

Adaptive Fractional Polynomial Modeling in SAS®

George J. Knafl, University of North Carolina at Chapel Hill

ABSTRACT

Regression predictors are usually entered into a model without transformation. However, it is not unusual for regression relationships to be distinctly nonlinear. Fractional polynomials account for nonlinearity through real-valued power transformations of primary predictors. Adaptive methods have been developed for searching through alternative fractional polynomials based on one or more primary predictors. A SAS macro called `genreg` (for general regression) is available from the author for conducting such analyses. It supports adaptive linear, logistic, and Poisson regression modeling of expected values and/or variances/dispersions in terms of fractional polynomials. Fractional polynomial models are compared using k-fold likelihood cross-validation scores and adaptively selected through heuristic search. The `genreg` macro supports adaptive modeling of both univariate and multivariate outcomes. It also supports adaptive moderation analyses based on geometric combinations, that is, products of transforms of primary predictors with possibly different powers, generalizing power transforms of interactions. Example analyses and code for conducting them are presented demonstrating adaptive fractional polynomial modeling.

INTRODUCTION

Regression predictors are usually entered into a model without transformation. However, it is not unusual for regression relationships to be distinctly nonlinear (examples are provided below). Fractional polynomials (Royston & Sauerbrei, 2008) account for nonlinearity through real-valued power transformations of primary predictors. Adaptive methods (Knafl et al., 2010; Knafl & Ding, in press) have been developed for searching through alternative fractional polynomials based on one or more primary predictors. A SAS macro called `genreg` for conducting such analyses and documentation for its parameters are available at <http://www.unc.edu/~gknafl/AFPM.html> (accessed 7/25/2015). It supports adaptive linear, logistic, and Poisson regression modeling of expected values and/or variances/dispersions in terms of fractional polynomials. The `genreg` macro supports adaptive modeling of both univariate and multivariate outcomes. Fractional polynomial models are compared using k-fold likelihood cross-validation (LCV) scores and adaptively selected through heuristic search based on those scores. The search process systematically expands the model by adding transforms of predictor variables and then contracts the expanded model by removing extraneous transforms, if any, and adjusting the powers of remaining transforms. LCV ratio tests, extensions of likelihood ratio tests, are used to determine if models are substantially better or worse. These are expressed in terms of the cutoff for a substantial (or distinct) percent decrease in the LCV score that changes with the sample size. If the smaller of two LCV scores generates a percent decrease compared to the larger score that is larger than the cutoff for the data, then the model with the larger score provides a substantial improvement over the model with the lower score. On the other hand, if the percent decrease is less than the cutoff, the model with the lower score is a competitive alternative. Moreover, if it is also less complex, then it is a parsimonious, competitive alternative, and so the preferable choice for modeling the data. An overview of the macro is presented and its use demonstrated through an adaptive analysis of data on mercury levels in fish (also available at the above URL).

The `genreg` macro only addresses power transformation of predictors. For a positive-valued continuous outcome, there can be a benefit to transforming that outcome as well as its predictors. A SAS macro called `ypower` is also available at the above URL for addressing this issue by computing, for each requested power p , the power-adjusted LCV score, denoted by $LCV(p)$, under the assumption that the power transform y^p of a positive-valued outcome y is normally distributed. The special case $p=0$ corresponds to the natural log transform. The `ypower` macro calls the `genreg` macro once for each requested power transform of y to generate the requested adaptive model for that transform and then computes its associated $LCV(p)$ score. A grid search can be conducted using `ypower` to identify an appropriate power transform for y .

All results were generated using SAS for Windows, version 9.4.

SETUP

Assume that the `genreg` code has been stored in the file called `genreg.20150720.sas` (version numbers are indicated by dates in the file name) located in the `c:\macros` folder (assuming the Windows operating system). This code can then be loaded with the following `%include` command.

```
%include "c:\macros\genreg.20150720.sas";
```

Assume that output has been formatted with the following options command.

```
options linesize=76 pagesize=53 pageno=1 nodate;
```

Assume as well that the mercury level data have been loaded into a SAS data set called MERCURY in the SAS default library. These data consist of variables WEIGHT (with weights in kg) and MERCURY (with mercury levels in parts per million (ppm)) for a total of $m=169$ measurements. Fish were caught in either the Lumber or Waccamaw Rivers in North Carolina. So there is also an indicator variable RIVER for being caught in the Waccamaw River (i.e., $RIVER=1$ if caught in the Waccamaw River and $RIVER=0$ if caught in the Lumber River). The data contain as well a variable LENGTH (with lengths in cm), but that is not used in reported analyses. The cutoff for a substantial percent decrease in LCV scores for these data with $m=169$ measurements is 1.13% (generated as part of the genreg output).

