

Paper 214-29

Assessing Model Fit and Finding a Fit Model

Pippa Simpson, University of Arkansas for Medical Sciences, Little Rock, AR
Robert Hamer, University of North Carolina, Chapel Hill, NC
ChanHee Jo, University of Arkansas for Medical Sciences, Little Rock, AR
B. Emma Huang, University of North Carolina, Chapel Hill, NC
Rajiv Goel, University of Arkansas for Medical Sciences, Little Rock, AR
Eric Siegel, University of Arkansas for Medical Sciences, Little Rock, AR
Richard Dennis, University of Arkansas for Medical Sciences, Little Rock, AR
Margaret Bogle, USDA ARS, Delta NIRI Project, Little Rock, AR

ABSTRACT

Often linear regression is used to explore the relationship of variables to an outcome. R^2 is usually defined as the proportion of variance of the response that is predictable from (that can be explained by) the independent (regressor) variables. When the R^2 is low there are two possible conclusions. The first is that there is little relationship of the independent variables to the dependent variable. The other is that the model fit is poor. This may be due to some outliers. It may be due to an incorrect form of the independent variables. It may also be due to the incorrect assumption about the errors; the errors may not be normally distributed or indeed the independent variables may not truly be fixed and have measurement error of their own. In fact the low R^2 may be due to a combination of the above aspects.

Assessing the cause of poor fit and remedying it is the topic of this paper. We will look at situations where the fit is poor but dispensing with outliers or adjusting the form of the independents improves the fit. We will talk of transformations possible and how to choose them. We will show how output of SAS[®] procedures can guide the user in this process. We will also examine how to decide if the errors satisfy normality and if they do not what error function might be suitable. For example, we will take nutrition data where the error function can be taken to be a gamma function and show how PROC GENMOD may be used.

The objective of this paper is to show that there are many situations where using the wide range of procedures available for general linear models can pay off.

INTRODUCTION

Linear regression is a frequently used method of exploring the relationship of variables and outcomes. Choosing a model, and assessing the fit of this model, are questions which come up every time one employs this technique. One gauge of the fit of the model is the R^2 , which is usually defined as the proportion of variance of the response that can be explained by the independent variables. Higher values of this are generally taken to indicate a better model. In fact, though, when the R^2 is low there are two possible conclusions. The first is that the independent variables and the dependent variables have very little relation to each other; the second is that the model fit is poor. Further, there are a number of possible causes of this poor fit – some outliers; an incorrect form of the independent variables; maybe even incorrect assumptions about the errors. The errors may not be normally distributed, or indeed the independent variables may not be fixed. To confound issues further, the low R^2 may be any combination of the above aspects.

Assessing the cause of poor fit and remedying it is the topic of this paper. We will look at a variety of situations where the fit is poor, including those where dispensing with outliers or adjusting the form of the independents improves the fit. In addition, we will discuss possible transformations of the data, and how to choose them based upon the output of SAS procedures. Finally we will examine how to decide whether the errors are in fact normally distributed, and if they are not, what error function might be suitable.

Our objective is to show how various diagnostics and procedures can be used to arrive at a suitable general linear model, if it exists. Specifically our aims are as follows.

Aim 1: To discuss various available diagnostics for assessing the fit of linear regression. In particular we will examine detection of outliers and form of variables in the equation.

Aim 2: To investigate the distributional assumptions. In particular, we will show that for skewed errors, it may be necessary to assume other distributions than normal.

To illustrate these aims, we will take two sets of data. In the first case, transformations of the variables and deletion of outliers allow a reasonable fit. In the second case, the error function can be taken to be a gamma function. We will show how we reached that conclusion and how GENMOD may be used. We will limit ourselves to situations where the response is continuous. However, we will indicate options available in SAS where that is not true.

Throughout a modeling process it should be remembered that *“An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem”*, John Tukey.

METHODS

ASSESSING FIT

Assessing the fit of a model should always be done in the context of the purpose of the modeling. If the model is to assess the predefined interrelationship of selected variables, then the model fit will be assessed and test done to check the significance of relationships. Often these days, modeling is used to investigate relationships, which are not (totally) defined and to develop prediction criteria. In that case, for validation, the model should be fit to one set of data and validated on another. If the model is for prediction purposes then, additionally, the closeness of the predicted values to the observed should be assessed in the context of the predicted values' use. For example, if the desire is to predict a mean, the model should be evaluated based on the closeness of the predicted mean to the observed. If, on the other hand, the purpose is to predict certain cases, then you will not want any of the expected values to be “too far” from the observed. *“The only relevant test of the validity of a hypothesis is comparison of its predictions with experience”*, Milton Friedman.

Outliers will affect the fit. There are techniques for detecting them – some better than others- but there is no universal way of dealing with them. Plots, smoothing plots such as PROC LOESS and PROC PRINCOMP help detect outliers. If there is a sound reason why the data point should be disregarded then the outlier may be deleted from analysis. This approach, however, should be taken with caution. If there is a group of outliers there may be a case for treating them separately. Robust regression will downplay the influence of outliers. PROC ROBUSTREG is software which takes this approach (Chen, 2002).

