

STAT 430, Fall 2010

ssh UserID@glue.umd.edu, tap sas913, sas
https://www.statlab.umd.edu/sasdoc/sashtml/onldoc.htm

Analysis of Covariance
=====

$$y_{ij} = \mu + a_i + b \cdot x_{ij} + e_{ij}, \quad i=1, \dots, k, \quad j=1, \dots, n$$

Ex: k=2, n=2

$$\begin{array}{rcllcl} y_{11} & & 110 & \mu & x_{11} & e_{11} \\ y_{12} & = & 110 & a_1 & + & x_{12} & b & + & e_{12} \\ y_{21} & & 101 & a_2 & x_{21} & e_{21} \\ y_{22} & & 101 & & x_{22} & e_{22} \end{array}$$

Ex 1. Math-IQ Relationship

Group A		Group B	
Math Score	IQ	Math Score	IQ
260	105	325	126
325	115	440	135
300	122	425	142
400	125	500	140
390	138	600	160

Two groups A,B
Math Score = Dependent variable
IQ = Covariate
i=A,B; j=1,2,3,4,5

We want to run Analysis of Covariance Model:

$$y_{ij} = \mu + \alpha_i + \beta \cdot IQ_{ij} + \epsilon_{ij}$$

But first we need to see if indeed the slope beta is the same for A,B.

```
OPTIONS PS=35 LS=70;
```

```
DATA COVAR;  
LENGTH GROUP $ 1;  
INPUT GROUP $ MATH IQ @@;  
DATALINES;  
A 260 105 A 325 115 A 300 122 A 400 125 A 390 138  
B 325 126 B 440 135 B 425 142 B 500 140 B 600 160  
;
```

First run simple regression for each group separately.

```
PROC REG DATA=COVAR;  
by group;  
MODEL MATH = IQ/;  
RUN;
```

```
----- GROUP=A -----  
The REG Procedure  
Model: MODEL1  
Dependent Variable: MATH  
  
Number of Observations Read      5  
Number of Observations Used      5  
  
Analysis of Variance  
  
Source                DF      Sum of Squares      Mean Square      F Value      Pr > F  
Model                  1      9793.31104      9793.31104      6.67      0.0816  
Error                  3      4406.68896      1468.89632  
Corrected Total        4          14200  
  
Root MSE              38.32618      R-Square          0.6897  
Dependent Mean        335.00000      Adj R-Sq          0.5862  
Coeff Var              11.44065
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-154.66555	190.41317	-0.81	0.4761
IQ	1	4.04682	1.56727	2.58	0.0816

----- GROUP=B -----

The REG Procedure
 Model: MODEL1
 Dependent Variable: MATH

Number of Observations Read 5
 Number of Observations Used 5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	35839	35839	20.71	0.0199
Error	3	5190.66110	1730.22037		
Corrected Total	4	41030			

Root MSE 41.59592 R-Square 0.8735
 Dependent Mean 458.00000 Adj R-Sq 0.8313
 Coeff Var 9.08208

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-608.23171	235.01043	-2.59	0.0812
IQ	1	7.58344	1.66624	4.55	0.0199

So, one slope is 4.04682 the other 7.58344. Since we have only 5 observations in each group, these could be "equal". To test equality of the two slopes we use PROC GLM.

```
PROC GLM DATA=COVAR;
CLASS GROUP;
MODEL MATH = IQ GROUP IQ*GROUP/SS3;
RUN;
```

Note: IQ*GROUP will test $H_0: \beta_1 = \beta_2$

The GLM Procedure

Class Level Information

Class	Levels	Values
GROUP	2	A B
Number of Observations Read		10
Number of Observations Used		10

The SAS System

The GLM Procedure

Class Level Information

Class	Levels	Values
GROUP	2	A B
Number of Observations Read		10
Number of Observations Used		10

The GLM Procedure

Dependent Variable: MATH

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	83455.14993	27818.38331	17.39
Error	6	9597.35007	1599.55834	
Corrected Total	9	93052.50000		

Source	Pr > F
Model	0.0023
Error	

Corrected Total

R-Square	Coeff Var	Root MSE	MATH Mean
0.896861	10.08688	39.99448	396.5000

Source	DF	Type III SS	Mean Square	F Value
IQ	1	41278.21495	41278.21495	25.81
GROUP	1	3634.41141	3634.41141	2.27
IQ*GROUP	1	3816.96372	3816.96372	2.39

Source	Pr > F
IQ	0.0023
GROUP	0.1824
IQ*GROUP	0.1734<--- Not Significant. Slopes are same.

Do not reject H₀: Slopes are equal.

So, now we can run
 $y_{ij} = \mu + \alpha_i + \beta \cdot IQ_{ij} + \epsilon_{ij}$, $i=1,2$, $j=1,\dots,5$.

We wish to test $H_0: \alpha_1 = \alpha_2$.

