifeSite device (test) and 34 with the Tesio-Cath catheter (control). Hypotension and cramping are common adverse events during hemodialysis treatment, frequently occurring together. Our goal is to compare the rates of hypotensions and crampings in two treatment groups during the 3-month follow up interval. The following statements input the data, which are arranged as one observation per subject.

```
data Hypo_Cramp;
input  id  device $  count  fu_time;
logt=log(fu_time);

datalines;
1 Control  2 2.5
2 Control  0 3
3 Control  1 1.8
…………..
68 Test  1 3
69 Test  3 3
70 Test  0 2.7
;
run;
```

The variable DEVICE represents treatment assignment, the variable COUNT contains the number of hypotensions and crampings for each subject during the follow up period and the variable FU_TIME represents follow up time for each subject measured in months.

POISSON REGRESSION

The most widely used regression model for multivariate count data is the log-linear model (see McCullagh and Nelder, 1989):

$$\log(E(Y_i)) = \log t_i + \beta' \mathbf{x}_i$$

where β is a vector of regression coefficients, $\mathbf{x}_i$ is a vector of covariates for subject i, so called offset variable $\log t_i$ is needed to account for possible different observation periods (t_i) for different subjects.

The popular measures of the adequacy of the model fit are deviance and Pearson Chi-Square (X^2). If statistical model is correct then both quantities are asymptotically distributed as χ^2 statistics with n-p degrees of freedom (df); where n is number of subjects and p is the number of fitted parameters (two in our case: intercept and regression coefficient for device variable). Thus if the regression model is adequate, the expected value of both the deviance and Pearson Chi-Square is equal (or close) to n-p (both the scaled deviance ≅ 1 or the scaled Pearson Chi-Square: $X^2/\text{df} \cong 1$), otherwise there could be doubt about validity of the model.

If the hypotensions and crampings counted on a subject were independent, we would expect to be able to use a Poisson distribution as the basis for our model. One important characteristic of counts is that the variance tends to increase with the average size of the counts. The main feature of the Poisson model is that expected value of the random variable Y_i (counts of hypotensions and crampings) for subject i is equal to its variance:

$$\mu = E(Y_i) = \mathrm{Var}(Y_i)$$

The Poisson regression is a member of a class of generalized linear models, which is an extension of traditional linear models that allows the mean of a population to depend on a linear predictor through a nonlinear link function and allows the response probability distribution to be any member of an exponentional family of distributions (McCullagh and Nelder, 1989). The PROC GENMOD of SAS can fit wide range of generalized linear models. The following SAS statements use PROC GENMOD to fit the Poisson regression

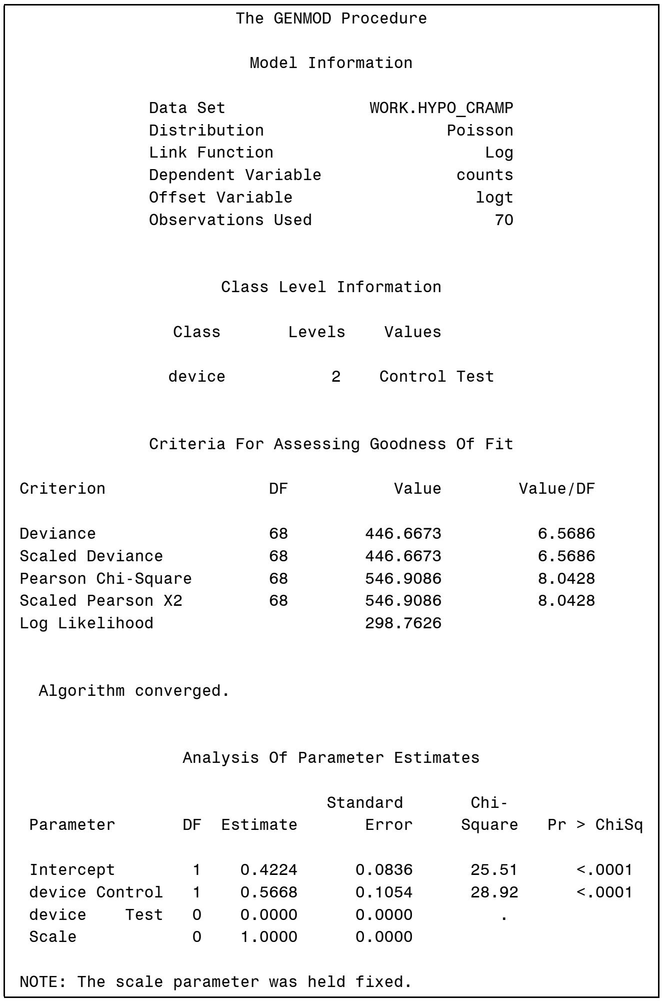
$$\log(\mu_i) = \log t_i + \beta_0 + \beta_1 \,\text{device}_i$$