DATA

Gene Data

We fit an example from a clinical study. The genetic dataset consists of a pilot study of young men who had muscle biopsies taken after a standard bout of resistance exercise. The mRNA levels of three genes were evaluated, and five SNPs in the IL-1 locus were genotyped, in order to examine the relationship between inflammatory response and genetic factors.

The dependent variable is gene expression. Independent variables are demographics, baseline and exercise parameters. The purpose of the modeling is to explain physiologic phenomena. One of the objectives of this study is to develop criteria for people at risk based on their genetic response. Thus exploratory modeling and ultimately prediction of evaluation of expression (not necessarily as a quantity but as a 0-1 variable) would be the goal. This data set is too small to use a holdout sample. For the nutrition data we show our results for a sample of data. For confidence the relationships should be checked on another sample.

Nutrition data

Daily caffeine consumption is the dependent variable and independents considered included demographics and health variables and some nutrient information.

We fit a nutrition model which has caffeine consumption as the dependent variable and, as its explanatory variables, race, sex, income, education, federal assistance and age. Age is recorded in categories and therefore can be considered as an ordinal variable or, by taking the midpoint of the age groupings, can be treated as a continuous variable. Alternatively all variables can be treated as categorical. In SAS we would specify them as CLASS variables. The objective of this study is to explore the relationship of caffeine intake to demographics.

LINEAR REGRESSION MODELING, PROC REG

Typically, a first step is to assume a linear model, with independent and identical normally distributed error terms. Justification of these assumptions lies in the central limit theorem and the vast number of studies where these assumptions seem reasonable. We cannot definitively know whether our model adequately fits assumptions. A test applied to a model does not really indicate how well a model fits or indeed, if it is the right model. Assessing fit is a complex procedure and often all we can do is examine how sensitive our results are to different modeling assumptions. We can use statistics in conjunction with plots. *"Numerical quantities focus on expected values, graphical summaries on unexpected values."* John Tukey.

SAS procedure REG fits a linear regression and has many diagnostics for normality and fit. These include

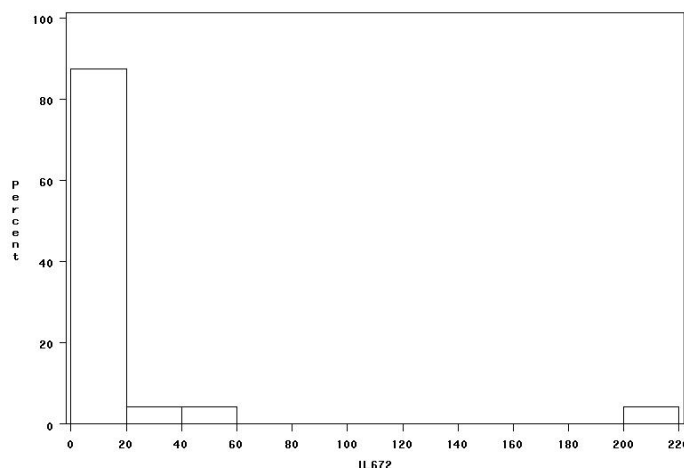
- Test statistics which can indicate whether a model is describing a relationship (somewhat). Typically we use R^2 . Other statistics include the log likelihood, F ratio and Akaike's criterion which are based on the log likelihood.
- Collinearity diagnostics which can be used to assess randomness of errors.
- Predicted values, residuals, studentized residuals, which can be used to assess normality assumptions.
- Influence statistics, which will indicate outliers.
- Plots which will visually allow assessment of the normality, randomness of errors and possible outliers. It is possible to produce
 - normal quantile-quantile (Q-Q) and probability-probability(P-P) plots and
 - for statistics, such as residuals display the fitted model equation, summary statistics, and reference lines on the plot .

INTERPRETING R^2

R^2 is usually defined as the proportion of variance of the response that is predictable from (that can be explained by) the regressor variables; that is, the variability explained by the model. It may be easier to interpret the square root of $1 - R^2$, which is approximately the factor by which the standard error of prediction is reduced by the introduction of the regressor variables. Nonrandom sampling can greatly distort R^2 . A low R^2 can be suggestive that the assumptions of linear regression are not satisfied. Plots and diagnostics will substantiate this suspicion.

PROC GLM can also be used for linear regression, Analysis of variance and covariance and is overall a much more general program. Although GLM has most of the means for assessing fit, it lacks collinearity diagnostics, influence diagnostics, or scatter plots. However, it does allow use of the class variable. Alternatively, PROC GLMMOD can be used with a CLASS statement to create a SAS dataset containing variables corresponding to matrices of dependent and independent variables, which you can then use with REG or other procedures with no CLASS statement. Another useful procedure is ANALYST. This will allow fitting of many models and diagnostics.

