METHODS QUALIFYING EXAM

August 2006

INSTRUCTIONS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER LEFT HAND CORNER of EACH PAGE of your exam.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. Be sure to attempt all parts of every question. It may be possible to answer a later part of a question without having solved the earlier parts.
5. Be sure to hand in all of your exam. No additional material will be accepted once the exam has ended and you have left the exam room.

Problem I.

A researcher is developing a commerical shrimp farming operation. She has sought your help in designing and analyzing a study to investigate the influence of three factors on the growth rate of shrimp raised in aquaria. The three factors are:

T = Water Temperature ($25°C$, $35°C$)

S = Water Salinity (10%, 25%, 40%)

D = Density of shrimp in the aquarium (2 shrimp/liter, 4 shrimp/liter)

The response variable is the four week weight gain on a per shrimp basis.

A. Two possible experimental designs are listed below. For each design, discuss the advantages and disadvantages of the design. In addition, give a brief description of how you would assign the levels of the three factors, or combinations of factor levels to the experimental material. Suppose there are 36 aquaria available for the study. If it helps, you can sketch the experimental layout.

   D1. Each aquarium can be partitioned into two sections, but the water in the two sections is common to both sections (i.e., the water in one section circulates through the entire aquarium).

   D2. It is not possible to partition the aquaria into sections.

For the following questions, assume that the experiment is run using Design D2:

B. The researcher asks if 36 aquaria will provide sufficient data points. How would you try to answer her question? Is there additional information needed? If so, what information?

C. Write a model for the experiment and provide an AOV table with just the source of variation and degrees of freedom (**do not calculate any sum of squares**).

D. The cell means for the experiment are given on the next page. Sketch the profile plot showing overall S*D interaction and then sketch the profile plots showing the S*D interactions separately for each level of Water Temperature.

E. Using the sketches and assuming that the experimental data yielded $MSE = 3280$, which of the following main effects and interactions do you think are significant (**do not calculate any sum of squares**)? Provide justifications for your answers.

   i. T*S*D

   ii. S*D

   iii. S

Tables of Means

TEMP	DEN	SAL	$\bar{y}_{ijk.}$	$\bar{y}_{ij..}$
$25°C$	2	10%	70	
		25%	465	
		40%	359	298
$25°C$	4	10%	72	
		25%	333	
		40%	252	219
$35°C$	2	10%	408	
		25%	276	
		40%	243	309
$35°C$	4	10%	330	
		25%	312	
		40%	231	291

	TEMP		
SAL	$25°C$	$35°C$	$\bar{y}_{..k.}$
10%	71	369	220
25%	399	294	346.5
40%	305.5	237	271.25
$\bar{y}_{i...}$	258.5	300	

	DEN		
SAL	2	4	$\bar{y}_{..k.}$
10%	239	201	220
25%	370.5	322.5	346.5
40%	301	241.5	271.25
$\bar{y}_{.j..}$	303.5	255	

Problem 2

(a) A drug is tested at four equally spaced doses on a total of 20 patients, with 5 patients in each dose group. The response of interest for the i^{th} patient at the j^{th} dose is Y_{ij}, $i = 1, ..., 5$; $j = 1, ..., 4$. The responses Y_{ij} are independent and normally distributed with common variance σ^2. Use a sum of squares partition in ANOVA to test for linear, quadratic and cubic trends. (HINT: The relevant orthogonal polynomials lead to these contrast vectors: (-3, -1, 1, 3), (1, -1, -1, 1) and (-1, 3, -3, 1).

(b) We can also approach this problem using ordinary least squares regression. What design matrix would you use to estimate regression parameters associated with linear, quadratic and cubit trends? Describe a test statistic and its null distribution to test for linear trends in a model that includes an intercept and covariates for linear, quadratic and cubic trends.

(c) Suppose we have two drugs, A and B, tested at the same doses as in part (a) and we have two sets of responses, Y_{ij} and Z_{ij}, $i = 1, ..., 5$; $j = 1, ..., 4$, corresponding to the two drugs. Assume the Y_{ij} and Z_{ij} are independently and normally distributed with common variance σ^2. Describe a test statistic and its null distribution for testing the null hypothesis that the effects of the two drugs follow the same linear trend.

