METHODS QUALIFYING EXAM

January 2010

Student's Name _____

INSTRUCTIONS FOR STUDENTS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER RIGHT 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 **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_____

Problem I. Part A:

For the following experiment, 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

An industrial engineer is studying the hand insertion of electronic components on printed circuit boards in order to improve the speed of the assembly operation. She has designed three assembly fixtures (F_1,F_2,F_3) and two workplace layouts (L_1,L_2) that seem promising. Specialized operators are required to perform the assembly and it was initially decided to randomly select four operators from the many qualified operators at the plant. However, because the workplaces are in different locations within the plant, it is difficult to use the same operators for each layout. Therefore, the four operators randomly chosen for layout 1 are different individuals from the four operators randomly chosen for layout 2. Each of the operators assemblies four circuit boards for each of the three fixture types with the 12 circuit boards assembled in random order. The 96 assembly times are measured in seconds. The engineer is interested in the effects of Assembly Fixtures (F), Workplace Layout (L), and Operator (O) on the average time required to assemble the circuit boards.

Problem 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. **Place** your selection in the space to the **left** of each situation.

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 effects $F_1 * F_2$, $F_1 * F_3$, and $F_1 * F_2 * F_3$ were all found to be nonsignificant. 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. In 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

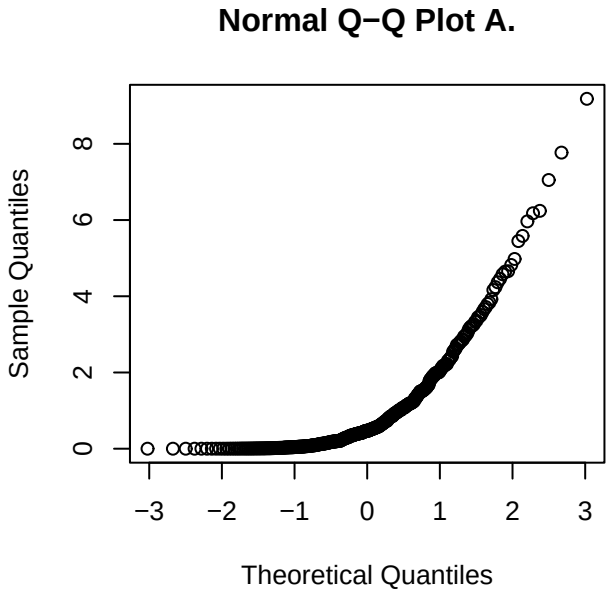
R. Nothing new is learned beyond the results of the F-tests from the AOV table.

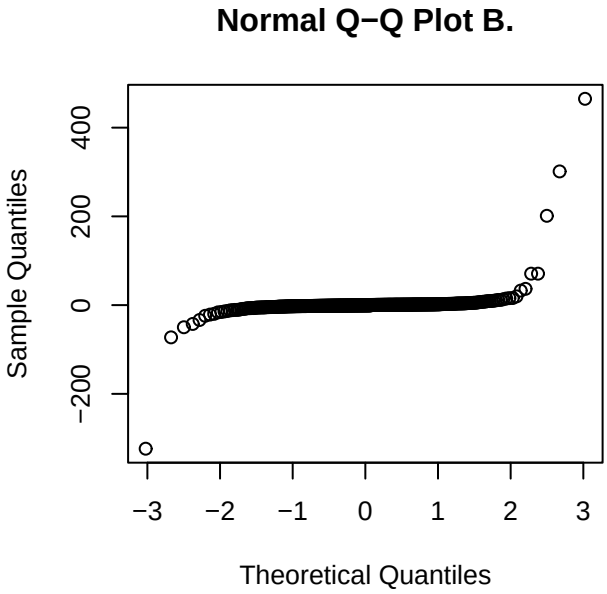
S. Comparison of marginal means is not appropriate.

