METHODS QUALIFYING EXAM

August 2007

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 1

A biologist conducted an experiment to investigate the impact of a photosynthesis-inhibiting herbicide on the plankton level in ponds. Water was obtained from three ponds. From each of the three ponds, four containers holding eight (8) gallons of well-mixed pond water were obtained. The four containers were randomly assigned to be dosed with one of the following rates of herbicide: 0, 0.1, 0.5, 1.0 mg/liter. The large containers were divided into eight 1-gallon glass jugs filled with the well-mixed, dosed pond water, sealed and suspended in the pond just below the water surface. Bottles were given labels so that two bottles of each dose could be removed at the end of the first day, at the end of the first week, the second week, and finally the third week. Rotifers are a major element of the plankton food change in this pond. The number of rotifers present in the bottle was counted immediately after the bottle was removed from the pond. The counts in numbers of rotifers/liter are given.

		Pond 1				Pond 2				Pond 3			
		Week				Week				Week			
Dose	Bottle	0	1	2	3	0	1	2	3	0	1	2	3
0	1	6718	5166	3815	2340	6222	4656	3351	1804	7081	5466	4151	2604
	2	6392	5039	3382	2881	5891	4527	2828	2318	6629	5393	3628	3118
0.1	1	5832	5233	4373	2453	5323	4722	3837	1935	6123	5533	4637	2735
	2	5569	4350	3333	2924	5071	3849	2833	2442	5896	4605	3633	3242
0.5	1	5924	3953	4912	2575	5423	3435	4421	2057	6242	4235	5221	2857
	2	6715	4295	4427	4211	6205	3759	3972	3711	7051	4559	4772	4511
1.0	1	6821	4849	4900	4451	6316	4394	4400	3915	7112	5194	5200	4715
	2	5871	4763	4067	4552	5368	4236	3576	4025	6117	5036	4376	4825

Answer each of the following questions.

a. Type of Randomization, for example, CRD, RCBD, LSD, BIBD, SPLIT-PLOT, SUBSAMPLING, etc.

b. Type of Treatment Structure, for example, Single Factor, Crossed, Nested, etc.

c. Identify each of the factors as being Fixed or Random.

d. Describe the Experimental Units and/or Measurement Units.

e. Provide a partial AOV Table with Source of Variation, Degrees of Freedom, and Expected Mean Squares.

f. Let y_{ijkl} be the count from the l^{th} bottle taken on the k^{th} week from the i^{th} dose from the j^{th} pond. Express the AOV-MOM estimate of the variance of $(\bar{y}_{1\bullet\bullet\bullet} - \bar{y}_{2\bullet\bullet\bullet})$ in terms of the means squares from the AOV given in part e).

g. Describe how you would obtain an estimate of the degrees of freedom of your estimate in f).

h. The researcher wants to compare the mean counts for the four doses.

   a. Under what relationship between the factors dose and week would these comparisons be valid?

   b. Provide the form of Tukey's HSD for making these comparisons. Make sure to identify all the terms in your formula.

Problem 2

The rate (R) of a metabolic reaction in humans is thought to be related to body temperature (t) by the following equation:

$$R_i = \beta_0 + \beta_1 t_i + \varepsilon_i \qquad \text{(I)}\ ,$$

where β_0 and β_1 are parameters, t_i is the body temperature for the i^{th} individual observed, and ε_i are assumed to be independent $N(0, \sigma^2)$ random variables.

At the standard body temperature of 98.6 $^{\text{o}}$F, thousands of observations have been made. Therefore, the expected value of R at $t^* = 98.6$ $^{\text{o}}$F is assumed to be known as $R^* = 73.8$ moles per minute, and the variance of R is known to be (0.08 moles/minute)^2. We observe the rate of reaction in 14 unhealthy individuals (i.e., individuals with elevated or depressed body temperature.) The following data are reported:

Temperature	**Number of Individuals Observed**	**Mean Reaction Rate**
103.8	2	84.2
100.1	3	75.9
97.4	4	71.3
96.9	5	70.5

a. Show that it suffices to consider the model:

$$Y_i = \beta x_i + \varepsilon_i \qquad \text{(II)}$$

where $Y_i = R_i - R^*$, $x_i = t_i - t^*$, and $\beta = \beta_1$.

b. Calculate the least squares estimate of β.

c. Test the hypothesis that $\beta = 0$ against the two sided alternative at the .01 level. Give the test statistic and the rejection rule (in terms of a commonly tabulated distribution.)

d. Test model (II) for lack of fit to the data at the .01 level. List the test statistic and the rejection rule (in terms of a commonly tabulated distribution.)

e. Can it occur that we reject the null hypothesis in both parts c) and d)? Explain briefly.

f. Do you suspect any of the observations are high leverage points? Why or why not?

