Paper AS15 - PHUSE US 2020

Importance of General Linear Models (GLM) in Clinical Research

Venkata Ikkurthy, Bayer, Whippany, USA
Ritesh Dhimmar, Bayer, Whippany, USA

ABSTRACT

A fully randomized clinical trial is necessary to conduct unbiased and objective research. However, randomization can also hide some relations between study subjects that could be interesting to the investigator. Classification and similarity techniques are a powerful tool to discover these lost relations. Among these, general linear models help to find the effects between the treatment groups and within treatment groups. This paper provides an example of General linear model, ANOVA (Analysis of Variance) for different treatment groups. Analyzing the data between the treatment groups and within the groups using Proc GLM. Keywords: ANOVA, Least squares, SAS, PROC GLM, Regression, Interaction effect, linear. Mathematical Approach

INTRODUCTION

This paper explains the importance and application of linear models in clinical trials. Most of the statistical analysis are based on linear models, and most of analysis of linear models can be performed by three SAS® procedures REG, ANOVA, and GLM. These Statistical procedures provide the power and flexibility for almost all linear model analysis. This paper explains to the basic statistical concepts to understand and using linear models that can be analyzed with these SAS (Statistical Procedures)

STATISTICAL BACKGROUND

Variation is inevitable, we need to identify the causes of variation and reason for variation, generally variation between the groups called it as assignable variation and within the groups as chance variation. Mathematically we need to identify the assignable variation and eliminate that variation, but we are not able to eliminate the chance variation, but we want to minimize this chance variation by using minima and maxima of differential calculus and principle of least squares.

Simple linear regression analysis is used to analyze the relationship between a response, Y, and a predictor, X. To assess the significance of the predictor in explaining the variability or behavior of the response. To predict the values of the response given the values of the predictor.

GLM PROCEDURE (PROC GLM)

LINEAR REGRESSION ANALYSIS

ANALYSIS OF VARIANCE (ANOVA)

ANALYSIS OF COVARIANCE (ANCOVA)