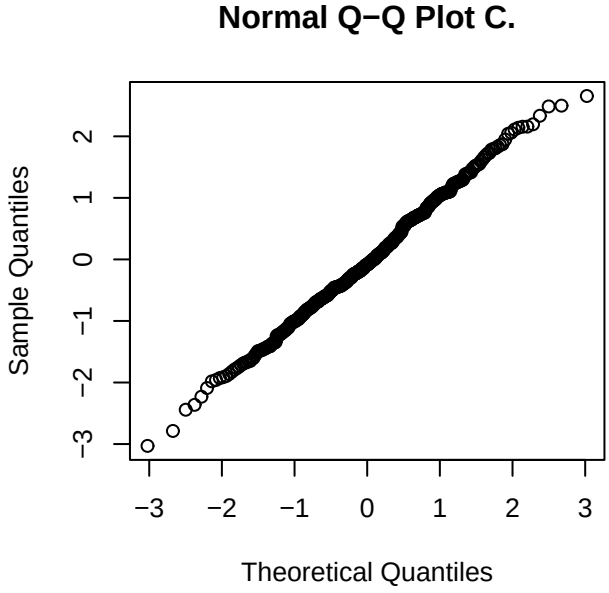
T. None of the above methods are appropriate.

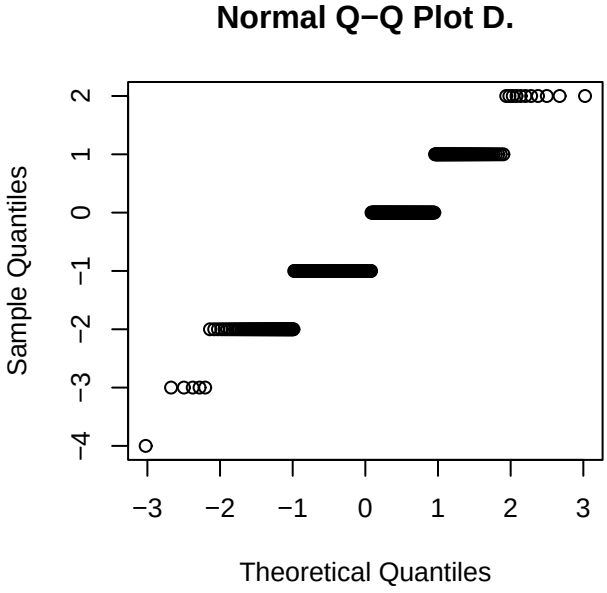
Problem II.

Normal Reference Distribution (QQ) plots are often used to assess distributional assumptions. The following normal quantile plots were produced in R using the command `qqnorm`. For each of the following four plots, discuss the assumption of normality for the pictured data. If the data are nonnormal, describe the manner in which the data are nonnormal and a transformation (if posssible) to make the data more normally distributed.









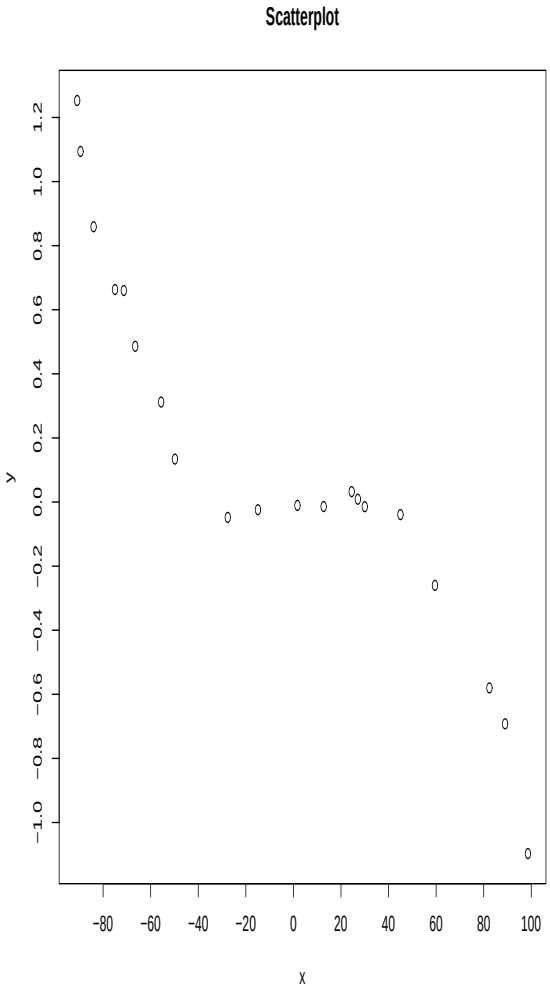
Problem III.

An engineer wants to evaluate the relationship between the standardized amount of additive, x, used in a chemical reaction and the deviation from the standard yield of a chemical reaction, y. Given the data in Table given below and the scatterplot of the data, answer the following questions. Please explain your answers. Do not make any calculations.

