Probability and Statistical Theory

MS Comprehensive Examination

April 17, 2004

Instructions:

Please answer all four questions.
Point Values: 10, 30, 15, 15

Record your answers in your blue books.

Show all of your computations.
Prove all of your assertions or quote the appropriate theorems.
Books, notes, and calculators *may be used*.

You have three hours.

Comprehensive Examination April 17, 2004

1. Three generalized linear models were fit to data from a 3-way factorial experiment involving factors A, B, C with number of levels 2, 3, and 4, respectively. The models and their associated maximum log-likelihood values are given below. The sample size in the experiment was 50.

Model	Terms in model	Maximum loglikelihood
M1	intercept+A+B	-1122.0
M2	intercept+A+B+A*B	-1118.5
M3	intercept+A+B+A*B+B*C	-1122.0

(a). According to an analysis of deviance, which model is most appropriate for these data?

(b). According to Schwarz's Bayesian Criterion (BIC), which model is most appropriate for these data?

2. An experiment was set up in two greenhouses to compare the effect of different temperatures, soil pHs, and calcium additives on the increase in trunk diameters for orange trees. The experiment was conducted in two greenhouses with 12 orange trees in each greenhouse. At the end of two-year period, two diameters were examined at each factor-level combination. The factors of interest were

A: temperature (3 levels),

B: soil pH (2 levels),

C: calcium additives (2 levels).

(a). Suppose 24 treatments are randomly assigned to the two greenhouses. Name the design of this experiment.

Suppose there are some differences in these two greenhouses. Therefore, the researcher decided to randomly assign 12 different treatments in each of the greenhouses, and obtained the following data

		Calcium low		Calcium high	
Temp	Greenhouse	pH low	pH high	pH low	pH high
1	1	21	12	13	1
	2	21	18	14	8
2	1	23	14	13	1
	2	23	17	16	11
3	1	17	20	16	14
	2	23	17	17	5

Based on Appendix One, answer the question (b)-(h).

(b). Name the design of this experiment. Write down an appropriate model for the analysis of these data.

(c). Explain if it is a better design compared with the design in (a) in this setting.

(d). If the design in (a) is used, what is the possible impact on your conclusion?

(e). Fill out the following table based on the appropriate SAS output.

	Sum of Square	Mean Square	F value	p-value
A				
B				
C				
Error			—	—

(f). Which of the main factors appear to have significant effects on diameter based on the table in (c)? Is the conclusion on the main effects meaningful?

(g). What conclusion can be obtained based on Plot 1, residuals vs. fitted values?

(h). Which treatment leads to the largest diameter based on Plot 2 and Plot 3?

3. Say that we wish to do a survey of house prices in a certain city. Here is some information that we know:

1) there are three areas, inexpensive (area 1), moderate (area 2), and expensive (area 3).
2) These areas have the following approximate statistics:
   a) number of homes: 100, 200, 300
   b) average prices: \$50K, \$100K, \$200K, where "K" stands for thousands of dollars
   c) standard deviations: \$25K, \$50K, \$100K.
3) The overall statistics for the (approximately) 600 homes for sale are:
   a) average = \$140K
   b) standard deviation = \$100K

a) If we sample n=60 from the population as a whole, what is the (approximate) standard deviation of the sample mean?
b) If we do a stratified sample of size 20 from each of the three areas (total sample size = 60), what is the (approximate) standard deviation of the stratified sampling estimate of the mean, i.e., the weighted average of the three area sample means, with weights equal to the fractions of homes in each area?
c) Find a sampling scheme that has smaller standard deviation than either of the two schemes in parts a & b. Clearly state the sampling method and the standard deviation.

4. Say that we have a six-sided die that we suspect is weighted so that the six outcomes are not equally likely.

a) Set up an experiment to test this suspicion using a chi-square test. Clearly define the model, its parameters, the hypotheses, the test statistic and the critical value. What minimum criteria would you establish in regard to the sample size and why? Use level of significance $\alpha = 0.05$.
b) Say that you rolled the die 60 times and obtained the following outcomes:

Number on the die:	1	2	3	4	5	6
Number of times rolled:	15	12	11	7	9	6

   Does this experiment's sample size satisfy your criteria in part a? Do the test with this data. What is your conclusion?
c) Use this data to find an approximate 90% confidence interval for the parameter p_1-p_6, i.e., the difference between the probability of a "1" minus the probability of a "6".

(Note: this data makes me suspect that there is a weight near the six, making it the least likely outcome, and the one the most likely outcome.)

