METHODS QUALIFYING EXAM

August 2009

Student's Name _____

INSTRUCTIONS FOR STUDENTS:

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 solutions.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Use only one side of each sheet of paper.
4. You must answer Questions I, II, and III but **select only ONE** of Questions IV and V to answer.
5. Be sure to attempt all parts of the four questions. It may be possible to answer a later part of a question without having solved the earlier parts.
6. 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.
7. You may use only a calculator, pencil or pen, and blank paper for this examination. No other materials are allowed.
8. You are to answer Questions I, II, and III and then select **ONE** of Questions IV and V in this exam.

I attest that I spent no more than 4 hours to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature_____

INSTRUCTIONS FOR PROCTOR:

Immediately after the student completes the exam, **fax** the student's solutions to **979-845-6060** or email to mspeed@stat.tamu.edu Do not send the questions, just send the student's solutions.

(1) I certify that the time at which the student started the exam was _____ and the time at which the student completed the exam was

_____.

(2) I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.

(3) I certify that the student's solutions were faxed to **979-845-6060** or

emailed to **mspeed@stat.tamu.edu**.

Proctor's Signature_____

Problem I. For the following two experiments, provide the following information:

1. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, Crossover, etc.;
2. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, Fractional, etc.;
3. Identify each of the factors as being Fixed or Random;
4. Describe the Experimental Units and Measurement Units.
5. Describe the Measurement Process: Response Variable, Covariates, SubSampling, Repeated Measures
6. An ANOVA Table Including: Sources of Variation and Degrees of Freedom Freedom

Experiment A:

An experiment studies the effect of insect infestation and weeds on the yield of cotton plants. The experiment will include four levels of infestation, three weed treatments (no additional weeds, addition of weed species 1, addition of weed species 2), and two herbivore treatments (clipping or no clipping). There are eight fields; each field has three plots of land. Each of the plots receive 200 cotton plants to start growth. The eight fields are randomly assigned to the four levels of insect infestation, with two fields to each level. Within each field, the three plots are randomly assigned to the three weed treatments. Each plot is then split into two subplots; with one subplot randomly assigned to be clipped and the other subplot is not clipped. At the end of 18 weeks, the researcher determines the total cotton yield of each subplot of land. The yields are given here with the following notation: Field (F), Infestation (I), weed treatment (W), and clipping (C).

		W1		W2		W3				W1		W2		W3	
F	I	C1	C2	C1	C2	C1	C2	F	I	C1	C2	C1	C2	C1	C2
F1	I1	83.2	81.8	67.4	79.7	75.9	80.6	F5	I1	78.2	80.5	65.1	68.3	65.3	66.6
F2	I2	77.5	78.2	69.2	71.5	75.9	78.2	F6	I2	79.8	85.2	57.6	61.4	58.5	61.6
F3	I3	72.7	69.3	70.1	71.2	75.9	81.3	F7	I3	82.4	83.1	50.5	54.0	51.6	54.7
F4	I4	75.3	78.9	72.7	74.6	75.9	82.8	F8	I4	75.5	78.7	39.0	43.9	41.9	45.1

Experiment B:

A human nutrition researcher conducted an experiment to determine the acceptability of cakes baked with sucrose substitutes as the sweetening agent. Specifically, there were 6 recipes formed by combinations of 3 sweeteners and 2 leavening agents:

S_1 : 100% sucrose S_2 : 75% corn syrup, 25% sucrose S_3 : 75% fructose, 25% sucrose

L_1 : Baking soda L_2 : Baking soda plus "additional acid"

A panel of 6 taste testers were used to evaluate various characteristics of the cakes. On each of three days, cakes were baked from all six recipes. On each day, the six tasters evaluated six cake samples, one from each of the six recipes. The tasters then assigned a taste evaluation score to each of the recipes. The following table provides the tasting regimen for the three days:

	Day 1						Day 2						Day 3					
	Order						Order						Order					
Taster	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
T1	R1	R2	R3	R4	R5	R6	R6	R3	R4	R2	R5	R1	R1	R3	R5	R2	R4	R6
T2	R2	R3	R4	R5	R6	R1	R1	R4	R5	R3	R6	R2	R3	R5	R2	R4	R6	R1
T3	R3	R4	R5	R6	R1	R2	R2	R5	R6	R4	R1	R3	R5	R2	R4	R6	R1	R3
T4	R4	R5	R6	R1	R2	R3	R3	R6	R1	R5	R2	R4	R2	R4	R6	R1	R3	R5
T5	R5	R6	R1	R2	R3	R4	R4	R1	R2	R6	R3	R5	R4	R6	R1	R3	R5	R2
T6	R6	R1	R2	R3	R4	R5	R5	R2	R3	R1	R4	R6	R6	R1	R3	R5	R2	R4

Problem II.

Two types of models for an experiment with a continuous response Y and four treatments are under consideration. The two models are given here:

"Dummy variable" model

$$(1) \quad \mu_i = E(Y_{ij}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3; \quad \text{for} \quad i = 1, 2, 3, 4; \quad j = 1, \ldots, n,$$

where $x_i = 1$ if treatment i and $x_i = 0$, otherwise, for $i = 1, 2, 3$.

"Design effects" model

$$(2) \quad \mu_i = E(Y_{ij}) = \alpha^* + \beta_1^* x_1 + \beta_2^* x_2 + \beta_3^* x_3; \quad \text{for} \quad i = 1, 2, 3, 4; \quad j = 1, \ldots, n,$$

where, for $i = 1, 2, 3$

$x_i = 1$ if treatment i, $x_i = -1$ if treatment 4, and $x_i = 0$, otherwise.

1. By considering the mean responses, $\mu_1, \mu_2, \mu_3, \mu_4$, of the four treatments, determine the relationship between $\alpha, \beta_1, \beta_2, \beta_3$ and $\alpha^*, \beta_1^*, \beta_2^*, \beta_3^*$.
2. Obtain expressions for the difference in mean responses of treatments 2 and 3 for both models (1) and (2).
3. Suppose that there are factors A and B, each with two levels, that were used to define the treatments as follows: Treatment 1= A_1B_1, Treatment 2= A_1B_2, Treatment 3= A_2B_1, Treatment 4= A_2B_2. Express each of the following null hypotheses in term of the coefficients for the dummy variable model (1):
   i. No interaction between A and B
   ii. No effect due to A
   iii. No effect due to B
4. Consider now the **"dummy variable"** model for the logarithm of the mean:

$$(3) \quad \mu_i^* = \log(E(Y_{ij})) = \alpha^{**} + \beta_1^{**} x_1 + \beta_2^{**} x_2 + \beta_3^{**} x_3; \quad \text{for} \quad i = 1, 2, 3, 4; \quad j = 1, \ldots, n,$$

where $x_i = 1$ if treatment i and $x_i = 0$, otherwise, for $i = 1, 2, 3$.

   i. Explain what the coefficients $\alpha^{**}, \beta_1^{**}, \beta_2^{**}, \beta_3^{**}$ represent in terms of the mean responses $\mu_i = E(Y_{ij})$ $i = 1, 2, 3, 4$.
   ii. Explain what the difference in coefficients, $\beta_1^{**} - \beta_2^{**}$ represents in terms of the mean responses, μ_i's, of the four treatments.

Problem III.

This is the "Nambeware Polishing Times" data file. It concerns the efforts of a metal tableware manufacturer (Nambe Mills, Santa Fe, N. M.) to plan its production schedule. Each case represents a different item in the product line. The diameter, polishing time, price, and product type (there are 5 product types) are recorded for each item. Price is the dependent variable. The model fit to the data was

$$\text{Price} = \beta_{0i} + \beta_{1i} * \text{Type} * \text{Time} + \beta_{2i} * \text{Type} * \text{Diameter} + \text{error} \quad \text{for} \quad i = 1, 2, 3, 4, 5$$

Using the output on this page and on Pages 5 and 6, answer the following questions: Do not make any additional assumptions.

1. Since Price is not normally distributed, what course of action should you take. Transform Price; transform both Price and the Predictors; do nothing. Explain you answer.