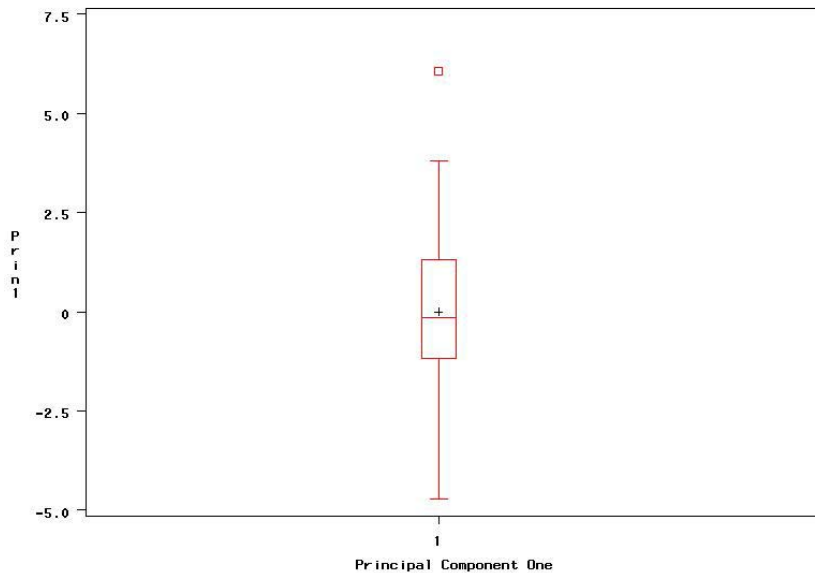
Note: If we use PROC REG or SAS/ANALYST® we will have to create dichotomized variables for age and other categorical variables or treat them as continuous.



**Genetic Data
Figure 1:**

For the genetic dataset, when we fit the quantity of IL-6 mRNA 72 hours after baseline on a strength variable and five gene variables, we had an R^2 of 0.286 and an adjusted R^2 of 0.034. The residuals for this model show some problems, since they appear to be neither normally distributed, nor in fact symmetric. The distribution, in figure 1. of the dependent variable also shows this skewness, perhaps indicating why the model fit fails.

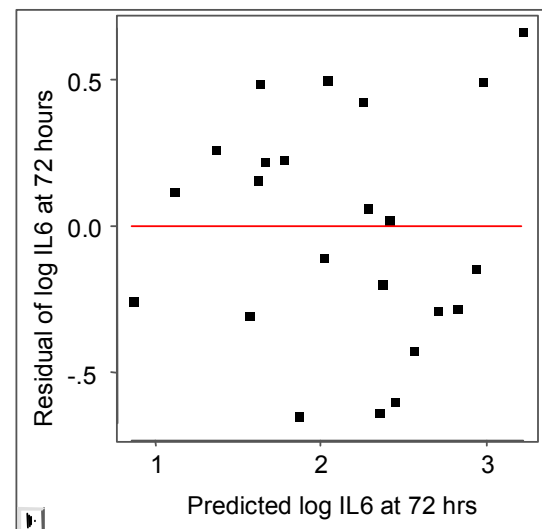
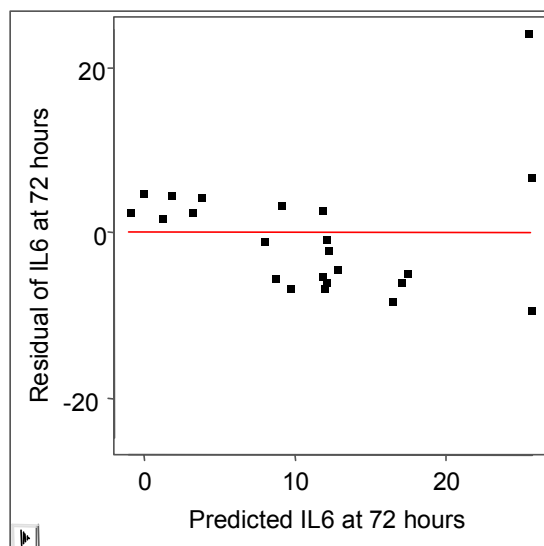
Figure 2: A Box plot of the first principal component



Procedure PROC PRINCOMP can be used to detect outliers and aspects of collinearity. The principal components are linear combinations of the variables specified which maximize variability while being orthogonal to (or uncorrelated with) each other. When it may be difficult to detect an outlier of many interrelated variables by traditional univariate, or multivariate plots, examination of the principal components may make the outlier clear. We use this method on the genetic dataset to notice that patient #2471 is an outlier. After running the procedure, we examine boxplots of the individual principal components to find any obvious outliers and immediately notice this subject. His removal from the dataset may improve our fit, as his values clearly do not match the rest of the data.

