SESUG Paper SD-29-2017

Using Categorical Variables in Regression Analysis

Jonas V. Bilenas, Barclays UK PCB

ABSTRACT

In this tutorial, we will review how to deal with categorical variables in regression models using SAS®. Techniques will show how categorical variables can test for changes in intercept and slope effects in regression models. We will review how to use the CLASS statement in a number of regression procedures and we will also review how to select the reference level for categorical variables.

INTRODUCTION

Categorical variables are often used to evaluate experimental designs which are often balanced and orthogonal. Categorical independent variables are also often used in regression models which are often observational studies. Including categorical variables in regression models can impact the intercept of the model and also have an effect on slopes of continuous variables via interaction.

A categorical variable, also known as a nominal variable, have a number of values or levels that describe the variable. A simple example would be SEX which takes on 2 values; “M” for male and “F” for female. Since regression requires numerical variables, we need to create one or more numeric variables to describe the levels of SEX. The most popular coding of categorical variables is to use “Dummy Variables” also known as binary variables. For regression models, you will need a k-1 dummy variables for each categorical variable where k is the number of levels of the categorical variable. For this example, we would need 1 dummy variable that has 1 for one category level and 0 for the other levels. We could have males coded as 1 and females coded as 0. Or we can have females coded as 1 and males coded as 0. Only 1 variable is required for the regression analysis in the example.

If we have 3 levels of SEX; “M”, “F”, and “X” for other values, we would need 2 dummy variables in regression to account for the 3 possible levels. For example, we could have the first variable called MALES where males are coded as 1 and other levels are coded with a 0. The second variable called FEMALES would have females coded as 1 and other levels are coded with a 0. Those records with SEX coded with an ‘X’ are called reference variables since they would have a 0 value for both binary variables; MALE and FEMALE

Having the number of dummy variables equal to the number of category levels in regression models will cause a perfect collinearity of the dummy variables with the intercept term which will result, for most regression procedures, a warning message due to singularity issues where X’X cannot be inverted. However, we will see some examples in this paper where we do have singularity coding of dummy variables (k variables for k levels), but these are more related to experimental design evaluations where the singularity is required for type1-type3 sum of square tests (see Baron, S. and Mengeling, M.A. 2015).

LET’S GENERATE SOME SIMULATED DATA

In this paper, we will be using the SASHELP.CLASS data. The number of observations is small at 19 so let’s have 10 observations for each row with some variability on height and weight using CALL STREAMINIT and RAND functions (see Wicklin, R. 2013):

```
data class;
  set sashelp.class;
  call streaminit(20171105);
  do enhance=1 to 10;
    height = height+rand("normal")*2;
    weight = weight+rand("Normal")*4;
    output;
  end;
run;
```

The data set has a categorical variable called SEX with values of 'M' or 'F'. We could have coded up the binary variable in the data set created in the previous code. For example, F=(SEX='F'). However, many new modelling procedures in SAS can handle the coding directly in the regression procedure using a CLASS statement. We will see examples in this paper.

SOME EXPLORATORY ANALYSIS

Let's do some graphical analysis of the data we created in the sample code. We are using SAS 9.4 with STAT 14.1. Some of the options may not be available in earlier versions:

```
proc sgplot data=class;
  pbspline x=height
           y=weight/group=sex
           CLM alpha=0.05 NKNOTS=5 NOLEGCLM;
  xaxis grid;
  yaxis grid;
  title PBSPLINE Fit of Weight as a Function of Height and Sex;
run;
```

The PBSPLINE statements generates a smoothed fit using penalized cubic splines. For more information on splines see (Bilenas, J. & Herat, N. 2016) and (Hastie, T., Tibshirani, and Friedman, J. 2009). The NKNOTS statement is optional with a default of 100 if not specified. Results with 5 knots are not much different since the smoothing algorithm of PBSPLINE determines the optimal number of knots with penalization to over testing.

CLM and alpha=0.05 are options for generating the confidence limits of the mean of the fit. NOLEGCLM removes the legend for the confidence limits. Note that the 2 spline lines and confidence limits are built individually for Males and Females. We will probably see different results when we model the entire data in a single regression model.

