MASTER'S DIAGNOSTIC EXAMINATION - JANUARY 7, 2014

Student's Name ____

INSTRUCTIONS FOR STUDENTS:

1. The exam is to be started at 1 pm (CDT) and completed by 5 pm (CDT) on January 7, 2014.
2. Put your name above but DO NOT put your NAME on the **SOLUTIONS** to the exam.
3. Place the NUMBER assigned to you on the

   UPPER RIGHT HAND CORNER of EACH PAGE of your SOLUTIONS.
4. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
5. Use only one side of each sheet of paper.
6. You must answer all four questions: Questions I, II, III and IV.
7. 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.
8. Be sure to hand in/send all of your pages to the solutions for the exam questions. No additional material will be accepted once the exam has ended and you have left the exam room or sent your solutions.
9. You may use the following:
   - Calculator which does not have capability to phone, text, or access the Web
   - Pencil or pen
   - Blank paper for the solutions for this examination
   - No other materials are allowed

- 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** cover page and solutions to **979-845-6060** or

Scan cover page and solutions into a **single** pdf file and **email to longneck@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 **longneck@stat.tamu.edu**.

Proctor's Signature____

QUESTION I. There are two parts to this Question.

Question I - Part A.

For the experiment described below, 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 with just the following information: Sources of Variation and Degrees of Freedom Freedom

Description of the Experiment:

An evaluation of the effectiveness of three weed treatments on the yield of wheat raised in the midwest of the U.S. was conducted. There were eight fields used in the study with each field containing three widely separated tracts of land. The eight fields are randomly assigned to the two varieties of wheat, V1, or V2, with four fields randomly assigned to each variety. Within each field, the three tracts are randomly assigned to the three weed treatments, W1, W2, or W3, with one weed treatment per tract. Finally, each tract is divided in half, with one half randomly assigned the L amount of weed treatment and the other half the H amount. At the end of the growing season, the total yield of wheat for each half of the twenty four tracts are recorded. The yields are given in the following table.

		WEED TREATMENT					
		W1		W2		W3	
FIELD	VARIETY	L	H	L	H	L	H
F1	V1	83.2	81.8	67.4	79.7	75.9	80.6
F2	V2	77.5	78.2	69.2	71.5	75.9	78.2
F3	V1	72.7	69.3	70.1	71.2	75.9	81.3
F4	V2	75.3	78.9	72.7	74.6	75.9	82.8
F5	V1	78.2	80.5	65.1	68.3	65.3	66.6
F6	V2	79.8	85.2	57.6	61.4	58.5	61.6
F7	V1	82.4	83.1	50.5	54.0	51.6	54.7
F8	V2	75.5	78.7	39.0	43.9	41.9	45.1

Question I - Part B.

For each of the following questions, select **ONE** letter from the list on the next page which is the **BEST** solution to each of the following situations. Provide justification for your selection.

SITUATION:

1. A CRD was conducted with Factor F_1 having four fixed quantitative levels and Factor F_2 with six randomly selected levels. The AOV table reveals that the interaction between F_1 and F_2 was significant. The researcher wanted to investigate the change in the mean of the response with increasing levels of factor F_1.

2. An experiment was designed to compare four techniques, the levels of F_1, for removing mercury contamination from drinking water. The researcher wanted to also evaluate the variability in the many devices, the levels of F_2, of measuring mercury levels in water. Five devices for detecting mercury were randomly selected from the list of all such devices. A specified amount of mercury was placed in 200 water samples. Ten of the 200 water samples were randomly assigned to each of the twenty combinations of a level of F_1 and a level of F_2. There was significant evidence of an interaction between factors F_1 and F_2. The researcher wants to determine which of the four techniques removed the greatest amount of mercury.

3. A three factor experiment is run with Factor F_1 having five fixed levels, Factor F_2 with six fixed selected levels and Factor F_3 with four fixed levels. The results from the AOV were

   F_2, $F_1 * F_2$, $F_1 * F_3$, $F_1 * F_2 * F_3$ were nonsignificant.

   F_1, F_3, and $F_2 * F_3$, were significant.

   The statistician wants to evaluate the pairwise differences in the levels of factor F_1.