However this still does not solve the problem. When we fit the data without the outlier we find that we have improved the R^2 but the residuals are still highly skewed. A plot of the predicted versus the residuals shows that there is a problem. We would expect a random scatter of the residuals around the 0 line. As the distribution of the dependent variable in the genetic example is heavily skewed, and the values are all positive, ranging from 1.85 to 216.12, we can consider a log-transformation. This will tend to symmetrize the distribution. Indeed, fitting a linear regression with log (IL-6 at 72 hours) as the dependent variable, and the previously mentioned predictors, improves the R^2 to 0.536, and the adjusted R^2 to 0.373. We try a log transformation, noting that there is some justification since we would expect expression to change multiplicatively rather than additively. The plots in figure 3 show an improvement in the residuals. They are not perfectly random around zero but the scatter is much improved.

Figure 3: Residual of IL6 at 72 hours by the predicted value before and after a log transformation of the gene



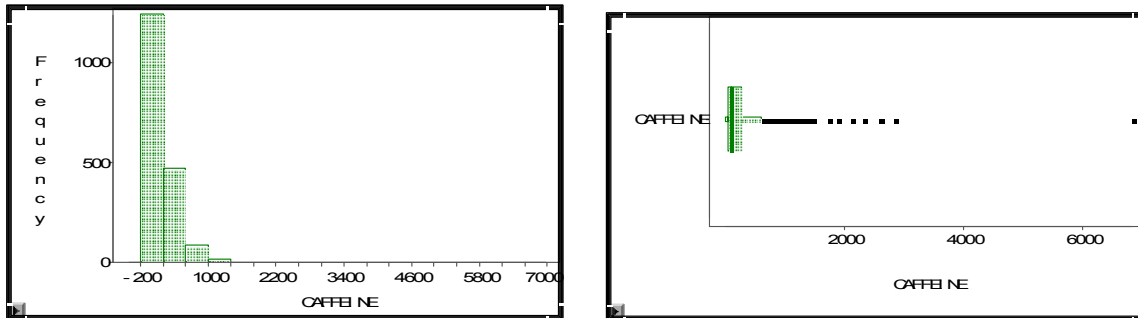
Nutrition Data

We use PROC GLM.

```
proc glm data =caff;
class Sex r_fed1 r_race r_inc r_edu;
model CAFFEINE= Sex r_agel r_fed1 r_race r_inc r_edu/ p clm;
  output out=pp p=caffpred r=resid;
  axis1 minor=none major=(number=5);
  axis2 minor=none major=(number=8);
  symbol1 c=black i=none v=plus;
proc gplot data=pp;
plot resid*caffeine=1/haxis=axis1
      vaxis=axis2;
run;
```

For caffeine the R^2 was 0.16. Using SAS/ ANALYST we look at the distribution of Caffeine and begin to suspect that our assumptions may have been violated

Figure 4: Plots of caffeine (frequency and boxplot)



Values extend to 2000, with an outlier close to 7000. For further analysis we choose to delete this outlier since it is hard to believe that someone consumed 6000+ milligrams of caffeine in a day. A large percentage are close to 0. The assumption of normal distribution for the error term may not be valid in this nutrient model. Residuals are highly correlated with observed values, the correlation is >0.9 . The (incorrect) assumption was that the errors were independently distributed normally.

PROC LOESS – ASSESSING LINEARITY

When there is a suspicion that the relationship is not completely linear, deviations may sometimes be accommodated by fitting polynomial terms of the regressors such as squares or cubic terms. LOESS often is helpful in suggesting the form by which variables may be entered. It provides a smooth fit of dependent to the independent variable. Procedure LOESS allows investigation of the form of the continuous variables to be included in the regression. By use of a smoothing plot it can be seen what form the relationship may take and a parametric model can then be chosen to approximate this form.

Gene Data

```
ods output OutputStatistics= genefit;

proc loess data=tmp3.red;
  model il672 = il60/ degree=2 smooth = 0.6;
run;
proc sort data= genefit;
  by il60;
run;
```

