Lecture 12: Analysis of Covariance

• 12. 1 Introduction

- A psychologist interested in music education designs five music training programs for young children labeled *Bach*, *Chopin*, *Mozart*, *Stravinsky*, and *Debussy*. She randomly assigns children to training programs but is concerned that differences in the children's musical abilities will obscure the differences between the programs. Therefore, she measures the children's musical abilities before the training program

- What can she say about the effectiveness of her music training programs when considering the children's initial musical ability?

Program	*Performance* (Y)	*Ability* (X)	
B	18	10	
B	17	20	$\Sigma_Y = 77$
B	23	15	$\Sigma_X = 57$
B	19	12	
C	40	22	
C	22	31	$\Sigma_Y = 121$
C	28	16	$\Sigma_X = 86$
C	31	17	
M	38	30	
M	40	31	$\Sigma_Y = 159$
M	41	18	$\Sigma_X = 101$
M	40	22	
S	25	35	
S	45	37	$\Sigma_Y = 171$
S	50	41	$\Sigma_X = 143$
S	51	30	
D	15	11	
D	17	16	$\Sigma_Y = 75$
D	20	19	$\Sigma_X = 71$
D	23	25	

- ANCOVA is a statistical technique that permits a post-hoc, statistical control for one or more simultaneous variables, removing their influence from the comparison of groups on the experimental factor(s)

- The simplest ANCOVA case has 2 IVs
    - One IV is categorical (e.g., teaching program)
    - One IV is continuous (e.g., music ability)
    - DV is continuous (e.g., performance)
    - Could use ANOVA for categorical and Regression for continuous. But both are part of the GLM
    - Many people call mixing categorical and continuous variables Analysis of Covariance (ANCOVA)

- ## 12. 2 Statistical Model

$Y_{ij} = \mu + \alpha_j + \beta_{Y \cdot X}(x_{ij} - \mu_X) + \varepsilon_{ij}$

Y_{ij} = score for person i in group j

μ = population mean

α_j = effect of treatment j

$\beta_{Y \cdot X}(x_{ij} - \mu_X)$ = value of x for person i in group j, independent from μ_X, and weighted by the linear regression coefficient $\beta_{Y \cdot X}$. This value is assumed to not depend on j

ε_{ij} = error or residual for score Y_{ij}

$\varepsilon_{ij} \sim NID(0, \sigma^2)$

Within each treatment population, the relationship between X and Y is assumed to be linear and has the regression coefficient $\beta_{Y \cdot X}$

• 12. 3 Partitioning Sums of Squares and Sums of Products

- The partitioning of variance is the same as in the ANOVA approach but we now need to consider partitions of variance for Y, X, and XY

$$SS_{Y\,\mathrm{TOTAL}} = \Sigma_j\Sigma_i\,(Y_{ij} - \overline{Y}.)^2 = SS_{Y\,\mathrm{BETWEEN}} + SS_{Y\,\mathrm{WITHIN}}$$

$$SS_{Y\,\mathrm{BETWEEN}} = \Sigma_j\, n_i\,(\overline{Y}_j - \overline{Y}.)^2$$

$$SS_{Y\,\mathrm{WITHIN}} = \Sigma_j\,\Sigma_i\,(Y_{ij} - \overline{Y}_j)^2$$

$$SS_{X\,\mathrm{TOTAL}} = \Sigma_j\Sigma_i\,(X_{ij} - \overline{X}.)^2 = SS_{X\,\mathrm{BETWEEN}} + SS_{X\,\mathrm{WITHIN}}$$

$$SS_{X\,\mathrm{BETWEEN}} = \Sigma_j\, n_i\,(\overline{X}_j - \overline{X}.)^2$$

$$SS_{X\,\mathrm{WITHIN}} = \Sigma_j\,\Sigma_i\,(X_{ij} - \overline{X}_j)^2$$

$$SP_{XY\,\mathrm{TOTAL}} = \Sigma_j\Sigma_i\,(X_{ij} - \overline{X}.)(Y_{ij} - \overline{Y}.)$$

$$SP_{XY\,\mathrm{BETWEEN}} = \Sigma_j\, n_i\,(\overline{X}_j - \overline{X}.)(\overline{Y}_j - \overline{Y}.)$$

$$SP_{XY\,\mathrm{WITHIN}} = \Sigma_j\,\Sigma_i\,(X_{ij} - \overline{X}_j)(Y_{ij} - \overline{Y}_j)$$