STANDARD REGRESSION MODELING

A linear regression model with expected values linear in WEIGHT and constant variances can be generated using the following command.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, xvars=WEIGHT, foldcnt=5);
```

The modtype parameter indicates the type of likelihood to be used. The setting "modtype=NORML" is used for linear regression analyses with likelihoods based on the normal distribution. The other possible settings are "modtype=LOGIS" for logistic regression and "modtype=P0ISS" for Poisson regression. The name of the source data set is provided by the datain parameter and the name of the outcome (response, dependent) variable by the yvar parameter. While parameter settings in the example code are in upper case, settings are not case sensitive.

The base model for the expected values in this case contains an intercept parameter since the default value for the xintrcpt parameter is "Y" (for yes), as opposed to "N" for a zero-intercept model. The xvars parameter setting requests that this model also includes the untransformed predictor WEIGHT. The base model for the variances is the default constant (i.e., intercept-only) model corresponding to the default settings "vintrcpt=Y" and "vvars=" (i.e., an empty setting). The xpowers (vpowers) parameter can be used to request transformation of xvars (vvars) variables. For example, "xvars=WEIGHT" with "xpowers=0.5" requests a square root transform of WEIGHT while the standard degree 2 polynomial model in WEIGHT is requested with "xvars=WEIGHT WEIGHT" and "xpowers=1 2". The variable RIVER can be added to the xvars list to control for a main effect to RIVER.

The "foldcnt=5" setting requests a 5-fold LCV. In other words, measurements (corresponding to fish in this case) are randomly partitioned into 5 disjoint subsets called folds. Likelihoods for data in folds are computed using maximum likelihood parameter estimates based on data in complements of folds, multiplied together, and normalized by taking the m^{th} root where m is the total number of measurements ($m=169$ in this case). A fixed initial seed is used in all analyses so that LCV scores are based on the same set of folds. For reported analyses, the fold sizes range from 27 to 38 so that parameter estimates for LCV scores are computed with subsets of the data of at least 131 or 77.5% of the 169 measurements. Cross-validation scores are based on likelihoods for all cases with "modtype=NORML", but can be based on likelihood-like functions in certain other situations. For example, logistic and Poisson regression models for univariate outcomes are based on likelihoods only if the dispersions (generalizing the variances of the linear regression case) are equal to 1, which is requested with "vintrcpt=N" and "vvars=" (meaning model the log of the dispersions as equal to zero). For non-unit dispersion models, the cross-validation scores are computed with extended quasi-likelihood functions (McCullagh & Nelder, 1999). The LCV score generated by the above code is 0.37343 as given in the genreg output.

ADAPTIVE REGRESSION MODELING

An adaptive model for expected values of mercury levels in terms of weight of the fish can be requested as follows.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, expand=Y,
        expxvars=WEIGHT, contract=Y);
```

An expansion is requested by "expand=Y", in this case with "expxvars=WEIGHT" specifying the one variable WEIGHT to consider for expansion of the model for the expected values. The base model for the search in this case is the default constant model (i.e., corresponding to the default settings of "xintrcpt=Y" and "xvars="), which can be changed if desired. An example is provided later. By default, "expttrans=Y" meaning consider power transforms of the expansion variables (as opposed to "expttrans=N" for generating linear models) and "multtrns=Y" meaning consider inclusion of multiple transforms of expansion variables (as opposed to "multtrns=N" for limiting the model to at most one transform per expansion variable). No variables are considered for expansion of the model for the variances since the expvvars parameter is not set and so has its default empty specification.

A contraction is requested by the setting "contract=Y" and always follows the expansion (unless "expand=N" and there is no expansion). Intercept terms of base models are considered for removal in the contraction, but all model terms in the base model for the expected values can be held fixed by setting "nocnxbas=Y". Alternately, the setting "nocnxint=Y" holds the intercept for the expected values fixed while allowing all other terms of the base model for the expected values to be considered for removal in the contraction. Terms added in the expansion are always considered for removal. By default, "cnret rns=Y", which requests retransformation of transforms remaining in the

model at each step of the contraction. When the contraction removes any transforms from the expanded model, the remaining transforms have their powers adjusted to improve the LCV score. However, when the contraction leaves the expanded model unchanged, it might be improved by having its powers adjusted. A conditional transformation occurs by default in such cases, due to the default setting "condtrns=Y".

The above macro invocation starts by generating a constant base model for expected values, expands it to include the two transform $\text{WEIGHT}^{0.2}$ and WEIGHT^{-10} , and then contracts this by first removing the second transform, then the intercept, and adjusts the remaining transform to $\text{WEIGHT}^{0.47}$. The 5-fold LCV score for the selected model is 0.38443. In comparison, the linear model in WEIGHT has LCV score 0.37343 (as reported above) with substantial percent decrease 2.86% (that is, larger than the cutoff of 1.13% for the data), indicating that expected mercury levels are distinctly nonlinear in WEIGHT.

The setting of the number of folds usually does not have much of an effect on the results. For the above analysis, the same model is generated with 10 folds as for 5 folds and almost the same model with power changed from 0.47 to 0.48 with 15 folds, suggesting that results for analyses of these data are likely to be reasonably robust to the setting of the number of folds. Moreover, the 5-fold LCV score is larger than the 10-fold and 15-fold scores, and so 5 folds are used in all subsequent analyses.