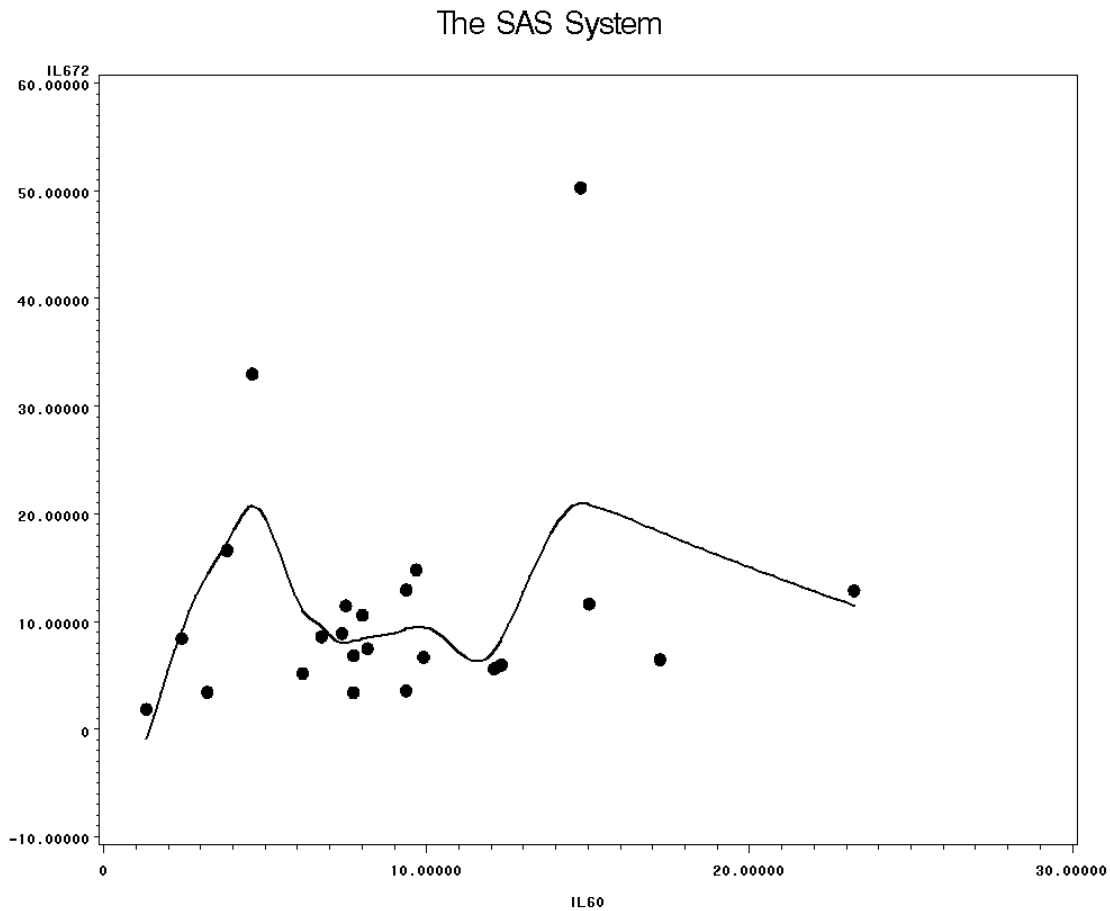
```

symbol1 color=black value=dot h=2.5 pct;
symbol2 color=black interpol=spline value=none width=2;
proc gplot data= genefit;
  plot (depvar pred)* il60/ overlay;

run; quit;

```

Figure 5: Loess plot to assess linearity



It can be seen that IL6 at 72 hours (IL672) is not linearly dependent on IL6 at 0 hours (IL60). The relationship does not take a polynomial form but it is not easy to see what is the relationship.

MODELS OTHER THAN LINEAR REGRESSION

In the last few years SAS has added several new procedures and options which aid in assessing the fit of a model and fitting a model with a minimal number of assumptions.

PROC ROBUSTREG® AND OUTLIERS

For example, to investigate outliers, the experimental procedure ROBUSTREG is ideal. By use of an optimization function which minimizes the effect of outliers, it fits a regression line. To achieve this, ROBUSTREG provides four methods: M estimation, LTS estimation, S estimation, and MM estimation.

M estimation is especially useful when the dependent variable seems to have outlying values. The other three methods will deal with all kinds of outliers leverage (independent variable) and influence points (dependent variable).

```
proc robustreg data =caff;
class Sex r_fed1 r_race r_inc r_edu;
model CAFFEINE= Sex r_age1 r_fed1 r_race r_inc r_edu/diagnostics;
output out=robout r=resid sr=stdres;
run;
```

On the nutrition data, the robust R^2 was very small, using the least trimmed square option. This is not really surprising given that our problem is the preponderance of zeros or near zeros.

EXTENSION OF LINEAR REGRESSION – ALLOWING THE DEPENDENCY TO BE NONLINEAR**Nutrition Data**

The consumption of caffeine is varied from 0 up to 6500mg so that a transformation of the variable can be considered. It is not clear what transformation is suitable. A log transformation will not work since we have a lot of zeros. A square root was tried. This is often a good first try when the residuals seem dependent on the outcome values. It did not work.

PROC GAM – A LINEAR COMBINATION OF FUNCTIONS OF VARIABLES

A general additive model (GAM) extends the idea of a general linear model to allow a linear combination of different functions of variables. It too can be used to investigate the form of variables should a linear regression model be appropriate. GAM has, as a subset, Loess smooth curves.

The PROC GAM can fit the data from various distributions including Gaussian and gamma distributions:

The Gaussian Model

With this model, the link function is the identity function, and the generalized additive model is the additive model.

```
proc gam data=caff;
  class Sex r_fed1 r_race r_inc r_edu;
  model CAFFEINE= param(Sex r_fed1 r_race r_inc r_edu)
                spline(r_age1,df=2);
run;
```

Some of the output is shown.

