METHODS QUALIFYING EXAM

January 2008

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.

Weight gain in the first 3 months after birth is important for new born infants. A pediatrician wishes to test a new feeding formula to determine if it will cause greater weight gain in new born infants than the standard formula.
From her records she finds that the first 3 months weight gains of single birth infants on the standard formula have the following characteristics:

$$\mu_S = 15oz. \quad \text{and} \quad \sigma_S = 6oz.$$

On the other hand, the first 3 months individual weight gains of identical twins on the standard formula have the following characteristics:

$$\mu_T = 12oz., \quad \sigma_T = 6 \text{ oz. , with}$$

$$\rho = .8 \text{ (correlation in individual weight gains of identical twins)}$$

She wants to run a 3 month experiment on a group of infants to test the new formula versus the standard formula. She has decided to use a 5% probability of Type I error, and wishes to be able to detect a 3 oz. increase in weight gain with 90% probability.

(a) Suppose the researcher conducts the experiment as a Completely Randomized Design of single birth infants with two treatments, standard and new. What is the required sample size? State any assumptions you are making in this calculation and show a detailed justification for your sample size.

(b) Suppose the researcher conducts the experiment as a Paired Design with two treatments, standard and new, randomly assigned within each set of twins. What is the required sample size? State any assumptions you are making in this calculation and show a detailed justification for your sample size.

(c) Discuss the relative merits of the two experiments in terms of practicality and the researcher's basic goal for the experiment.

PROBLEM II.

A tire manufacturer wants to study the relationship between tread density and traction. Four different tread densities are considered and for each density there are three tires tested. The data are summarized below. Let (Y_{ji} be the traction measured for the i'th tire having tread density x_j.)

x_j (treads/inch)	$\bar{Y}_j = \sum_{i=1}^{3} Y_{ji}/3$	$\sum_{i=1}^{3}(Y_{ji} - \bar{Y}_j)^2$
0.0	2.0	0.38
1.0	4.0	23.6
2.0	3.8	0.14
3.0	3.0	13.38

A simple linear regression relating traction to tread density was fit yielding $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with $\hat{\beta}_0 = 2.78$ and $\hat{\beta}_1 = 0.28$.

(a) Could these least-squares estimates be obtained using *only* the information in the above table. Why or why not?

(b) Calculate the expected traction at each tread density assuming that the simple linear regression model is correct.

(c) Based on a plot of $(x_j, \bar{Y}_j)$ do you feel the simple linear regression is appropriate? Why or why not?

(d) Do the errors appear to be homoscedastic? (Yes or No). If not, what transformation should be applied so that homoscedasticity would be a better assumption. Give the transformed $\tilde{Y}$ and $\tilde{X}$ as functions of Y and X.

(e) Consider the model that uses $\bar{Y}_j$ to estimate mean traction for tread density x_j, $j = 1, 2, 3, 4$. Are these estimates unbiased if the simple linear regression model holds?

(f) Are the estimates in (e) unbiased if the simple linear regression model does not hold?

(g) How are the estimates in (e) typically inferior to those in (b) to estimate mean traction for tread density x_j when the simple linear regression model does hold? Explain briefly.

PROBLEM III.

A poultry researcher is growing small chickens in an experiment conducted in an enclosed chicken house. The researcher has 12 pens available and has 12 different sources of protein that she wants to evaluate. There are 20 chickens in each of the pens and they are **pen fed**, that is, the 20 chickens share a common feeding trough. Therefore, the 12 sources of protein are randomly assigned to the 12 pens with 1 source for each pen. Some of the response variables are measured on a **pen** basis, for example, **feed conversion**, the amount of feed needed for an increase of 1 kg in body weight. Other response variables are measured on the individual chicken, for example, **average daily weight gain** (ADWF) and **percent body fat** (PDF).

1. If this experiment is not repeated elsewhere (either in time or space) is it a valid experiment? Why or why not?
   (a) If valid, explain how you would analyze the data.
   (b) If invalid, explain why it is invalid.
2. If the experiment was repeated in time, that is, a similar experiment was conducted 3 months later, explain how you would analyze the data. Include in your explanation, models, anova table, expected mean squares, testing procedures, and any other pertinent information.

PROBLEM IV.

Part (A)

Consider the simple linear regression model: $Y = \beta_o + \beta_1 x + e$. Analysis of Variance can be used to test

$$H_o : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0.$$

Analysis of Variance is based on the following three statistics:

The Total Corrected Sum of Squares of the Ys:

$$SST = SYY = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

The Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

The Regression Sum of Squares:

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

where $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 \bar{x}$, $\hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SXY}{SXX}$$

In this questions, we will show that $SST = SS_{reg} + RSS$. To do this we will show that $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$.

(a) Show that $(y_i - \hat{y}_i) = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})$.