to the HYPO_CRAMP data with DEVICE as the explanatory variable:

```
proc genmod data=Hypo_Cramp;
class device;
model count=device/ offset=logt dist=poisson
link=log;
run;
```

Here DIST= option specifies Poisson distribution, LINK= option specifies log-linear regression model (which is default for the Poisson distribution and can be omitted) and LOGT is an offset variable, which was defined in the data step. An intercept term (β_0) is included by default in the regression equation.

The output from these statements is displayed in Figure 1.



```
                         The GENMOD Procedure

                          Model Information

               Data Set                    WORK.HYPO_CRAMP
               Distribution                        Poisson
               Link Function                           Log
               Dependent Variable                   counts
               Offset Variable                        logt
               Observations Used                        70


                       Class Level Information

                    Class        Levels    Values

                    device            2    Control Test


               Criteria For Assessing Goodness Of Fit

 Criterion                 DF           Value        Value/DF

 Deviance                  68        446.6673          6.5686
 Scaled Deviance           68        446.6673          6.5686
 Pearson Chi-Square        68        546.9086          8.0428
 Scaled Pearson X2         68        546.9086          8.0428
 Log Likelihood                      298.7626


   Algorithm converged.


                  Analysis Of Parameter Estimates

                                 Standard       Chi-
  Parameter       DF   Estimate     Error     Square    Pr > ChiSq

  Intercept        1     0.4224    0.0836      25.51      <.0001
  device Control   1     0.5668    0.1054      28.92      <.0001
  device    Test   0     0.0000    0.0000        .
  Scale            0     1.0000    0.0000

 NOTE: The scale parameter was held fixed.
```

Figure 1. Output from Poisson regression

From ‘Analysis Of Parameter Estimates' table of output, we can see that variable DEVICE is highly significant (p<.0001), with higher rate of hypotensions and crampings for the control group as compared to the test group. But the ‘Criteria For Assessing Goodness Of Fit' section of output suggests that, because value/df for both deviance and Pearson Chi-Square statistics is much higher than 1, Poisson model is not quite adequate to describe the counts of hypotensions and crampings. It also suggests that there is a greater variability among counts than would be expected for Poisson distribution. Such extra-variability usually arises because the repeated events on a subject not may be independent. This is called overdispersion. One of the most common reason for data being over-dispersed is that experimental conditions are not perfectly under control and thus the unknown μ_i parameters vary not only with measured covariates but with latent and uncontrolled factors.

OVERDISPERSION

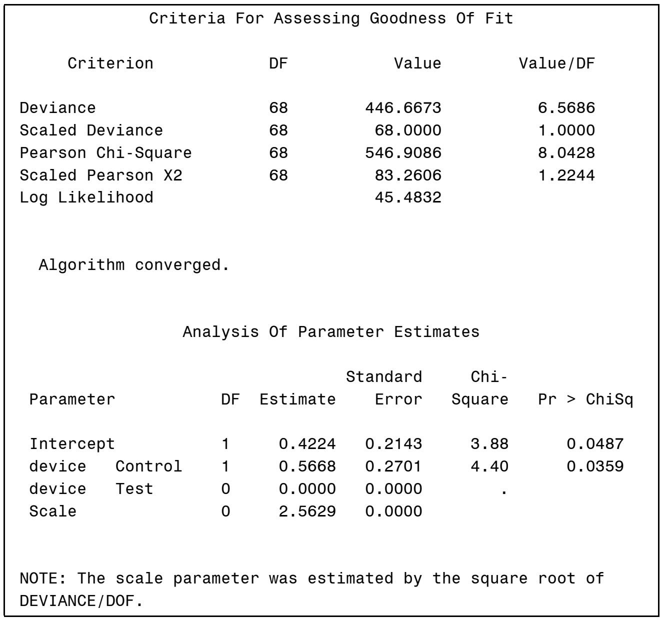
It is possible to account for overdispersion with respect to the Poisson model by introducing a dispersion parameter ϕ into the relationship between the variance and the mean