Problem 3

a. A statistics professor at Texas A&M University is involved in a collaborative research project with two entomologists. The statistics part of the project involves fitting regression models. Together they have written and submitted a manuscript to an entomology journal. The manuscript contains a number of scatter plots which each show an estimated regression line and associated individual 95% confidence intervals for the regression function at each *x* value, as well as the observed data. A referee has asked the following question:

> "I don't understand how 95% of the observations fall outside the 95% CI as depicted in the figures."

Prepare a response to this statement.

b. The Sunday April 15, 2007 issue of the *Houston Chronicle* included a section devoted to real estate prices in Houston. In particular, data are presented on the 2006 median price per square foot for 1922 subdivisions. Interest centers on developing a regression model to predict

Y_i = 2006 median price per square foot

from

x_{1i} = %NewHomes (i.e., of the houses that sold in 2006, the percentage that were built in 2005 or 2006)

x_{2i} = %Foreclosures (i.e., of the houses that sold in 2006, the percentage that were identified as foreclosures)

for the i = 1, … 1922 subdivisions.

The first model considered was

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \quad (1)$$

Model (1) was fit using weighted least squares with weights

$$w_i = n_i$$

where

n_i = the number of homes sold in subdivision i in 2006.

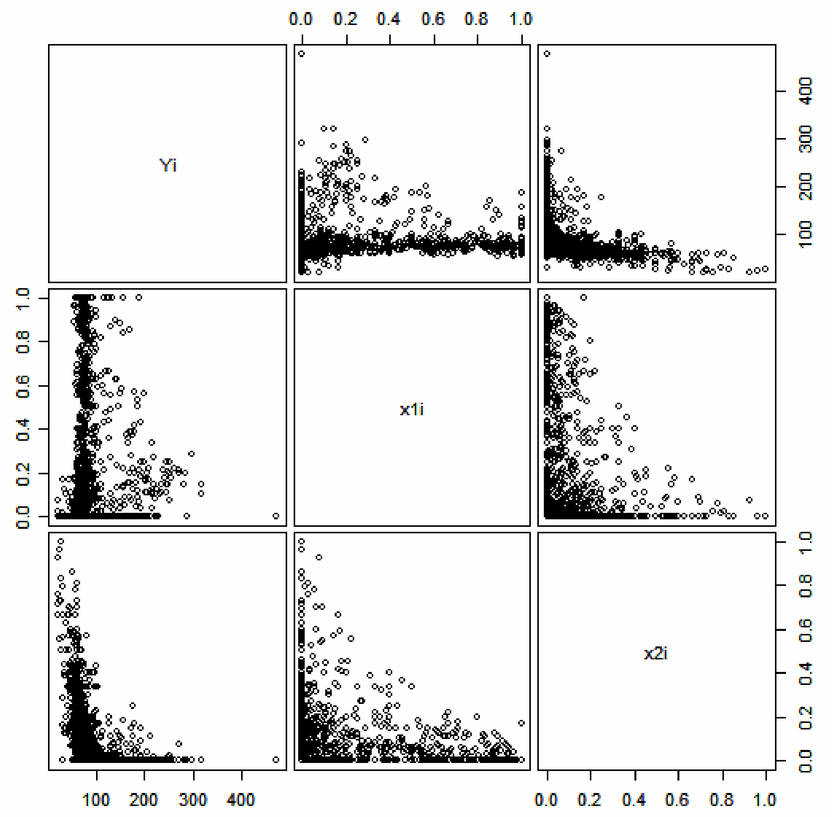
Output from model (1) appears on the following pages.

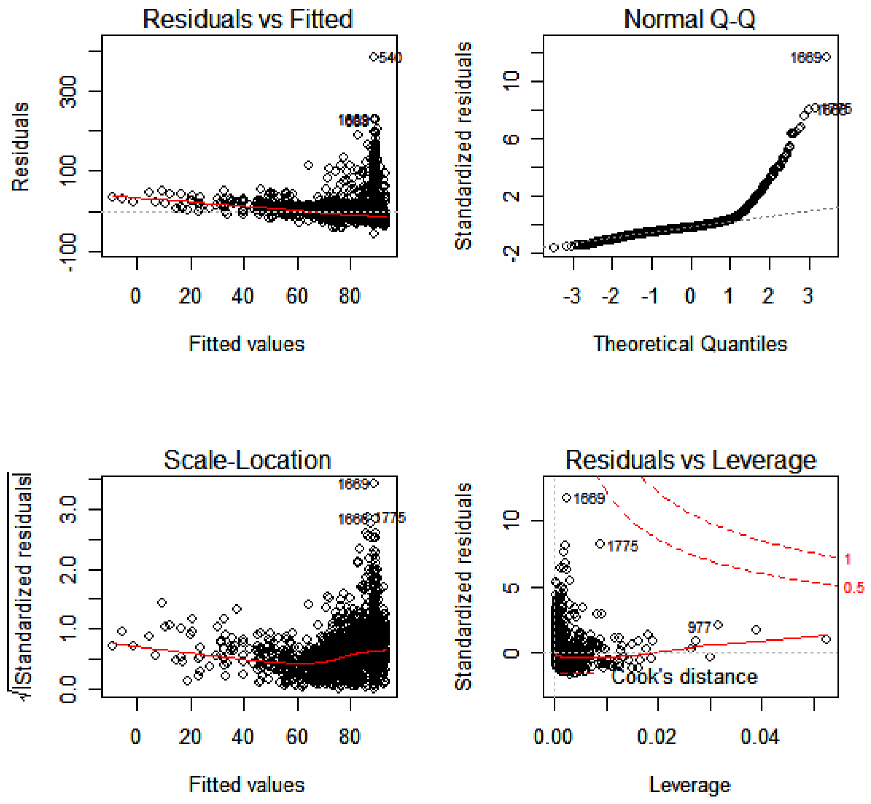
i. Explain why it is necessary to use weighted least squares to fit model (1) and why $w_i = n_i$ is the appropriate choice for the weights.