4. An experiment is conducted using a factorial treatment structure with factor F_1 having values $40°C$, $50°C$, $60°C$, $70°C$ crossed with factor F_2 having levels A, B, C in a CRD with three reps per treatment. There is not significant evidence of an interaction between F_1 and F_2. The researcher wants to determine the temperature that yields the maximum mean response.

5. In an experiment having the levels of factor F_1-qualitative and the levels of factor F_2-quantitative, there was significant evidence of an interaction between F_1 and F_2. The experimenter wants to compare the mean responses across the levels of factor F_1, averaged over the levels of factor F_2.

6. An experiment was designed to compare the performance of three new types of machine tools to the machine tool currently in use, factor F_1, with four levels. A random sample of five machinists, factor F_2, were randomly selected from the workforce. Each machinists produced ten units of product from each of the four types of machines. A quality rating was determined for each of the 200 units produced in the study. There was significant evidence of an interaction between factors F_1 and F_2. The company wants to know if any of the new types of machines have a higher mean quality rating than the type of machine the company is currently using.

TECHNIQUE:

A. Trend analysis using Scheffe contrasts

B. Trend analysis using Bonferroni contrasts

C. Trend analysis in the levels of F_1 averaged over levels of the other factors

D. Trend analysis in the levels of F_1 separately at each level of the other factors

E. Scheffe's test for contrast differences

F. Dunnett's comparison technique

G. Dunnett's comparison technique to all combinations of the factors

H. Dunnett's comparison technique applied to the levels of factor F_1 separately at each level of the other factors

I. Dunnett's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors

J. Tukey's comparison technique

K. Tukey's comparison technique to all combinations of the factors

L. Tukey's comparison technique applied to the levels of factor F_1 separately at each level of the other factors

M. Tukey's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors

N. Hsu's comparison technique

O. Hsu's comparison technique applied to the levels of factor F_1 separately at each level of the other factors

P. Hsu's comparison technique applied to the levels of factor F_1 averaged over the levels of the other factors

Q. Hsu's comparison technique applied to all combinations of the factors

R. Nothing new is learned beyond the results of the F-tests from the AOV table.

S. Comparison of marginal means is not appropriate.

T. None of the above methods are appropriate.

QUESTION II.

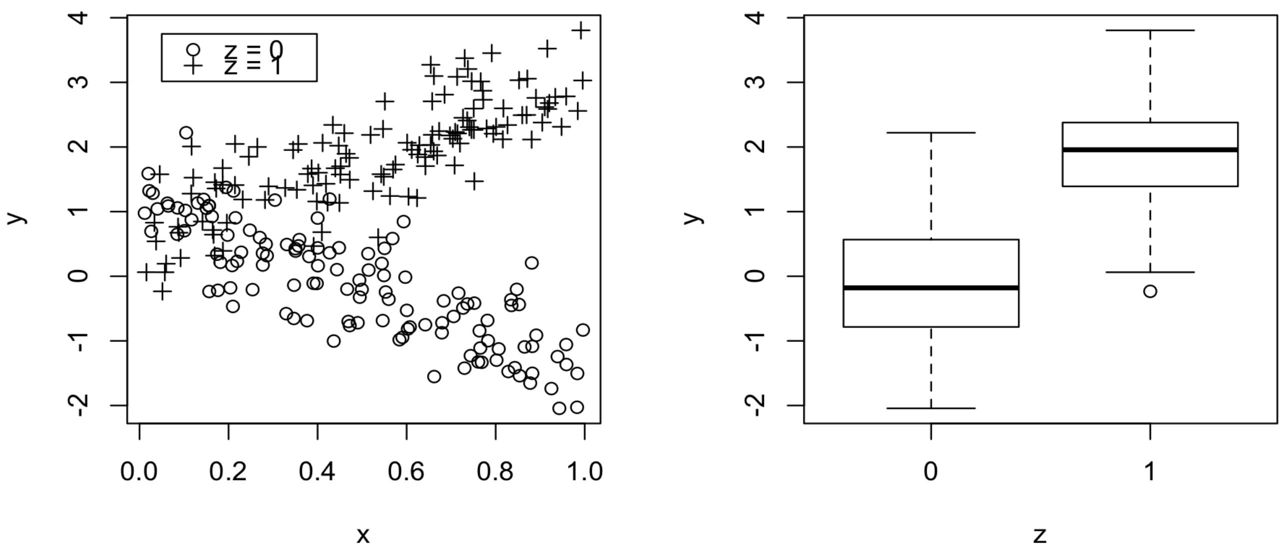
Consider the following linear model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \times z_i) + \epsilon_i$$