(b) Show that $(\hat{y}_i - \bar{y}) = \hat{\beta}_1 (x_i - \bar{x})$.

(c) Utilizing the fact that $\hat{\beta}_1 = \frac{SXY}{SXX}$, show that $\sum_{i=1}^{n} (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$.

Part (B)

This question deals with a generated multiple regression data set with outcome variable Y and predictors x_1 and x_2. There are a total of $n = 601$ cases. One aim is to develop a valid model for Y based on x_1 and x_2. The first model fit to the data was

$$Y_i = \beta_o + \beta_1 x_{1i} + x_{2i} + e_i \qquad (1)$$

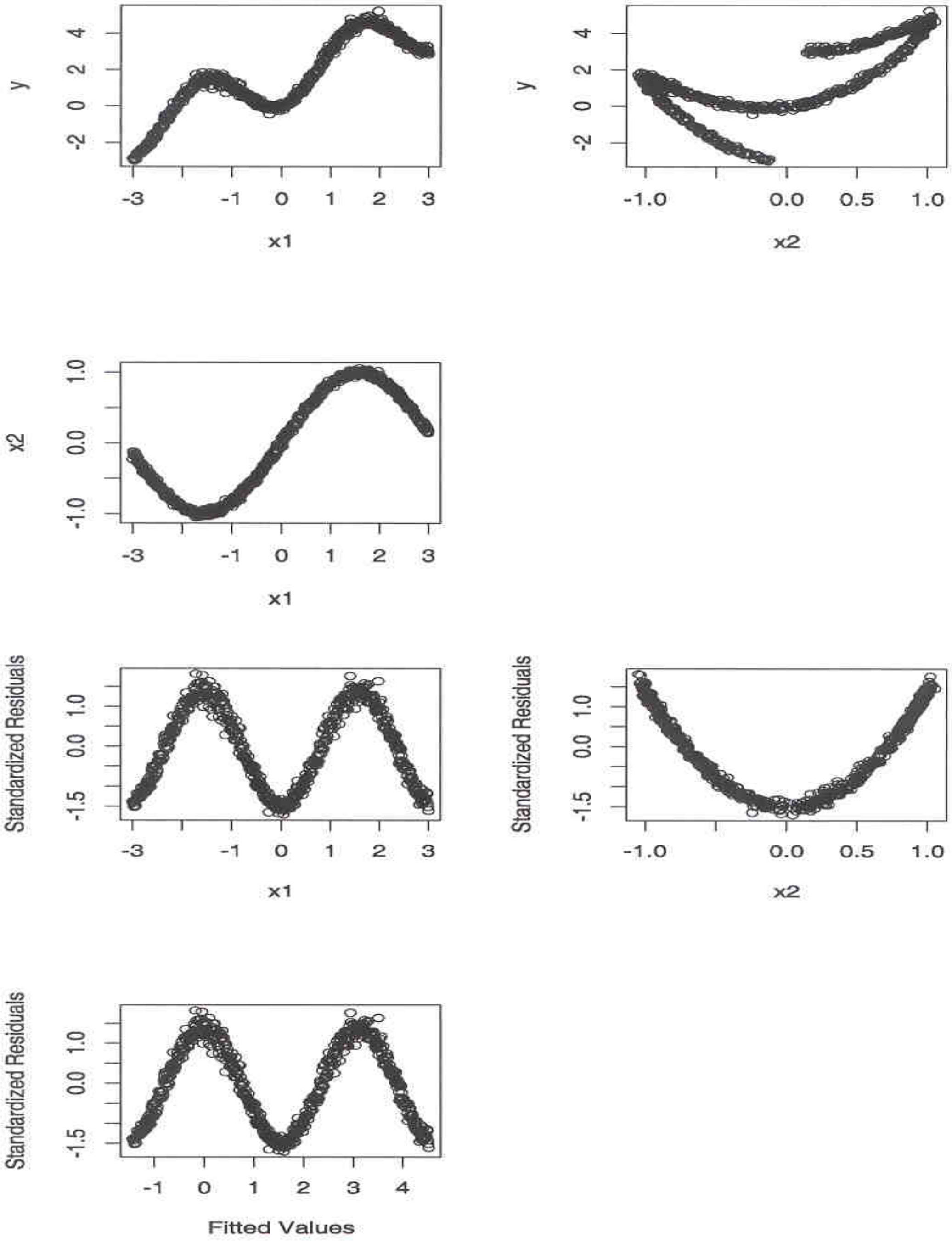
Plots associated with Model (1) appear on the following page.

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

b. Decide whether the plots of standardized residuals provide any direct information on how model (1) is misspecified. Give a reason to support your answer.

c. Describe what steps you would take to obtain a valid regression model.

Output from model (1)


y
x1
y
x2
x2
x1
Standardized Residuals
x1
Standardized Residuals
x2
Standardized Residuals
Fitted Values