1. Does a reasonable model for this data satisfy the requirement for multiple linear regression?
2. If the model $y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$ is fit to the data, how many distinct populations are involved in the modelling?
3. If you use the model in part 2. above and you test for equality of variance, what is the null hypothesis in terms of $H_o : \ldots$.?
4. If you use Anderson-Darling statistic or Shapiro-Wilks statistic to test H_o : y has a normal distribution and reject that null hypothesis, would this conclusion violate the assumptions for multiple linear regression?

Data Table

ID	x	y
1	-90.9	1.25300
2	-89.5	1.09400
3	-84.0	0.85900
4	-75.0	0.66300
5	-71.3	0.66000
6	-66.5	0.48600
7	-55.6	0.31200
8	-49.8	0.13400
9	-27.6	-0.04800
10	-14.9	-0.02400
11	1.7	-0.01050
12	12.7	-0.01400
13	24.5	0.03240
14	27.1	0.00871
15	30.0	-0.01460
16	45.0	-0.03930
17	59.5	-0.26000
18	82.4	-0.58000
19	89.0	-0.69200
20	98.6	-1.09700



Problem IV.

Consider the usual multiple linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{Y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is an $n \times p$ matrix of predictor variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of unobservable independent and identically distributed random variables, each with mean zero and variance σ^2. In what follows, you **may not** assume that the matrix $\mathbf{X}$ is of full column rank.

Four results are given below about least squares estimation of $\boldsymbol{\beta}$ for this multiple linear regression model. You are to prove any **three of the four results** that you choose. Clearly indicate **which three** results are to be graded.

Result 1:

> The normal equations, $\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$, are consistent, that is, they always have at least one solution.

Result 2:

> Any solution to the normal equations satisfies the least squares criterion, that is, if $\widehat{\boldsymbol{\beta}}$ satisfies:
>
> $$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y}$$
>
> then the minimum value $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ can attain is $(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})$.

Result 3:

> The least squares predicted values, $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}}$ is any solution to the normal equations, are unique.

Result 4:

> If the model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ contains an intercept parameter, the mean (that is, arithmetic average) of the least squares residuals that result from the fitting of this model is zero.

Problem V.

A book[1] on robust biostatistical methods published in 2009 considers a data set taken from Everitt (1994), *The Handbook of Statistical Analysis Using Splus*, Chapman & Hall. The data consist of the IQ scores (IQ) and behavioral problem scores (BP.score) of children of age five, labeled according to whether or not their mothers had suffered an episode of postnatal depression (state.mother = 1 if yes and 0 if no). We seek to model IQ as a function of BP.score and to determine whether the effect of BP.score differs significantly across the two groups of mothers.

The two models under consideration are as follows:

$$(1) \quad \text{IQ} = \beta_o + \beta_1 \text{BP.score} + \beta_2 \text{state.mother} + e$$

and

$$(2) \quad \text{IQ} = \beta_o + \beta_1 \text{BP.score} + \beta_2 \text{state.mother} + \beta_3 \text{BP.score} \times \text{state.mother} + e$$

A plot of the data and the two regression lines from model (2) along with a plot of the standardized residuals and leverage values from model (2) can be found below. In addition, numerical output from R for models (1) and (2) appears below.

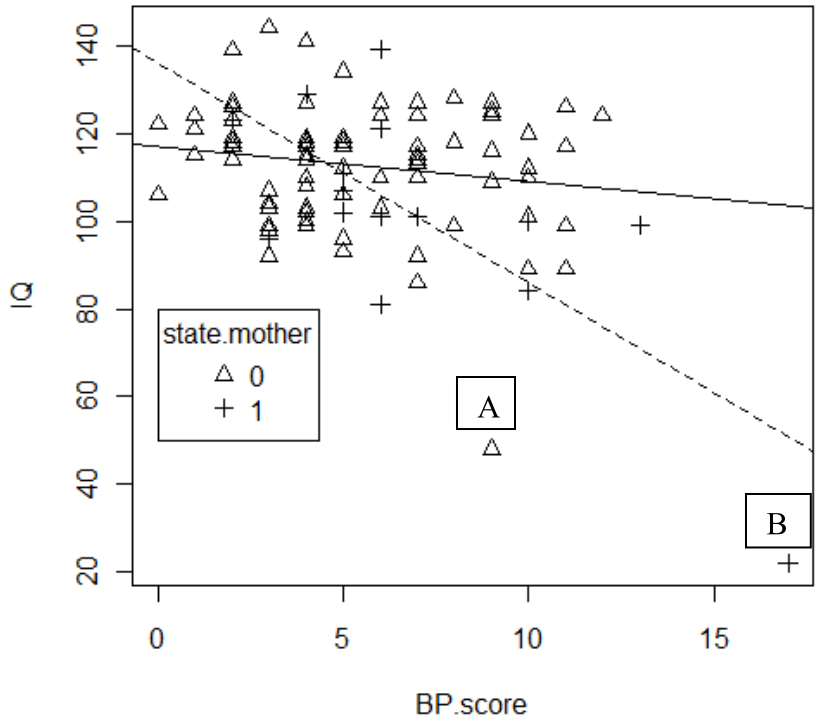
a) Two points are marked as "A" and "B" in the plot of IQ and BP.score. Two points are marked as "C" and "D" in the plot of the standardized residuals and leverage values from model (2). Match "A" and "B" with "C" and "D". Give reasons to support your choices.