Problem 3

A book[1] on robust statistical methods published in June 2006 considers regression models for a data set taken from Jalali-Heravi and Knouz (2002, Electronic Journal of Molecular Design, 1, 410-417). The aim of the modeling is to predict a physical property of chemical compounds called the Krafft point based on four potential predictor variables using a data set of size $n=32$. According to Maronna, Martin and Yohai (2006, p. 380):

> The Krafft point is an important physical characteristic of the compounds called surfactants, establishing the minimum temperature at which a surfactant can be used.

The authors of the original paper sought to find a regression model to predict:

Y = Krafft Point (KPOINT)

from

x_1 = Randic Index (RA)
x_2 = Heat of formation (HEAT)
x_3 = Reciprocal of volume of the tail of the molecule (VTINV)
x_4 = Reciprocal of Dipole Moment (DIPINV)

The first model considered by Jalali-Heravi and Knouz (2002) was

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e \qquad (1)$$

Output from model (1) appears on the following pages.

a) Decide whether (1) is a valid model. Give reasons to support your answer.

b) The plots of standardized residuals against RA and VTINV produce curved patterns. Describe what, if anything can be learnt about model (1) from these plots. Give a reason to support your answer.

c) Jalali-Heravi and Knouz (2002) give "four criteria of correlation coefficient (r), standard deviation (s), F value for the statistical significance of the model and the ratio of the number of observations to the number of descriptors in the equation" for choosing between competing regression models. Provide a detailed critique of this suggestion.

[1]Maronna, R.A., Martin. R.D. & Yohai, V.I. (2006) *Robust Statistics.* Wiley, New York

The second model considered by Jalali-Heravi and Knouz (2002) was:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + e \quad (2)$$

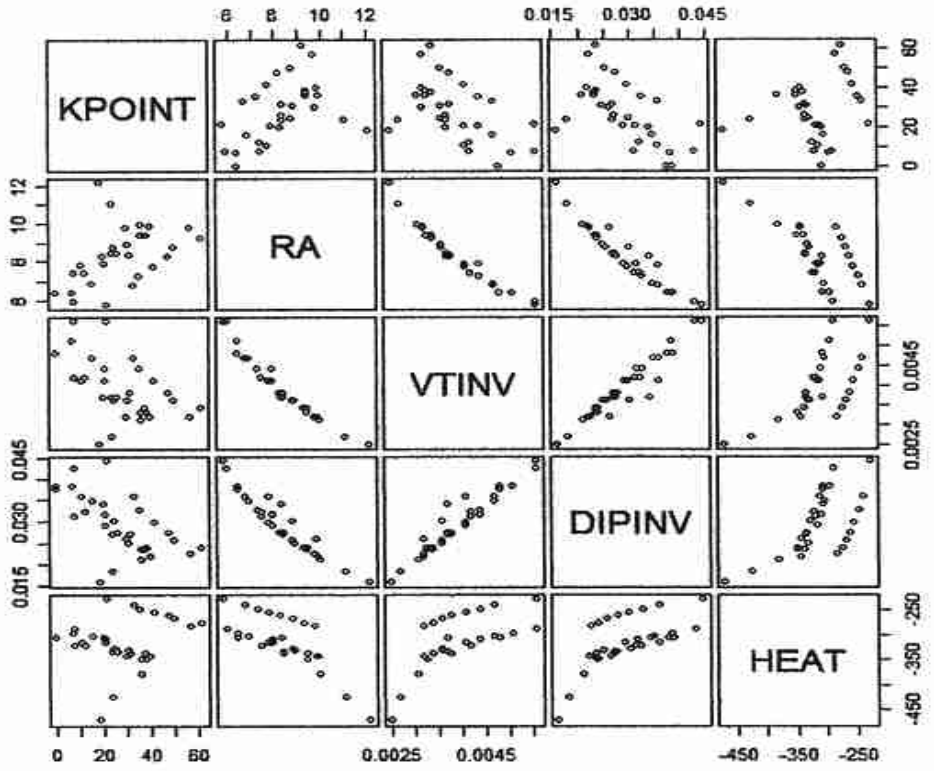
Output from model (2) appears on the following pages.