2. The R-Square is quite high. Does this mean that the model assumptions have been met. Explain.

3. In the following table, what null hypothesis is being tested by F = 151.95? Spell out the null hypothesis.

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr>F**
Model	15	583240.0866	38882.6724	151.95	< 0.0001

4. In the following table, what null hypothesis is being tested by F = 15.22?

   Spell out the null hypothesis.

Source	**DF**	**Type III SS**	**Mean Square**	**F Value**	**Pr>F**
Type	5	3891.92579	778.38516	3.04	0.0192
Time*Type	5	19470.94059	3894.18812	15.22	< 0.0001

Distribution analysis of: price

The UNIVARIATE Procedure
Variable: price (price)

Basic Statistical Measures			
Location		Variability	
Mean	86.38136	Std Deviation	51.57129
Median	75.00000	Variance	2660
Mode	99.00000	Range	238.50000
		Interquartile Range	64.00000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	12.86584	Pr > |t|	<.0001
Sign	M	29.5	Pr >= |M|	<.0001
Signed Rank	S	885	Pr >= |S|	<.0001

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.904181	Pr < W	0.0002
Kolmogorov-Smirnov	D	0.132164	Pr > D	0.0112
Cramer-von Mises	W-Sq	0.219789	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.472978	Pr > A-Sq	<0.0050

The GLM Procedure

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	583240.0866	38882.6724	151.95	<.0001
Error	44	11259.1634	255.8901		
Uncorrected Total	59	594499.2500			

R-Square	Coeff Var	Root MSE	price Mean
0.927010	18.51854	15.99656	86.38136

Source	DF	Type III SS	Mean Square	F Value	Pr > F
type	5	3891.92579	778.38516	3.04	0.0192
time*type	5	19470.94059	3894.18812	15.22	<.0001
diam*type	5	10339.64186	2067.92837	8.08	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
type Bowl	-25.43340388	11.96199181	-2.13	0.0391
type Casserole	-29.73215408	37.86173748	-0.79	0.4365
type Dish	-25.90176313	26.04032795	-0.99	0.3253
type Plate	-2.74650494	32.51765971	-0.08	0.9331
type Tray	-48.87508509	16.22377283	-3.01	0.0043
time*type Bowl	1.37132089	0.36671340	3.74	0.0005
time*type Casserole	2.18134783	0.32535774	6.70	<.0001
time*type Dish	1.78116744	0.71406240	2.49	0.0164
time*type Plate	-0.94400915	1.55388850	-0.61	0.5466
time*type Tray	1.81331318	0.55784223	3.25	0.0022
diam*type Bowl	5.89371075	1.42682164	4.13	0.0002
diam*type Casserole	3.94831670	2.76362585	1.43	0.1602
diam*type Dish	4.91642503	4.30069811	1.14	0.2592
diam*type Plate	7.54121327	1.86103147	4.05	0.0002
diam*type Tray	5.00293936	2.64552608	1.89	0.0652

Problem IV. Assume the two-part linear regression model:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

where $\mathbf{Y}$ is an $n \times p$ matrix of response variables, $\mathbf{X}_1$ and $\mathbf{X}_2$ are, respectively, $n_1 \times p_1$ and $n_2 \times p_2$ matrices of fixed (i.e., nonrandom) explanatory variables, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are. respectively, $p_1 \times 1$ and $p_2 \times 1$ vectors of unknown parameters, and $p_1 + p_2 = p$. Assume that the $n \times p$ matrix $\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2]$ is of full column rank. The $n \times 1$ vector $\boldsymbol{\epsilon}$ is assumed to be comprised of independent, identically distributed $(0, \sigma^2)$ random errors.

Hint: To answer parts a) through c) below you may wish to use the following result:

For any real symmetric matrix $\mathbf{Q}$*, the expectation of the quadratic form* $\mathbf{Y}^T\mathbf{Q}\mathbf{Y}$ *is*

$$E\left(\mathbf{Y}^T\mathbf{Q}\mathbf{Y}\right) = tr\left[\mathbf{Q}Var\left(\mathbf{Y}\right)\right] + \left[E\left(\mathbf{Y}\right)\right]^T\mathbf{Q}\left[E\left(\mathbf{Y}\right)\right].$$

1. Show that:

$$E\left(\mathbf{Y}^T\mathbf{P}_1\mathbf{Y}\right) = p_1\sigma^2 + \left(\boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2\right)^T\mathbf{X}_1{}^T\mathbf{X}_1\left(\boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2\right),$$

where $\mathbf{P}_1 = \mathbf{X}_1\left(\mathbf{X}_1{}^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1{}^T$ and $\mathbf{A} = \left(\mathbf{X}_1{}^T\mathbf{X}_1\right)^{-1}\mathbf{X}_1{}^T\mathbf{X}_2$.