Thus, the computation for analysis of covariance is based on:

1. The *SS* values for an ANOVA carried out on *X*
2. The *SS* values for an ANOVA carried out on *Y*
3. The *SP* values for an ANCOVA carried out on *XY* products

These computations can be put in matrix form as

$$\mathbf{SSCP_{TOTAL}} = \begin{bmatrix} SS_{X\ total} & SP_{XY\ total} \\ SP_{YX\ total} & SS_{Y\ total} \end{bmatrix}$$

$$\mathbf{SSCP_{BETWEEN}} = \begin{bmatrix} SS_{X\ between} & SP_{XY\ between} \\ SP_{YX\ between} & SS_{Y\ between} \end{bmatrix}$$

$$\mathbf{SSCP_{WITHIN}} = \begin{bmatrix} SS_{X\ within} & SP_{XY\ within} \\ SP_{YX\ within} & SS_{Y\ within} \end{bmatrix}$$

Thus, in matrix form, $\mathbf{SSCP_{TOTAL}} = \mathbf{SSCP_{BETWEEN}} + \mathbf{SSCP_{WITHIN}}$

• 12. 4 Computations in ANCOVA

- A simple one-way ANCOVA (i.e., one IV and one covariate) requires three sets of calculations, for Y, X, and XY

$$SS_{X\ \text{total}} = \Sigma_j \Sigma_i\, x_{ij}^2 - \frac{T_x^2}{N}, \text{ where } T_x = \Sigma_j \Sigma_i\, x_{ij}$$

$$SS_{X\ \text{between}} = \Sigma_j\, \frac{T_{xij}^2}{n_j} - \frac{T_x^2}{N}$$

$$SS_{X\ \text{within}} = SS_{X\ \text{total}} - SS_{X\ \text{between}}$$

$$SS_{Y\ \text{total}} = \Sigma_j \Sigma_i\, y_{ij}^2 - \frac{T_y^2}{N}$$

$$SS_{Y\ \text{between}} = \Sigma_j\, \frac{T_{yij}^2}{n_j} - \frac{T_y^2}{N}$$

$$SS_{Y\ \text{within}} = SS_{Y\ \text{total}} - SS_{Y\ \text{between}}$$

$$SP_{XY\ \text{total}} = \Sigma_j \Sigma_i\, x_{ij}\, y_{ij} - \frac{(T_x)(T_y)}{N}$$

$$SP_{XY\ \text{between}} = \Sigma_j\, \frac{(T_{xj})(T_{yj})}{n_j} - \frac{(T_x)(T_y)}{N}$$

$$SP_{XY\ \text{within}} = SP_{XY\ \text{total}} - SP_{XY\ \text{between}}$$

- Now we need to find the adjusted sums of squares, in order to remove the regression of Y on the covariate X

$$SS_{Y\ \text{adjusted means}} = SS_{Y\ between} + \frac{(SP_{XY\ within})^2}{SS_{X\ within}} - \frac{(SP_{XY\ \text{total}})^2}{SS_{X\ \text{total}}}$$

$$SS_{Y\ \text{adjusted error}} = SS_{Y\ within} - \frac{(SP_{XY\ within})^2}{SS_{X\ within}}$$

- Source Table for ANCOVA

Source	*SS*	*df*	*MS*	*F*
Adjusted Means	$SS_{Y\ adjusted\ between}$	$J-1$	$\frac{SS_{Y\ adjusted\ means}}{J-1}$	$\frac{MS_{Y\ adjusted\ means}}{MS_{Y\ adjusted\ error}}$
Adjusted Error	$SS_{Y\ adjusted\ within}$	$N-J-1$	$\frac{SS_{Y\ adjusted\ error}}{N-J-1}$	
Adjusted Total		$N-2$		