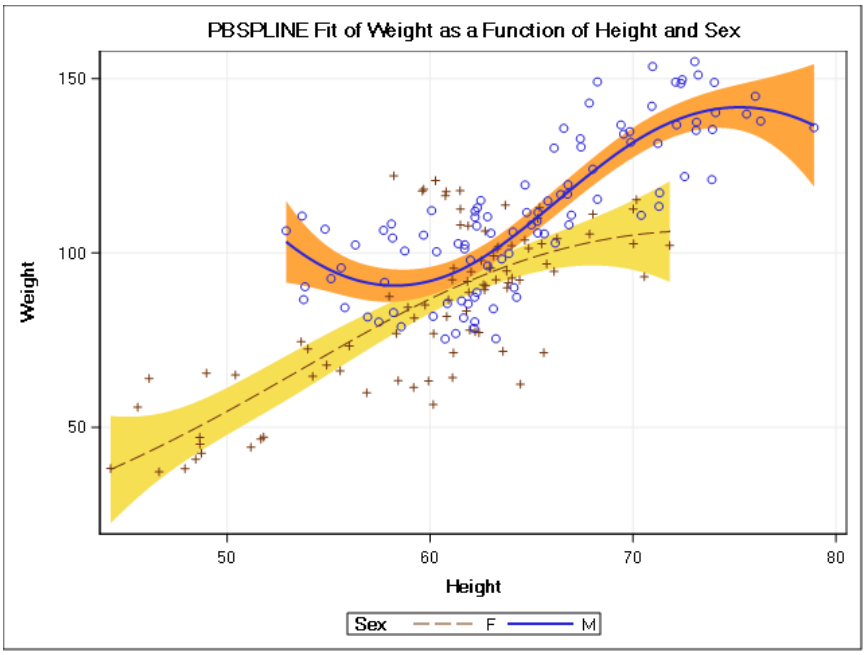

Figure 1. PBSPLINE

We can also look at a linear regression fit for each category of SEX (Male and Female):

```
proc sgplot data=class;
  REG x=height y=weight
      /group=sex
       CLM alpha=0.05;
  xaxis grid;
  yaxis grid;
  title Regression Fit of Weight as a Function of Height and Sex;
run;
```

Figure 2 shows the plot. Keep in mind that each regression is done on each category. In the next section, we will run many regression examples on the entire data set which includes both male and female records. This plot does not test to see if there are any differences between males and females in terms of intercept and HEIGHT slope so, we may see different results in the regression model.

