Table of Contents

Introduction of General Linear Model Procedure (PROC GLM)

SAS has several procedures for analyze of variance models, for example, PROC ANOVA, PROC GLM and PROC MIXED. PROC GLM also known as the "flagship" procedure for analysis of variance. PROC GLM is short for Generalized Linear Model Procedure in SAS. It fit general linear models by using the lease squares method. It can read the data with multiple independent and dependent variables and fit with the generalized regression to play with the data. For example, it can used to test the association of smoking and pneumonia. The model is sensitive to both continuous variables and classification variables such as discrete groups. The GLM procedure have cluster functions including simple, multiple, weighted, polynomial regressions. The interactive feature of GLM allows users to execute further statements without recomputing the model parameters or sum of squares. GLM also includes analytical tools such as ANOVA, Multivariate ANOVA for users to obtain the different parameters of the models. Moreover, GLM can find out the partial correlation between specified variables. Comparing to other procedures in SAS, GLM has collective analyzing functions within the framework of general linear models. The Output Delivery System of GLM procedure also comes with the corresponding statistical graphics. In this report, we will introduce the use of PROC GLM and fit an application by use a real-world dataset.

Data Description

The data was collected by Ronny Kohavi and Barry Becker in 1994. (Dua & Graff, 2019) The adult.data was collected from 48842 participant's personal information. There are approximately 32561 observations with 14 variables as shown in Table 1. Each observation measures one participant's personal information.

Preprocessing

1. Modifying the adult.data to convert the type of variable "earning" from categorical to numerical named "numeric_earning" variable. There are two groups of earning level in "earning" variable: "<=50K" and ">50K". We assign "0" to "numeric_earning" variable indicate "<=50K" group, assign "1" to "numeric_earning" variable indicate ">50K" group.
2. Modifying the adult.data to create a new numeric variable "age_group" that represents age groups as: age_group1 represent participant's age less or equal to 30. The different group's lag is 10. (If age is missing value, then group to age_group8)

Syntax of PROC GLM Fitting to the Data

One-way ANOVA

One of the crucial features of PROC GLM is One-way analysis of variance (ANOVA). One-way ANOVA is a statistical method used for testing for differences in the means of three or more groups. First, we need to set null hypothesis (H_0) that more than three population means are equal, and alternative hypothesis (H_a) that at least one mean is different.

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$
$$H_a: not\ all\ means\ are\ equal$$

Here μ_i is the mean of the i-th level of the factor.

The syntax of do one-way ANOVA by PROC GLM:

```
PROC GLM
      DATA = ds;
      CLASS <categorical_variable>;
      MODEL <variable>=<categorical_variable>;
      MEANS <categorical_variable> / HOVTEST = <test methods> <other
options>;
```

Details:

1. MODEL Statement: names the dependent variables and independent effect variable. Note that only one MODEL statement can be specified in the GLM procedure.
2. CLASS Statement: names the classification variables used in the model. Note that it must be placed before the MODEL statement
3. MEANS Statement: PROC GLM calculates the arithmetic mean and standard deviation of all continuous variables (dependent and independent) in the model. You can specify only categorical effects in the MEANS statement. Note: HOVTEST means test for homogeneity of variances test. The test methods can be Levene test (default) or a Bartlett test on the data. Either one is fine, and we should put only one of them.
4. <other options>

a) “welch” option: Welch’s test in all cased when the hypothesis group variances are not equal. Welch's ANOVA is robust to violation of the assumption of equal variance for one-way models.

One-way ANOVA and application

In the following example, we analyze “numeric_earning” variable and “education_num” variable to investigate the association of an individual's level of education and income. In this example, we set the null hypothesis: Different levels of the education have the same effect on average income. Set the alternative hypothesis: At least one education level has different effect on the income than others.

Following “education_num” variable is indicating different levels of education; “numeric_earning” variable is indicating different levels income per year, represent by binary number.

example: adult.data

```
proc glm
      data=proj.adult;
      class education_num;
      model numeric_earning = education_num;
      means education_num / hovtest=levene;
run;
```

From output of One-way ANOVA by Levene’s Test for equal variance (Table 2), we obtain that p-value<.0001. This implies that we reject null hypothesis. We have strong evidence that at least one education level has different effect on the income than other education level.
From Box-plot of different education level’s income (Figure 1) shows the distribution of numeric_earning by education, we observe that the average of income of participants increases with education advancement. Besides, from 14 education level onwards, more than half of the participants have an average income greater than 50k.