```
PROC GLM DATA=COVAR;
CLASS GROUP;
MODEL MATH = IQ GROUP/ SS3;
LSMEANS GROUP;
RUN;
```

The SAS System

The GLM Procedure

Class Level Information

Class	Levels	Values
GROUP	2	A B
Number of Observations Read		10
Number of Observations Used		10

The GLM Procedure

Dependent Variable: MATH

Source	DF	Sum of Squares	Mean Square	F Value
Model	2	79638.18621	39819.09311	20.78
Error	7	13414.31379	1916.33054	
Corrected Total	9	93052.50000		

Source	Pr > F
Model	0.0011

Error

Corrected Total

R-Square	Coeff Var	Root MSE	MATH Mean
0.855841	11.04058	43.77591	396.5000

Source	DF	Type III SS	Mean Square	F Value
IQ	1	41815.68621	41815.68621	21.82
GROUP	1	96.59793	96.59793	0.05

Source	Pr > F
IQ	0.0023 <-- Significant slope.
GROUP	0.8288 <-- Not significant

Do not reject H_0 : $\alpha_1 = \alpha_2$

So, when adjusted for IQ there are no longer differences between the groups on math scores.

The GLM Procedure
Least Squares Means

GROUP	MATH LSMEAN
A	392.345889
B	400.654111

Note: LSMEANS gives the means of the math scores adjusted for IQ. As is, the math scores means are A: 335, B: 458. To get LSMEANS we need the regression estimates.

If we also want the estimates themselves: use SOLUTION:

```

PROC GLM DATA=COVAR;
CLASS GROUP;
MODEL MATH = IQ GROUP/ SOLUTION SS3;
LSMEANS GROUP;
RUN;

```

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-364.7379627 B	177.2121809	-2.06	0.0786
IQ	5.8516214	1.2526848	4.67	0.0023
GROUP A	-8.3082214 B	37.0049116	-0.22	0.8288
GROUP B	0.0000000 B	.	.	.

NOTE: SAS replaces one effect by 0.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Now we can compute the LSMEANS as follows:

model applied to Average(All IQ_{ij} from both groups)

Use $\mu + \alpha_i + \beta \cdot \text{Average}(\text{of all IQ}_{ij} \text{ from both groups})$

Let IQ be the vector of all IQ's from both A and B.

IQ = (105, 115, 122, 125, 138, 126, 135, 142, 140, 160)

MATH LSMEAN For GROUP A:

> -364.7379627 - 8.3082214 + 5.8516214*mean(IQ)

[1] 392.3459 <-----OK. NOTE: Mean(Math Scores for A)=335

MATH LSMEAN For GROUP B:


```
> -364.7379627 + 0.0000000 + 5.8516214*mean(IQ)
[1] 400.6541 <-----OK.    NOTE: Mean(Math Scores for B)=458
```

```
=====
Ex 2. Fish Diet
-----
```

Three groups of guppy fish (*poecilia reticulatia*) were fed different diets. the resulting WEIGHT y and Initial weight IW are given below.

```
OPTIONS PS=35 LS=70;
```

```
DATA FISH;
```

```
LENGTH GROUP $ 1;
```

```
INPUT GROUP $ WEIGHT IW @@;
```

```
DATALINES;
```

```
A 49 35 A 61 26 A 55 29 A 69 32 A 51 23 A 38 26 A 64 31
```

```
B 68 33 B 70 35 B 60 28 B 53 29 B 59 32 B 48 23 B 46 26
```

```
C 59 33 C 53 36 C 54 26 C 48 30 C 54 33 C 53 25 C 37 23
```

```
;
```

We want to run Analysis of Covariance Model:

$$\text{WEIGHT}_{ij} = y_{ij} = \mu + \alpha_i + \beta \cdot \text{IW}_{ij} + \epsilon_{ij},$$
$$i=1,2,3, \quad j=1,\dots,7$$

But first we need to see if indeed the slope beta is the same for A,B,C. To actually get estimates of the slopes we first run simple regression for each group separately.

