Offset Techniques for Predictive Modeling for Insurance
Matthew Flynn, Ph.D, ISO Innovative Analytics, W. Hartford CT
Jun Yan, Ph.D, Deloitte & Touche LLP, Hartford CT

ABSTRACT
This paper presents the “how-to” of the application of several “offset” techniques that can be used to build Generalized Linear Models. Modelers can use techniques to eliminate the impact of the variables that will not be included in the model but will influence modeling results. In the paper, we give some detailed SAS® codes for the techniques.

More theoretical background and Insurance applications are discussed in a paper to be presented at the 2007 CAS Predictive Modeling Seminar and accepted for the 2009 Casualty Actuarial Society Ratemaking Discussion Paper Program.

INTRODUCTION
Property and casualty insurance is a complex and dynamic business. There exist many risk factors that affect the outcomes of the business. For example, a typical automobile insurance rating plan contains more than 20 variables, including a wide range of driver, vehicle, and territorial characteristics. In addition, the business is greatly influenced by the dynamic changes from the external economic and legal environment, i.e. underwriting cycle.

Due to the dynamic and complex nature associated with the business, it is practically impossible for modelers to conduct “comprehensive” multivariate modeling by including all the possible variables. Such issue will become even worse if the modelers deal with highly non-ideal insurance data, such as low volume of data or dirty data.

The major risk when certain important factors cannot be included in a multivariate modeling analysis is that the analysis results can be highly “biased”. Commonly known factors which will bias the property and insurance predictive modeling results include underwriting cycle and external environmental change (i.e., year), loss maturity, state/territory, exposure, and distribution channel, to name a few.

Offset techniques are simple, straightforward, and versatile in dealing with data bias issue. Intuitively, it is a simple method used to run a model against the residual of a set of selected factors, of which the modelers believe that they will influence the modeling results but will not be included in the models.

A SIMPLE OFFSET EXAMPLE AND ITS SAS CODES
Offsets can be used in a number of ways. This section will demonstrate a simple example - on a “count” basis instead of on a “rate” basis using an offset statement in a Poisson GLM.

Suppose we are building a model to estimate claim frequency of personal auto Comp coverage. The claim frequency is defined as claim count over adjusted exposure (adjexp), where the adjusted exposure is the exposure after removing the bias of territory and vehicle symbol as we described in Introduction.

The GLM frequency model is usually set up as:

$$\log(\text{claim}_{\text{count}}/\text{adj exp}) = f(X).$$

Where “log” is the link function, $f(X)$ is a linear combination of the selected predictive variables. The above equation is equivalent to:

$$\log(\text{claim}_{\text{count}}) - \log(\text{adj exp}) = f(X).$$

$$\log(\text{claim}_{\text{count}}) = f(X) + \log(\text{adj exp})$$
This shows why one can build up the frequency model on a count basis using the log of exposure as the "offset" when coding Proc GENMOD in SAS.

Suppose there are three predictive variables, policy type (type), driver age group (driver_age_group) and vehicle use (Vehicle_Use) Let's further define ladjexp= log(adjexp), then the above GLM model with the offset factor can be coded in SAS as follows:

**MODEL 1**

```sas
proc genmod data=indata;
   class type driver_age_group Vehicle_Use;
   model claim_count = type driver_age_group Vehicle_Use / dist=poisson link=log offset=ladjexp;
run;
```

**OFFSET TECHNIQUE FOR CONSTRAINED, MULTIPLE FACTORS ANALYSIS**

In this section, we will give another example by expanding the technique to multiple factors with constraints. The constraint issue can arise due to market competitions or regulation. For example, despite what the analysis results suggest, the insurance market will constrain the discount for multi-car or home-auto package policies no more than 20%. Another example is that the insurance regulation will suppress the rates for youthful drivers or certain disadvantage territories.