Unbalanced Two-way ANOVA

Two-way ANOVA is a statistical method used for analyzing for differences in the effects of two independent variables (factors) on a dependent variable. It will find out the two explanatory variables effect on the dependent variable plus the joint effect of the two variables. Besides, unbalanced Two-way ANOVA means there exists unequal size of sample between each group.

Firstly, Let’s test the interaction effect. Set the model1:

$$Y = \mu + \tau_i + \beta_j + \epsilon$$

Here $\mu\ is\ the\ grand\ mean$

$\tau_i\ is\ the\ effect\ of\ treatment\ i, i = 1, \ldots., T$

$\beta_j\ is\ the\ effect\ of\ block\ j, j = 1, \ldots., B$

$\epsilon\ is\ the\ error$

The next step is checking does exists interaction effect between two factors by the interaction plot. If the different lines (group's overall trend) are parallel, means the differences are equal, there is no interaction. If there are some crossing parts between different lines or the lines are not even close to parallel, means the differences are different, there might exist interaction between two factors. If the interaction is observed, we need to add the interaction to the model.

The third step: set the model2:

$$Y = \mu + \tau_i + \beta_i + \gamma_{ij} + \epsilon$$

Here $\mu\ is\ the\ grand\ mean$

$\tau_i\ is\ the\ effect\ of\ level\ i\ of\ the\ first\ factor, i = 1, \ldots., T$

$\beta_j\ is\ the\ effect\ of\ level\ j\ of\ the\ second\ factor, j = 1, \ldots., B$

$\gamma_{ij}\ is\ the\ interaction\ effect\ for\ thee\ (i,j)th\ combination\ of\ factors$

$\epsilon\ is\ the\ error$

There are there sets of hypotheses with the two-way ANOVA:

Set 1: test for interaction

$$H_0: \gamma_{ij} = 0\ for\ all\ i,j$$

$$H_a: at\ least\ one\ \gamma_{ij}\ not\ 0$$

Set2: test for one factor of the main effects

$$H_0: \alpha_i = 0\ for\ all\ i$$

$$H_a: at\ least\ one\ \alpha_i\ not\ 0$$

Set3: test for another of the main effect

$$H_0: \beta_j = 0\ for\ all\ j$$

$$H_a: at\ least\ one\ \beta_j\ not\ 0$$

The syntax of do one-way ANOVA by PROC GLM:

```
PROC GLM
      DATA = ds;
      CLASS <categorical_variable>;
      MODEL <variable>=<categorical_variable1> <categorical_variable2>
<interaction terms>;
RUN;
```

(There is detail description of the syntax in one-way ANOVA)

Note:

- Two-way ANOVA, main effects only: `MODEL Y = A B;`
- Two-way ANOVA, factorial with interaction: `MODEL Y = A B A*B;` or `MODEL A|B;`

Two-way ANOVA and application

In the following example, we set "numeric_earning" variable as the response variable, and two factors: "occupation" variable and "age_group" variable as the predictor variable. We analyze the association between the two factors and the response variable. In this example, we firstly set the null hypothesis: there is no interaction between participant's age and occupation. Set the alternative hypothesis: the participant's age and occupation have interaction effect.

Following “age_group” variable is indicating different groups of age; “occupation” variable is indicating 14 different types of occupation; “numeric_earning” variable is indicating different levels income per year, represent by binary number.

```
ods graphics on;
      proc glm
      data=proj.adult plots=IntPlot plots(MAXPOINTS=10000000);
      class occupation age_group;
      model numeric_earning = occupation age_group occupation*age_group;
      run;
ods graphics off;
```

Note:

1. ODS GRAPHICS statement used to control many aspects of the graphics.
   a) Here we use `ods graphics on` to enables ODS Graphics processing, use `ods graphics off` to disables ODS Graphics processing
2. PLOTS is one of the PROC GLM statement options. It used to control the plots produces through ODS Graphics.
   a) Here we use `plots=IntPlot` to tell SAS provide the interaction plot.
   b) Since ODS graphics with more than 5000 points have been suppressed, we use `plots(MAXPOINTS=10000000)` option to change or override the cutoff.

From output of Overall test by two-way ANOVA (Table 3), we obtain that the overall F-test is significant (p-value<.0001), indicating that we have strong evidence that the participant’s means of earning for the different age and different occupation are different.

From output of Two-way ANOVA table with interaction (Table 4), we obtain that the occupation*age_group interaction is significant (p-value<.0001), indicating that we have strong evidence that the effect of groups of age depend on the occupation and vice versa. We can also say we have strong evidence the interaction effect is significant after controlling for age_group and occupation. Therefor, the tests for the individual effects are invalid.