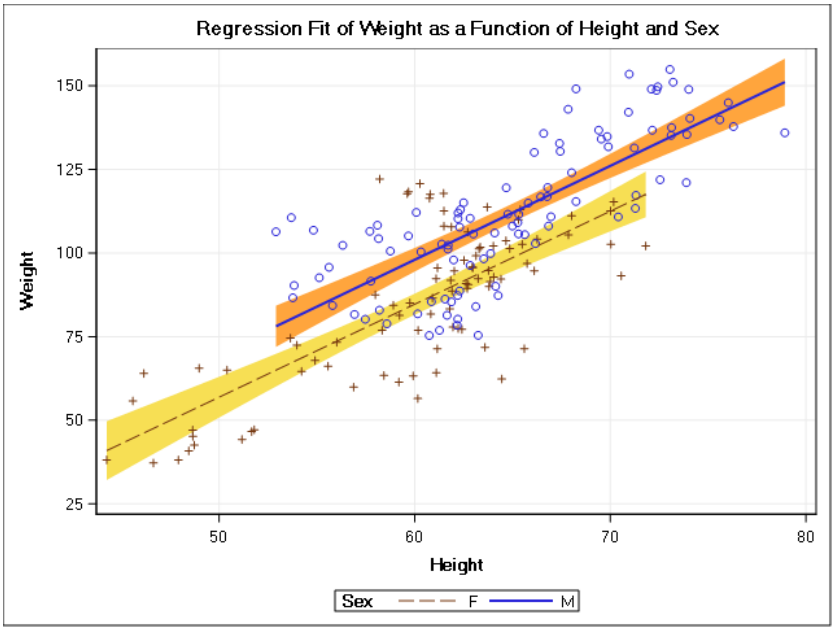

Figure 2. LINEAR REGRESSION FIT Using SGPLOT

USING GENMOD PROCEDURE FOR REGRESSION MODELS

The GENMOD procedure in SAS is very handy in that it can handle many different distributions and transformation links. It also has the CLASS statement to include the coding up of categorical variables. There are many CLASS specifications that are available in PROC GENMOD with the DEFAULT set to GLM which is typically used in experimental designs. In this paper, we will be looking at GLM and REF coding of categorical variables. There are many other coding methods, mostly dealing with experimental designs. Here is the listing from SAS® support

http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_genmod_syntax05.htm#statug.genmod.classstmtdefault for CLASS coding for GENMOD is SAS/STAT 14.1:

EFFECT	Effect coding
GLM	Less-than-full-rank reference cell coding (this *keyword* can be used only in a global option)
ORDINAL THERMOMETER	Cumulative parameterization for an ordinal CLASS variable
POLYNOMIAL POLY	Polynomial coding
REFERENCE REF	Reference cell coding
ORTHEFFECT	Orthogonalizes PARAM=EFFECT coding
ORTHORDINAL ORTHOTHERM	Orthogonalizes PARAM=ORDINAL coding
ORTHPOLY	Orthogonalizes PARAM=POLYNOMIAL coding
ORTHREF	Orthogonalizes PARAM=REFERENCE coding

Table 1. CLASS Options for PROC GENMOD

LINEAR MODEL WITH INTERACTION BETWEEN SEX AND HEIGHT USING GENMOD

Figure 1 showed a nonlinear relationship between WEIGHT as a function of HEIGHT but let's look at a linear model first. Code is shown here:

```
proc genmod data=class;
  CLASS sex/ missing;
  model weight=height | sex
        /dist=NOR wald type3;
run;
```

I should have used a different name for the data set, other than class. Some items to point out in the code:

- The default CLASS option in GENMOD is the same as GLM; GLM. This results in a singularity since we will have 2 dummy variables instead of 1. However, using the GLM class options will converge since one of the dummy variables will have their regression coefficient to 0. The GLM classing is required to calculate TYPE1 and TYPE3 analysis of effects similar to using PROC GLM.
- The missing options in the CLASS statement requests that missing values for SEX be included as level in the categorical variable SEX. This example has no missing values.
- dist=NOR specifies that this is a linear regression model. As mentioned above, GENMOD can be used for many linear models including LOGISTIC, BINOMIAL, POISSON, and many others.

- WALD and TYPE3 are used to compute Type 3 tests for each factor after removing the variance explained by all other variables. This is typically done for experimental designs and may not be reliable for unbalanced factors in observational linear regression models. A TYPE 1 analysis can also be requested which evaluates the model on a sequential basis starting with the intercept and adding each variable in the MODEL. This will not be illustrated in the paper.
- Also check what other modelling procedures use as the default CLASS coding. Not knowing can make an impact on interpretation of the output and implementation of the model.

Here is some of the output form the code above:

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	186	38672.2962	207.9156
Scaled Deviance	186	190.0000	1.0215
Pearson Chi-Square	186	38672.2962	207.9156
Scaled Pearson X2	186	190.0000	1.0215
Log Likelihood		-774.6045	
Full Log Likelihood		-774.6045	
AIC (smaller is better)		1559.2090	
AICC (smaller is better)		1559.5351	
BIC (smaller is better)		1575.4442	

OUTPUT 2. GENMOD FIT CRITERIA

Output 1 provides goodness of fit tests. R-Square is not provided but that is ok since that metric is overrated. AIC, AICC, and BIC offer other metrics of fit adjusting for the number of observations and variables in the model (Burnham & Anderson 2004). The smaller the metric the better the model. The only drawback is that it cannot be compared to models generated in different data sets. BIC seems to be most popular at the moment.

Also note that GENMOD does not have stepwise or other variable selection methods. That is good since STEPWISE results are not that reliable (see Flom, P. L and Cassell, D.L., 2007).