- In addition to the *SS*, we can obtain the Pearson correlation between the covariate *X* and the dependent variable *Y* over the entire sample

$$r_{xy} = \frac{SP_{XY\ total}}{\sqrt{(SS_{X\ total})(SS_{Y\ total})}}$$

- **12. 5 Example (a)**

- Based on the example of music performance, teaching program, and ability

$$SS_{X \text{ total}} = \Sigma_j\Sigma_i\, x_{ij}^2 - \frac{T_x^2}{N} = 12{,}066 - \frac{(458)^2}{20} = 1{,}577.8 \text{ , where } T_x = \Sigma_j\Sigma_i\, x_{ij}$$

$$SS_{X \text{ between}} = \Sigma_j \frac{T_{xij}^2}{n_j} - \frac{T_x^2}{N} = \frac{46{,}336}{4} - \frac{(458)^2}{20} = 1{,}095.8$$

$$SS_{X \text{ within}} = SS_{X \text{ total}} - SS_{X \text{ between}} = 1{,}577.8 - 1{,}095.8 = 482.00$$

For *Y*,

$$SS_{Y \text{ total}} = 20{,}851 - \frac{(603)^2}{20} = 2{,}670.55$$

$$SS_{Y \text{ between}} = \frac{80{,}717}{4} - \frac{(603)^2}{20} = 1{,}998.8$$

$$SS_{Y \text{ within}} = 2{,}670.55 - 1{,}998.8 = 671.75$$

For X*Y*,

$$SP_{XY \text{ total}} = \Sigma_j\Sigma_i\, x_{ij}\, y_{ij} - \frac{(T_x)(T_y)}{N} = 15{,}140 - \frac{458 \text{ x } 603}{20} = 1{,}331.3$$

$$SP_{XY \text{ between}} = \Sigma_j \frac{(T_{xj})(T_{yj})}{n_j} - \frac{(T_x)(T_y)}{N} = \frac{60{,}632}{4} - \frac{458 \text{ x } 603}{20} = 1{,}349.3$$

$$SP_{XY \text{ within}} = SP_{XY \text{ total}} - SP_{XY \text{ between}} = 1{,}331.3 - 1{,}349.3 = -18$$

And the adjusted *SS* are

$$SS_{Y\, adjusted \text{ total}} = 2{,}670.55 - \frac{(1{,}331.3)^2}{1{,}577.8} = 1{,}547.24$$

$$SS_{Y\, adjusted \text{ error}} = 671.75 - \frac{(-18)^2}{482} = 671.08$$

$$SS_{Y\, adjusted \text{ means}} = 1{,}547.24 - 671.08 = 876.16$$

- Summary Table

Source	*SS*	*df*	*MS*	*F*	*p*
Adjusted Means	876.16	4	219.1	4.6	< .05
Adjusted Error	671.08	14	47.9		
Adjusted Total	1,547.24	18			

The observed *F* > critical *F*, thus we reject the null hypothesis of no difference between the *adjusted means* (music teaching program after controlling for initial music ability)

The *F* value if the covariate *X* (music ability) had not been introduced would be 11.6, with degrees of freedom (4, 15). Therefore, introducing the covariate in the model did reduce the discrepancies among music teaching programs.

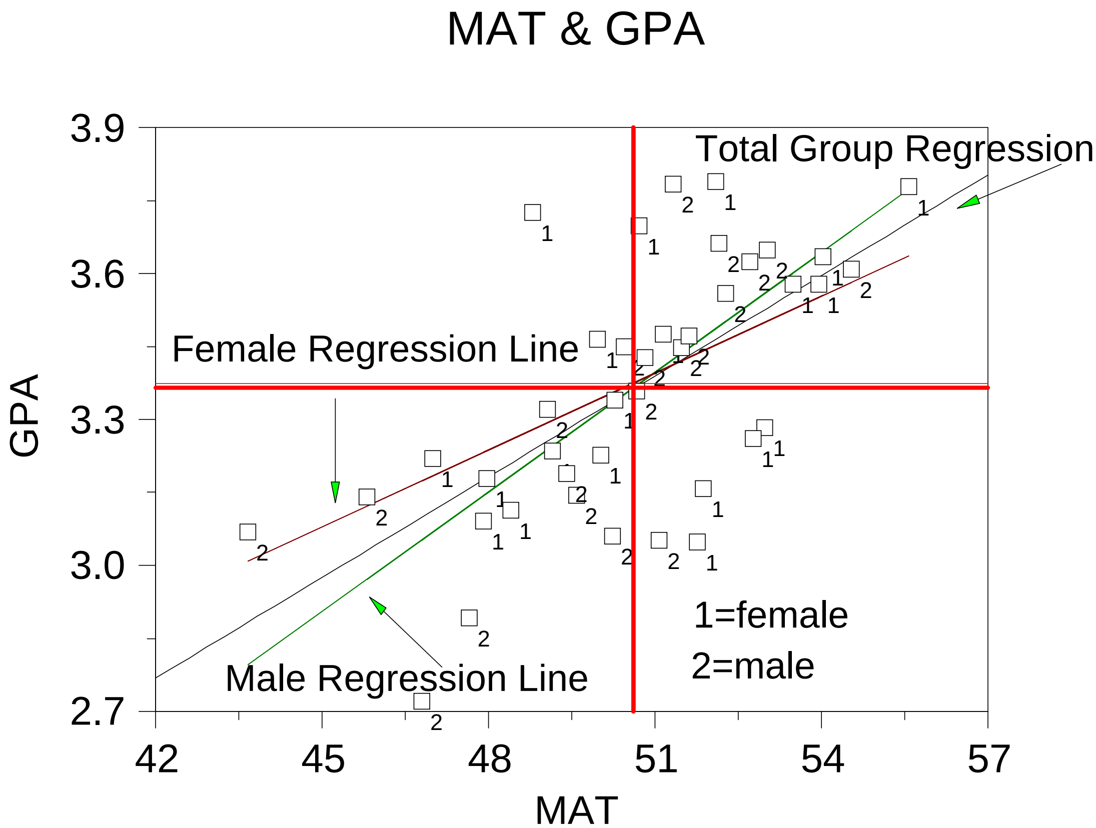
The average correlation within groups (or factor levels) can be calculated as