From interaction plot for numeric_earning by different occupation and different age groups (Figure 2) we can observe that:

1. Overall. All 15 types of occupation have different trend on 6 different groups of age.
2. For the third age_group, it has the highest earning at Armed-Forces occupation.
3. All the line crossing each other in different occupation. This suggests the age group may depend on the occupation.

Multiple linear regression

Multiple linear regression is a method that can predict of new data by establishing a linear relationship between multiple predictor variables and one response variable. The multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{ip} + \epsilon_i$$

$$\text{Here } \beta_0 \textit{ is the intercept of the line}$$

$$\beta_i \textit{is the coefficient of different preddictor variables}$$

$$p \textit{ is the number of predictors}$$

$\epsilon_i is\ the\ error\ term$

Multiple linear regression and application

In the following example, we set "numeric_earning" variable as the response variable, and predictor variables are three different education levels. We fit these variables into the linear regression model.

Following "education_num" variable is indicating 16 different levels of education; "numeric_earning" variable is indicating different levels income per year, represent by binary number.

```
proc glm data=proj.adult;
      class  education_num;
      model numeric_earning = education_num / solution clparm;
run;
```

Note:

1. 'solution' option in proc glm to get a regression equation predicting earning using dummy variables to represent different education levels. (Here the baseline is education_num 16)
2. 'clparm' option to get confidence limits of these regression parameters.

From output of Multiple linear regression output (Table 5), we obtain that the model as:

Numeric_earning = 0.7409 – 0. 7409 * education_num 1 – 0. 7052 * education_num 2 – 0.6928* education_num 3 - 0. 6790 * education_num 4 – 0. 6883 * education_num 5 – 0.6745* education_num 6 - 0. 6899 * education_num 7 – 0. 6647 * education_num 8 – 0.5814* education_num 9 - 0. 5507 * education_num 10 – 0. 4797 * education_num 11 – 0.4926* education_num 12 – 0. 3262 * education_num 13 – 0.1843* education_num 14 - 0. 0065 * education_num 15 - 0 * education_num 16 + ϵ_i

Base on the equation above, if there is a future person belongs in education_num10, the numeric_earning = 0.7409 - 0.5507=0.1902

Conclusion

PROC GLM analyzes data within the framework of a general linear model. Besides, PROC GLM also deals with models that relate one or more continuous dependent variables to one or more independent variables. The independent variables do not restrict by the types of variables, it can be categorical, dividing the observations into discrete groups, or they can be continuous. Thus, GLM programs can be used for many different analyses. Instead of using several specific procedures, we can just use PROC GLM to obtain lots of the feedbacks.

This article is only a short introduction to the capabilities of PROC GLM, which are by no means limited to the above. For example, we can also use PROC GLM to fit different regression models.

Appendix

Table 1: Variables of adult.data

Variable	Description	Example
“age”	Participant’s age	20,21
“workclass”	Participant’s type of work	Private, self-emp-inc
“fnlwgt”	Final weight: It represents how many participant with same chrematistic in the real world	10,11
“education”	Participant’s education level	Bachelors, Some-college
“education_num”	Represent Participant’s education level by number	13, 10
“marital_status”	Participant’s marital status	Divorced, Never-married
“occupation ”	Participant’s occupation	Tech-support, Craft-repair
“relationship”	Participant’s social relationship	Wife, Own-child
“race”	Participant’s race	White, Asian-Pac-Islander
“sex”	Participant’s sex	Female, Male
“capital_gain”	Participant’s increase in capital assets	20, 21
“capital_loss”	Participant’s loss in capital assets	20, 21
“hours_per_week”	Participant’s work how many hours per week	20, 21
“native_country”	Participant’s native country	United-States, Cambodia
“Earning”	Participant’s income per year	<=50K, >50K
“numeric_earning”	Represent Participant’s earning level by binary number	0, 1
“age_group”	Group different age range	1, 2

Table 2: One-way ANOVA by Levene’s Test for equal variance

The SAS System

The GLM Procedure

Levene's Test for Homogeneity of numeric_earning Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
education_num	15	108.5	7.2345	161.31	<.0001
Error	32545	1459.6	0.0448		

Table 3: Overall test by two-way ANOVA

Dependent Variable: numeric_earning

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	99	1196.674327	12.087619	82.50	<.0001
Error	32461	4756.137933	0.146519		
Corrected Total	32560	5952.812260			

Table 4: Two-way ANOVA table with interaction

Source	DF	Type I SS	Mean Square	F Value	Pr > F
occupation	14	737.1268059	52.6519147	359.35	<.0001
age_group	6	358.3797940	59.7299657	407.66	<.0001
occupation*age_group	79	101.1677271	1.2806041	8.74	<.0001