Same as the examples used in the previous section, suppose we are building a model to estimate claim frequency of personal auto Comp coverage. Same two predictive variables, policy type (type) and driver age group (driver_age_group) are selected to predict claim frequency that is defined as claim count over adjusted exposure (adjexp).

As we described in Introduction, there are four values for "driver_age_group", from 1 to 4, and there are two values for "type", "S" and "M". In the process, the two variables of age and type can be transformed into 6 separate binary variables.

Let us further assume that the analysis is to focus on the relativity for "driver_age_group" 1 and 2, and for "type" S and M. The relativity for "driver_age_group" 3 and 4 will not be included, and they are assumed to be 1.05 and 1.25, respectively.

The following section is the SAS code for the computation.

**MODEL 2**

```sas
data freq_data;
set input;
claim_freq=claim_count / exposure;
offset_factor=1;
   if driver_age_group =3 then offset_factor=1.05;
   if driver_age_group =4 then offset_factor=1.25;
   logoffset=log(offset_factor);
   if driver_age_group in (1,2) then driver_age_group_new= driver_age_group;
   else driver_age_group_new =9;
run;

proc genmod data=freq_data;
   class driver_age_group_new type;
   model claim_freq = driver_age_group_new type / dist=poisson link=log offset=logoffset;
run;
```
The final frequency relativity table should be based on both of the restricted values and the model estimates after taking the inverse link function which is exponential.

**TWEEDIE COMPOUND POISSON MODELS**

For insurance ratemaking predictive modeling, pure premium, loss cost per exposure, is a frequently used target variable. Pure premium is compound with claim frequency and claim severity. In general, claim frequency follows Poisson distribution and claim severity follows Gamma distribution, so the most proper distribution for pure premium is Tweedie, a Poisson and Gamma compound distribution.

Unfortunately, the current SAS statistical package does not cover Tweedie model, many readers may not understand the Tweedie distribution well and how they can be applied in SAS. Therefore, below, we show how the SAS codes can be used for the Tweedie model with Proc NLMIXED, where the offset option we described above is available.

The parameter “p” in “Tweedie” is chosen between 1 and 2 based on input data and the target variable.

**Tweedie Compound Poisson Model and Corresponding SAS Code**

Following Smyth & Jorgenson (2002), section 4.1, page 11, the Tweedie Compound Poisson joint likelihood function defined as:

\[
f(n,y; \varphi/w,p) = a(n,y; \varphi/w,p) \exp \left\{ \frac{\alpha}{\varphi} t(y,\mu,p) \right\}
\]

with

\[
a(n,y; \varphi/w,p) = \left\{ \left( \frac{w/\varphi}{p-1} \right)^{\alpha+1} y^{\alpha} \right\}^{n} \frac{1}{n! \Gamma(n\alpha)y}
\]

and

\[
t(y,\mu,p) = y \frac{\mu^{1-p}}{1-p} - \frac{\mu^{2-p}}{2-p}.
\]

The SAS codes using “Proc NLMIXED” to fit the above maximum likelihood regression model for the cross coverage tiering score example is as follows:

```sas
proc nlmixed data=the_appended_dataset;
   parms p=1.5;
   bounds 1<p<2;
   eta_mu = b0 +
            c1*(credit_grp=1)+c2*(credit_grp=2)+c3*(credit_grp=3)+c4*(credit_grp=4)
            + naf1*(naf_pol=1)+naf2*(naf_pol=2)
            + coverage_COMP*(coverage='COMP');
   mu = exp(eta_mu + current_factor);

   eta_phi = phi0 +
             phi_c1*(credit_grp=1)+ phi_c2*(credit_grp=2)+ phi_c3*(credit_grp=3)+
             phi_c4*(credit_grp=4)+ phi_naf1*(naf_pol=1)+ phi_naf2*(naf_pol=2)
             + phi_coverage_COMP*(coverage='COMP');
   phi = exp(eta_phi);
   n = claims;
```