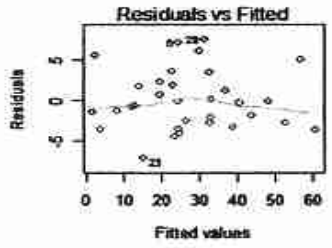
d) Decide whether (2) is a valid model. Give reasons to support your answer.

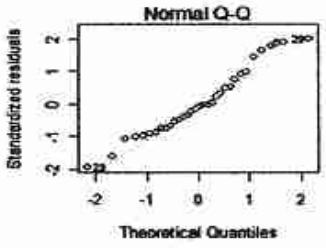
e) Jalali-Heravi and Knouz (2002) "believe that model (2) is superior to model (1)" since it produces a lower value of AIC, and since VTINV shows a "high correlation with" RA and DIPINV" and they recommend this model or use in practice. Do you agree with their conclusion and recommendation? Give reasons to support your answer.

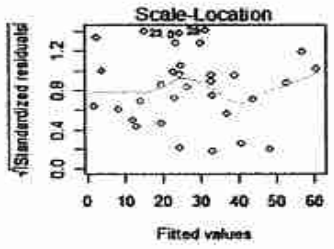
f) A statistics professor from Texas A&M University discovered while reading the paper by Jalali-Heravi and Knouz (2002) that there are 4 groups in the data set corresponding to four different surfactants. The last page contains a scatter plot matrix of the data with the four different groups marked by different plotting symbols. Also given in each plot are the least squares fits for each group. Explain carefully the steps you would take to obtain a final model allowing for the different groups.

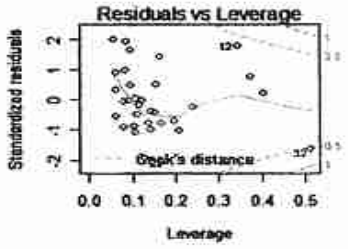
Output from model (1)