$$r_w = \frac{SP_{XY\ within}}{\sqrt{(SS_{X\ within})(SS_{Y\ within})}} = \frac{-18}{\sqrt{(482)(671.75}} = -.032$$

- **12. 5 Example (b)**

- Effect of Gender and MAT score on GPA in law school

N	Sex	MAT	GPA	N	Sex	MAT	GPA
1	1	51	3.7	21	-1	47	2.72
2	1	53	3.28	22	-1	53	3.62
3	1	52	3.79	23	-1	51	3.45
4	1	50	3.23	24	-1	51	3.78
5	1	54	3.58	25	-1	46	3.14
6	1	50	3.34	26	-1	48	2.89
7	1	52	3.05	27	-1	51	3.36
8	1	56	3.78	28	-1	51	3.05
9	1	49	3.23	29	-1	53	3.65
10	1	52	3.16	30	-1	55	3.61
11	1	50	3.46	31	-1	50	3.45
12	1	51	3.47	32	-1	51	3.43
13	1	49	3.73	33	-1	52	3.56
14	1	54	3.63	34	-1	50	3.14
15	1	48	3.09	35	-1	49	3.19
16	1	48	3.18	36	-1	49	3.32
17	1	53	3.58	37	-1	50	3.06
18	1	53	3.26	38	-1	52	3.47
19	1	48	3.11	39	-1	44	3.07
20	1	47	3.22	40	-1	52	3.66


MAT & GPA
Total Group Regression
Female Regression Line
Male Regression Line
1=female
2=male
GPA
MAT
3.9
3.6
3.3
3.0
2.7
42
45
48
51
54
57

- Testing Sequence:

  - 1 Construct vectors X, G and XG

    X is continuous

    G is group (categorical)

    XG is the product of the two

  $Y = a + b_1G + b_2X + b_3GX$

  where a is the intercept for common group, b_1 represents the difference in groups, b_2 represents the common slope, and b_3 represents the interaction term (difference in group slopes). That is, there are two common terms and two difference terms

  - 2 Estimate 3 slopes (and intercept)

    Examine R^2 for model. If R^2 is significant and large enough:

    Examine b_3. If significant, there is an interaction. Then, estimate separate regressions for different groups. If b_3 is not significant, re-estimate the model without XG. Examine and interpret b_1 and b_2

  - Heuristic Table

	Is b_1 significant? (G, categorical)	
Is b_2 significant? (X, cont)	Yes	No
Yes	Parallel slopes, different intercepts	Identical regressions
No	Mean diffs only; slopes are zero	Only possible with severe confounding; ambiguous story.