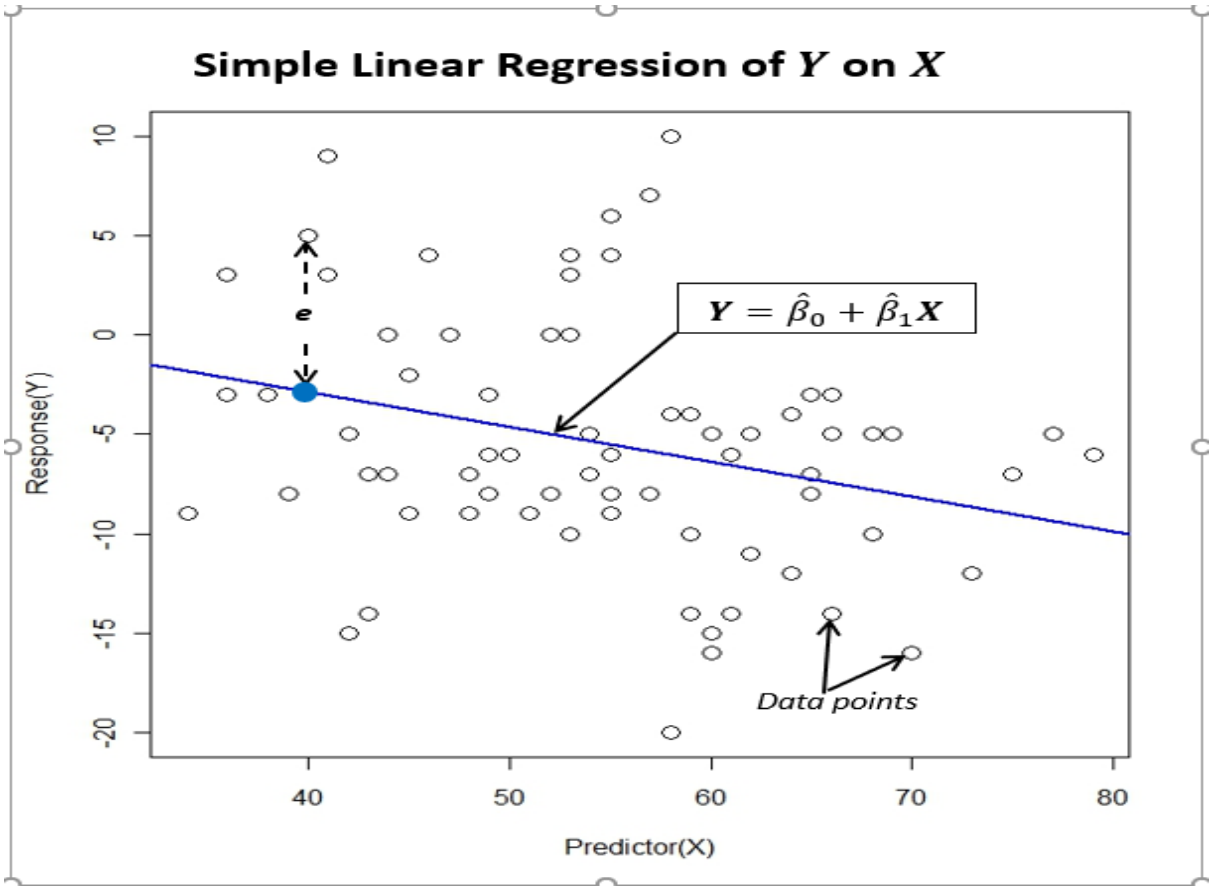
SIMPLE LINEAR REGRESSION ANALYSIS

Simple linear regression analysis is used to analyze the relationship between a response, $\boldsymbol{Y}$, and a predictor, $\boldsymbol{X}$.

Objectives

To assess the significance of the predictor in explaining the variability or behavior of the response. To predict the values of the response given the values of the predictor.

Ω



- Hypothesis Test: $Y = \beta_0 + \beta_1 X + \varepsilon$

null hypothesis: $H_0: \beta_1 = 0$
alt. hypothesis: $H_a: \beta_1 \neq 0$
test statistic: $t = \frac{\hat{\beta}_1}{s/\sqrt{S_{xx}}}$
decision rule: reject H_0 if $|t| > t_{n-2,1-\alpha/2}$

Study design:

A multi-center, randomized, double-blind, parallel study conducted at four study centers. 75 patients in 4 centers are randomized in equal numbers to one of three treatment groups (i.e., High Dose of Anti-hypertensive medication (HI), Low Dose of Anti-hypertensive medication (LO) and Placebo (PB)).

Objectives: To compare the efficacy of three treatment groups in terms of the mean decreases in diastolic blood pressure (DBP).

To compare the safety of three treatment groups by evaluating their Adverse Drug Reaction (ADR,) rates (i.e., diarrhea) rates.

Obs	pat	trt	age	bpdia0	bpdiach	gender	diarrhea	center
1	101	HI	55	102	-6	F	NO	1
2	103	HI	68	115	-10	F	YES	1
3	109	HI	54	115	-7	F	NO	1
4	202	HI	58	99	-4	F	NO	2
5	203	HI	62	119	-11	M	NO	2
6	204	HI	51	107	-9	F	YES	2
7	208	HI	61	98	-6	F	NO	2
8	216	HI	64	119	-12	M	NO	2
9	219	HI	54	97	-5	F	NO	2
10	223	HI	39	95	-8	F	YES	2
11	304	HI	59	116	-10	M	YES	3
12	307	HI	42	108	-5	F	YES	3
...	...	...	...	...	...	...	...	...
66	310	PB	66	101	-5	M	YES	3
67	311	PB	68	102	-5	F	NO	3
68	405	PB	53	117	4	M	NO	4
69	406	PB	53	117	0	F	YES	4
70	407	PB	46	96	4	F	NO	4
71	408	PB	55	96	6	F	NO	4
72	415	PB	41	109	3	M	NO	4
73	417	PB	53	109	3	M	NO	4
74	424	PB	52	103	0	F	NO	4
75	425	PB	41	95	9	M	NO	4

Variable	Descriptions
pat	Patient's ID
trt	Treatment (HI/LO/PB)
age	Patient's Age
bpdia0	Baseline of diastolic blood pressure
bpdiach	Diastolic blood pressure change from baseline to endpoint
gender	Patient's gender (M/F)
diarrhea	Whether patients got diarrhea during the clinical trial (YES/NO)
center	Center ID (1/2/3/4)

QUESTION:

DO THE DATA SHOW ANY LINEAR RELATIONSHIP BETWEEN MEAN CHANGE OF DIASTOLIC BLOOD PRESSURES (BPDIACH) AND PATIENTS' AGES (AGE)?

```
ODS GRAPHICS ON;
PROC GLM DATA = BP PLOTS = ALL;
  MODEL BPDIACH = AGE;
  TITLE1 'SIMPLE LINEAR REGRESSION';
  TITLE2 'DBP CHANGES FROM BASELINE TO ENDPOINT VS. AGE';
 QUIT;
ODS GRAPHICS OFF;
```

The GLM Procedure

Dependent Variable: bpdiach

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	241.198811	241.198811	6.68	0.0118
Error	73	2637.547856	36.130793		
Corrected Total	74	2878.746667			

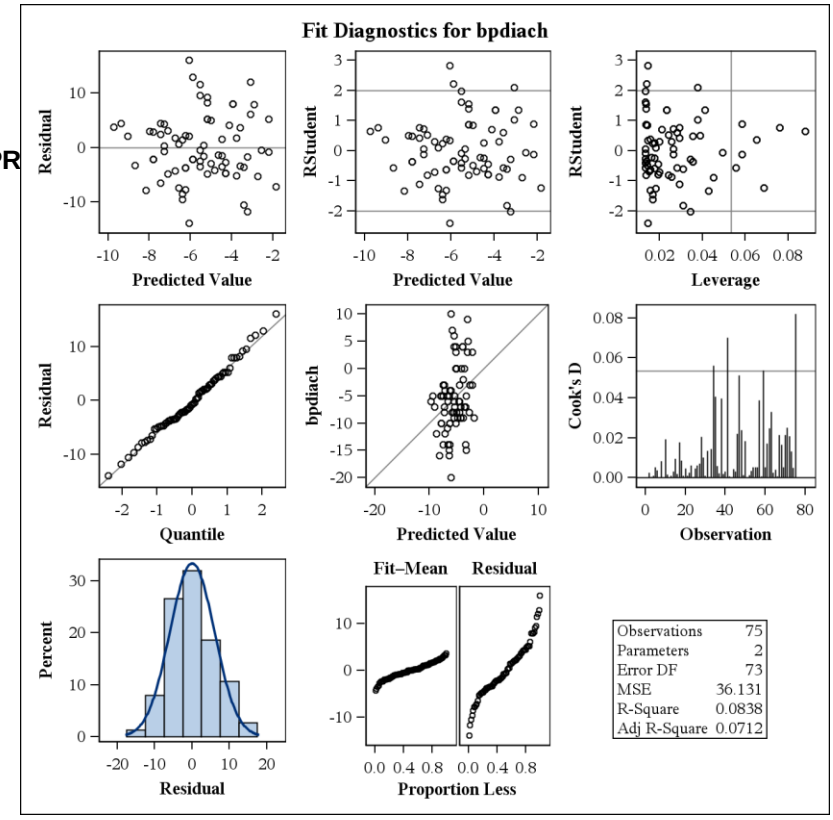
R-Square	Coeff Var	Root MSE	bpdiach Mean
0.083786	-109.4215	6.010889	-5.493333

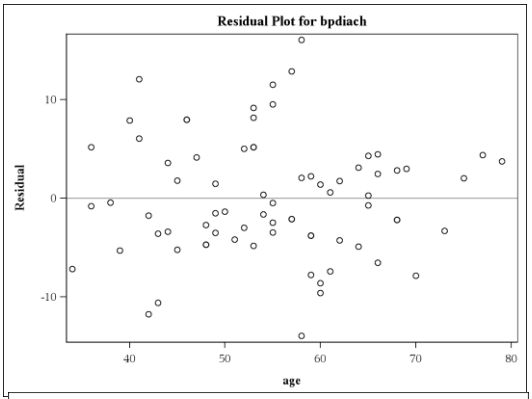
Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	241.1988106	241.1988106	6.68	0.0118

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	241.1988106	241.1988106	6.68	0.0118

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	4.122527237	3.78585057	1.09	0.2798
age	-0.175429225	0.06789739	-2.58	0.0118

PR





Simple Linear Regression syntax :

proc glm data = SAS-data-set;

model Y = X;

quit;

Multiple Linear Regression syntax:

proc glm data = SAS-data-set;

model Y = X Z;

quit;