$Var(Y_i)=\phi \mu$

This method based on a quasi-likelihood approach, which permits estimation of parameters and inferential testing without full knowledge of the probability distribution of the data (Wedderburn, 1974, McCullagh and Nelder, 1989).
The scale parameter in the ‘Analysis Of Parameter Estimates' table of output is equal $\sqrt{\phi}$. When ϕ=1 we have the ordinary Poisson model (scale is fixed to 1 in the Figure 1), and when ϕ>1 we have the overdispersed Poisson model. The introduction of the dispersion parameter, however, does not introduce a new probability distribution, but just gives a correction term for testing the parameter estimates under the Poisson model. The models are fit in the usual way, and the parameter estimates are not affected by the value of ϕ, but the estimated covariance matrix is inflated by this factor. This method produces an appropriate inference if overdispersion is modest (Cox, 1983) and it has become the conventional approach in Poisson regression analysis.
McCulagh and Nelder (1989) suggested to estimate the dispersion parameter ϕ as a ratio of the deviance or the Pearson Chi-Square to its associated degrees of freedom. SAS implemented this approach by introducing an option SCALE= in the model statement of PROC GENMOD. One can estimate dispersion (scale) parameter by either specifying SCALE=DEVIANCE (=D, or just DSCALE) or SCALE=PEARSON (=P, or just PSCALE) and then appropriately adjust standard errors of regression coefficients. For example, the code below uses deviance to account for overdispersion in the Hypo_Cramp data

```
proc genmod data=Hypo_Cramp;
class device;
model count=device/ offset=logt dist=poisson
link=log dscale;
run;
```

The results of the fitting the model are displayed in Figure 2.



```
                Criteria For Assessing Goodness Of Fit

     Criterion              DF          Value        Value/DF

Deviance                    68       446.6673          6.5686
Scaled Deviance             68        68.0000          1.0000
Pearson Chi-Square          68       546.9086          8.0428
Scaled Pearson X2           68        83.2606          1.2244
Log Likelihood                        45.4832


  Algorithm converged.


                   Analysis Of Parameter Estimates

                                      Standard      Chi-
 Parameter            DF   Estimate      Error    Square    Pr > ChiSq

 Intercept             1     0.4224     0.2143      3.88        0.0487
 device    Control     1     0.5668     0.2701      4.40        0.0359
 device    Test        0     0.0000     0.0000        .
 Scale                 0     2.5629     0.0000


NOTE: The scale parameter was estimated by the square root of
DEVIANCE/DOF.
```

Figure 2. Results of Poisson regression, corrected for overdispersion

We can see from this output that the scaled deviance is now held fixed to 1 and scale parameter ($\sqrt{\phi}$) is estimated as 2.5629

(= $\sqrt{\frac{Deviance}{df}} = \sqrt{6.5686}$). The parameter estimates for intercept and treatment have not been changed, but their standard errors are now inflated by the value of the scale parameter and although the treatment effect is still significant (p=0.0359), the confidence intervals are much wider, p-values are now much higher and significance tests are more conservative than those based on the Poisson distribution before adjustment for overdispersion.