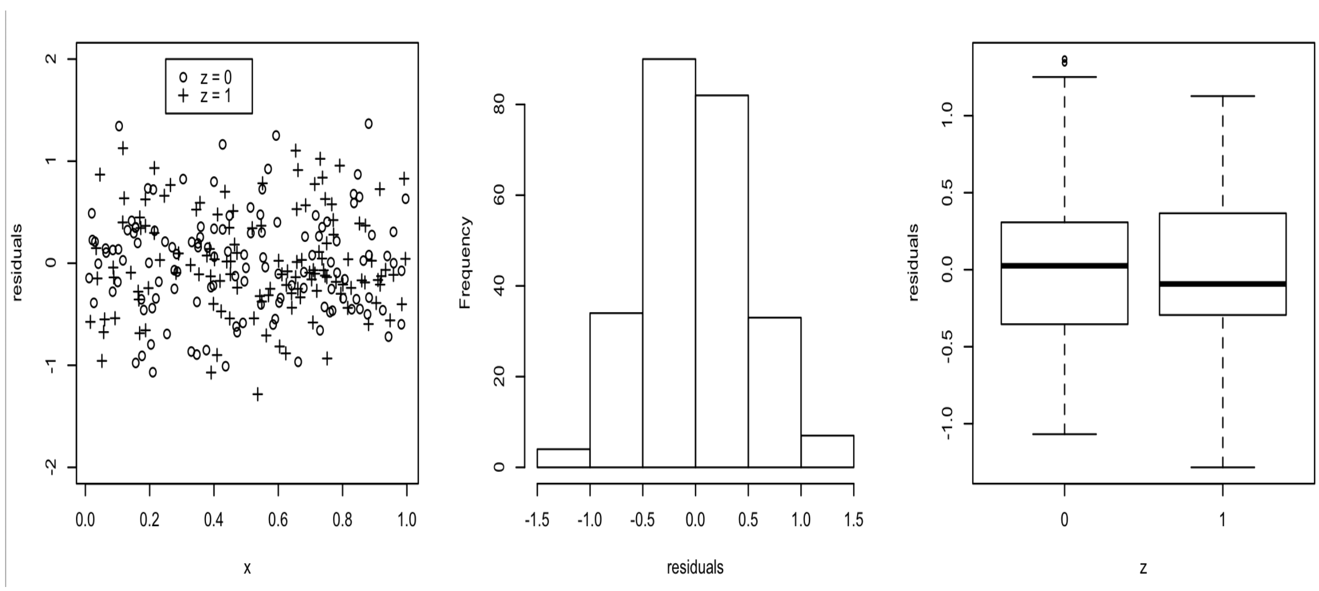
where x is continuous and z is binary. The output from fitting this model to a sample of size $n = 250$ is shown below:

```
Call:
lm(formula = y ~ x + z + x * z)

Residuals:
     Min       1Q   Median       3Q      Max
-1.28335 -0.34073 -0.04031  0.33622  1.36754

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15235    0.08933  12.900  < 2e-16 ***
x           -2.62637    0.16003 -16.412  < 2e-16 ***
z           -0.55213    0.13378  -4.127 5.03e-05 ***
x:z          5.02318    0.23094  21.751  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.5035 on 246 degrees of freedom
Multiple R-squared: 0.8554,Adjusted R-squared: 0.8536
F-statistic: 485.1 on 3 and 246 DF,  p-value: < 2.2e-16
```

1. Interpret each of the model coefficients (β_0, β_1, β_2, β_3) in terms of expected values.
2. Report a 95% confidence interval for the mean change in y associated with a one-unit increase in x, when $z = 0$.
3. What is the slope parameter (mean change in y associated with a one-unit increase in x) for x when $z = 1$?
4. What does the adjusted R squared measure?
5. What null and alternative hypotheses are tested using the F statistic at the bottom of the `R` output?
6. The figures below show diagnostic plots for the above model. Which of your model assumptions, if any, appear to not be met? And why?
7. One of the assumptions of our model is that the ϵ_i are *i.i.d.* realizations from the Normal distribution with mean 0 and constant variance σ^2. What does the model report as an estimate of σ?



