Paper PO 15

Overview of Analysis of Covariance (ANCOVA) Using GLM in SAS®

Abbas S. Tavakoli, DrPH, MPH, ME, University of South Carolina, Columbia, SC

ABSTRACT

Analysis of covariance (ANCOVA) is a more sophisticated method of analysis of variance. Analysis of covariance is used to compare response means among two or more groups **(Categorical variables**) adjusted for a quantitative variable (**Covariate**), thought to influence the outcome (**Dependent**). A covariate is a continuous variable that can be used to reduce the Sum Square Error (SSE) and subsequently increase the statistical power of an ANOVA design. There may be more than one covariate. The purpose of this paper is to overview of Analysis of Covariance (ANCOVA) using GLM with two examples in SAS with interpretation to use for publication.

Keywords: ANCOVA

University of South Carolina, College of Nursing.

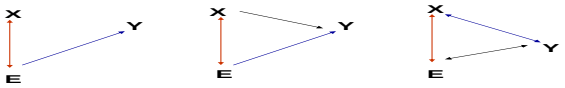
INTRODUCTION

Analysis of covariance (ANCOVA) is a more sophisticated method of analysis of variance. Analysis of covariance is used to compare response means among two or more groups **(Categorical variables**) adjusted for a quantitative variable (**Covariate**), thought to influence the outcome (**Dependent**). A covariate is a continuous variable that can be used to reduce the Sum Square Error and subsequently increase the statistical power of an ANOVA design. There may be more than one covariate. The purpose of this paper is to overview of Analysis of Covariance (ANCOVA) using GLM with two examples in SAS with interpretation to use for publication.

Overview ANCOVA

Extraneous variables (**E**) are those that have some relationship to the variables (X**= independent and Y=dependent**) included in the study. The extraneous variables can be categorized into five categories:

1. Confounder: When a third variable (E) is related to **X (independent) i**n a non-causal manner and is associated to **Y (dependent)** either causally or correlation. There are three possible confounding situations can be shown as following (a bi-directional arrow indicates a correlation while a unidirectional arrow represents a causal relationship):

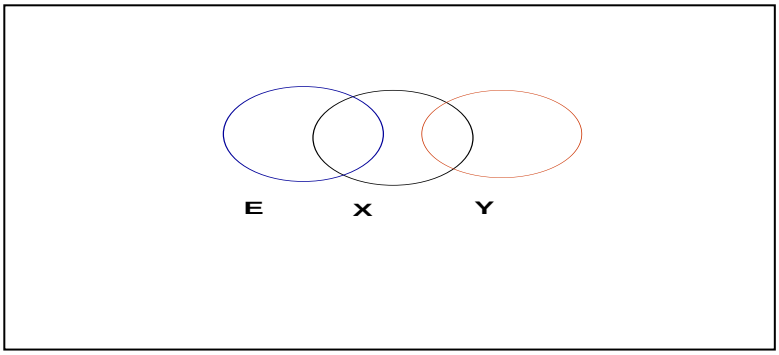


2. Mediator: The third variable (E) may be the cause of any relationship between the original two variables X and Y. Two possibilities of this are diagrammed below:

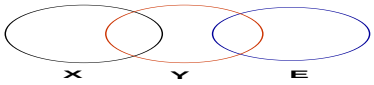
3. Moderator: A moderator is a variable (**E**) whereby **X** and **Y** have a different relationship between each other at the

various levels of **E**. Note that this is essentially what is entailed in an interaction.

4. Suppressor: The correlation between **Y** and **E** equals zero, while the correlation between **X** and **E** is greater than zero. The third variable removes variance from **X**.



5. Covariate: The correlation between **Y** and **E** is greater than zero, while the correlation between **X** and **E** equals zero.



Analysis of Covariance (ANCOVA) use to examine and adjust the effect of covariate. If there are no covariates, ANOVA must be used instead of ANCOVA, and if there are covariates, ANCOVA is used instead of ANOVA.