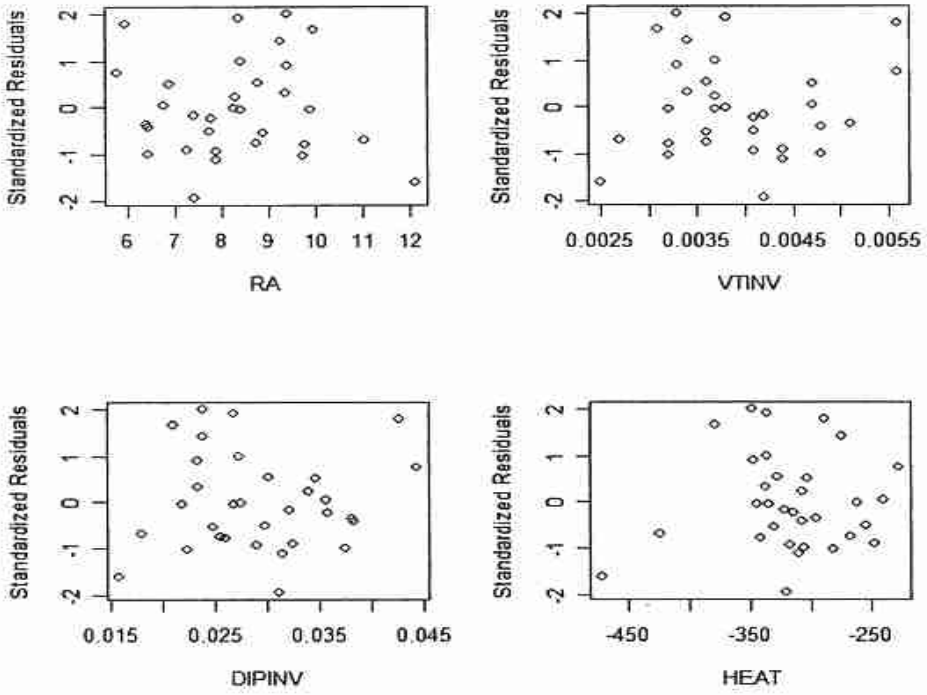








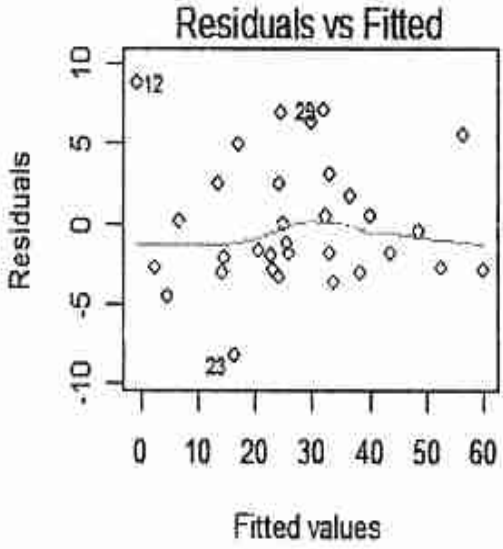
Output from model (1)

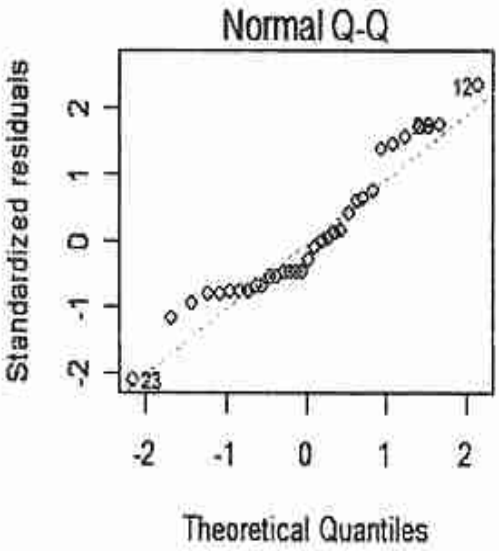


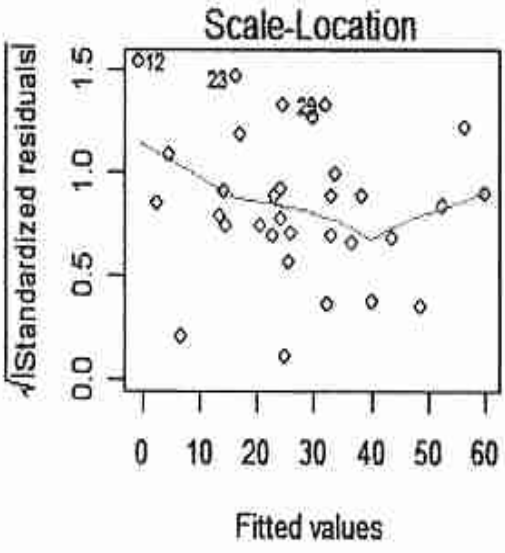
Output from R

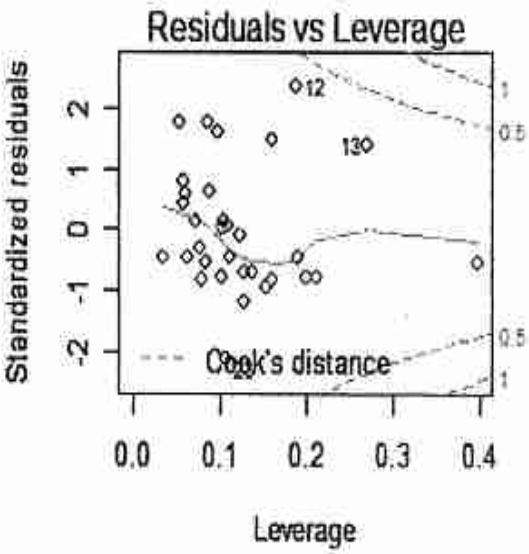
```
Call:
lm(formula = KPOINT ~ RA + VTINV + DIPINV + HEAT)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.031e+01  3.368e+01   2.088 0.046369 *
RA           1.047e+01  2.418e+00   4.331 0.000184 ***
VTINV        9.038e+03  4.409e+03   2.050 0.050217 .
DIPINV      -1.826e+03  3.765e+02  -4.850 4.56e-05 ***
HEAT         3.550e-01  2.176e-02  16.312 1.66e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.919 on 27 degrees of freedom
Multiple R-Squared: 0.9446,     Adjusted R-squared: 0.9363
F-statistic:   115 on 4 and 27 DF,  p-value: < 2.2e-16

vif(m1)
       RA     VTINV    DIPINV      HEAT
25.792770 22.834190 13.621363  2.389645
```