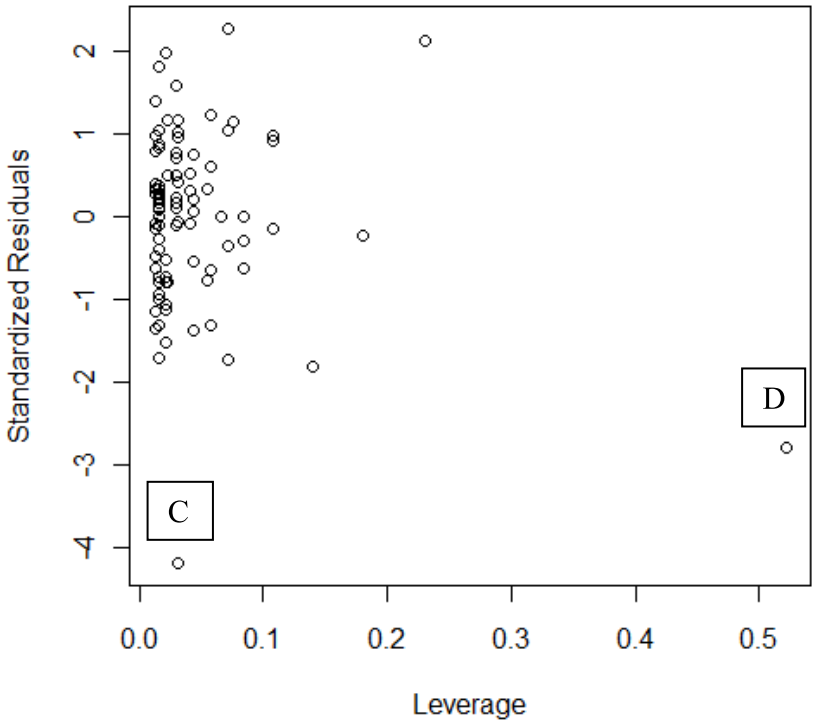
b) Using the output from R provided below, calculate the value of the F-statistic for testing the null hypothesis, $H_o : \beta_3 = 0$.

c) Briefly describe the steps you would follow in order to obtain a final model for the data on IQ and BP.score , labelled according to whether or not their mothers had suffered an episode of postnatal depression.

[1] Heritier, S., E. Cantoni, S. Copt, & M.-P. Victoria-Feser (2009) *Robust Methods in Biostatistics.* Wiley, New York

Plots associated with model (2)





Output from R for model (1)

```
Call:
lm(formula = IQ ~ BP.score + state.mother)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.5832     3.3674  36.403  < 2e-16 ***
BP.score      -1.8171     0.5281  -3.441 0.000878 ***
state.mother  -8.7970     4.5782  -1.922 0.057797 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.9731 on 91 degrees of freedom
Multiple R-squared: 0.1699,     Adjusted R-squared: 0.1516
F-statistic: 9.312 on 2 and 91 DF,  p-value: 0.0002093
```

Edited Output from R for model (2)

```
Call:
lm(formula = IQ ~ BP.score + state.mother + BP.score:state.mother)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           117.1334     3.5034  33.434  < 2e-16 ***
BP.score               -0.8064     0.5693  -1.417 0.160063
state.mother           18.9970     8.8000   2.159 0.033531 *
BP.score:state.mother  -4.2027     ??????   ????? 0.000486 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.0061 on 90 degrees of freedom
Multiple R-squared: 0.2754,     Adjusted R-squared: 0.2513
F-statistic:  11.4 on 3 and 90 DF,  p-value: 2.084e-06
```