NEGATIVE BINOMIAL REGRESSION

Another count model, which allows for overdispersion, is the negative binomial model (NB). The negative binomial distribution can be derived from the Poisson when the mean parameter is not identical for all members of the population, but itself is distributed with gamma distribution. In other words, the occurrence of hypotensions and crampings in each patient is a Poisson process with its own parameter μ_i, but the joint distribution of these Poisson processes is no longer Poisson. Thus NB distribution provides one way of modeling heterogeneity in a population. The relationship between variance and mean for NB distribution has the form

$$Var(Y_i)= \mu + k\,\mu^2$$

where k is an additional distribution parameter that must be estimated or set to a fixed value. The NB model is only an exponential family when k is known.

One important characteristic of the NB distribution is that it naturally accounts for overdispersion due to its variance is always greater (k>0) than the variance of a Poisson distribution with the same mean μ.

For this reason the NB model has greater flexibility in modeling the relationship between the expected value and the variance of Y_i than the highly restrictive Poisson model. Note that, for small k, the NB model approaches the Poisson model.

Although NB distribution is not in the exponential family, starting from version 7 of SAS, PROC GENMOD includes the possibility to run NB regression, by specifying option DIST=NB in the model statement. SAS is accounting for possible “residual” overdispersion by including additional scale parameter ϕ to the relationship $Var(Y_i)=\phi(\mu + k\,\mu^2)$.

The NB regression model for Hypo_Cramp data is produced in Figure 3.

```
                      The GENMOD Procedure

                        Model Information

           Data Set                  WORK.HYPO_CRAMP
           Distribution            Negative Binomial
           Link Function                         Log
           Dependent Variable                 counts
           Offset Variable                      logt
           Observations Used                      70


                     Class Level Information

               Class        Levels     Values

              device            2    Control Test
```

Figure 3. Negative binomial regression results

```
                 Criteria For Assessing Goodness Of Fit

  Criterion                 DF           Value        Value/DF
  Deviance                  68         79.1674          1.1642
  Scaled Deviance           68         68.0000          1.0000
  Pearson Chi-Square        68         77.1031          1.1339
  Scaled Pearson X2         68         66.2269          0.9739
  Log Likelihood                      365.5844

  Algorithm converged.

                   Analysis Of Parameter Estimates

                                    Standard    Chi-
Parameter          DF    Estimate      Error   Square    Pr > ChiSq

Intercept           1      0.4923     0.2168     5.16        0.0232
device     Control  1      0.5162     0.3068     2.83        0.0925
device        Test  0      0.0000     0.0000      .
Dispersion          1      1.1731     0.2643

NOTE: The covariance matrix was multiplied by a factor of
DEVIANCE/DOF.
```

Figure 3. (continued)

We can see that in the case of NB regression, in the ’Analysis Of Parameter Estimates’ table PROC GENMOD reports the dispersion parameter k (=1.1731), instead of scale parameter ($\sqrt{\phi}$) for the ordinary and overdispersed Poisson regressions as it is shown in Figures 1 and 2. The dispersion parameter can be set to a fixed value, by using both NOSCALE and SCALE=’number’ options in the model statement.

From ‘Analysis Of Parameter Estimates’ table we can see that the treatment effect became non-significant, (p=0.0925) which reflects a combination of a decrease of the value of the parameter estimate for DEVICE variable and the increase of its standard error. From ‘Criteria For Assessing Goodness Of Fit’ table we can see that the NB model fits the data very well (the deviance is 79.1674 with 68 degrees of freedom) and almost no over-dispersion is seen (φ=1.1642), compared to the ordinary Poisson model.

STATISTICAL INFERENCE

For visual evaluation of the fit, the estimated cumulative probability distributions for Poisson and NB models can be compared to the observed one. Figure 4 clearly shows that the NB model catches the features of the Hypo_Cramp data, whereas the Poisson model is inferior.

Comparison of p-values from outputs 1-3 suggests that our conclusions about the associations between the treatment and rates of hypotensions and crampings are greatly affected by the choice of the model. Clearly, ignoring over-dispersion in the analysis would lead to underestimation of standard errors, and consequent over-statement of significance in hypothesis testing. Thus we can conclude that using inappropriate model for count data can dramatically change a statistical inference. The overdispersion must be accounted for by the analysis methods appropriate to the data. In the particular case of Hypo_Cramp data the Goodness-of-Fit test suggests that the NB model, provides a better account of the probability distribution of the individual responses, than the simple Poisson model or Poisson model with correction for overdispersion.