Comprehensive Exam Applied Statistics 2004

Appendix One

****

Model 1 2

The GLM Procedure

Dependent Variable: response

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	679.6666667	169.9166667	15.21	<.0001
Error	19	212.2916667	11.1732456		
Corrected Total	23	891.9583333			

R-Square	Coeff Var	Root MSE	response Mean
0.761994	22.59813	3.342641	14.79167

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	2	27.5833333	13.7916667	1.23	0.3133
B	1	260.0416667	260.0416667	23.27	0.0001
C	1	392.0416667	392.0416667	35.09	<.0001

****

Model 2

The GLM Procedure

Dependent Variable: response

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	726.4583333	66.0416667	4.79	0.0059
Error	12	165.5000000	13.7916667		
Corrected Total	23	891.9583333			

R-Square	Coeff Var	Root MSE	response Mean
0.814453	25.10679	3.713713	14.79167

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	2	27.5833333	13.7916667	1.00	0.3966
B	1	260.0416667	260.0416667	18.85	0.0010
A*B	2	16.5833333	8.2916667	0.60	0.5639
C	1	392.0416667	392.0416667	28.43	0.0002
A*C	2	10.0833333	5.0416667	0.37	0.7013
B*C	1	15.0416667	15.0416667	1.09	0.3169
A*B*C	2	5.0833333	2.5416667	0.18	0.8340

****

The GLM Procedure

Dependent Variable: response

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	705.7083333	141.1416667	13.64	<.0001
Error	18	186.2500000	10.3472222		
Corrected Total	23	891.9583333			

R-Square	Coeff Var	Root MSE	response Mean
0.791190	21.74677	3.216710	14.79167

Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	2	27.5833333	13.7916667	1.33	0.2885
B	1	260.0416667	260.0416667	25.13	<.0001
C	1	392.0416667	392.0416667	37.89	<.0001
greenhouse	1	26.0416667	26.0416667	2.52	0.1301

****

The GLM Procedure

Dependent Variable: response

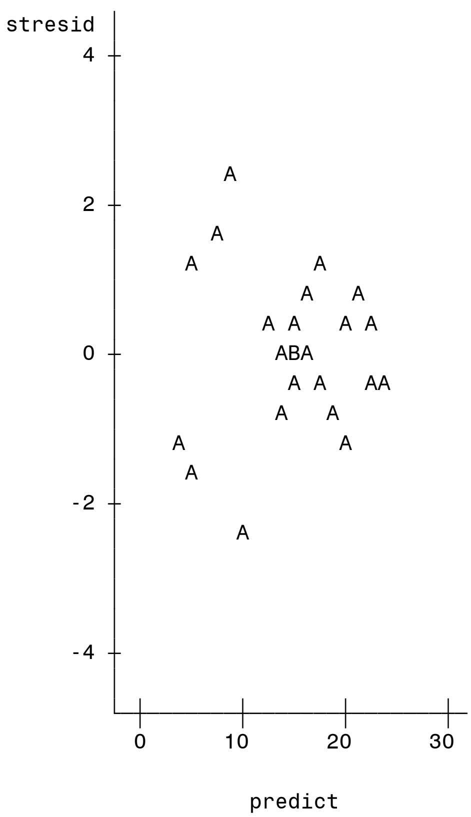
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	752.5000000	62.7083333	4.95	0.0063
Error	11	139.4583333	12.6780303		
Corrected Total	23	891.9583333			

R-Square	Coeff Var	Root MSE	response Mean
0.843649	24.07181	3.560622	14.79167

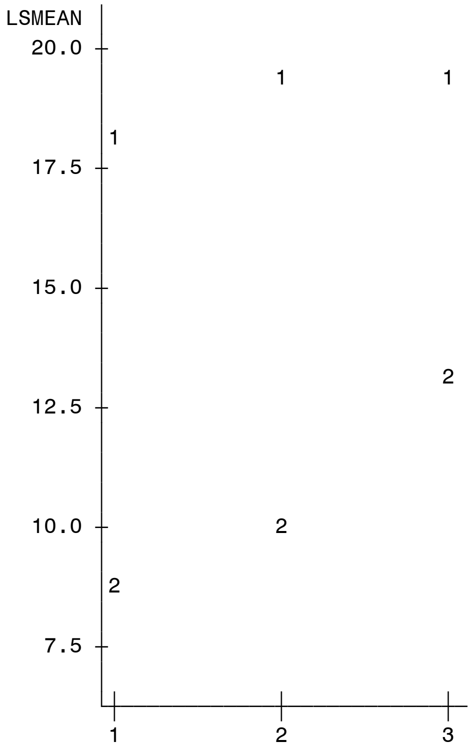
Source	DF	Type III SS	Mean Square	F Value	Pr > F
A	2	27.5833333	13.7916667	1.09	0.3706
B	1	260.0416667	260.0416667	20.51	0.0009
A*B	2	16.5833333	8.2916667	0.65	0.5390
C	1	392.0416667	392.0416667	30.92	0.0002
A*C	2	10.0833333	5.0416667	0.40	0.6812
B*C	1	15.0416667	15.0416667	1.19	0.2993
A*B*C	2	5.0833333	2.5416667	0.20	0.8213
greenhouse	1	26.0416667	26.0416667	2.05	0.1796

****

stresid*predict. A=1, B=2, etc.



Plot of LSMEAN*A=C.



****

Plot of LSMEAN*A=B.