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$

Obs	pat	trt	age	bpdia0	bpdiach	gender	diarrhea	center
1	101	HI	55	102	-6	F	NO	1
2	103	HI	68	115	-10	F	YES	1
3	109	HI	54	115	-7	F	NO	1
4	202	HI	58	99	-4	F	NO	2
5	203	HI	62	119	-11	M	NO	2
6	204	HI	51	107	-9	F	YES	2
7	208	HI	61	98	-6	F	NO	2
8	216	HI	64	119	-12	M	NO	2
9	219	HI	54	97	-5	F	NO	2
10	223	HI	39	95	-8	F	YES	2
11	304	HI	59	116	-10	M	YES	3
12	307	HI	42	108	-5	F	YES	3
...	...	...	...	...	...	...	...	...
66	310	PB	66	101	-5	M	YES	3
67	311	PB	68	102	-5	F	NO	3
68	405	PB	53	117	4	M	NO	4
69	406	PB	53	117	0	F	YES	4
70	407	PB	46	96	4	F	NO	4
71	408	PB	55	96	6	F	NO	4
72	415	PB	41	109	3	M	NO	4
73	417	PB	53	109	3	M	NO	4
74	424	PB	52	103	0	F	NO	4
75	425	PB	41	95	9	M	NO	4

Variable	Descriptions
pat	Patient's ID
trt	Treatment (HI/LO/PB)
age	Patient's Age
bpdia0	Baseline of diastolic blood pressure
bpdiach	Diastolic blood pressure change from baseline to endpoint
gender	Patient's gender (M/F)
diarrhea	Whether patients got diarrhea during the clinical trial (YES/NO)
center	Center ID (1/2/3/4)

Question:
Do the data show any linear relationship of mean change of diastolic blood pressure (bpdiach) versus patients' ages (age) and baseline of diastolic blood pressures (bpdia0)?

```
ods graphics on;
proc glm data  = bp plots = all;
   model bpdiach = age bpdia0;
   title1 'Multiple Linear Regression';
   title2 'DBP changes from baseline to endpoint vs. Age & DBP baseline';
quit;
ods graphics off;
```

$$\mathbf{bpdiach} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} \cdot \mathbf{age} + \boldsymbol{\beta_2} \cdot \mathbf{bpdia0} + \boldsymbol{\varepsilon}$$

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	28.72936509	9.66834425	2.97	0.0040
age	-0.12521294	0.06756928	-1.85	0.0680
bpdia0	-0.25691148	0.09357203	-2.75	0.0076

- **Fail to reject $\beta_1 = 0$ at .05 level, i.e., we cannot conclude that there is a linear relationship between DBP change from baseline to endpoint and age.**
- **Succeed to reject $\beta_2 = 0$ at .05 level, i.e., we can conclude that there is a linear relationship between DBP change from baseline to endpoint and baseline DBP.**

Analysis of Variance (Oneway classification)

Goal: To simultaneously compare two or more group means based on independent samples from each group.
Assumptions

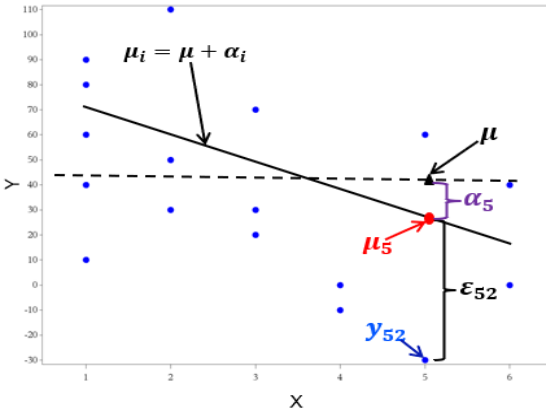
- Normality: the samples are from normally distributed populations.
- Independency: samples are independent within group.

- Variance homogeneity the within group variance is constant across groups

In clinical trials, ANOVA method might be appropriate for comparing mean responses among a number of parallel-dose groups or among various strata based on patients' background information, such as race, age group, or disease severity.

- Typical layout for the one-way ANOVA

Group 1	Group 2	⋯	Group g
y_{11}	y_{21}	⋯	y_{g1}
y_{12}	y_{22}	⋯	y_{g2}
⋯	⋯	⋯	⋯
y_{1n_1}	y_{2n_2}	⋯	y_{gn_g}



- Statistical model:

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

μ_i is the mean response of group i
ε_{ij} is a random error of the j-th data in group i.