ANCOVA Model

$$y_{ij} = \mu + \alpha_1 + \beta (x_{ij} - x) + \varepsilon_{ij}$$

Y_{ij} = jth replicate observation of response variable

μ = mean value of response variable

$\alpha_1 = \mu_1 - \mu$

β = combined regression coefficient

X_{ij} = covariate value for the jth replicate observation from the ith level of factor A

x = mean value of covariate

ε_{ij} = unexplained error assoc. with jth replicate observation from the ith level of factor A

Assumptions of ANCOVA:

At least one categorical and at least one interval independent: The independent variable(s) may be categorical, except at least one must be a covariate (interval level). Likewise, at least one independent must be categorical.

Interval dependent: The dependent variable is continuous and interval level.

Linearity: Since ANCOVA is a general linear model procedure with much in common with multiple regression model. It is also assumed that the covariate has a linear relationship with the dependent variable.

Homogeneity of Variance: ANCOVA assumes homogeneity of variance. In other words, the variance of group one is equal to the variance of group 2 and so on.

Homogeneity of Regression: The correlation between Y and E is equal for all levels of X. In other words, for each level of the independent variable, the slope of the prediction of the dependent variable from the covariate must be equal (important assumption).

Good reliability of the covariate. The covariate variables are continuous and interval level, and are assumed to be measured without error. As a rule of thumb, covariates should have a reliability coefficient of .80 or higher.

No high multicollinearity of the covariates: ANCOVA is sensitive to multicollinearity among the covariates and also loses statistical power as redundant covariates are added to the model.

Independence of the error term: The error term is independent of the covariates and the categorical independents. Randomization in experimental designs assures this assumption will be met.

Normal distribution within groups: The dependent variable should be normally distributed within groups formed by the factors.

The ANOVA Table for ANCOVA

The following table contains the notation for a one-way analysis of covariance.

Table1:ANOVA Summary Table for a simple ANCOVA Model

Source	Degrees of Freedom (df)	Sum of Square (SS)	Mean Square (MS)	F Ratio
Group(Between)	**(k – 1)**	**SSG**	**MSG=SSG/(k-1)**	**MSG/MSE**
X (Covariate)	**1**	**SSX**	**MSX=SSX/1**	**MSX/MSE**
Error	**n –(k -1)**	**SSE**	**MSE=SSE/n-(k-1)**	
Total	**n - 1**	**SST**		

One of important assumptions is the slope is the same for all groups. In order to examine this assumption, you run the model included group, covariate, and group*covariate interaction. If the P-value for the interaction term from type III SS is non-significant, then you can conclude that the slope may be equal for all group means. Then, you can run the same model without the interaction term and look for differences among group means. If the P-value for the interaction term from type III SS is significant, then you cannot assume parallel slopes among all groups. These results indicate that the covariate influences the response variable and that it needs to be taken into account. Therefore, a comparison of group means depends on the value of the covariate.

Examples of Analysis of Covariance:

Example1: Assumption of equal slopes are met.

Analysis of covariance is like ANOVA, except in addition to the categorical predictors you have continuous predictors as well. For example, the one way ANOVA example used partner abuse (tbpast) as the dependent variable and race (dq3g) as the independent variable. Let's add depression (tcesd) as a continuous variable to this model. Is there any different in race group with partner abuse after controlling for depression?

SAS Syntax:

```
ods rtf;
ods listing close;
ods graphics on;
proc glm data=two;
     class   dq3g;
     model tbpast = dq3g tcesd dq3g*tcesd  /solution ;
     title ' glm' ; Run;
  proc glm data=two;
  class   dq3g;
  model tbpast = dq3g tcesd   /solution ;
     lsmeans dq3g / tukey line;
     title ' glm' ;  run;
```

```
ods graphics off;
ods rtf close;
ods listing;
quit; run;
```