Parameter estimates are computed by genreg directly, solving associated equations using the matrix language of PROC IML, rather than by invoking standard SAS procedures. Results for the associated SAS procedure, in this case PROC REG (and PROC LOGISTIC for binary outcomes and PROC GENMOD for count outcomes), can be requested by the setting "procmod=Y". This can be used to generate p-values for tests of zero coefficients. However, those p-values are of questionable value for adaptively generated models. Coefficients for those models are typically significant as a consequence of the heuristic search process and often highly significant.

ADAPTIVE MODERATION

There is possibly an effect to RIVER on mercury levels. This can be addressed by adding RIVER to the expxvars list, but that only address a possible additive effect to RIVER not an interaction effect. The consideration of interactions addresses the issue called moderation (Baron & Kenny, 1986). A more general approach is to request consideration of geometric combinations using the setting "geomcmbn=Y" (overriding the default setting "geomcmbn=N") as in the following macro code.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, expand=Y,
        expxvars=WEIGHT RIVER, geomcmbn=Y, contract=Y);
```

In this case, the expanded model is based on the five transforms: $\text{WEIGHT}^{0.2}$, $\text{RIVER} \cdot \text{WEIGHT}^{1.4}$, $(\text{WEIGHT}^{-2} \cdot \text{RIVER})^{0.8}$, $\text{RIVER} \cdot \text{WEIGHT}$, and $(\text{RIVER} \cdot \text{WEIGHT}^2)^{-10}$ with LCV score 0.40775. Note that individual powers are first generated for the primary predictors in a geometric combination, but then the whole geometric combination can be adjusted using a single power rather than adjusting each of the individual powers. The contracted model is based on the two transforms: $\text{WEIGHT}^{0.36}$ and $\text{RIVER} \cdot \text{WEIGHT}^{1.4}$ without an intercept. The LCV score has the smaller value 0.40720, but the percent decrease compared to the score for the expanded model is insubstantial at 0.13%. Consequently, the contracted model is a parsimonious, competitive alternative to the expanded model with larger LCV score, but also with five transforms and an intercept compared to only two transforms without an intercept.

The above model contains a geometric combination, suggesting that there is substantial moderation of the effect of the weight of fish by the river in which they are caught. However, this is only the case if the model with geometric combinations provides a substantial improvement over the additive model in WEIGHT and RIVER. This model can be generated by changing to the setting "geomcmbn=N" in the above code (or removing "geomcmbn=Y" to request the default setting). The generated model is based on the two transforms $\text{WEIGHT}^{0.52}$ and RIVER without an intercept. Since its LCV score is 0.39114 with substantial percent decrease 3.94%, there is a substantial moderation effect.

The adaptively generated model based on geometric combinations can be generated directly as follows.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, xintrcpt=N,
        xvars=WEIGHT, xpowers=0.36, xgcs=river 1 weight 1.4);
```

The xgcs macro parameter is used to generate geometric combinations for modeling the expected values. Each geometric combination is determined by a list of primary predictors with associated powers (with $\text{RIVER} \cdot \text{WEIGHT}^{1.4}$ represented by "river 1 weight 1.4" in the above code). If there are two or more geometric combinations, their specifications are separated by colons (:). An example is provided later. The xgcpowers macro parameter assigns powers to use in transforming geometric combinations. The default setting "xgcpowers=" means do not transform any of the geometric combinations (or equivalently, transform them with default power 1). The vgcs and vgcpowers parameters are used in the same way to generate geometric combinations for modeling variances/dispersions.

The commonly used linear moderation model can be generated as follows.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, xvars=WEIGHT RIVER,
        xgcs=RIVER 1 WEIGHT 1, procmod=Y);
```

The setting "procmod=Y" is used to invoke PROC REG to generate significance tests for this model. The LCV score is 0.40182 with substantial percent decrease 1.32% compared to the adaptive moderation model. Consequently, the moderation effect is distinctly nonlinear. However, the interaction term in the linear moderation model is significant ($p < 0.001$), and so linear moderation also holds in this case.

ADAPTIVE VARIANCE MODELING

The genreg macro also supports adaptive modeling of the log of the variance (or the dispersion for logistic and Poisson regression) in terms of fractional polynomial models. At each step of the expansion, a transform is identified for expanding the expected value part of the model and also for the variance part of the model. The choice with the better LCV score is added next to the model. Similarly, at each step of the contraction, the next term removed is the one for either the expected value or the variance that generates the better LCV score. An adaptive model for both expected values and variances in WEIGHT, RIVER, and geometric combinations is requested as follows.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, expand=Y,
        expxvars=WEIGHT RIVER, expvvars=WEIGHT RIVER, geomcmbn=Y, contract=Y);
```

The generated model has expected values based on the three transforms: $WEIGHT^{0.26}$, $(RIVER \cdot WEIGHT^{1.5})^{0.8}$, and $(RIVER \cdot WEIGHT^{-1.3})^{0.7}$ without an intercept and variances based on the single transform: $WEIGHT^{0.53}$ also without an intercept. The LCV score is 0.43919, which is a substantial improvement over the associated constant variance model with LCV score 0.40720 (as reported above) and substantial percent decrease 7.28%. Consequently, the variances are distinctly non-constant in this case, indicating the importance of considering non-constant variances in general. The adaptive model with both expected values and variances depending only additively on WEIGHT and RIVER (using the above code with the revised setting "geomcmbn=N") has expected values based on $WEIGHT^{0.49}$ without an intercept, variances based on $WEIGHT^{-0.61}$ without an intercept, LCV score 0.41622, and substantial percent decrease 5.23%. Consequently, there is distinct moderation of the effect of WEIGHT on the expected values by RIVER when the variances are modeled as well, but not on the variances since they do not depend on RIVER.