ii. Explain why (1) is not a valid regression model.

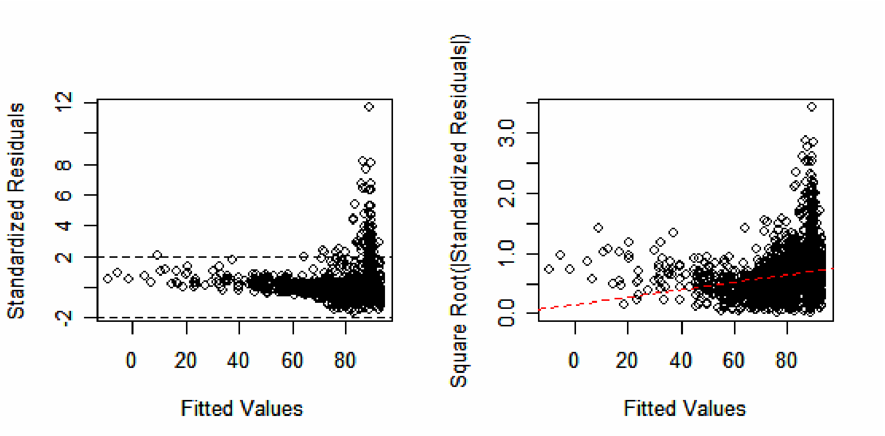
iii. Describe what steps you would take to obtain a valid regression model.

Output from model (1)





(continued)


Standardized Residuals
12
8
6
4
2
-2
0
20
40
60
80
Fitted Values
Square Root(|Standardized Residuals|)
3.0
2.0
1.0
0.0
0
20
40
60
80
Fitted Values

Problem 4

A running shoe company produces a new model of running shoe that includes a harder material for the insert that corrects for overpronation. Two studies were carried out to see if the proportion of runners with heel tenderness is affected by the new shoe. Researchers asked the runners whether they experienced occasional heel tenderness.

a. Two random samples of 87 runners were assigned the two types of running shoes. The first sample of runners was given ordinary running shoes and the second sample was given the shoes with the new insert. At the end of the study, each runner was asked whether he or she experienced occasional heel pain. Write out the hypotheses, formula for the test statistic, and the rejection region (in terms of a commonly tabulated distribution) to test whether the proportion of runners that experience occasional heel pain differs between the two types of shoes.

	Heel Pain	
Shoe Type	Yes	No
Ordinary	63	24
New	53	34

b. Now suppose that two random samples of eight (8) runners were assigned the two types of running shoes. The first sample of runners was given ordinary running shoes and the second sample was given the shoes with the new insert. At the end of the study, each runner was asked whether he or she experienced occasional heel pain. Discuss the difficulty in using the data in the following table to test whether the proportion of runners that experience occasional heel pain differs between the two types of shoes. What procedure should you use?

	Heel Pain	
Shoe Type	Yes	No
Ordinary	6	2
New	5	3

c. Now suppose a group of 87 runners uses both types of shoes. The runner indicated whether heel tenderness was experienced after using an ordinary shoe and the same runner was asked whether heel tenderness was experienced after using the new shoe. Explain why the method used in part a) cannot be used to test whether the proportion of runners that experience occasional heel pain differs between the two types of shoes.

Ordinary	**New Shoes**	
Shoes	Yes	No
Yes	48	15
No	5	19

d. Suppose we are back in the situation of part a). Define the variables $Y = 1$ if "yes" and $Y = 0$ if "no" and $x = 1$ if new type and $x = 0$ if ordinary type. Consider the logistic regression model

$$\text{logit}(\pi(x)) = \alpha + \beta x.$$

Explain how to test whether the proportion of runners experiencing heel pain differs using the logistic regression model. Write out the hypotheses, formula for the test statistic, and the rejection region (in terms of a commonly tabulated distribution.)