- Illustration

$R^2 = .44;\ p < .05$			
Y' = -.0389 +.75G +.0673X-.0146GX			
Term	Estimate	SE	t
G (b1; Sex)	.75	.6856	1.0567
X (b2; MAT)	.0673	.0125	4.9786*
GX (b3; Int)	-.0146	.0135	-1.0831

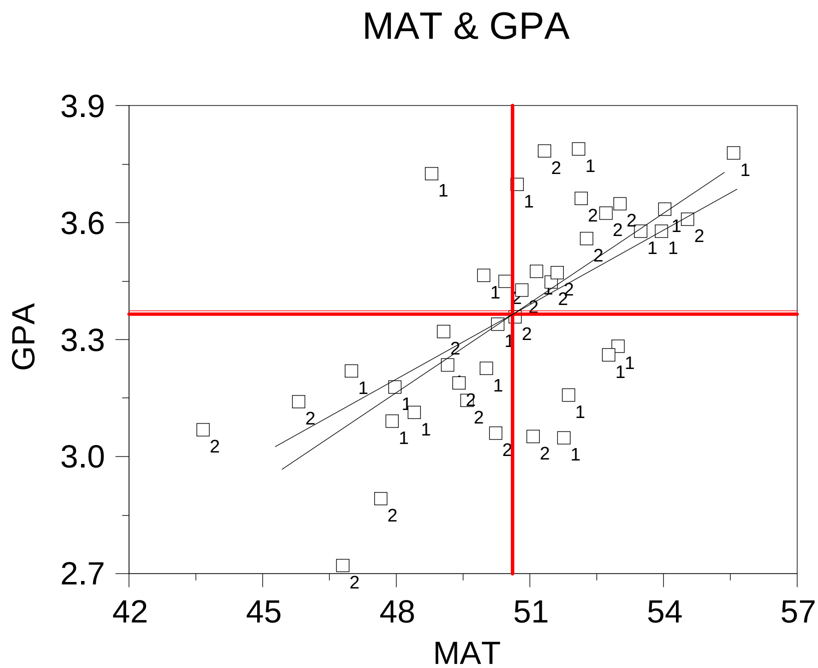
Step 1. R^2 is large and significant

Step 2. Slope for interaction (b_3) is non-significant

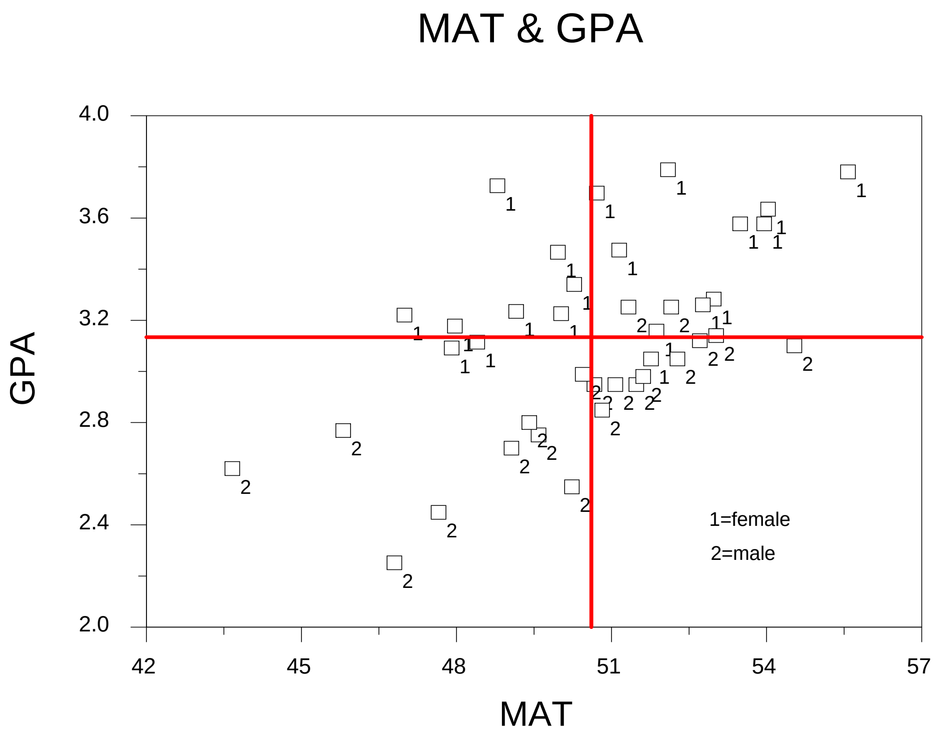
Step 3. Drop *GX* and re-estimate the model

$R^2 = .42;\ p < .05$			
Y' = .1154+.0045G+.0687X			
Term	Estimate	SE	t
G (b1; Sex)	.0045	.0833	.1365
X (b2; MAT)	.0687	.0135	5.0937*

Step 4. Examine slopes (*b* weights). The only significant slope is for MAT. We can conclude that both gender groups have identical regression lines and the slight apparent differences are due to sampling error


MAT & GPA
GPA
MAT
3.9
3.6
3.3
3.0
2.7
42
45
48
51
54
57

- Now, suppose that the data look like this, should we expect any differences in the results?