Table1. Model with Interaction

Class Level Information		
Class	**Levels**	**Values**
dq3g	2	black white

Number of Observations Read	246
Number of Observations Used	218

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8547.59696	2849.19899	15.60	**<.0001**
Error	214	39086.11404	182.64539		
Corrected Total	217	47633.71101			

R-Square	Coeff Var	Root MSE	tbpast Mean
0.179444	164.6837	13.51464	8.206422

Source	DF	Type I SS	Mean Square	F Value	Pr > F
dq3g	1	1830.826085	1830.826085	10.02	0.0018
tcesd	1	6229.416514	6229.416514	34.11	<.0001
tcesd*dq3g	1	487.354365	487.354365	2.67	0.1038

Source	DF	Type III SS	Mean Square	F Value	Pr > F
dq3g	1	30.016282	30.016282	0.16	**0.6856**
tcesd	1	6710.075602	6710.075602	36.74	**<.0001**
tcesd*dq3g	1	487.354365	487.354365	2.67	**0.1038**

Table2. Model without Interaction

Dependent Variable: tbpast

Class Level Information		
Class	**Levels**	**Values**
dq3g	2	black white

Number of Observations Read	246
Number of Observations Used	218

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8060.24260	4030.12130	21.90	**<.0001**
Error	215	39573.46841	184.06264		
Corrected Total	217	47633.71101			

R-Square	Coeff Var	Root MSE	tbpast Mean
0.169213	165.3214	13.56697	8.206422

Source	DF	Type I SS	Mean Square	F Value	Pr > F
dq3g	1	1830.826085	1830.826085	9.95	0.0018
tcesd	1	6229.416514	6229.416514	33.84	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
dq3g	1	1089.687564	1089.687564	5.92	**0.0158**
tcesd	1	6229.416514	6229.416514	33.84	**<.0001**

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-4.363925809	B	2.12222362	-2.06	0.0410
dq3g	**black**	4.660413333	B	1.91538581	2.43	0.0158
dq3g	**white**	0.000000000	B	.	.	.
tcesd		0.470873325		0.08094001	5.82	<.0001

dq3g	tbpast LSMEAN	H0:LSMean1=LSMean2 Pr > |t|
black	**11.1352139**	**0.0158**
white	**6.4748006**	