Table5: Multiple linear regression output

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	0.7409200969	B	0.01956097	37.88	<.0001	0.7025798799	0.7792603139
education_num 1	-.7409200969	B	0.05900166	-12.56	<.0001	-.8565655251	-.6252746686
education_num 2	-.7052058111	B	0.03637674	-19.39	<.0001	-.7765055586	-.6339060636
education_num 3	-.6928720488	B	0.02927774	-23.67	<.0001	-.7502575083	-.6354865893
education_num 4	-.6790005922	B	0.02504506	-27.11	<.0001	-.7280898299	-.6299113546
education_num 5	-.6883909140	B	0.02626931	-26.21	<.0001	-.7398797263	-.6369021017
education_num 6	-.6744677925	B	0.02349482	-28.71	<.0001	-.7205184965	-.6284170884
education_num 7	-.6898562671	B	0.02274032	-30.34	<.0001	-.7344281399	-.6452843943
education_num 8	-.6647076257	B	0.02734208	-24.31	<.0001	-.7182991118	-.6111161397
education_num 9	-.5814114786	B	0.01994192	-29.16	<.0001	-.6204983772	-.5423245800
education_num 10	-.5506855611	B	0.02010735	-27.39	<.0001	-.5900967154	-.5112744069
education_num 11	-.4797044673	B	0.02229300	-21.52	<.0001	-.5233995707	-.4360093639
education_num 12	-.4925602093	B	0.02303769	-21.38	<.0001	-.5377149357	-.4474054830
education_num 13	-.3261675292	B	0.02030127	-16.07	<.0001	-.3659587675	-.2863762908
education_num 14	-.1843327492	B	0.02177952	-8.46	<.0001	-.2270214121	-.1416440863
education_num 15	-.0065450969	B	0.02563169	-0.26	0.7985	-.0567841529	0.0436939592
education_num 16	0.0000000000	B	.	.	.	.	.