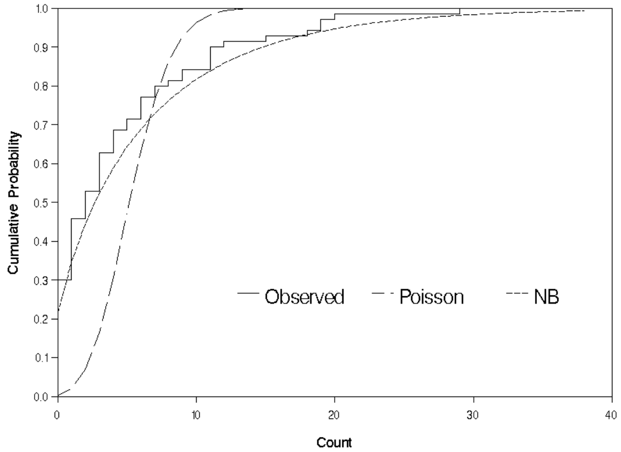


Figure 4. The observed and estimated cumulative distribution function for Hypo_Cramp data

TREND OVER TIME

There is big temptation in a longitudinal study to summarize the results in a count. This could describe the question of interest well. But it discards all available information about reason of overdispersion, or even about any simple trend in the rate of events. If the timing of each event is recorded, then, such a trend could be taken into account, by choosing an appropriate time unit for observation and to count the number of events by each unit in each successive period. The idea is to select this time interval small enough to reflect a possible time trend. In this way, each subject provides a series of counts that can be analyzed by Poisson or NB regressions with some function of time as one of the covariate. Correlation between consecutive counts on a subject could be taken into account by using the GEE approach (see, Diggle, Liang and Zeger (1994)). In PROC GENMOD such analysis could be done, by invoking a REPEATED statement with a correspondingly specified covariance structure of the correlated counts.

MIXED MODEL

The mixed model with count data could be fit by using either the %GLIMMIX macro or the PROCEDURE NLMIXED.

%GLIMMIX is described in Little, Milliken, Stroup and Wolfinger (1996) and its latest version could be found at the SAS web site. Unfortunately, it does not support negative binomial regression yet.

The PROC NLMIXED can fit both Poisson and, starting from SAS version 8.1, NB models. Example 46.4 of the SAS/STAT User's Guide, Version 8 (1999) describes how to fit the Poisson model. The PROC NLMIXED statements to fit NB model are as follows

```
   proc nlmixed data=Hypo_Cramp df=67;
   parms beta0 -3 beta1 0.5 log_sig 0.8;
   group=(device='Test');
   log_mu=beta0 + beta1*(1-group) + logt + u;
   p=exp(-log_mu)/(exp(-log_mu) + 1);
   model count ~ negbin(1,p);
   random u~normal(0,exp(2*log_sig)) subject=id;
   run;
```

The PARMS statement identifies the unknown parameters and their starting values. The next two statements are SAS programming statements specifying the non-linear model and the MODEL statement defines the dependent variable and its conditional NB distribution given the random effect. There are two undocumented features, which should be mentioned here. First, p should be expressed as a probability and not just as an exp(-log_mu), which should be expected for log-linear model. Second, negbin(1,p) defines the NB distribution with fixed dispersion parameter *k* =1. No other value of *k* is allowed at the present time.

The RANDOM statement defines the single random effect to be u, and the SUBJECT=id defines the clustering variable. The only distribution currently available for the random effect is normal(*m*,*v*), with mean *m* and variance *v*.