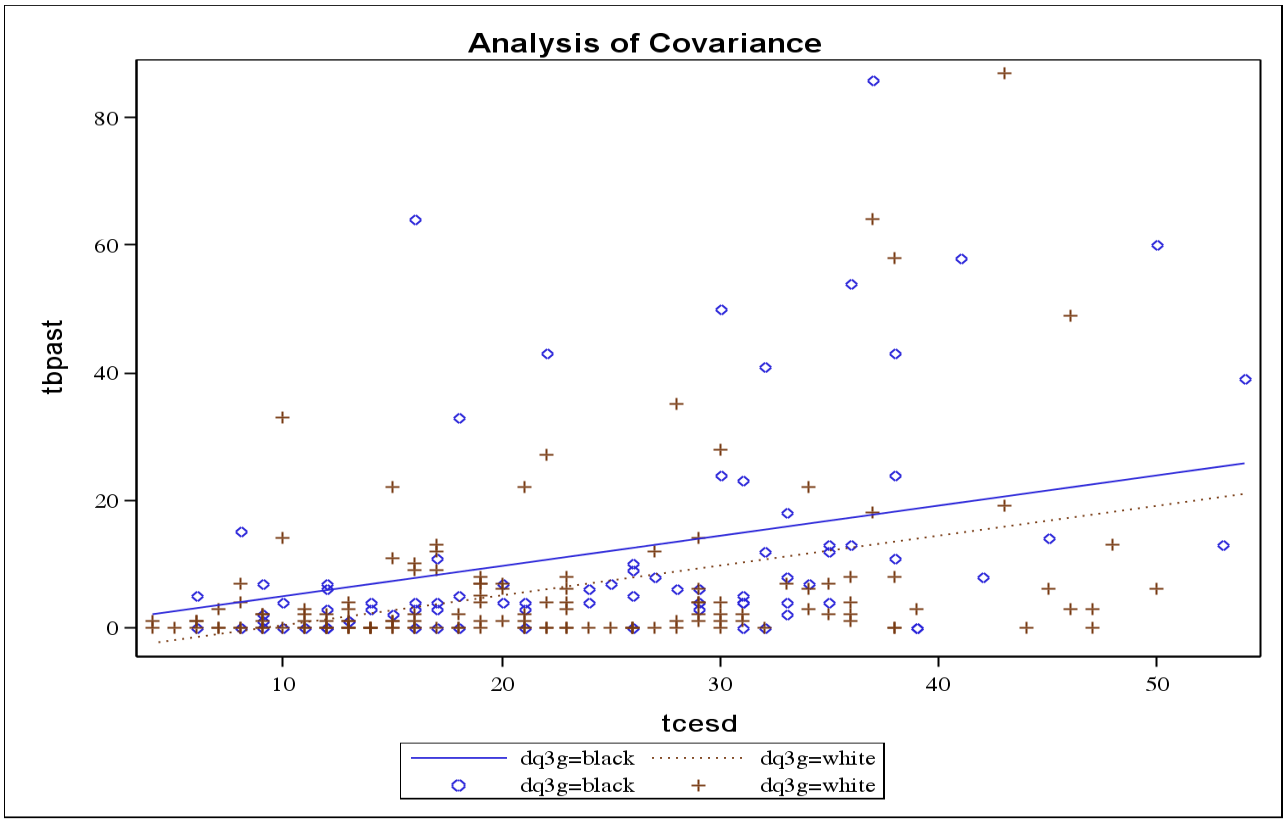


Figure 1. Scatter plot of tbpast by tcesd with separate regression lines for each group (black=blue, white=red).

Example of table to present for paper

Table3. ANCOVA Summary Table examining Age group and Social Support

Source	DF	Sum of Squares	Mean Square	F Value
Model	**2**	**8060.24**	**4030.12**	**21.90**
Error	**215**	**39573.47**	**184.06**	
Corrected Total	**217**	**47633.71**		

Source	DF	Type III SS	Mean Square	F Value
Dq3g[a]	**1**	1089.69	1089.69	5.92
Tcesd[b]	**1**	6229.42	6229.42	33.84

a. Dq3g= race , P-value = 0.0158 b. tcesd=depression, P-value=0.0001

Interpretation:
To check the assumptions of the slope is the same for all groups, we run the model included race (dq3g), tcesd (depression as covariate), and dq3g*tcesd interaction on partner abuse (tbpast). The result did not indicate there is

interaction effect between covariate (depression) and race (P-value=0.103) which means slope are the same for all group (figure 1). Then we run the model without the interaction effect. The results revealed significant effect for both covariate (Tcesd; p-value=0.0158) and group (p-value=0.0001) with partner abuse. The Tukey multiple comparison results indicate that there are significant difference in terms of partner abuse adjusted means for black (11.13) and white (6.47).

Example2: Assumption of equal slopes is not met.

Let us now look at partner abuse (tbpast) as the dependent variable and have you seen parents hit or beat (dq9) as the independent variable. Let's add depression (tcesd) as a continuous variable to this model. Is there any different for have you seen groups with partner abuse, after controlling depression?

SAS Syntax:

```
ods rtf;
ods listing close;
ods graphics on;
proc glm data=two;
     class   dq9;
              model tbpast = dq9 tcesd dq9*tcesd  /solution ;
       title ' glm' ;  run;
ods graphics off;
ods rtf close;
ods listing; quit; run;
```

Table4. Dependent Variable: tbpast

Class Level Information		
Class	**Levels**	**Values**
DQ9	2	no yes

Number of Observations Read	246
Number of Observations Used	225

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	8444.78509	2814.92836	16.67	**<.0001**
Error	221	37321.21491	168.87428		
Corrected Total	224	45766.00000			