μ is the overall mean response
α_i is the average additive effect of group i
ε_{ij} is a random error of the j-th data in group i

- ANOVA summary table for the one-way ANOVA

Source	df	SS	MS	F
Group	$g-1$	SSG	$MSG = SSG/(g-1)$	$F = MSG/MSE$
Error	$N-g$	SSE	$MSE = SSE/(N-g)$	
Total	$N-1$	SS_{TOT}		

Note: **MSG** is an estimate of the variability among groups, and **MSE** is an estimate of the variability within groups.

- Hypothesis Test:

null hypothesis: $H_0: \mu_1 = \mu_2 = \cdots = \mu_g$
alt. hypothesis: H_a: NOT H_0
test statistic: $F = MSG/MSE$
decision rule: reject H_0 if $F > F_{g-1,N-g,1-\alpha}$

Example

Data set (bp):

Obs	pat	trt	age	bpdia0	bpdiach	gender	diarrhea	center
1	101	HI	55	102	-6	F	NO	1
2	103	HI	68	115	-10	F	YES	1
3	109	HI	54	115	-7	F	NO	1
4	202	HI	58	99	-4	F	NO	2
5	203	HI	62	119	-11	M	NO	2
6	204	HI	51	107	-9	F	YES	2
7	208	HI	61	98	-6	F	NO	2
8	216	HI	64	119	-12	M	NO	2
9	219	HI	54	97	-5	F	NO	2
10	223	HI	39	95	-8	F	YES	2
11	304	HI	59	116	-10	M	YES	3
12	307	HI	42	108	-5	F	YES	3
...	...	...	...	...	...	...	...	...
66	310	PB	66	101	-5	M	YES	3
67	311	PB	68	102	-5	F	NO	3
68	405	PB	53	117	4	M	NO	4
69	406	PB	53	117	0	F	YES	4
70	407	PB	46	96	4	F	NO	4
71	408	PB	55	96	6	F	NO	4
72	415	PB	41	109	3	M	NO	4
73	417	PB	53	109	3	M	NO	4
74	424	PB	52	103	0	F	NO	4
75	425	PB	41	95	9	M	NO	4

Variable	Descriptions
pat	Patient's ID
trt	Treatment (HI/LO/PB)
age	Patient's Age
bpdia0	Baseline of diastolic blood pressure
bpdiach	Diastolic blood pressure change from baseline to endpoint
gender	Patient's gender (M/F)
diarrhea	Whether patients got diarrhea during the clinical trial (YES/NO)
center	Center ID (1/2/3/4)

Question:
Whether the means of changes of diastolic blood pressure(bpdiach) are significantly different among treatment groups (trt)?

```
ods graphics on;
proc glm data  = bp plots = all;
   class trt;
   model bpdiach = trt;
       lsmeans trt / stderr pdiff cl;
       estimate 'High Dose - Placebo '    trt 1     0    -1;
       estimate 'Active       - Placebo '    trt  0.5  0.5 -1;
quit;
ods graphics off;
```

The GLM Procedure

Class Level Information

Class	Levels	Values
trt	3	HI LO PB

Number of Observations Read 75
Number of Observations Used 75

Dependent Variable: bpdiach

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1729.169286	864.584643	54.15	<.0001
Error	72	1149.577381	15.966353		
Corrected Total	74	2878.746667			

R-Square	Coeff Var	Root MSE	bpdiach Mean
0.600667	-72.73893	3.995792	-5.493333

Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	2	1729.169286	864.584643	54.15	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	1729.169286	864.584643	54.15	<.0001

Analysis of variance (ANOVA) Two way classification

- Typical layout for the two-way ANOVA

	Factor i = 1	Factor i = 2	⋯	Factor i = g
Factor j = 1	$y_{111}, y_{112}, \cdots, y_{11n_{11}}$	$y_{211}, y_{212}, \cdots, y_{21n_{21}}$	⋯	$y_{g11}, y_{g12}, \cdots, y_{g1n_{g1}}$
Factor j = 2	$y_{121}, y_{122}, \cdots, y_{12n_{12}}$	$y_{221}, y_{222}, \cdots, y_{22n_{22}}$	⋯	$y_{g21}, y_{g22}, \cdots, y_{g2n_{g2}}$
⋯	⋯	⋯	⋯	⋯
Factor j = b	$y_{1b1}, y_{1b2}, \cdots, y_{1bn_{1b}}$	$y_{2b1}, y_{2b2}, \cdots, y_{2bn_{2b}}$	⋯	$y_{gb1}, y_{gb2}, \cdots, y_{gbn_{gb}}$

- Model: $y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$

μ_{ij} is the mean response of factor i and factor j
ε_{ijk} is a random error.