Figure 5 shows the part of the output from PROC NLMIXED. From 'Parameter Estimates' table we can see that in the case of NB regression with dispersion parameter *k* =1 and normal random effect the treatment effect (beta1) is non-significant (p=0.0699).

```
                         The NLMIXED Procedure
                            Specifications

     Data Set                                   WORK.HYPO_CRAMP
     Dependent Variable                         count
     Distribution for Dependent Variable        Negative Binomial
     Random Effects                             u
     Distribution for Random Effects            Normal
     Subject Variable                           id
     Optimization Technique                     Dual Quasi-Newton
     Integration Method                         Adaptive Gaussian
                                                Quadrature


                          Parameter Estimates

                          Standard
  Parameter   Estimate       Error     DF   t Value     Pr > |t|

  beta0        -3.0384      0.2371     67    -12.82       <.0001
  beta1         0.5621      0.3051     67      1.84       0.0699
  log_sig      -0.7947      0.5840     67     -1.36       0.1781
```

Figure 4. Output from PROC NLMIXED

SPARSE DATA

If the total number of the events in any group is small the large-sample inference, based on maximizing of the likelihood function or use quasi-likelihood equation, may not be valid. In those cases, exact Poisson regression is a better approach to get regression estimates and p-values that are statistically valid. In this method, an exact inference is based on the permutation distribution of the sufficient statistic for β (regression coefficient for device variable), unlike asymptotic inference, which is based on a large sample distribution of estimated $\hat{\beta}$. LogXact-4.1® software, Cytel Software Corporation (2000), provides this capability currently.

The special case is zero number of events in the both treatment groups. In this case PROC GENMOD could exit abnormally. Even if it does not exit abnormally PROC GENMOD will not estimate parameters in the case of Poisson regression and SAS will issue the following messages in the LOG window:

```
WARNING: The specified model did not converge.
ERROR: The mean parameter is either invalid or at a limit of its
range for some observations.
```

In the case of Negative Binomial regression parameters are estimated, but they should be disregarded because of the inability for model to converge and the mentioned above inappropriateness of the large-sample inference in this case. SAS will issue the following messages:

```
WARNING: The negative of the Hessian is not positive definite.
The convergence is questionable.
WARNING: The procedure is continuing but the validity of the
model fit is questionable.
WARNING: The specified model did not converge.
WARNING: Negative of Hessian not positive definite.
```

COUNTS MACRO

The %COUNTS macro for SAS is found in the Appendix. The macro produces tables of event rates and statistics summarizing results of a two-group comparison. It provides a skeletal structure which then can be used in a clinical trial setting. It may easily be expanded to provide more extensive and sophisticated output. The %COUNTS program is structured as follows:

1. PROC MEANS is used to calculate rates of count events for two groups and create a corresponding data set. Then in the case of zero counts in both of the groups CALL SYMPUT is used to create a global character macro variable &cancel with the value ‘CANCEL’ .
2. PROC GENMOD is used to calculate appropriate p-value for a two-group comparison. Output delivery System (ODS) then is used to output this p-value (p_value variable) to the data set ‘pvalue’. In the case of zero counts in both of the groups, the macro variable &cancel is used to cancel PROC GENMOD run (to avoid possible abnormal exit of PROC GENMOD) and to a create data set ‘pvalue’ with a missing value of the p_value variable.
3. combine rates and p-value in one observation
4. append this observation to the previously created data set
5. export resulting data set to the excel spreadsheet for easy formatting and incorporating to the report

To use the macro, first to invoke the code (e.g., by using %include statement or just copy the entire macro to your SAS program) and then issue

%counts(dataset,outcome,dist,scale,offset);

with all the parameters properly substituted. For example to use NB regression to compare rates of hypotension and cramping in the test and control groups from data set Hypo_Cramp we will need to submit the next statement

%counts(Hypo_Cramp,count,nb,dscale,logt);

Example of the 5-fold application of the macro for 5 different adverse events is shown below.

```
                   Test                Control          p-value

Fever      0.94(128/136.6)      0.86(134/155.9)      0.8687
Pain        0.12(17/136.6)       0.17(26/155.9)      0.5841
Edema       0.27(37/136.6)       0.38(60/155.9)      0.3119
Malaise     0.31(43/136.6)       0.28(43/155.9)      0.7836
Nausea      0.10(14/136.6)       0.23(36/155.9)      0.1456
```