Scatterplot of y vs. x, and side-by-side boxplots comparing y to z.



Residual plots: scatterplot of residuals vs. x; histogram of residuals; and side-by-side boxplots of residuals vs. z

QUESTION III.

Question III - Part A.

Let $X_1, \ldots, X_{20}$ be a random sample from the normal distribution with mean 20 and standard deviation 5. and $Y_1, \ldots, Y_{25}$ be a random sample from the normal distribution with mean 24 and standard deviation 4. Assume that $X_1, \ldots, X_{20}, Y_1, \ldots, Y_{25}$ are mutually independent.

1. Identify completely the distribution of $\bar{X} = \sum_{i=1}^{20} X_i/20$ and the distribution of $\bar{Y} = \sum_{j=1}^{25} Y_i/25$.

2. Identify the distribution of $W = \bar{X} - \bar{Y}$ and obtain an expression for $P(\bar{X} < \bar{Y})$ in terms of the standard normal cumulative distribution function, Φ.

3. Let $U = (X_1 - \bar{X})^2 + \cdots + (X_{20} - \bar{X})^2$.

   Derive an expression for $P(U > 50)$ in terms of the cumulative distribution function of a chi-squared distribution.

4. For what values of K and m is it true that the quantity

$$T = \frac{K(\bar{X} - 20)}{\sqrt{U}}$$

   has a t distribution with m degrees of freedom.

Question III - Part B.

Let $X \sim N(-2, 6)$, $Y \sim N(3, 4)$, and $Z \sim N(0, 1)$ be independent normal random variables. (Note: The notation $N(a, b)$ indicates a normal distribution with mean a and variance b.)

1. Let $U = 2X + 3Y + Z - 5$ and $V = X - 2Y - Z$.

   Identify the distributions of U and V.

2. Let $W = C_1(X + C_2)^2 + C_3(Y + C_4)^2$.

   Find values of C_1, C_2, C_3, C_4, and C_5 (with $C_1 \neq 0$ and $C_3 \neq 0$) so that W has a chi-squared distribution with C_5 degrees of freedom.

3. Let

$$V = \frac{C_1(X + C_2)^{C_3}}{(Y + C_4)^{C_5}}.$$

   Find values of C_1, C_2, C_3, C_4, C_5, C_6 and C_7 so that V has an F distribution with (C_6, C_7) degrees of freedom.

QUESTION IV.

Consider the usual linear regression model, written either in non-matrix notation as:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + e_i, \quad i = 1, 2, \ldots, n, \tag{A}$$

where $e_1, e_2, \cdots, e_n$ are independently and identically distributed as $N(0,\ \sigma^2)$ random variables or, in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{B}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times p$ matrix of predictor variables (with $p = k + 1$), $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\mathbf{e}$ is an $n \times 1$ vector of unobservable independent and identically distributed random $N(0, \sigma^2)$ variables. In what follows, you may assume that the matrix $\mathbf{X}$ is of full column rank. Using whichever notation above (A or B) that makes you more comfortable, answer the following parts to this problem. Please be concise with your answers - highly irrelevant statements may be counted against you!

1. The above model is called a **linear** regression model even though, for example, it encompasses polynomial (in the predictor variables) regression models. Explain what is **linear** about the above model.

2. Define the **least squares criterion**. That is, state what property must be satisfied for estimates of the unknown parameters of the above model to be called **least squares estimates**. Use formulas as part of your definition.

3. Specify which of the above assumptions made about the e_i's, $i = 1, 2, \cdots, n$, need **not** be true for the least squares estimators of the β_j's, $j = 0, 1, \cdots, k$, to be unbiased estimators. If all the above assumptions about the e_i's need to be true, then state so.

4. Specify which of the above assumptions made about the e_i's, $i = 1, 2, \cdots, n$, need **not** be true for the least squares estimator of σ^2 to be an unbiased estimator. If all the above assumptions about the e_i's need to be true, then state so.

5. Specify which of the above assumptions made about the e_i's, $i = 1, 2, \cdots, n$, need **not** be true for the usual least squares t tests and F tests of hypotheses about the β_j's, $j = 0, 1, \cdots, k$, to be statistically valid. If all the above assumptions about the e_i's need to be true, then state so.