μ is the overall mean response
α_i is the average additive effect of factor i
β_j is the average additive effect of factor j
γ_{ij} is the average additive effect of $i-j$ interaction
ε_{ijk} is a random error.

- ANOVA summary table for the two-way ANOVA

Source	*df*	*SS*	*MS*	*F*
Factor i (G)	$g-1$	SSG	MSG	$F_G = \text{MSG}/\text{MSE}$
Factor j (B)	$b-1$	SSB	MSB	$F_B = \text{MSB}/\text{MSE}$
Factor $i \times j$	$(g-1)(b-1)$	SSG	MSGB	$F_{GB} = \text{MSGB}/\text{MSE}$
Error	$N-gb$	SSE	MSE	
Total	$N-1$	SS_T		

- Hypothesis Test:

null hypothesis: $H_0: \mu_1 = \mu_2 = \cdots = \mu_g$
alt. hypothesis: H_a: NOT H_0
test statistic: $F_G = \frac{\text{MSG}}{\text{MSE}}$
decision rule: reject H_0 if $F > F_{g-1,N-gb,1-\alpha}$

One-way ANOVA syntax:

```
proc glm data  = SAS-data-set;
   class   var1;
   model y = var1;
quit;
```

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Two-way ANOVA syntax:

```
proc glm data  = SAS-data-set;
   class   var1 var2;
   model y = var1 var2;
quit;
```

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

- μ is the overall mean response
- α_i is the average additive effect of factor i
- β_j is the average additive effect of factor j
- γ_{ij} is the average additive effect of interaction.

One-way ANOVA syntax:

```
proc glm data  = SAS-data-set;
   class   var1;
   model y = var1;
quit;
```

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

Two-way ANOVA syntax with interaction:

```
proc glm data  = SAS-data-set;
   class   var1 var2;
   model y = var1 var2 var1*var2;
quit;
```

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

- μ is the overall mean response
- α_i is the average additive effect of factor i
- β_j is the average additive effect of factor j
- γ_{ij} is the average additive effect of interaction.

Question:
Whether the means of changes of diastolic blood pressure (bpdiach) are significantly different among treatment groups (trt) & centers (center)?

```
ods graphics on;
proc glm data  = bp plots = all;
   class trt center;
   model bpdiach = trt center trt*center;
      lsmeans trt / stderr pdiff cl;
```

```
      estimate 'High Dose - Placebo '    trt 1    0   -1;
      estimate 'Active      - Placebo '    trt  0.5  0.5 -1;
 quit;
ods graphics off;
```

The GLM Procedure

Class Level Information

Class	Levels	Values
trt	3	HI LO PB
center	4	1 2 3 4

Number of Observations Read 75
Number of Observations Used 75

Dependent Variable: bpdiach

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1915.413117	174.128465	11.39	<.0001
Error	63	963.333550	15.291009		
Corrected Total	74	2878.746667			