Let's look at the analysis of parameters:

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		**DF**	**Estimate**	**Standard Error**	**Wald 95% Confidence Limits**		**Wald Chi-Square**	**Pr > ChiSq**
Intercept		1	-70.3166	15.6507	-100.991	-39.6419	20.19	<.0001
Height		1	2.8052	0.2403	2.3343	3.2761	136.32	<.0001
Sex	F	1	-11.7455	21.5941	-54.0691	30.5781	0.30	0.5865
Sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Height*Sex	F	1	-0.0258	0.3441	-0.7002	0.6486	0.01	0.9402
Height*Sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	14.2667	0.7319	12.9020	15.7757		

Output 2. GENMOD Analysis of Maximum Likelihood Parameter Estimates

Output 2 shows that SEX and SEX*HEIGHT are not significant in the linear model. When we include both SEX populations there is no evidence that SEX has an effect on Weight in a LINEAR model. The only variable significant is HEIGHT.

Recall Figure 2 which did show 2 regression fits that did not have confidence intervals overlapping. However, that model was done for each SEX group independently. When we run 1 model with both SEX groups we have a larger N and based on the variability of data in the linear model we conclude that there is no difference between MALES and FEMALES. But wait, a linear relationship does not look appropriate. This will be cleaned up later in the paper.

Note that the Scale parameter is not part of the model parameters. For normal linear regression, it is the estimated standard deviation of the errors.

Let's look at the TYPE 3 analysis:

Wald Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
Height	1	263.40	<.0001
Sex	1	0.30	0.5865
Height*Sex	1	0.01	0.9402

Output 3. GENMOD TYPE 3 Analysis

Output 3 shows similar analysis as Output 2. The Chi-Square values are a bit different since each effect is tested after taking into account of the contribution to the other effects. So, HEIGHT after adjusting for SEX and HEIGHT*SEX is still significant which makes sense since output 2 showed no other parameters were significant at the 0.05 level, if that is your cut-off.

Let's look at the same model but now use the non-singular coding of the categorical variable. Only one dummy variable for the categorical variable, SEX. Here is the code:

```
proc genmod data=class;
  class sex/order=freq param=ref ref=first missing;
  model weight=height | sex
        /dist=NOR wald type3;
run;
```

The CLASS statement changed a bit. PARAM=REF specifies we want 0, 1 coding and, in this example, only 1 dummy variable for SEX. The ORDER=FREQ option orders the frequency by level from highest to lowest. The data has more males than females so males are the reference level and get the 0 code. Output 4 illustrates the GENMOD coding for the categorical variable, sex:

Class Level Information		
Class	**Value**	**Design Variables**
Sex	**M**	0
	F	1

Output 4. Class Level Information for SEX using PARAM=REF

Why would you want the level with the largest frequency in the data at the reference level of 0? The main reason is that you don't want the dummy variable to be correlated with the intercept term which could result in a collinearity issue when inverting X'X. For example, if 99% of the sample were males and had received a 1 value in the dummy variable it would be collinear with the intercept term. If the males received a 0 the dummy variable would not be collinear with the intercept. I have also heard that coding this way will also improve processing time but I haven't tested that out.

The Goodness of Fit test matches to Output 1 so it is not shown. The next output is the analysis of parameters using the REF coding:

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-70.3166	15.6507	-100.991	-39.6419	20.19	<.0001
Height		1	2.8052	0.2403	2.3343	3.2761	136.32	<.0001
Sex	F	1	-11.7455	21.5941	-54.0691	30.5781	0.30	0.5865
Height*Sex	F	1	-0.0258	0.3441	-0.7002	0.6486	0.01	0.9402
Scale		1	14.2667	0.7319	12.9020	15.7757		

Output 5. Analysis of Parameter Estimates for PARAM=REF CLASS Coding

The regression estimates in output 5 are the same along with Chi-Square and p-values. What's the difference, other than the dummy coding? Let us examine the WALD TYPE3 output in output 6:

Wald Statistics For Joint Tests			
Source	DF	Chi-Square	Pr > ChiSq
Height	1	136.32	<.0001
Sex	1	0.30	0.5865
Height*Sex	1	0.01	0.9402