APPENDIX

%macro ***counts***(dataset,outcome,dist,scale,offset);

```
ods listing close; ** Turn off output **;

proc means data=&dataset noprint;
var &outcome.;
output out=check00 sum=&outcome.;
run;

%let cancel=;
data _null_ pvalue (drop=&outcome.);
set check00;

if &outcome.^=0 then output _null_ ;
if &outcome.=0 then do;

call symput('CANCEL','CANCEL');

P_value=.;

output pvalue ;
end;
run;

ods output
ParameterEstimates=pvalue(where=(Parameter='device' and
DF=1) rename=(probchisq=P_value));
proc genmod data=&dataset. ;
    class device;
    model &outcome = device /maxiter=1000 dist=&dist. &scale.
offset=&offset.;
run &cancel.;

proc means data=&dataset noprint; by group;
   var &varname MonthatRisk;
   output out=countsum sum=&varname Time_At_Risk;
run;

data &outcome._out;
retain outcome Test Control Pvalue;
length outcome $15 Pvalue $13;
merge
countsum(where=(group=2) rename=(&outcome =Total2
Time_At_Risk=TimeAtRisk2))
countsum(where=(group=1) rename=(&outcome =Total1
Time_At_Risk=TimeAtRisk1))
pvalue (Keep=P_value);

Pvalue=substr(input(P_value,$13.),1,6);

if Total1=. then Total1=0;
if Total2=. then Total2=0;

Variable="&outcome.";

if .<P_value<0.0001 then Pvalue='<.0001';
if P_value=1 then Pvalue='1.0000';

EventsMonth1=round(Total1/TimeAtRisk1,0.01);
EventsMonth2=round(Total2/TimeAtRisk2,0.01);
TimeAtRisk1=round(TimeAtRisk1,0.1);
TimeAtRisk2=round(TimeAtRisk2,0.1);
Control=compress(EventsMonth1||'('||Total1||'/'||TimeAtRisk1||')');
Test=compress(EventsMonth2||'('||Total2||'/'||TimeAtRisk2||')');

keep Variable Test Control Pvalue;

proc append force base=countsall data=&outcome._out;
proc export data=&outcome._out
     outfile="C:\Analysis\counts.xls"
     dbms=excel2000 replace;
run;

proc datasets;
delete &outcome._out pvalue check00;

quit;

ods listing;

%mend counts;
```

TRADEMARKS

SAS and SAS/STAT are registered trademarks or trademarks of SAS Institute Inc in the USA and other countries. LifeSite, Vasca and Vasca, Inc are registered trademarks of Vasca Inc. and/or its affiliates. ® indicates USA registration.
Other brand and product names are registered trademarks of their respective companies.

REFERENCES

Cox D. R. (1983), “Some Remarks on Overdispersion. ,” *Biometrika*, 70, 269-274.

Diggle, P.J.., Liang, K.Y., and Zeger, S.L. (1994), *Analysis of Longitudinal Data*, Oxford:Claredon Press.

Gardner, W., Mulvey E.P. and Shaw E.C. (1995), “Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models,” *Psyhological Bulletin*, 118, 392-404.

Glynn, R.J. and Buring J.E. (1996), “Ways of Measuring Rates of Recurrent Events,” BMJ, 312, 364-367.

Little, R.C., Milliken, G.A., Stroup, W.W. and Wolfinger R.D. (1996), *SAS System for Mixed Models*, Cary: SAS Institute, Inc.

McCullagh, P. and Nelder, J.A. (1989), *Generalized Linear Models*, Second Edition, Longon: Chapman and Hall.

SAS Institute Inc (1999). SAS/STAT User’s Guide, Version 8, Cary, NC: SAS Institute Inc.

Wedderburn, R. W. M. (1974), “Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method,” *Biometrika*, 61, 439-447.

CONTACT INFORMATION

Your comments and questions are valued and encouraged.
Contact the author at:

Alex Pedan, Ph.D.
Vasca Inc.
3 Highwood Drive
Tewksbury, MA 01876
Work Phone: 978-863-4442
Fax: 978-863-4401
Email: apedan@vasca.com
Web: www.vasca.com