Statistics and Data Analysis
NESUG 2007
w = insured;
y = pp;

$$t = \frac{((y\cdot\mu^2 \cdot (1 - p))}{(1 - p))} - \frac{((\mu^2 \cdot (2 - p))}{(2 - p))};$$

$$a = \frac{(2 - p)}{(p - 1)};$$

if (n = 0) then
  loglike = \(\frac{w}{\phi}\) \(t\);
else
  loglike = \(n\cdot(a + 1)\ln\left(\frac{w}{\phi}\right) + a\cdot\ln(y) - a\cdot\ln(p - 1) - \ln(2 - p)\)
          \(= \ln\gamma(n + 1) - \ln\gamma(n\cdot a) - \ln(y) + \ln\gamma(n + 1) - \ln\gamma(n\cdot a) - \ln(y) + \ln\gamma(n + 1) - \ln\gamma(n\cdot a)
          - (\frac{w}{\phi})^t;$$

model \(y \sim \text{general(loaglike)}\);
replicate adjexp;
estimate 'p' p;
run;

The above codes can be broken down into the following major sections:

- First we call the Proc NLMIXED, addressing the desired input dataset:

  \texttt{proc nlmixed data=the_appended_dataset;}

  The PARMS statement provides a starting value for the algorithm's parameter search. Multiple starting values are allowed, as well as input from datasets (from prior model runs, for example). With some domain knowledge we anticipate this parameter to be in the neighborhood of 1.5.

  \texttt{parms p=1.5;}

  Parameters can also be easily restricted to ranges, such as to be positive, and here we require the estimated Tweedie power parameter to fall between one and two.

  \texttt{bounds 1<p<2;}

- Next we specify the linear model/predictor for the mean response. Proc NLMIXED does not have the convenient CLASS statement of some of the other regression routines, like Proc GENMOD or Proc LOGISTIC. However, the design matrix can be created "on-the-fly", so to speak, by effectively including programming statements in the Proc NLMIXED code. Here, we create dummy variables by coding the linear model with logical statements. For example, the phrase, \((\text{credit_grp}=1)\) resolves to either true (1) or false (0) at runtime, creating our desired indicator variables to test discrete levels of right-hand side variables. As a reminder, for a GLM, the linear predictor is required to be \textit{linear in the estimated parameters}, so non-linear effects such as powers of covariates or splines can be accommodated.

  \(\eta_{\mu} = b0 + c1\cdot(\text{credit_grp}=1) + c2\cdot(\text{credit_grp}=2) + c3\cdot(\text{credit_grp}=3)
  + c4\cdot(\text{credit_grp}=4) + naf1\cdot(\text{naf_pol}=1) + naf2\cdot(\text{naf_pol}=2)
  + \text{coverage_COMP}\cdot(\text{coverage}='\text{COMP}');\)

- Next we create a log link that maps the linear predictor to the mean response. That log link on the left hand side, becomes an exponential as the inverse link (on the right-hand side). The "current_factor" variable is one place where an offset can be used. Log(current_factor) could be include in the line above as part of the linear predictor.

  \(\mu = \exp(\eta_{\mu} + \text{current_factor});\)

- A great feature of using Proc NLMIXED is it's flexibility. Here we are specifying what Smyth & Jorgenson (2002) refer to as a double GLM. Instead of a single constant dispersion constant as in the code block above, below we can fit an entire second linear model with log link for the dispersion factor.

  \(\eta_{\phi} = \phi0 + \phi_{c1}\cdot(\text{credit_grp}=1) + \phi_{c2}\cdot(\text{credit_grp}=2) + \phi_{c3}\cdot(\text{credit_grp}=3) + \phi_{c4}\cdot(\text{credit_grp}=4) + \phi_{naf1}\cdot(\text{naf_pol}=1) + \phi_{naf2}\cdot(\text{naf_pol}=2) + \phi_{comp}\cdot(\text{coverage}='\text{COMP}');\)
phi_c3*(credit_grp=3) + phi_c4*(credit_grp=4) +
phi_naf1*(naf_pol=1) + phi_naf2*(naf_pol=2) +
phi_coverage_COMP*(coverage='COMP');

phi = exp(eta_phi);