Explain why it is <u>not</u> necessary that $\boldsymbol{\beta}_1 = 0$ in order for $E\left[\left(\mathbf{Y}^T\mathbf{P}_1\mathbf{Y}\right)/(p_1\sigma^2)\right] = 1$.

Give simple sufficient conditions (that $\mathbf{X}_1$ and $\mathbf{X}_2$ satisfy)

in order for $E\left[\left(\mathbf{Y}^T\mathbf{P}_1\mathbf{Y}\right)/(p_1\sigma^2)\right] = 1$ to imply that $\boldsymbol{\beta}_1 = 0$.

2. Show that:

$$E\left[\mathbf{Y}^T\left(\mathbf{P} - \mathbf{P}_1\right)\mathbf{Y}\right] = p_2\sigma^2 + \boldsymbol{\beta}_2^T\mathbf{X}_2{}^T\left(I - \mathbf{P}_1\right)\mathbf{X}_2\boldsymbol{\beta}_2,$$

where $\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$. Give sufficient conditions (that $\mathbf{X}_1$ and $\mathbf{X}_2$ satisfy) in order

for $E\left[\mathbf{Y}^T\left(\mathbf{P} - \mathbf{P}_1\right)\mathbf{Y}/(p_2\sigma^2)\right] = 1$ to imply that $\boldsymbol{\beta}_2 = 0$.

3. Show that:

$$E\left[\mathbf{Y}^T\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y}\right] = (n - p)\sigma^2.$$

4. Suppose that we are interested in testing:

$$H_o : \boldsymbol{\beta}_1 = 0 \ \text{ versus } \ H_1 : \boldsymbol{\beta}_1 \neq 0.$$

Would the statistic $F = \left(\mathbf{Y}^T\mathbf{P}_1\mathbf{Y}\right)/\left(p_1 s^2\right)$, where $s^2 = \left[\mathbf{Y}^T\left(\mathbf{I} - \mathbf{P}\right)\mathbf{Y}\right]/(n - p)$ is the

least squares estimator of σ^2, be a useful test statistic in general? Justify your answer heuristically using the results in parts (1.) and (3.).

5. Assume $\mathbf{X}$ is of full rank and suppose that we are interested in testing:

$$H_o : \boldsymbol{\beta}_2 = 0 \ \text{ versus } \ H_1 : \boldsymbol{\beta}_2 \neq 0.$$

Would the statistic $F = \left[\mathbf{Y}^T\left(\mathbf{P} - \mathbf{P}_1\right)\mathbf{Y}\right]/\left(p_2 s^2\right)$ ever be a reasonable test statistic? Again, justify your answer heuristically, now using the results in parts (2.) and (3.).

Problem V.

Part 1:

Suppose that Y has mean equal to μ and variance equal to μ^4 show that the appropriate transformation of Y for stabilizing variance is the reciprocal transformation (i.e., 1/Y).

Part 2:

Chu (1996, Diamond ring pricing using linear regression *Journal of Statistics Education*, **4**, http://www.amstat.org/publications/jse/v4n3/datasets.chu.html) discusses the development of a regression model to predict the price of diamond rings from the size of their diamond stones (in terms of their weight in carats). Data on both variables were obtained from a full page advertisement placed in the Straits Times newspaper by a Singaporebased retailer of diamond jewelry. Only rings made with 20 carat gold and mounted with a single diamond stone were included in the data set. There were 48 such rings of varying designs. (Information on the designs was available but not used in the modeling.) The weights of the diamond stones ranged from 0.12 to 0.35 carats (a one carat diamond stone weighs 0.2 gram) and were priced between \$223 and \$1086.