NOTE: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests.
The joint test for an effect is a test that all the parameters associated with that effect are zero.
Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Output 6. TYPE3 Analysis with PARAM=REF CLASS Coding

The output in table 6 warns that the TYPE3 tests with this dummy coding is not the same as in GLM coding. The Chi-square value is different than the one in table 3 for the HEIGHT EFFECT. It is a joint test that jointly tests for all parameters associated with HEIGHT are 0 under the null hypotheses. Each joint test effect on parameters can be interpreted as follows:

- `SOURCE=HEIGHT.` H_0: `HEIGHT=0 and HEIGHT*SEX=0.` `Reject the null`
- `SOURCE=SEX.` H_0: `SEX=0 and HEIGHT*SEX=0.` `Failed to reject the null`
- `SOURCE=HEIGHT*SEX.` H_0: `HEIGHT*SEX=0.` `Failed to reject the null`

LINEAR MODEL WITH CURVATURE TERMS ADDED TO THE MODEL

Figure 1 showed that the relationship between HEIGHT and WEIGHT is not linear and changes by SEX. We could add some spline terms to the model but let's keep it simple and add a square term for height directly in the GENMOD code. Note that you can add square terms in the MODEL statement as opposed to the data set; height*height. Same for any term raised to any integer value.

We will first look at GLM coding and then compare to REF coding. Here is the GLM code:

```
proc genmod data=class namelen=38;
  class sex/order=freq param=GLM ref=first missing;
  model weight=height | sex
              height*height
              height*height*sex
     /dist=NOR wald type3;
 run;
```

Output 7 will compare some of the metrics we observed in Output 1 to see if there is an improvement in fit:

Criteria For Assessing Goodness Of Fit			
Criterion	**DF**	**Value**	**Value/DF**
Deviance	184	36690.2858	199.4037
Scaled Deviance	184	190.0000	1.0326
Pearson Chi-Square	184	36690.2858	199.4037
Scaled Pearson X2	184	190.0000	1.0326
Log Likelihood		-769.6064	
Full Log Likelihood		-769.6064	
AIC (smaller is better)		1553.2129	
AICC (smaller is better)		1553.8282	
BIC (smaller is better)		1575.9420	

Output 7. Goodness of Fit with Square Terms

The BIC increased slightly over the first model at 1575.4442 due to added terms. But remember the first model had 2 non-significant terms and was not the optimal model in explaining the relationship of WEIGHT as a function of HEIGHT and SEX. Let's look at the parameter estimates:

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	322.8391	146.1902	36.3115	609.3667	4.88	0.0272
Height		1	-9.3572	4.5039	-18.1846	-0.5298	4.32	0.0377
Sex	F	1	-584.252	180.2199	-937.477	-231.028	10.51	0.0012
Sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Height*Sex	F	1	18.4377	5.8132	7.0440	29.8313	10.06	0.0015
Height*Sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Height*Height		1	0.0933	0.0345	0.0257	0.1609	7.31	0.0068
Height*Height*Sex	F	1	-0.1479	0.0469	-0.2399	-0.0560	9.94	0.0016
Height*Height*Sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	13.8963	0.7129	12.5670	15.3661		

Output 8. Analysis of Parameters with Square Terms

All regression parameters are all significant at 0.05 level. What about TYPE 3 which may not be valid if the data is not balanced:

Wald Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Height	1	0.00	0.9620
Sex	1	10.51	0.0012
Height*Sex	1	10.06	0.0015
Height*Height	1	7.31	0.0068
Height*Height*Sex	1	9.94	0.0016

Output 9. WALD TYPE 3 Analysis with Square Terms

In output 9 we see that the HEIGHT term after accounting for other terms is not significant. However, typical regression practice states that if an interaction effect is significant then the main effect should be retained even if the main effect term is not significant in the analysis.

Here is the CLASS=REF version which maybe more valid in observational regression models. Code:

```
proc genmod data=class namelen=38;
  class sex/order=freq param=ref ref=first missing;
  model weight=height | sex
              height*height
              height*height*sex
     /dist=NOR wald type3;
  output out=pred p=p l=Lower u=Upper;
run;
```

Note the **NAMLEN=38** option. Sometimes variable names, especially with interactions, don't fit on the SAS listing and are truncated. You can extend the length using the handy NAMELEN option.