Figure 1: Box-plot of different education level's income

The GLM Procedure

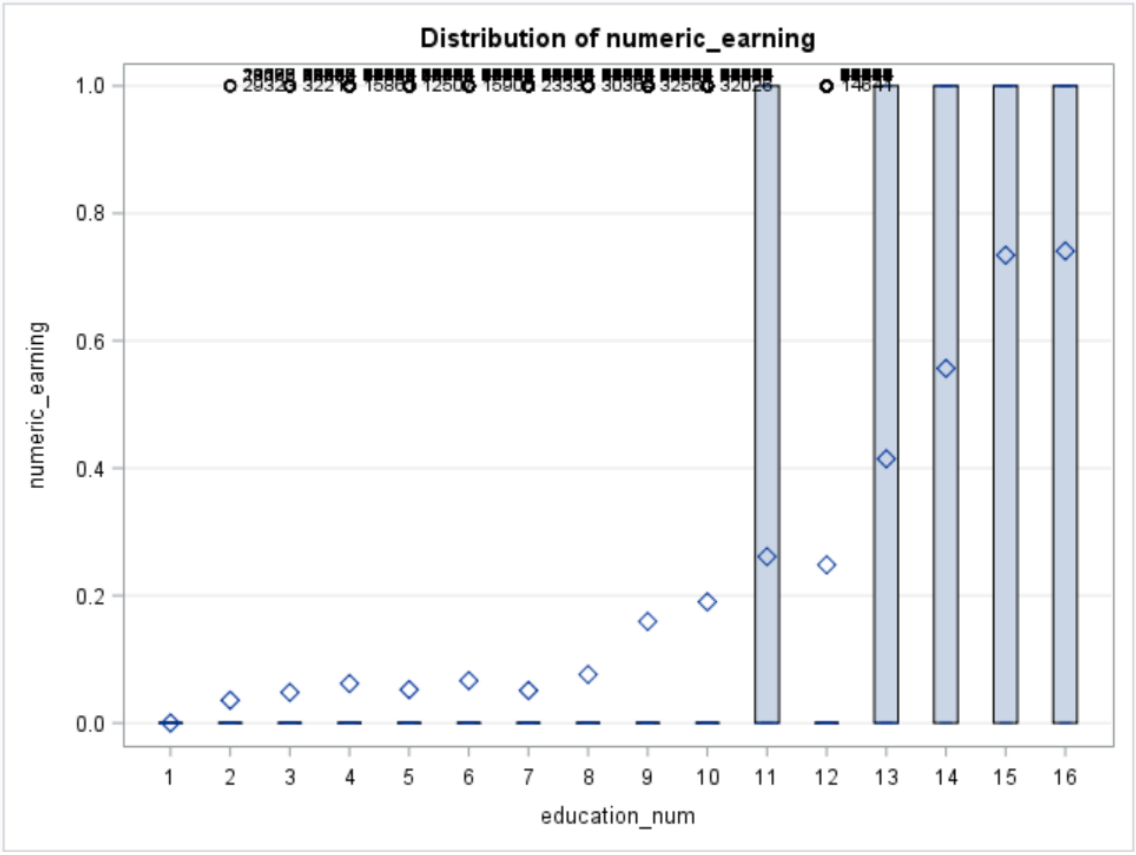
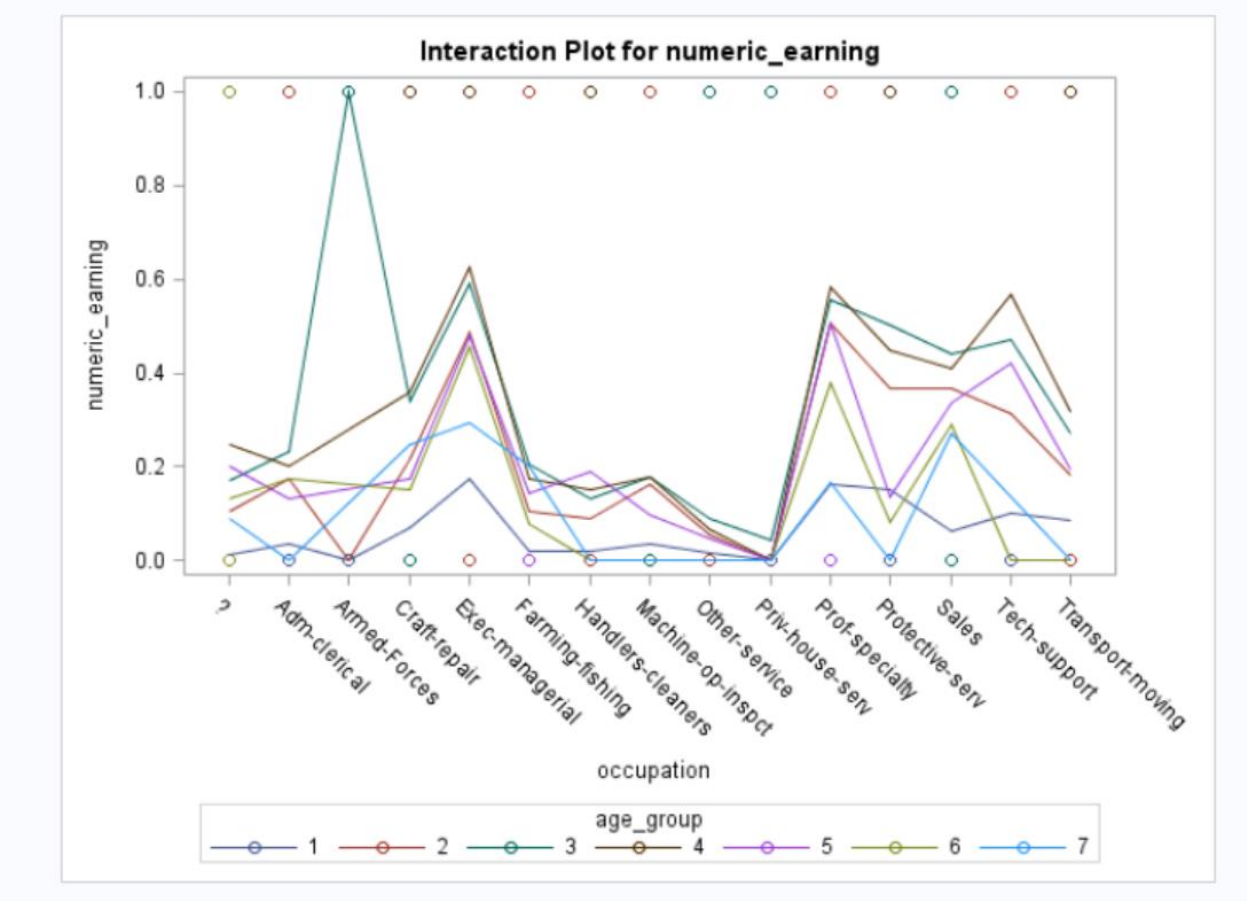


Figure 2: interaction plot for numeric_earning by different occupation and different age groups



Reference:

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.