An analyst fit two models to the data. The first model fit to the data was

$$\text{Price} = \beta_o + \beta_1 \text{Size} + e \qquad (1)$$

On Page 9 is some output from fitting model (1) as well as some plots.

The second model fit to the data was

$$\text{Log(Price)} = \beta_o + \beta_1 \text{Log(Size)} + e \qquad (2)$$

Output from model (2) and plots appear on Page 10.

(a) Based on the output for models (1) and (2) the analyst concluded the following:

Since model (1) has a higher R^2 than model (2), model (1) is a more effective model for producing prediction intervals for Price.

Provide a detailed critique of this conclusion.

(b) Carefully describe any shortcomings evident in model (1). For any shortcoming, describe the steps needed to overcome the shortcoming.

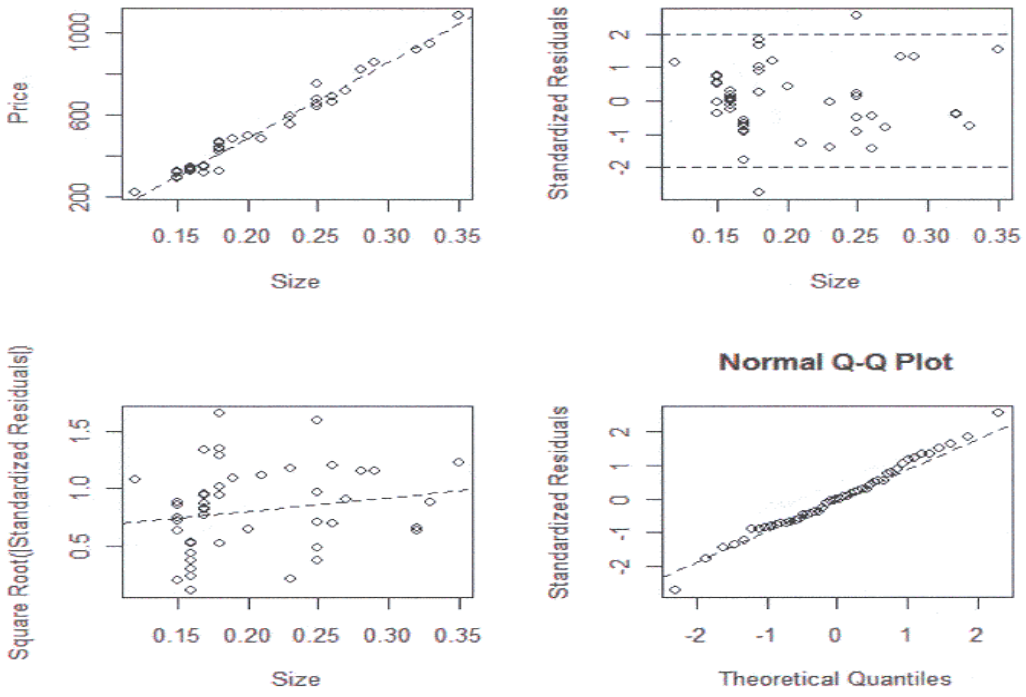
(c) Is model (2) an improvement over model (1) in terms of predicting Price? If so, please describe all the ways in which it is an improvement.

(d) Interpret the estimated coefficient of log(Size) in model (2).

(e) List any weaknesses apparent in model (2).

Output from R for model (1)

```
Call:

lm(formula = Price ~ Size)


Coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -258.05      16.94  -15.23   <2e-16 ***
Size         3715.02      80.41   46.20   <2e-16 ***

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 31.6 on 47 degrees of freedom

Multiple R-squared: 0.9785,     Adjusted R-squared: 0.978

F-statistic:  2135 on 1 and 47 DF,  p-value: < 2.2e-16
```



Output from R for model (2)

```
Call:

lm(formula = log(Price) ~ log(Size))


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.56317    0.06221  137.65   <2e-16 ***
log(Size)    1.49566    0.03772   39.65   <2e-16 ***

---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


Residual standard error: 0.06796 on 47 degrees of freedom

Multiple R-squared: 0.971,      Adjusted R-squared: 0.9704

F-statistic:  1572 on 1 and 47 DF,  p-value: < 2.2e-16
```