Output 10 had the same Goodness of Fit tests and Parameter estimate analysis shown in output 8:

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	322.8391	146.1902	36.3115	609.3667	4.88	0.0272
Height		1	-9.3572	4.5039	-18.1846	-0.5298	4.32	0.0377
Sex	F	1	-584.252	180.2199	-937.477	-231.028	10.51	0.0012
Height*Sex	F	1	18.4377	5.8132	7.0440	29.8313	10.06	0.0015
Height*Height		1	0.0933	0.0345	0.0257	0.1609	7.31	0.0068
Height*Height*Sex	F	1	-0.1479	0.0469	-0.2399	-0.0560	9.94	0.0016
Scale		1	13.8963	0.7129	12.5670	15.3661		

Output 10. Analysis of Parameters with Square Terms and CLASS=REF

Let's take a look at the TYPE3 analysis in output 11:

Wald Statistics For Joint Tests			
Source	DF	Chi-Square	Pr > ChiSq
Height	1	4.32	0.0377
Sex	1	10.51	0.0012
Height*Sex	1	10.06	0.0015
Height*Height	1	7.31	0.0068
Height*Height*Sex	1	9.94	0.0016

NOTE: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests.
The joint test for an effect is a test that all the parameters associated with that effect are zero.
Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Output 11. WALD TYPE 3 Analysis with Square Terms and CLASS=REF

The joint tests in output 11 are significant. All main effects and interactions are significant in both output 11 and 10 so let's keep the model.

In the previous code we added an output statement. We can now run code to get the graphical look at the fit:

```
proc sort data=pred tagsort noequals force;
  by sex height weight;
run;

proc sgplot data=pred;
  band x=height lower=Lower upper=Upper / group=sex;

  scatter y=weight x=height / group=sex;
  series y=p x=height
    /  lineattrs=GraphPrediction group=sex;

  xaxis grid; yaxis grid;
  KEYLEGEND;
run;
```

Figure 3 has the results:

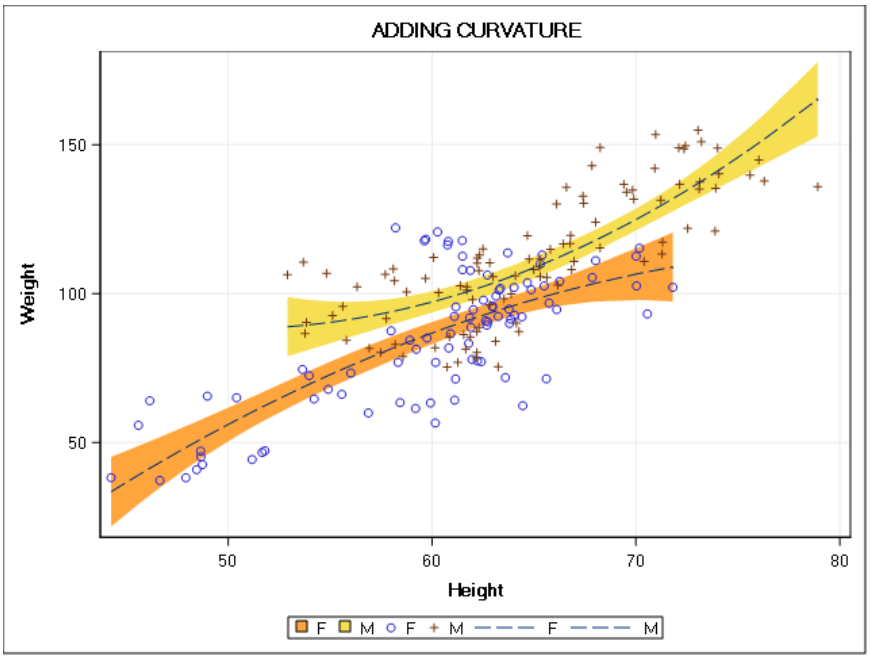

Figure 3. Fit of the Final Model.

You can try adding a cubic term (HEIGHT*HEIGHT*HEIGHT) along with interactions but the model failed to converge. Are you ready for a categorical variable with 3 levels?