Even though the variances are distinctly non-constant, the model for the expected values might not have changed much compared to the model based on constant variances. This can be assessed with the following code.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, xintrcpt=N,
        xvars=WEIGHT, xpowers=0.36, xgcs=RIVER 1 WEIGHT 1.4, expand=Y,
        expvvars=WEIGHT RIVER, geomcmbn=Y, contract=Y, nocnxbas=Y, notrxbas=Y);
```

The base model for the expected values is the one generated with constant variances. This is not changed by the expansion since the expxvars parameter has its default empty setting, but the model for the variances is changed since the expvvars parameter has a non-empty setting. However, by default the contraction can change the base model. The setting "nocnxbas=Y" means do not contract the base model for expected values. However, its powers could still be adjusted as part of retransforming contracted models. This is avoided with the setting "notrxbas=Y". The generated model for the variances is based on the single transform $WEIGHT^{-0.5}$ without an intercept. The LCV score is 0.42921 with substantial percent decrease 2.27% compared to the fully adaptive model. Consequently, consideration of non-constant variances has a distinct effect on the model for the expected values, indicating the importance of considering non-constant variances over the conventional use of constant variances models.

The adaptive non-constant variances model can be generated directly as follows.

```
%genreg(modtype=NORML, datain=MERCURY, yvar=MERCURY, foldcnt=5, xintrcpt=N,
        xvars=WEIGHT, xpowers=0.26,
        xgcs=RIVER 1 WEIGHT 1.5 : RIVER 1 WEIGHT -1.3, xgcpowers=0.8 0.7,
        vintrcpt=N, vvars=WEIGHT, vpowers=-0.53);
```

Table 1 contains part of the genreg output. Note that this is in listing format not in the default HTML format of SAS version 9.3 or later, and so a "ods listing;" command needs to have been executed earlier for results to appear in the SAS output window. Geometric combinations for the expected values (variances/dispersions) are given names with prefix "XGC_" ("VGC_") followed by an index number. The LCV score is described as the "mth root of the likelihood using deleted predictions" and rounds to 0.43919 as reported earlier. Figures 1-2 displays the estimates for the expected values and standard deviations, respectively, for this model. Estimated expected mercury levels start out around the same value for the two rivers at a low weight of 0.5 kg, increase nonlinearly with increasing weights after that, but to a distinctly higher level for fish caught in the Waccamaw River than in the Lumber River. Estimated standard deviations increase nonlinearly with weight in the same way for fish caught in the two rivers. These plots were generated in Excel using the powers and slope estimates given in Table 1.

```

base model

geometric combination expectation variables:
  XGC_1 RIVER*WEIGHT**(1.5)
  XGC_2 RIVER*WEIGHT**(-1.3)
...

base expectation component

predictor      power      estimate

WEIGHT         0.26    1.0959948
XGC_1          0.8     0.3071072
XGC_2          0.7    -0.115155

base log variance component

predictor      power      estimate

WEIGHT        -0.53    -1.098512
...