$R^2 = .72$; $p < .05$				
Y' = -11.54+.8268G+.0643X-.0117GX				
Term	Estimate	SE	t	p
G (b1; Sex)	.8268	.6627	1.2476	.22
X (b2; MAT)	.0643	.0131	4.2947	.0001
GX (b3; Int)	-.0117	.0131	-.8945	.3770

- Is there any interaction?

$R^2 = .72$; $p < .05$				
Y' = -.1805+.2346G+.0655X				
Term	Estimate	SE	t	p
G (b1; Sex)	.2346	.0320	7.34	.0001
X (b2; MAT)	.0655	.0130	5.05	.0001

• 12. 6 More Complex Designs

- With more complex designs, logic and sequence of tests remain the same

- The categorical variables may have more than 2 levels

- We may have several continuous IVs

- If multiple categories, create multiple (G – 1) interaction terms. If multiple Xs, create products for each. Test the terms as a block using hierarchical regression

• 12. 7 Should One Categorize Continuous IVs?

- This is sometimes used as the median split (e.g., personality, stress, BEM sex-role scales)

- This procedure is not desirable because

    - Loss of power and information – treat IQs of 100 and 140 as identical

    - Loss of replication (median changes by sample)

    - Arbitrary value of split - "high stress" group may not be very stressed

- An alternative is to throw out middle people – but this is also a problem because of range enhancement bias

- Example (c)

Model1: Dependent Variable: relsat1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25.7996156	25.7996156	64.46	<.0001
Error	260	104.0612930	0.4002357		
Corrected Total	261	129.8609086			

R-Square	Coeff Var	Root MSE	relsat1 Mean
0.198671	10.22429	0.632642	6.187637

Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	7.194924963	0.13140697	54.75	<.0001
avoid1	-0.400713850	0.04990978	-8.03	<.0001

Model2: Dependent Variable: relsat1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	78.2943706	39.1471853	196.62	<.0001
Error	259	51.5665380	0.1990986		
Corrected Total	261	129.8609086			

R-Square	Coeff Var	Root MSE	relsat1 Mean
0.602909	7.211228	0.446205	6.187637

Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	6.464038986	0.10303387	62.74	<.0001
avoid1	-0.128104762	0.03900012	-3.28	0.0012
grp	0.498013567	0.03067025	16.24	<.0001

Model3: Dependent Variable: relsat1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	79.5625424	26.5208475	136.04	<.0001
Error	258	50.2983662	0.1949549		
Corrected Total	261	129.8609086			

R-Square	Coeff Var	Root MSE	relsat1 Mean
0.612675	7.135793	0.441537	6.187637

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.507570668	0.10337482	62.95	<.0001
avoid1	-0.132106261	0.03862402	-3.42	0.0007
grp	0.245976747	0.10337482	2.38	0.0181
avoid1*grp	0.098509800	0.03862402	2.55	0.0113

By Group

-- grp=-1 --

Number of Observations Read 119
Number of Observations Used 119

Dependent Variable: relsat1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.33945822	3.33945822	9.52	0.0025
Error	117	41.03240434	0.35070431		
Corrected Total	118	44.37186255			

R-Square	Coeff Var	Root MSE	relsat1 Mean
0.075261	10.58136	0.592203	5.596664

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.261593922	0.22221399	28.18	<.0001
avoid1	-0.230616061	0.07473467	-3.09	0.0025

-- grp=1 --

Number of Observations Read 143
Number of Observations Used 143

Dependent Variable: relsat1

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.07687511	0.07687511	1.17	0.2813
Error	141	9.26596187	0.06571604		
Corrected Total	142	9.34283698			

R-Square	Coeff Var	Root MSE	relsat1 Mean
0.008228	3.837925	0.256351	6.679427

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.753547415	0.07180506	94.05	<.0001
avoid1	-0.033596461	0.03106250	-1.08	0.2813