MODELLING WITH A 3 LEVEL CATEGORICAL VARIABLE

The data set is generated off the original data:

```
data class2;
  set class;
  call streaminit(20171105);
  output;
  if sex='F' & rand("uniform") <= 0.5 then do;
    sex='O';
    weight=weight+50;
    output;
  end;
run;
```

We will focus only on the REF= class coding. Code for the model and model fit is shown next:

```
proc genmod data=class2 namelen=38;
  class sex/order=freq param=ref ref=first missing;
  model weight=height | sex
              height*height
              height*height*sex
     /dist=NOR wald type3;
  output out=pred p=p l=Lower u=Upper;
run;


proc sort data=pred tagsort noequals force;
  by sex height weight;
run;

proc sgplot data=pred;
  band x=height lower=Lower upper=Upper / group=sex;

  scatter y=weight x=height / group=sex;
  series y=p x=height
    /  lineattrs=GraphPrediction group=sex;

  xaxis grid values=(45 to 80 by 5);
  yaxis grid values=(30 to 180 by 20);
  KEYLEGEND / DOWN=3;
run;
```

Results are shown in output 12 and figure 4:

Class Level Information			
Class	Value	Design Variables	
Sex	M	0	0
	F	1	0
	O	0	1

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	322.8391	147.7331	33.2875	612.3907	4.78	0.0289
Height		1	-9.3572	4.5514	-18.2778	-0.4366	4.23	0.0398
Sex	F	1	-584.252	182.1220	-941.205	-227.300	10.29	0.0013
Sex	O	1	-586.764	205.1913	-988.932	-184.597	8.18	0.0042
Height*Sex	F	1	18.4377	5.8746	6.9237	29.9516	9.85	0.0017
Height*Sex	O	1	20.4421	6.7391	7.2336	33.6506	9.20	0.0024
Height*Height		1	0.0933	0.0349	0.0250	0.1616	7.16	0.0075
Height*Height*Sex	F	1	-0.1479	0.0474	-0.2408	-0.0550	9.73	0.0018
Height*Height*Sex	O	1	-0.1663	0.0554	-0.2749	-0.0578	9.02	0.0027
Scale		1	14.0430	0.6450	12.8340	15.3658		

Wald Statistics For Joint Tests			
Source	DF	Chi-Square	Pr > ChiSq
Height	1	4.23	0.0398
Sex	2	11.77	0.0028
Height*Sex	2	12.52	0.0019
Height*Height	1	7.16	0.0075
Height*Height*Sex	2	12.83	0.0016

NOTE: Under full-rank parameterizations, Type 3 effect tests are replaced by joint tests.
The joint test for an effect is a test that all the parameters associated with that effect are zero.
Such joint tests might not be equivalent to Type 3 effect tests under GLM parameterization.