mth root of the likelihood using deleted predictions:      0.4391878

```

Table 1. Example Output for the Adaptive Model for the Expected Values and Variances in WEIGHT, RIVER, and Geometric Combinations

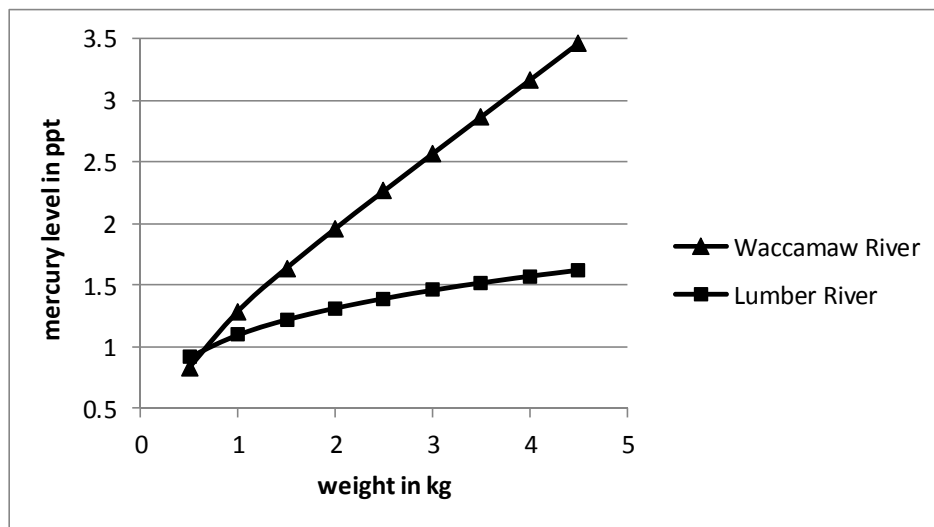


Figure 1. Estimated Expected Values for Mercury Levels in Fish Depending on Weight of the Fish Caught in the Lumber River Compared to Fish Caught in the Waccamaw River Based on the Model for Untransformed Mercury

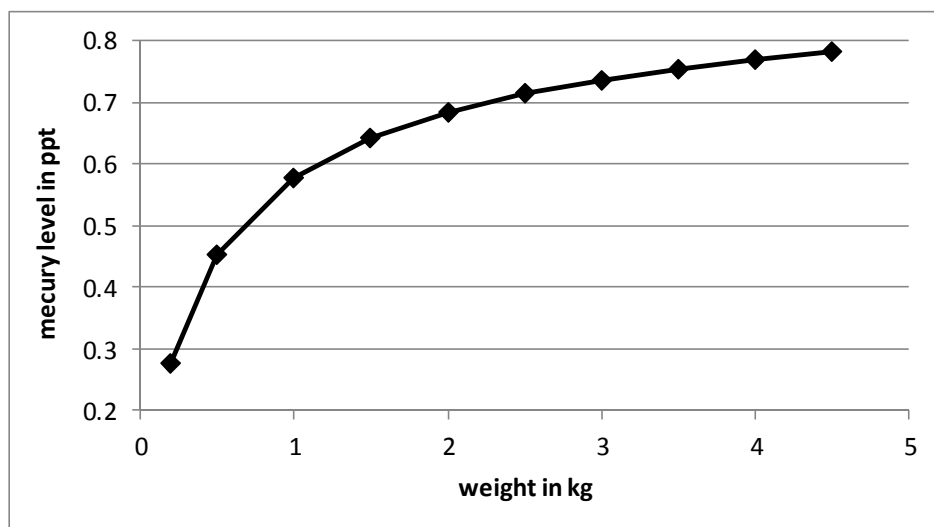


Figure 2. Estimated Standard Deviations for Mercury Levels in Fish Depending on Weight of the Fish Caught in Either the Lumber or Waccamaw Rivers Based on the Model for Untransformed Mercury

The genreg macro generates several output data sets. The dataout parameter is used to name a data set containing a copy of the datain data set along with several generated output variables describing the results for the selected model. Macro parameters are also available for setting the names of these output variables. For example, the stdrsvar parameter with default value "stdres" is used to name a variable containing standardized residuals for conducting residual analyses. Also, the nscrevar parameter with default value "nscore" is used to name a variable containing normal scores for generating normal (probability) plots. Figure 3 displays this plot for the adaptive non-constant variances model. It is distinctly curved, there is a very outlying observation with standardized residual 4.19, and the Shapiro-Wilk test for normality of the standardized residuals is highly significant ($p < 0.001$), suggesting that linear regression models with their assumption of normality are questionable for these data.

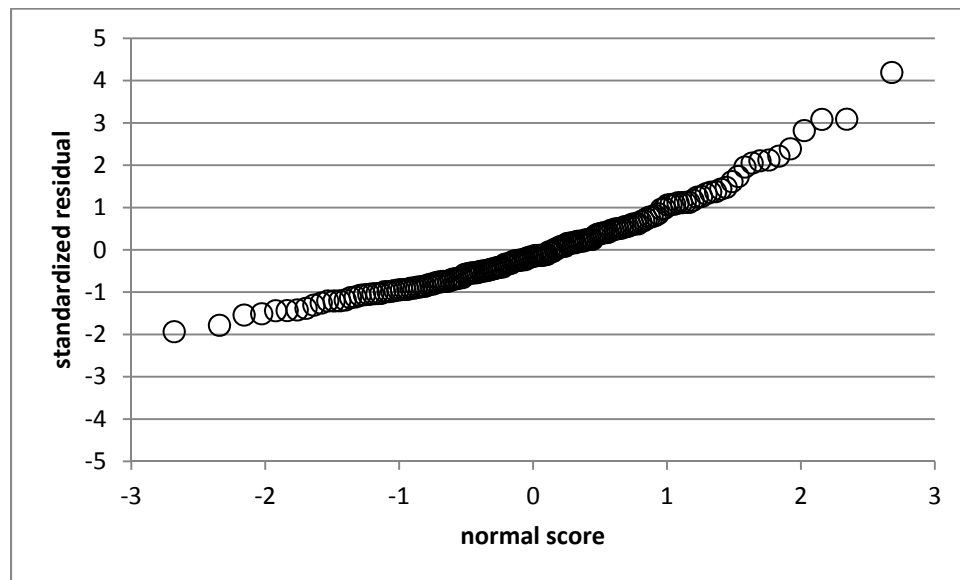


Figure 3. Normal Plot for the Standardized Residuals for the Adaptive Non-Constant Variances Model for Untransformed Mercury Levels Depending on the Weight of Fish and the River Caught In

CATEGORIZING MERCURY LEVELS

One way to circumvent the model assumption problems for the continuous outcome MERCURY is to categorize it into two or more levels. The binary outcome HIGHMERC with value 1 when mercury levels exceed 1 ppm and value 0 otherwise can be generated using "HIGHMERC=(MERCURY>1);" within a data step changing the MERCURY data set. There are 80 or 47.3% of the fish with $MERCURY > 1$, so the choice of 1 ppm is close to choosing a median split. That choice is also the Federal Drug Administration limit for human consumption

(http://www.fda.gov/OHRMS/DOCKETS/ac/02/briefing/3872_Advisory%207.pdf, accessed 7/25/15).

The adaptive model for HIGHMERC in WEIGHT, RIVER, and geometric combinations for expected values and dispersions can be generated as follows.