Output from model (2)

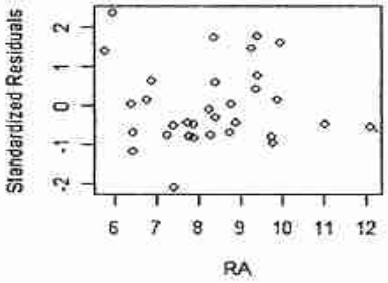


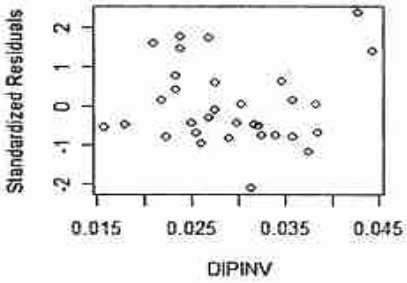


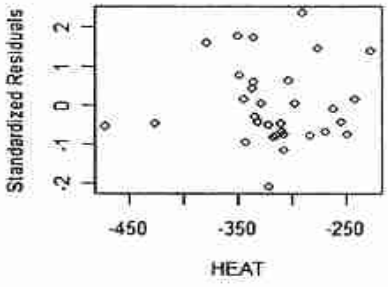




Output from model (2)



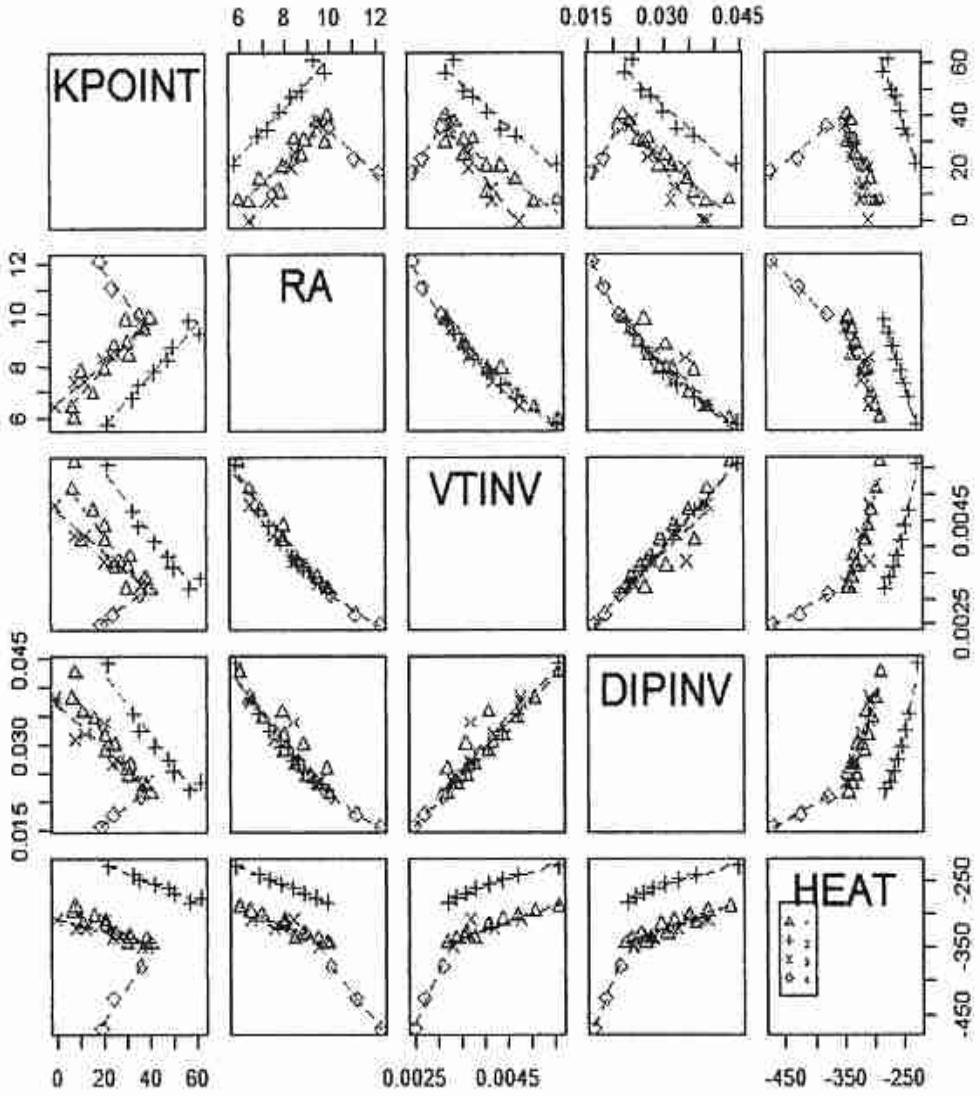




Output from R

```
Call:
lm(formula = KPOINT ~ RA + DIPINV + HEAT)

Residuals:
    Min      1Q  Median      3Q     Max
-8.304 -2.762 -1.491   2.408   8.775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.214e+02  2.388e+01   5.086 2.19e-05 ***
RA           7.147e+00  1.893e+00   3.776 0.000764 ***
DIPINV      -1.508e+03  3.621e+02  -4.165 0.000270 ***
HEAT         3.465e-01  2.255e-02  15.364 3.58e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.137 on 28 degrees of freedom
Multiple R-Squared: 0.9359,     Adjusted R-squared: 0.9291
F-statistic: 136.4 on 3 and 28 DF,  p-value: < 2.2e-16

vif(m2)
       RA    DIPINV      HEAT
14.178611 11.304448  2.302705
```

Scatter plot matrix of the data with the four different groups marked by different plotting symbols



Problem 4

Background:

The following questions have been posted to a discussion board. Assume that you are the statistical consultant. Please provide answers to the client's questions.

A.

Hello everybody.

I am doing a regression analysis. I made the K-S test on the dependent variable and saw that it is not normally distributed (Asymp. Sig. = 0.000). Is there a data transformation or some magical incantation that exists somewhere to transform this data into a normally distributed set?

Bye,

Marc.

B.

Hi all:

I have question (a couple maybe) about the GLM procedure. I have run a simply two-way ANOVA, but the cell sizes are unequal (i.e., I have an unbalanced design). When I run the basic procedure (and ask for descriptive statistics) I get weighted and unweighted marginal means. This means are fairly different.