Summary of Input Data Set	
Number of Observations	1479
Distribution	Gaussian
Link Function	Identity

Iteration Summary and Fit Statistics	
Final Number of Backfitting Iterations	5
Final Backfitting Criterion	3.692761E-11
The Deviance of the Final Estimate	92889266.839

The deviance is very large

Regression Model Analysis Parameter Estimates				
Parameter	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	205.64771	98.94328	2.08	0.0378
sex 1	49.42032	13.84611	3.57	0.0004
sex 2	0	.	.	.
r_fed1 1	-8.35210	41.66707	-0.20	0.8412
r_fed1 2	-39.71518	40.35210	-0.98	0.3252
r_fed1 3	0	.	.	.
r_race 1	82.80921	66.48645	1.25	0.2131
r_race 2	-134.09739	66.78547	-2.01	0.0448
r_race 3	-17.10858	81.23195	-0.21	0.8332
r_race 4	0	.	.	.
r_inc 1	2.52953	29.95996	0.08	0.9327
r_inc 2	28.55041	28.11922	1.02	0.3101
r_inc 3	0	.	.	.
r_edu 1	45.91183	67.07355	0.68	0.4938
r_edu 2	48.40920	66.45607	0.73	0.4665
r_edu 3	0	.	.	.
Linear(r_age1)	-5.76201	7.88958	-0.73	0.4653

Gender and race seem significant ($P < 0.05$)

Smoothing Model Analysis Fit Summary for Smoothing Components				
Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(r_age1)	0.948717	2.000000	141976	5

Smoothing Model Analysis Analysis of Deviance				
Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(r_age1)	2.00000	1555182	24.7117	<.0001

The Gamma Model

Similar to the Poisson model, the Gamma model assumes a log link between dependent variables and the mean.

```
proc gam data=caff;
  class Sex r_fed1 r_race r_inc r_edu;
  model CAFFEINE= param(Sex r_fed1 r_race r_inc r_edu)
              spline(r_age1,df=2) /dist = gamma;
run;
```

Iteration Summary and Fit Statistics	
Number of local score iterations	7
Local score convergence criterion	1.0622751E-9
Final Number of Backfitting Iterations	3
Final Backfitting Criterion	2.908349E-10
The Deviance of the Final Estimate	1851.9195211

Note that the deviance is much reduced

Regression Model Analysis Parameter Estimates				
Parameter	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	-0.00599	0.00256	-2.34	0.0195
sex 1	0.00067902	0.00023575	2.88	0.0040
sex 2	0	.	.	.
r_fed1 1	-0.00008154	0.00078358	-0.10	0.9171
r_fed1 2	-0.00063327	0.00074953	-0.84	0.3983
r_fed1 3	0	.	.	.
r_race 1	0.00091289	0.00118	0.77	0.4389
r_race 2	-0.00555	0.00125	-4.45	<.0001
r_race 3	-0.00048189	0.00151	-0.32	0.7498
r_race 4	0	.	.	.

Regression Model Analysis Parameter Estimates				
Parameter	Parameter Estimate	Standard Error	t Value	Pr > t
r_inc 1	-0.00003993	0.00063791	-0.06	0.9501
r_inc 2	0.00052028	0.00055370	0.94	0.3476
r_inc 3	0	.	.	.
r_edu 1	0.00203	0.00223	0.91	0.3644
r_edu 2	0.00214	0.00222	0.96	0.3356
r_edu 3	0	.	.	.
Linear(r_age1)	-0.00007709	0.00015799	-0.49	0.6257

Race and gender are much more significant (P < 0.005).

Smoothing Model Analysis Fit Summary for Smoothing Components				
Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(r_age1)	0.973331	2.000000	0.000127	5

Smoothing Model Analysis Analysis of Deviance				
Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(r_age1)	2.00000	31.453618	25.0689	<.0001

Age is significant

PORC TRANSREG – TRANSFORMATIONS OF VARIABLES, INCLUDING THE DEPENDENT

TRANSREG extends the idea of regression by providing optimal transformed variables after iterations. In TRANSREG, a general Box-Cox transformation of the dependent variable allows for an investigation of various parametric forms. Transformations of the following form can be fit:

$$\begin{aligned} & ((y + c)^\lambda - 1) / (\lambda g), & \lambda \neq 0 \\ & \log(y + c) / g, & \lambda = 0. \end{aligned}$$

Transformations linearly related to square root, inverse, quadratic, cubic, and so on are all special cases. By default, $c = 0$. The parameter c can be used to rescale y so that it is strictly positive. By default, $g = 1$. Alternatively, g can be $\bar{y}^{\lambda-1}$ where \bar{y} is the geometric mean of y . As Box said "All Models Are Wrong But Some Are Useful".