```
%genreg(modtype=LOGIS, datain=MERCURY, yvar=HIGHMERC, foldcnt=5, expand=Y,
        expxvars=WEIGHT RIVER, expvvars=WEIGHT RIVER, geomcmbn=Y, contract=Y,
        nocnxint=Y);
```

Since HIGHMERC has two distinct values (0 and 1), a standard logistic regression analysis is conducted using the canonical logit link function, but adjusted to account for non-unit dispersions with parameters estimated using maximum extended quasi-likelihood estimation (McCullagh & Nelder, 1999) and adaptive search based on extended quasi-likelihood cross-validation (Q^+LCV) scores. The setting "nocnxint=Y" means do not contract the intercept from the model for the expected values. This guarantees that the log odds when predictors have value zero are always estimated and never set to zero. The generated model has expected values based on the two transforms $WEIGHT^{-0.6}$ and $(RIVER \cdot WEIGHT^{6.9})^{0.9998}$ with an intercept (as guaranteed by the above code) and dispersions based on the two transforms $(RIVER \cdot WEIGHT^4)^{0.9101}$ and $WEIGHT^{-2.989} \cdot RIVER$ also with an intercept. The Q^+LCV score is 0.69976. The associated adaptive additive model (generated by changing to "geomcmbn=N" in the above code) has expected values based on the single transform $WEIGHT^{-0.4}$ and constant variances (based on only an intercept). The Q^+LCV score is 0.56005 with substantial percent decrease 19.97%. As for the continuous outcome MERCURY, there is distinct moderation of the effect of weight on the expected values by the river in which they were caught, but now that also holds for dispersions.

Mercury values could have been categorized into three or more ordered levels rather than just two. When "modtype=LOGIS" and the outcome variable has more than two values, by default a multinomial regression model is generated using generalized logits as is appropriate for discrete outcomes with nominal values. By default the lowest outcome value is the reference value (due to the default setting "refyval=MIN") but this can be changed if desired. However, ordinal regression models based on proportional odds can be more appropriate for ordinal outcomes. This is controlled by the propodds parameter. Its default value is "N" meaning generate a multinomial model (which is the same as a standard logistic regression when the outcome has only two distinct values). The setting "propodds=Y" requests an ordinal regression instead.

TRANSFORMING MERCURY LEVELS

Another alternative that might circumvent the model assumption problems for the continuous outcome MERCURY is to transform it. MERCURY is positive-valued (with values ranging from 0.11 ppm to 3.6 ppm) and so power transforms $MERCURY^p$ are well-defined for all real valued powers p .

Power for Transforming MERCURY	Power-Adjusted LCV Score
-2.5	0.0145993
-2.0	0.0438313
-1.5	0.1188705
-1.0	0.2482369
-0.5	0.3831880
0.0	0.4544766
0.5	0.4484637
1.0	0.3844301
1.5	0.2968758
2.0	0.2118910
2.5	0.1426923

Table 2. Power-Adjusted LCV Scores for Adaptive Models in WEIGHT for the Expected Values of Power Transformed Mercury Levels with Constant Variances

The following code request generation of the power-adjusted LCV scores $LCV(p)$ for powers p ranging from -2.5 to 2.5 by steps of size 0.5 for modeling the expected values of $MERCURY^p$ as a possibly nonlinear function of WEIGHT with constant variances.

```
%ypower(datain=MERCURY, yvar=MERCURY, foldcnt=5, yfst=-2.5, ycnt=11, ystp=0.5, expand=Y,
        expxvars=WEIGHT, contract=Y);
```

The datain, yvar, foldcnt, expand, expxvars, and contract parameters have the same meanings as for genreg. The yfst parameter specifies the first power, -2.5 in this case, ycnt the number of powers to generate, 11 in this case, and

ystp the step size or increment, 0.5 in this case. The output is displayed in Table 2. The power $p=0$ corresponding to $\log(\text{MERCURY})$ generates the largest power-adjusted LCV score of Table 2 with $\text{LCV}(0)=0.45448$ (after rounding to 5 digits). A search around the power 0 over steps of size 0.1 can be generated by changing to the settings "yfst=-0.4", "ycnt=9", and "ystp=0.1". The largest power-adjusted LCV score for this second search is $\text{LCV}(0.2)=0.46173$, which is a substantial improvement over the associated model for untransformed MERCURY with $\text{LCV}(1)=0.38443$ (the same as its power-unadjusted LCV score reported above) with percent decrease 16.74%.

By default, ypower turns off generation by genreg of results in the log and output windows. These results can be turned back on in order to identify the model generated for the selected power 0.2 as follows.