My first question is: to which of these means do the statistical tests for the main effects in the ANOVA Summary table correspond to, weighted or unweighted?

My second question is: how do I get an ANOVA summary table for the "other" set of means given that the default is either weighted or unweighted?

Kris

(Note: To answer Kris, give the model using the "cell means" model and give the null hypothesis for testing "main effect A.")

C.

Dear All,

Here is my question:

I have three Groups (Experimental 1, Experimental 2, Control); a dependent variable Y, and a covariate X in my data file. I am running GLM treating Group as a fixed effect and using X as the covariate. I got the following output. It seems to me that since the p-value of X is greater than .05, there is no effect to the covariate X. Is this correct? If not, what is being tested by the term X?

Tests of Between-Subjects Effects

Dependent Variable: y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	2142.958[a]	5	428.592	587.859	.000
Intercept	10.840	1	10.840	14.869	.002
group	1.594	2	.797	1.093	.362
x	.182	1	.182	.249	.625
group * x	131.341	2	65.671	90.074	.000
Error	10.207	14	.729		
Total	2257.412	20			
Corrected Total	2153.165	19			

a. R Squared = .995 (Adjusted R Squared = .994)

Best.

Enis

D.

Hi Group,

My boss asked me to explain what is meant by "testing for equal variances" in regression. He learned that "equal variances" was an important assumption in regression analysis. I told him that we were testing a null hypothesis which is:

Ho: variance = a constant (this value comes from the MSE from the ANOVA table).

He said "No Way. The term "testing for equal variances" means more than one variance." Am I correct and, if not, he wants to know how many variances we are testing?

Please help.

Mary