Nutrition Data

We first ran the Box- Cox transformation. The default values for λ of -3 incremented by 0.25 are used up to 3.

```
proc transreg data=caff;
model Boxcox(CAFFEINE)= class(sex r_age1 r_fed1 r_race r_inc r_edu);
output out=oneway;
run;
proc glm data=oneway;
class sex r_age1 r_fed1 r_race r_inc r_edu;
model tCAFFEINE = sex r_age1 r_fed1 r_race r_inc r_edu;
run;
```

We also ran the monotone transformation. They both gave similar results.

```
proc transreg data=caff;
model monotone(CAFFEINE)= class(sex r_age1 r_fed1 r_race r_inc r_edu);
output out=oneway;
run;
proc glm data=oneway;
class sex r_age1 r_fed1 r_race r_inc r_edu;
model tCAFFEINE = sex r_age1 r_fed1 r_race r_inc r_edu;
run;
```

We show some of the output for the monotone transformation.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	31599399.9	2872672.7	51.78	<.0001
Error	1467	81387994.1	55479.2		
Corrected Total	1478	112987394.0			

R-Square	Coeff Var	Root MSE	TCAFFEINE Mean
0.279672	104.5031	235.5402	225.3908

Compared to ordinary regression, the R-square increased from 0.16 to 0.28 by TRANSREG.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	2164614.36	2164614.36	39.02	<.0001
Tr_age1	1	3013848.14	3013848.14	54.32	<.0001
r_fed1	2	2421031.75	1210515.87	21.82	<.0001

Source	DF	Type I SS	Mean Square	F Value	Pr > F
r_race	3	23551830.56	7850610.19	141.51	<.0001
r_inc	2	402034.75	201017.38	3.62	0.0269
r_edu	2	46040.36	23020.18	0.41	0.6605

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	607720.20	607720.20	10.95	0.0010
Tr_age1	1	2261444.88	2261444.88	40.76	<.0001
r_fed1	2	276197.09	138098.54	2.49	0.0833
r_race	3	21657703.83	7219234.61	130.13	<.0001
r_inc	2	385200.30	192600.15	3.47	0.0313

Age, race and income came out significant in a Type III test.

Note: It is not surprising that the R^2 is increased since we are using the data to decide what is a better transformation.

Genetic Data

As a check on our use of the log transformation previously, we used the Box-Cox transformation in PROC TRANSREG on the gene dataset, with the same variables as used previously. This actually found that a transformation with $\lambda = -0.5$ was the best choice overall, but the R^2 for this model was 0.59, which is not a major improvement over the log transformation ($\lambda = 0$) where we found an R^2 of 0.54. The monotone transformation actually produced much better results than either of these, with a final R^2 after convergence of 0.77. Again, for this the data was used to develop the best transformation, so these values might not apply to another dataset. However, the log or Box-Cox transformations may be simpler to interpret and hence preferable.

GENMOD - ALLOWING THE ERROR TERM TO BE NONNORMAL

If the normal assumption is satisfied the error terms may be expected to be symmetric around zero. If they are highly skewed then other distributions that would give this kind of distribution can be considered. PROC GENMOD offers a selection of distributions from the exponential family, including some for discrete and continuous variables. . The Gamma is a general form for continuous outcomes which has the form of a peak close to zero or no peak (the negative exponential) and decreasing from zero. From Figure 4 it can be seen that the distribution of caffeine might be approximated by a gamma. If there is data with a zero value PROC GENMOD will delete the zeros with a gamma link function. We are only modeling the consumers of caffeine. The assumption of normal distribution for error term may not be valid in this nutrient model. Thus, we tried gamma distribution for original values of caffeine.

The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter(s). There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. The GENMOD procedure estimates the parameters of the model numerically through an iterative fitting process. One biggest advantage of PROC GENMOD is that error distribution can be specified for our problem. A number of popular link functions and probability distributions are available in the GENMOD procedure. In our example, we used the default canonical link function, which is power (-1). In addition, we choose our distribution as gamma.

```
proc genmod data = caff;
  class Sex r_age1 r_fed1 r_race r_inc r_edu;
```

```

model CAFFEINE= sex r_age1 r_fed1 r_race r_inc r_edu /
dist=gamma type1 type3;
run;

```

All the independent variables are specified as CLASS variables so that PROC GENMOD automatically generates the indicator variables associated with variables. Some of the output from the preceding statements is shown.

Model Information		
Data Set	WORK.CAFFN0	
Distribution	Gamma	
Link Function	Power(-1)	
Dependent Variable	CAFFEINE	Caffeine (mg)
Observations Used	1479	

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	1467	1883.3725	1.2838
Scaled Deviance	1467	1725.4732	1.1762
Pearson Chi-Square	1467	1587.9359	1.0824
Scaled Pearson X2	1467	1454.8056	0.9917
Log Likelihood		-9264.8876	

The deviance seems reasonable with the value divided by the degrees of freedom close to 1.