R-Square	Coeff Var	Root MSE	tbpast Mean
0.184521	165.1928	12.99516	7.866667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DQ9	1	2791.046276	2791.046276	16.53	<.0001
tcesd	1	4745.247102	4745.247102	28.10	<.0001
tcesd*DQ9	1	908.491715	908.491715	5.38	0.0213

Source	DF	Type III SS	Mean Square	F Value	Pr > F
DQ9	1	98.095365	98.095365	0.58	**0.4468**
tcesd	1	4856.339565	4856.339565	28.76	**<.0001**
tcesd*DQ9	1	908.491715	908.491715	5.38	**0.0213**

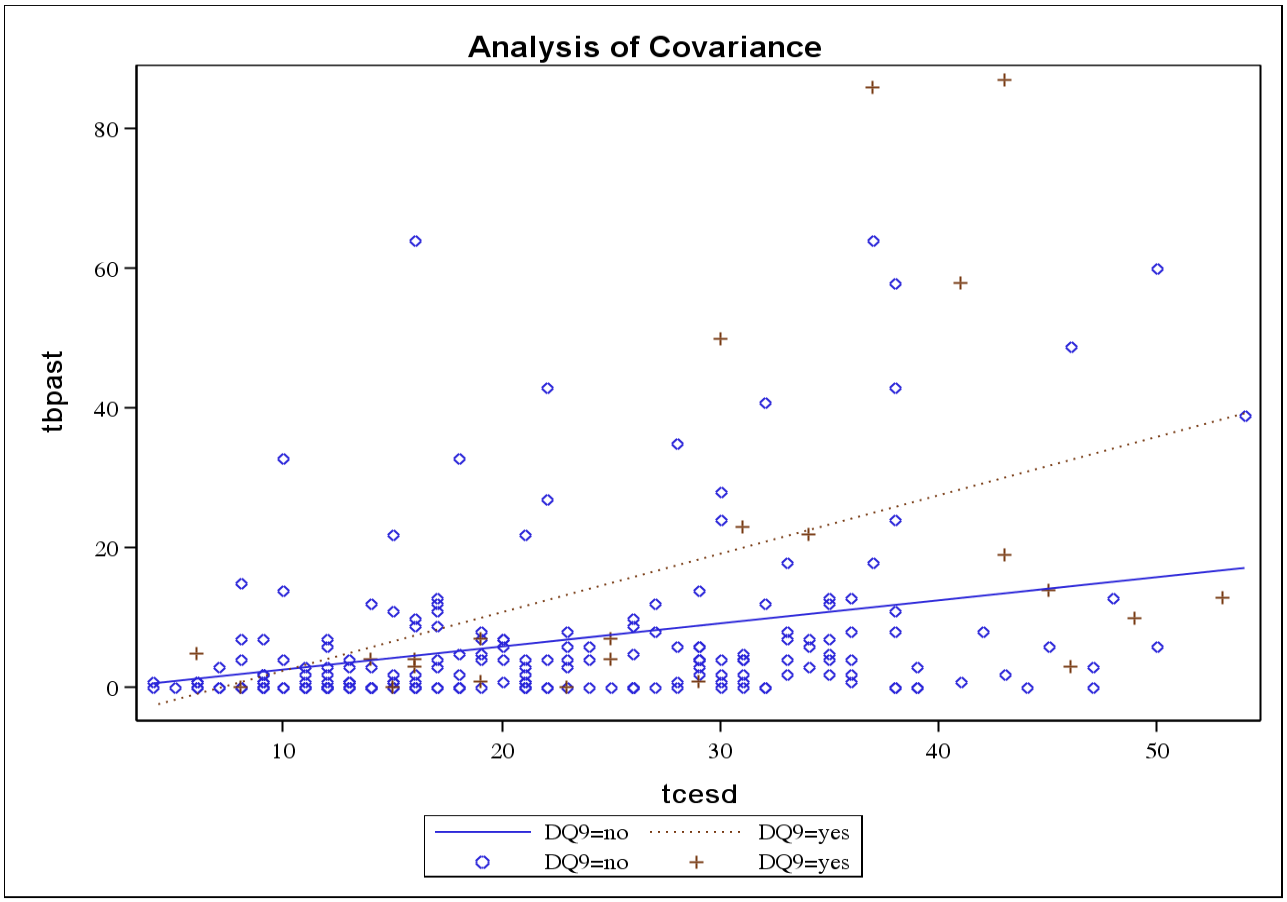


Figure2. Scatter plot of tbpast by tcesd with separate regression lines for each group (no=blue, yes=red).