Output 12. GENMOD LINEAR REGRESSION RUN with a 3-LEVEL Categorical Variable

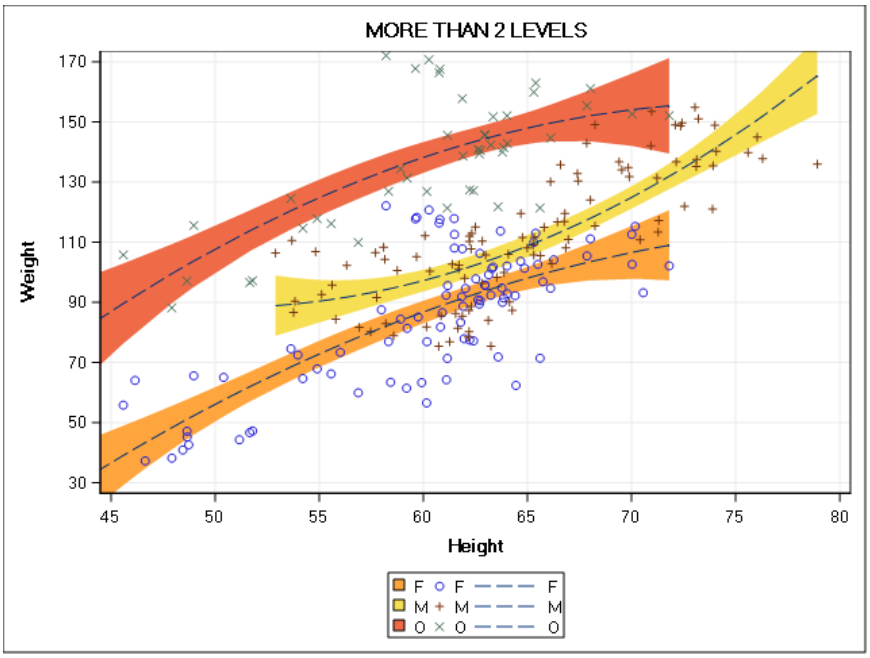

Figure 4. 3 level categorical variable with interactions.

All terms are significant fit results look good.

CONCLUSION

Building regression models that include categorical variables are easy to do with many of the new SAS regression procedures. Including categorical variables in regression models with the REF parameterization can improve your model prediction accuracy. Also including interaction effects between the added categorical variables with continuous independent variables may provide more insight into your model's prediction as well as improve prediction accuracy.

One thing to be careful when using CLASS statements in different regression procedures is to be familiar with the default parameterization. For example, GENMOD has the default set to GLM while LOGISIC uses EFFECT parameterization. Not knowing what the default parameterization could impact on the interpretation and implementation of the model.

Try the PARAM=EFFECT with the simulated data using the 3-level categorical variable. Do you get similar results as shown with the REF parameterization? The prediction plots are the same but the regression coefficients and p-values associated with those coefficients are different due to the fact that the coding of categorical variables have changed from binary (0, 1) to a balanced coding of (-1, 0, 1).

REFERENCES

Barron, Sheila and Mengeling, Michelle A. 2015. “Sum of Squares: The Basics and a Surprise.” Available at https://support.sas.com/resources/papers/proceedings15/1521-2015.pdf

Bilenas, Jonas V. and Herat, Nish 2016. “Using Regression Splines in SAS® STAT Procedures” SouthEast SAS Users Group 2016 Conference, Bethesda, Maryland. Available at http://analytics.ncsu.edu/sesug/2016/BF-140_Final_PDF.pdf

Burnhamm, Kenneth P. and Anderson, David R. Multimodel Inference: Understanding AIC and BIC in Model Selection” Sociological Methods & Research. Available at: http://journals.sagepub.com/doi/abs/10.1177/0049124104268644 and http://journals.sagepub.com/doi/abs/10.1177/0049124104268644

Flom, Peter L. and Cassell, David L. 2007. “Stopping Stepwise: Why Stepwise and Similar Selection Methods are bad, and what Should You Use.” North East SAS Users Group (NESUG) 2007 conference in Baltimore, Maryland. Available at http://www.lexjansen.com/pnwsug/2008/DavidCassell-StoppingStepwise.pdf

Hastie, Trevor, Tibshirani, Robert and Friedman, Jerome 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, 2nd Edition. Springer.

Wicklin, Rick 2013. “Six reasons you should stop using the RANUNI function to generate random numbers.” Available at https://blogs.sas.com/content/iml/2013/07/10/stop-using-ranuni.html

SAS CLASS CODING in PROC GENMOD: http://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_genmod_syntax05.htm#statug.genmod.classstmtdefault for CLASS coding for GENMOD is SAS/STAT 14.1:

ACKNOWLEDGMENTS

Thanks to all the SAS statisticians and SAS gurus over the last 30+ years including, in no particular order; Ming Wang, Melissa Dietz, Nish Herat, Richard Hixson, Peter Flom, WenSui Liu, Sonia Sun, George Soulellis, Pini Ben-Or, Anita Blanchard, David Cassell, Art Carpenter, David Corliss, Ronald Fehd, Mike Zdeb, Julia Menichilli, Ian Whitlock, Paul Dorfman, Sunil Gupta, David Horvath, Michael Davis, Wendi L. Wright, many more who I may have forgotten or will run out of space to include so many! I also thank my family for their understanding of the time I spend on improving my SAS skills.

CONTACT INFORMATION

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

Jonas V. Bilenas
Jonas.Bilenas@gmail.com