• Proc NLMIXED allows a number of datastep style programming statements. Here we are assigning input
dataset variables claims, insured, and pp as new variables (n, w and y) to be used subsequently in building
out our likelihood equation. That way, one can easily adapt pre-existing code to a particular input dataset,
without requiring modifications to the “guts” of the log-likelihood equation (it is complicated enough already).

n = claims;
w = insured;
y = pp;

• Now one can begin to specify the loglikelihood. Here, for clarity, we build it out in several steps. Simply
refer to the Tweedie Compound Poisson likelihood described above from Smyth & Jorgensen (2002), and
lay it out.

\[ t = \frac{(y * mu^*(1 - p))}{(1 - p)} - \frac{(mu^*(2 - p))}{(2 - p)}; \]
\[ a = \frac{(2 - p)}{(p - 1)}; \]
\[ \text{if } (n = 0) \text{ then } \loglike = \frac{(w/\phi)*t}; \]
\[ \text{else } \loglike = n*(a + 1)*\log(w/\phi) + a*\log(y) - a*\log(p - 1) - \log(2 - p) - \loggamma(n + 1) - \loggamma(n*a) - \log(y) + (w/\phi)*t; \]

Proc NLMIXED includes several pre-specified likelihoods, for example, Poisson and Gamma, the GENERAL
specification allows the great flexibility to specify one’s desired model specification.

model y ~ general(loglike);

Weights can either be included directly in the loglikelihood above, or with the handy REPLICATE statement.
Each input record in the dataset represents an amount represented by the input variable “adjexp”.

replicate adjexp;

The ESTIMATE statement can easily calculate and report a variety of desired statistics from one’s model
estimation. Here, we are interested in the Tweedie Power parameter.

estimate 'p' p;

Without using any of the additional “Mixed” modeling power, Proc NLMIXED performs as a great Maximum
Likelihood Estimator using a variety of numeric integration techniques.

The following graph displays Tweedie Compound Poisson densities for varying values of the power parameter, "p".
As P varies between one (1) and two (2), the unusual characteristic of the Tweedie density, the spike at zero grows
larger.

The graphs below were generated in the sample Program Tweedie_densities.sas. Individual graphs were gathered
together using Proc GREPLAY. (Thanks to the help of the good folk @ SAS Tech Support).
How to best estimate the appropriate value for the power parameter? We can query the data. The following graphs generated in Tweedie_mean_var.sas grouped the data into bins, then plotted the mean-variance relationship within the bins – providing an estimate of 1.75 for the Tweedie power parameter. This can be one way to provide your Proc NLMIXED regression code starting values for estimation.
Alternatively, because the power parameter is defined over a limited range, one can in NLMIXED, provide a series of starting values across the range [1,2] of the parameter. The next two graphs plot profile log-likelihoods across that range, then focusing in to the narrow portion of the range.
Profile log-likelihood for the variance power. The horizontal line crosses the 95% likelihood interval.
For more information re: Proc NLMIXED: a great resource is a class offered by SAS Education titled: *Advanced Statistical Modeling Using the NLMIXED Procedure* (Live Web) See: [http://support.sas.com/training/us/crs/lwnlm.html](http://support.sas.com/training/us/crs/lwnlm.html)

**REFERENCES:**


**CONTACT INFORMATION**

Matthew Flynn | Jun Yan  
ISO Innovative Analytics | Advanced Quantitative Services  
545 Washington Blvd. | Deloitte Consulting LLP  
Jersey City, JC 07320-1686 | City Pl 185 Asylum St 33rd Fl  
Phone: (860) 570-0015 | Hartford, CT 06103-3402  
Fax: (201) 748-1900 | Phone (860) 725-3029  
Email: mflynn@iso.com | Email juyan@deloitte.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. © indicates USA registration.