SAS Syntax:

```
ods rtf;
ods listing close;
ods graphics on;
PROC SORT DATA=two;   BY dq9;
PROC GLM DATA=two;
  BY dq9 ;
  MODEL tbpast = tcesd/ SOLUTION; run;
ods graphics off;
ods rtf close;
ods listing;  quit; run;
```

Table5. Regression model by group

Dependent Variable: tbpast

(Seen partner or beat each other=no)

Number of Observations Read	220
Number of Observations Used	202

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2749.24484	2749.24484	21.93	<.0001
Error	200	25072.83931	125.36420		
Corrected Total	201	27822.08416			

R-Square	Coeff Var	Root MSE	tbpast Mean
0.098815	167.6587	11.19662	6.678218

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tcesd	1	2749.244844	2749.244844	21.93	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tcesd	1	2749.244844	2749.244844	21.93	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-.8015639976	1.78094817	-0.45	0.6531
tcesd	**0.3314866010**	0.07078578	4.68	**<.0001**

Dependent Variable: tbpast

(Seen partner or beat each other=yes)

Number of Observations Read	24
Number of Observations Used	23

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2904.49397	2904.49397	4.98	**0.0367**
Error	21	12248.37559	583.25598		
Corrected Total	22	15152.86957			

R-Square	Coeff Var	Root MSE	tbpast Mean
0.191679	131.9397	24.15069	18.30435

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tcesd	1	2904.493973	2904.493973	4.98	0.0367

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tcesd	1	2904.493973	2904.493973	4.98	0.0367

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-5.962527777	11.98386922	-0.50	0.6240
tcesd	**0.836788814**	0.37498183	2.23	**0.0367**

Interpretation:
To check the assumptions of the slope is the same for all groups, we run the model included seen parents hit or beat (dq9), tcesd (depression as covariate), and dq9*tcesd interaction on partner abuse (tbpast). The result indicated that there is interaction effect between covariate (depression) and parents hit or beat (P-value=0.0213) which means slope are not the same for all group (figure 5). Therefore, we cannot run the model without the interaction effect. These results indicate that the covariate influences the response variable and that it needs to be taken into account. Therefore, a comparison of group means depends on the value of the covariate. Because the slopes for the two groups are not the same, we should not use a traditional ANCOVA model that assumes the slopes for the two groups are the same. Instead, we can use a model that estimates separate slopes for all three group groups. The results of regression indicated that both slopes are positive. However, the slope for yes (0.84) was greater than no (0.33).

Conclusion

This paper showed overview of Analysis of Covariance (ANCOVA) using two examples in SAS. The paper indicated how to test the equal slopes assumption. Also, the paper showed how to present and interpret the result for publication. Proc GLM in SAS used to examine the data.

References

Freund, R. L., Ramonl. (1991.). SAS System for Linear Model, SAS Inc.
Muller, K. F., Bethel . (2002.). Regression and ANOVA, SAS Inc.
Agresti, A. F., Barbara. (1997). Statistical methods for the Social Science., Prentice Hall.
Khattree, R. n., Dayanad (2006). Applied multivariate Statistics with SAS, SAS Inc.

Contact Information

Abbas S. Tavakoli, DrPH, MPH, ME
College of Nursing
University of South Carolina
1601 Greene Street
Columbia, SC 29208-4001
Fax: (803) 777-5561
E-mail: abbas.tavakoli@sc.edu