Analysis Of Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept		1	0.0064	0.0024	0.0016	0.0112	6.93	0.0085
sex	1	1	-0.0007	0.0002	-0.0011	-0.0003	10.21	0.0014
sex	2	0	0.0000	0.0000	0.0000	0.0000	.	.
r_fed1	1	1	0.0002	0.0007	-0.0012	0.0017	0.09	0.7701
r_fed1	2	1	0.0007	0.0007	-0.0007	0.0021	0.97	0.3252
r_fed1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
r_race	1	1	-0.0010	0.0011	-0.0031	0.0012	0.75	0.3872

Analysis Of Parameter Estimates								
Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq	
r_race	2	1	0.0055	0.0012	0.0032	0.0078	21.93	<.0001
r_race	3	1	0.0004	0.0014	-0.0024	0.0032	0.08	0.7726
r_race	4	0	0.0000	0.0000	0.0000	0.0000	.	.
r_inc	1	1	0.0001	0.0006	-0.0011	0.0013	0.02	0.8884
r_inc	2	1	-0.0006	0.0005	-0.0017	0.0004	1.47	0.2261
r_inc	3	0	0.0000	0.0000	0.0000	0.0000	.	.
r_edu	1	1	-0.0022	0.0021	-0.0064	0.0020	1.09	0.2965
r_edu	2	1	-0.0024	0.0021	-0.0066	0.0018	1.29	0.2563
r_edu	3	0	0.0000	0.0000	0.0000	0.0000	.	.
r_age1		1	0.0001	0.0001	-0.0002	0.0003	0.30	0.5825
Scale		1	0.9162	0.0294	0.8603	0.9757		

It is usually of interest to assess the importance of the main effects in the model. Type I and Type III analyses generate statistical tests for the significance of these effects. One can request these analyses with the TYPE I and TYPE III options in the MODEL statement. A Type 1 analysis consists of fitting a sequence of models, beginning with a simple model with only an intercept term, and fitting one additional effect on each step. Likelihood ratio statistics, that is, twice the difference of the log likelihoods, are computed between successive models. The order in which the terms are added therefore can affect the significance. To fully investigate models the order of entry should be varied. Clearly unless maximum likelihood is used for estimation, type I analysis will not be available.

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	-18901.363			
sex	-18873.365	1	28.00	<.0001
r_fed1	-18858.602	2	14.76	0.0006
r_race	-18539.857	3	318.74	<.0001
r_inc	-18532.134	2	7.72	0.0210
r_edu	-18530.077	2	2.06	0.3575
r_age1	-18529.775	1	0.30	0.5825

Sex, age race and federal assistance were significant, by a type I analysis.

Type III analysis is similar to the type III sum of squares analysis performed in PROC GLM. Generalized score tests for Type III contrasts are computed when General estimating equations are used. The results of type III analysis do not depend on the order in which the terms are specified in the MODEL statement.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
sex	1	10.20	0.0014
r_fed1	2	3.44	0.1789
r_race	3	288.23	<.0001
r_inc	2	5.07	0.0791
r_edu	2	1.99	0.3702
r_age1	1	0.30	0.5825

By type III analysis federal assistance and income appear to no longer be significant.

The "Criteria For Assessing Goodness Of Fit" table contains statistics that summarize the fit of the specified model. These statistics are helpful in judging the adequacy of a model and in comparing it with other models under consideration. We use this statistic to compare our model to the other model with normal errors. We compare the normal and gamma distribution assumptions we see that the gamma assumption is better but not ideal.

Cumulative residual plots (Johnston and So, 2003) show the value of these in assessing fit. This is available in PROC GENMOD and PROC PHREG procedures in the version 9.1 release.

SUMMARY

For the gene expression, we find that deletion of an outlier, a transformation of the dependent variable and a transformation of some independents improved the fit, so that the model was satisfactory.

For the nutrition data, procedures REG, ROBUSTREG, LOESS, and GAM were conducted and the R-square's are reported as low as 0.16. Compared to ordinary regression, the R-square increased from 0.16 to 0.28 by TRANSREG. The GENMOD procedure was used in a generalized linear model when the assumption of normal distribution for error term may not be valid. The gam model seems better but not ideal. Although we never really found an ideal model, consistently age, race and income came out significant. Consistently age and race and gender seem to play a role in caffeine consumption.

ACKNOWLEDGEMENTS

This work was mainly funded under the Lower Mississippi Delta Nutrition Intervention Research Initiative, USDA ARS grant # 6251-53000-003-00D.

REFERENCES

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

SAS Institute Inc. (1999), *SAS OnlineDoc®, Version 8*, Cary, NC: SAS Institute Inc.

Chen, C. (2002), "Robust Regression and Outlier Detection with the ROBUSTREG Procedure", *Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Johnston, G. and So, Y. (2003), "Let the Data Speak: New Regression Diagnostics Based on Cumulative Residuals", *Proceedings of the Twenty-Eighth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Contact the author at:

Pippa Simpson
UAMS Pediatrics / Section of Biostatistics
1120 Marshall St
Slot 512-43
Little Rock, AR 72202
Work Phone: (501) 364-6631
Work Fax: (501) 364-1552
Email: SimpsonPippaM@uams.edu

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. □