```
%ypower(datain=MERCURY,yvar=MERCURY,foldcnt=5,yfst=0.2,ycnt=1,expand=Y,
        expxvars=WEIGHT,contract=Y,noglog=n,nogprint=n);
```

The setting "noglog=n" means do not have genreg turn off generation of its results in the Log window while the setting "nogprint=n" has the same meaning but for the Output window. The generated model for the expected values of $\text{MERCURY}^{0.2}$ is based on $\text{WEIGHT}^{0.1}$ without an intercept.

The model for $\text{MERCURY}^{0.2}$ with both the expected values and variances adaptively modeled in terms of WEIGHT, RIVER, and geometric combinations is generated as follows.

```
%ypower(datain=MERCURY,yvar=MERCURY,foldcnt=5,yfst=0.2,ycnt=1,expand=Y,
        expxvars=WEIGHT RIVER,expvvars=WEIGHT RIVER,geomcmbn=y,contract=Y,
        noglog=n,nogprint=n);
```

The expected values are based on the three transforms: $\text{WEIGHT}^{0.5}$, $(\text{RIVER} \cdot \text{WEIGHT}^{1.1})^{0.6}$, and $(\text{RIVER} \cdot \text{WEIGHT}^{-1.5})^{0.8}$ without an intercept and the variances on the single transform: $\text{WEIGHT}^{-0.01}$ also without an intercept. The $\text{LCV}(0.2)$ score is 0.49466, which is a substantial improvement over the associated model for untransformed MERCURY with $\text{LCV}(1)=0.43919$ (as reported above) and percent decrease 11.21%.

The model for $\text{MERCURY}^{0.2}$ with only the expected values adaptively modeled in terms of WEIGHT, RIVER, and geometric combinations and with constant variances is obtained by removing "expvvars=WEIGHT RIVER" from the above code so expvvars has its default empty setting. The expected values are based on the three transforms: $\text{WEIGHT}^{0.5}$, $(\text{RIVER} \cdot \text{WEIGHT}^{1.1})^{0.5}$, and $(\text{RIVER} \cdot \text{WEIGHT}^2)^{-0.5}$ without an intercept. The $\text{LCV}(0.2)$ score is 0.49474, a little larger than the score for the non-constant variances model. Thus, appropriately power transformed MERCURY is reasonably considered to satisfy the usual assumption of constant variances.

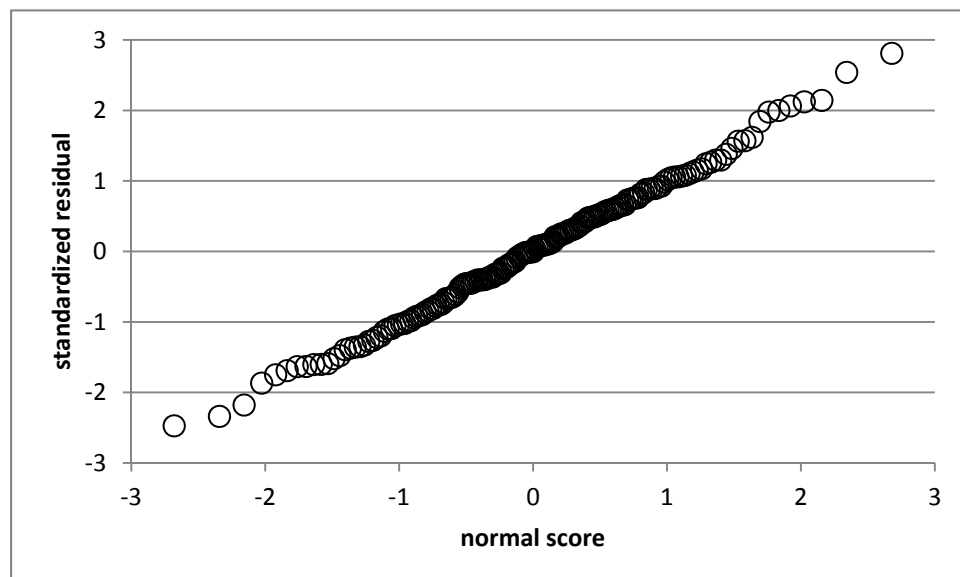


Figure 4. Normal Plot for the Standardized Residuals for the Adaptive Constant Variances Model for Transformed Mercury Levels Depending on the Weight of Fish and the River Caught In

Figure 4 displays the normal plot for the standardized residuals for the constant variances model generated for $\text{MERCURY}^{0.2}$. It is reasonably close to linear. Moreover, the Shapiro-Wilk test for normality of the standardized residuals is nonsignificant ($P=0.939$). Consequently, appropriately power transformed MERCURY is reasonably considered to satisfy the normality assumption as well as the constant variances assumption.

Figure 5 displays estimated expected values for untransformed mercury levels induced by the constant variances model for transformed mercury levels (that is, the estimated expected values for $\text{MERCURY}^{0.2}$ raised to the power $1/0.2=5$). The estimated expected values for the model for transformed MERCURY start out for both rivers at about the same place and increase to a higher level for fish caught in the Waccamaw River compared to those caught in the Lumber River. The estimates of Figure 5 are similar to those of Figure 1 except for high values of WEIGHT, for which the estimates are lower in Figure 5 than in Figure 1.