R-Square	Coeff Var	Root MSE	bpdiach Mean
0.665364	-71.18396	3.910372	-5.493333

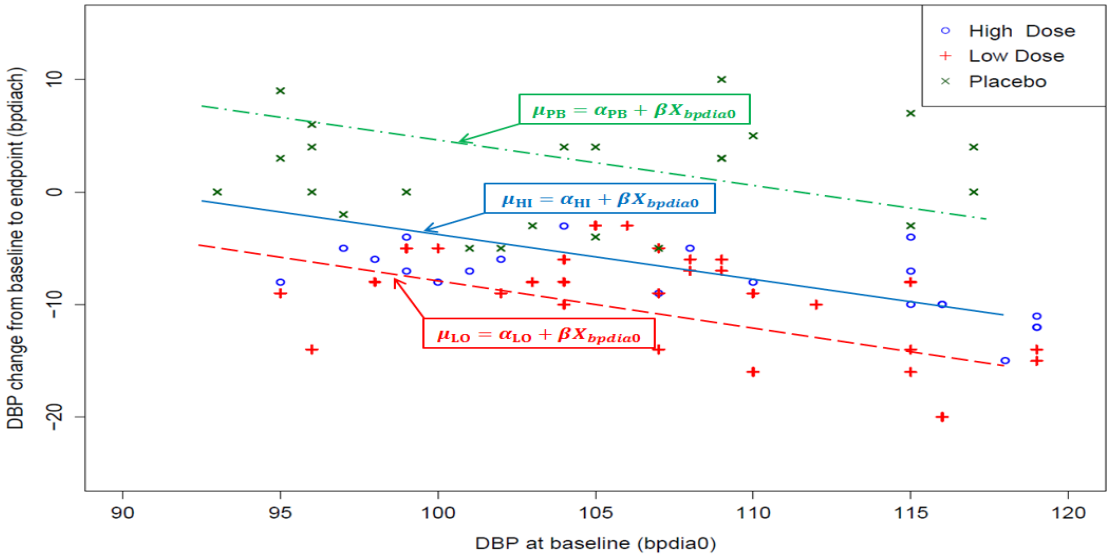
Source	DF	Type I SS	Mean Square	F Value	Pr > F
trt	2	1729.169286	864.584643	56.54	<.0001
center	3	56.972540	18.990847	1.24	0.3026
trt*center	6	129.271291	21.545215	1.41	0.2253

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	1165.777295	582.888648	38.12	<.0001
center	3	48.924177	16.308059	1.07	0.3697
trt*center	6	129.271291	21.545215	1.41	0.2253

Analysis of Covariance (ANCOVA)

Goal: To compare response means among two or more groups adjusted for a quantitative concomitant variable or 'covariate', thought to influence the response. i-th group mean as $\mu_i = \alpha_i + \beta X$

ANCOVA combines regression and ANOVA methods by fitting simple linear regression models within each group and comparing regressions among groups.



- Hypothesis Test:

null hypothesis: $H_0: \mu_1 = \mu_2 = \cdots = \mu_g$
alt. hypothesis: H_a: NOT H_0
test statistic: $F = \frac{MSG}{MSE}$
decision rule: reject H_0 if $F > F_{g-1,N-g,1-\alpha}$

Note: **MSG** is an estimate of the variability among groups, and **MSE** is an estimate of the variability within groups.

One-way ANOVA Syntax

```
proc glm data  = SAS-data-set;
   class   var1;
   model y = var1;
quit;
```

ANCOVA Syntax

```
proc glm data  = SAS-data-set;
   class   var1;
   model y = var1 var2;
quit;
```

covariate (continuous /categorical)

Analysis of Covariance (ANCOVA)

Example Syntax for ANCOVA

```
ods graphics on;
proc glm data  = bp plots = all;
   class trt;
   model bpdiach = trt bpdia0;
       lsmeans trt / stderr pdiff cl;
       estimate 'High Dose - Placebo '    trt     1      0    -1;
       estimate 'Active    - Placebo '    trt   0.5     0.5 -1;
quit;
ods graphics off;
```

REFERENCES

1) **FUNDAMENTALS OF APPLIED STATISTICS BY S.C GUPTA AND V.K.KAPOOR.**
2) **SAS SYSTEM FOR LINEAR MODELS BY SAS INSTITUTE.**
3) **GENERAL LINEAR MODELS BY SAS INSTITUTE.**

ACKNOWLEDGMENTS

The authors wish to thank to colleague, Iraj Mohebalian, for his valuable comments to this paper.
Management at Bayer Joerg Guettner and Michael Badlani, who have long been encourage and supporting me to present this paper.

CONTACT INFORMATION

Name : Venkata Ikkurthy
Enterprise: Bayer U.S. LLC
Address: 100 Bayer Boulevard, P.O. Box 915
City, State ZIP: Whippany, NJ 07981-0915
Phone:+1 732 662 0832
E-mail: venkata.ikkurthy@bayer.com

Name : Ritesh Dhimmar
Enterprise: Bayer U.S. LLC
Address: 100 Bayer Boulevard, P.O. Box 915
City, State ZIP: Whippany, NJ 07981-0915
Phone: +1 551 998 9086
E-mail: ritesh.dhimmar@bayer.com