```
PROC REG DATA=FISH;
```

```
by group;
```

```
MODEL WEIGHT = IW/;
```

```
RUN;
```

The SAS System

----- GROUP=A -----

The REG Procedure
Model: MODEL1
Dependent Variable: WEIGHT

Number of Observations Read 7
Number of Observations Used 7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	64.23829	64.23829	0.55	0.4935
Error	5	589.19028	117.83806		
Corrected Total	6	653.42857			

Root MSE 10.85532 R-Square 0.0983
Dependent Mean 55.28571 Adj R-Sq -0.0820
Coeff Var 19.63495

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	32.48056	31.15854	1.04	0.3450
IW	1	0.79028	1.07035	0.74	0.4935

----- GROUP=B -----

The REG Procedure
Model: MODEL1
Dependent Variable: WEIGHT

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	416.59479	416.59479	20.66	0.0061
Error	5	100.83378	20.16676		
Corrected Total	6	517.42857			

Root MSE	4.49074	R-Square	0.8051
Dependent Mean	57.71429	Adj R-Sq	0.7662
Coeff Var	7.78099		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-0.70541	12.96505	-0.05	0.9587
IW	1	1.98514	0.43677	4.55	0.0061

----- GROUP=C -----

The REG Procedure
 Model: MODEL1
 Dependent Variable: WEIGHT

Number of Observations Read	7
Number of Observations Used	7

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	104.29162	104.29162	2.74	0.1590
Error	5	190.56552	38.11310		
Corrected Total	6	294.85714			

Root MSE 6.17358 R-Square 0.3537
 Dependent Mean 51.14286 Adj R-Sq 0.2244
 Coeff Var 12.07125

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	25.89718	15.43893	1.68	0.1543
IW	1	0.85786	0.51860	1.65	0.1590

So, slopes are 0.79028, 1.98514, 0.85786. Since we have only 7 observations in each group, these could be "equal". To test equality of the three slopes we use PROC GLM. Use IW*GROUP to test slopes equality.

```

PROC GLM DATA=FISH;
CLASS GROUP;
MODEL WEIGHT = IW GROUP IW*GROUP/ SS3;
RUN;

```

The GLM Procedure

Class Level Information

Class	Levels	Values
-------	--------	--------

GROUP 3 A B C

Number of Observations Read 21
Number of Observations Used 21

Dependent Variable: WEIGHT

Source	DF	Sum of Squares	Mean Square	F Value
Model	5	739.696129	147.939226	2.52
Error	15	880.589586	58.705972	
Corrected Total	20	1620.285714		

Source Pr > F

Model 0.0757

Error

Corrected Total

R-Square	Coeff Var	Root MSE	WEIGHT Mean
0.456522	14.00362	7.661982	54.71429

Source	DF	Type III SS	Mean Square	F Value
IW	1	503.1109877	503.1109877	8.57
GROUP	2	75.8994724	37.9497362	0.65
IW*GROUP	2	98.8557067	49.4278534	0.84

The GLM Procedure

Dependent Variable: WEIGHT

Source	Pr > F
IW	0.0104
GROUP	0.5379
IW*GROUP	0.4502 <---Same slopes.

Do not reject H_0 : Slopes are equal.

So, now we can run

$$y_{ij} = \mu + \alpha_i + \beta \cdot IQ_{ij} + \epsilon_{ij}$$

We wish to test:

$$H_0: \alpha_1 = \alpha_2 = \alpha_3$$

```

PROC GLM DATA=FISH;
CLASS GROUP;
MODEL WEIGHT = GROUP IW/ SOLUTION SS3;
LSMEANS GROUP;
RUN;

```

The SAS System 10
09:49 Thursday, November 15, 2007

The GLM Procedure

Class Level Information

Class	Levels	Values
GROUP	3	A B C

Number of Observations Read	21
Number of Observations Used	21

Dependent Variable: WEIGHT

Source	DF	Sum of Squares	Mean Square	F Value
Model	3	640.840422	213.613474	3.71
Error	17	979.445292	57.614429	
Corrected Total	20	1620.285714		

Source	Pr > F
Model	0.0322
Error	
Corrected Total	

R-Square	Coeff Var	Root MSE	WEIGHT Mean
0.395511	13.87282	7.590417	54.71429

Source	DF	Type III SS	Mean Square	F Value
GROUP	2	162.0256063	81.0128031	1.41
IW	1	486.2689932	486.2689932	8.44

Dependent Variable: WEIGHT

Source	Pr > F
GROUP	0.2722 <-- Not significant.
IW	0.0099

Do not reject H₀: alpha₁=alpha₂=alpha₃

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		16.46947099 B	12.27500014	1.34	0.1973
GROUP	A	4.81612678 B	4.06386179	1.19	0.2523
GROUP	B	6.57142857 B	4.05724850	1.62	0.1237
GROUP	C	0.00000000 B	.	.	.
IW		1.17822186	0.40555935	2.91	0.0099

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The GLM Procedure
Least Squares Means

GROUP	WEIGHT LSMEAN
A	55.7345607
B	57.4898625
C	50.9184339

Check LSMEANS: model applied to Average(All IW_ij)

Use: $\mu + \alpha_i + \beta \cdot \text{Average}(\text{all IW}_{ij})$:

IW = (35, 26, 29, 32, 23, 26, 31,
33, 35, 28, 29, 32, 23, 26,
33, 36, 26, 30, 33, 25, 23) <---All Initial Weights
from all the groups.

GROUP A: $16.46947099 + 4.81612678 + 1.17822186 \cdot \text{mean}(\text{IW}) = 55.73456$, OK
GROUP B: $16.46947099 + 6.57142857 + 1.17822186 \cdot \text{mean}(\text{IW}) = 57.48986$, OK
GROUP C: $16.46947099 + 0.00000000 + 1.17822186 \cdot \text{mean}(\text{IW}) = 50.91843$, OK