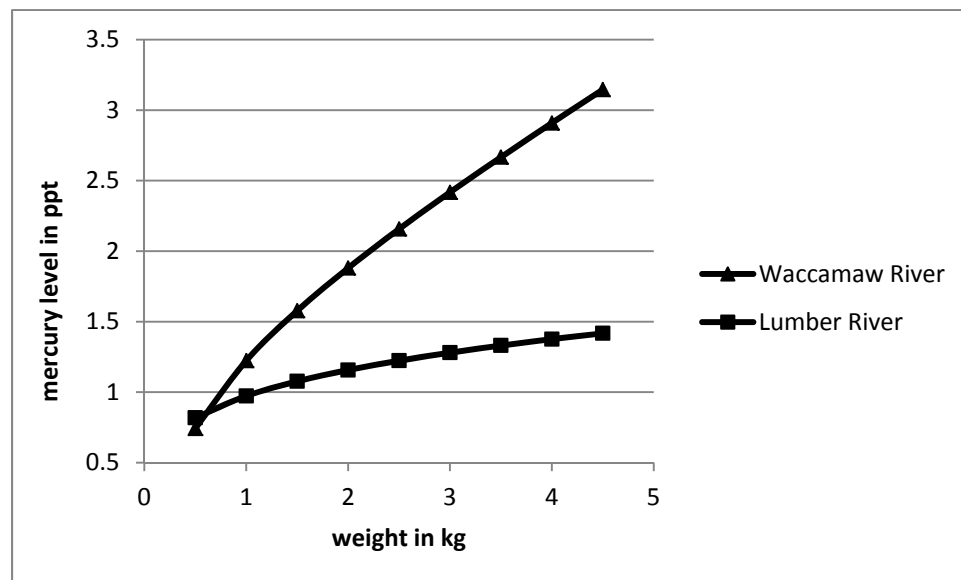


Figure 5. Estimated Expected Values for Mercury Levels in Fish Depending on Weight of the Fish Caught in the Lumber River Compared to Fish Caught in the Waccamaw River Based on the Model for Transformed Mercury Levels

CONCLUSION

The reported analyses of mercury levels in fish demonstrate the use of the genreg macro for adaptive modeling of univariate continuous and discrete outcomes in terms of fractional polynomial transforms of predictors. The results demonstrate the need for adaptive modeling of variances/dispersions as well as of expected values. The macro also supports adaptive modeling of univariate count outcomes using adaptive Poisson regression as well as of multivariate outcomes of these three types. The analyses also demonstrate the use of the ypower macro for adaptive power transformation of positive-valued continuous outcomes as well as their predictors. Power transformation of such outcomes can resolve problems with the standard assumptions of normality and constant variances.

REFERENCES

- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality & Social Psychology*, 51, 1173-1182.
- Knafl, G. J., Delucchi, K. L., Bova, C. A., Fennie, K. P., & Williams, A. B. (2010). A systematic approach for analyzing electronically monitored adherence data. In B. Ekwall & M. Cronquist (Eds.), *Micro electro mechanical systems (MEMS) technology, fabrication processes and applications*, Chapter 1, pp. 1-66, Hauppauge, NY: Nova Science Publishers.
- Knafl, G. J., & Ding, K. (in press). *Adaptive regression for modeling nonlinear relationships*. New York, NY: Springer Verlag.
- McCullagh, P., & Nelder, J. A. (1999). *Generalized linear models*, 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.
- Royston, P., & Sauerbrei, W. (2008). *Multivariable model-building: A practical approach to regression analysis based on fractional polynomials for modelling continuous variables*. Hoboken, NJ: John Wiley & Sons.

ACKNOWLEDGMENTS

The development of the genreg macro has been supported in part by Award Number R01 AI57043 from the National Institute of Allergy and Infectious Diseases (NIAID) and by Award Number R03 MH086132 from the National Institute of Mental Health (NIMH). The content is solely the responsibility of the author and does not necessarily represent the official views of the NIAID, the NIMH, or the National Institutes of Health.

CONTACT INFORMATION

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

George J. Knafl
Professor, School of Nursing
Carrington Room 5014, Campus Box 7460
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-7460
Work Phone: 919-843-9686
Work Fax: 919-943-9969
E-mail: gknafl@unc.edu
Web: <http://www.unc.edu/~gknafl/>

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.

Other brand and product names are trademarks of their respective companies.