Qualifying Exam on Topics in Statistics: Spring 2011

1. This question has three parts.

(a) Prove the Bonferroni inequality for two events A and B, given by $P(A \cap B) \geq 1 - P(A) - P(B)$.

(b) Extend the Bonferroni inequality you proved in part **1(a)** to a finite number, say p, of events.

(c) For first-order normal linear models (usual multiple regression) with p parameters including the Y intercept, use part **1(b)** and mathematical arguments to obtain simultaneous confidence intervals for all the parameters with family confidence coefficient $1 - \alpha$.

2. This question has three independent parts.

(a) In a study in which the response is the occurrence of preterm birth, 2000 pregnant women are enrolled and possible predictors include age of the woman, smoking, socioeconomic status, body mass index, bleeding during pregnancy, serum level of dde, and several dietary factors. Formulate the problem of selecting the important predictors of the occurrence of preterm birth in a GLM framework. Show the components of the GLM, including the link function and distribution (in exponential family form).

(b) For the model $\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $E(\boldsymbol{\epsilon}) = 0$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a known $n \times n$ matrix, derive the weighted least squares estimator of $\boldsymbol{\beta}$.

(c) Suppose that $y_1, \ldots, y_n$ are a random sample from $N(\log\beta, \sigma^2)$ where σ^2 is known. Obtain the score equation. Verify the iterative weighted least squares equation in this case.

3. This question has two independent parts.

(a) Explain how the odds ratio in logistic regression is connected with the odds ratio in the following 2×2 contingency table.

Response	$x_1 = 0$, Active Drug	$x_1 = 1$, Placebo
$y = 0$, not infected	n_{00}	n_{01}
$y = 1$, infected	n_{10}	n_{11}

(b) A multiple regression model is fitted with response y, and predictors x_1 and x_2. The ANOVA Table and some sums are given below:

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	p-value
Regression	651,996.1	2	325,983	267.2	4.74E-10
Residual	13,421,1	11	1220.1		
Total	665,387.2	13			

$$\sum_{i=1}^{n} x_{i1} = 3,125, \qquad \sum_{i=1}^{n} x_{i2} = 60.72, \qquad \sum_{i=1}^{n} x_{i1}^2 = 702,205, \qquad \sum_{i=1}^{n} x_{i2} = 264.258,$$

$$\sum_{i=1}^{n} x_{i1}x_{i2} = 13,563.5, \qquad \sum_{i=1}^{n} x_{i1}y_i = 3,952,040, \qquad \sum_{i=1}^{n} x_{i2}y_i = 75,738.3.$$

Suppose it is of interest to investigate the contribution of x_2 to the original first-order model. Write down the hypotheses you want to test. Use the extra sum of squares approach to show clearly whether or not x_2 contributes significantly to the model.

4. An experiment was run to determine the effect of CO2 pressure (A), CO2 temperature (B), peanut moisture (C), CO2 flow rate (D), and peanut particle size (E) on the yield of peanut oil (y) as shown below:

Pressure	Temperature	Moisture	Flow	Particle Size	Yield
415	25	5	40	4.05	63
550	25	5	40	1.28	21
415	95	5	40	1.28	36
550	95	5	40	4.05	99
415	25	15	40	1.28	24
550	25	15	40	4.05	66
415	95	15	40	4.05	71
550	95	15	40	1.28	54
415	25	5	60	1.28	23
550	25	5	60	4.05	74
415	95	5	60	4.05	80
550	95	5	60	1.28	33
415	25	15	60	4.05	63
550	25	15	60	1.28	21
415	95	15	60	1.28	44
550	95	15	60	4.05	96

The average yield for all runs is 54.25.
The effects significant at $\alpha = 0.01$ and the sum of squares are shown below:

Factor	Effect	Sum of Squares
B	19.75	1,560.25
E	44.50	7,921.00
Total		10,363.00

(a) What type of design has been used in this experiment?

(b) Identify the defining relation and aliases?

(c) What is the resolution of this design?

(d) Can you improve the resolution by using the same number of runs? Explain why or why not.

(e) Prepare ANOVA for the above experiment and draw conclusions.

(f) Fit a model for predicting yield of peanut oil.

5. An engineer is deigning a new circuit pack (CP) for use in electronic equipment. He needs to select the supplier to obtain the longest lifetime. There are 3 suppliers: S1, S2, and S3. The components can be assembled in one of 3 ways: A1, A2, and A3. The base of the CP can be chosen in 3 ways: B1, B2, and B3. The response variable is the lifetime of CP in accelerated testing of CPs in the manufacturing location.

(a) What design do you suggest for this experiment?

(b) What is the minimum number of CPs that need to made and tested for this experiment?

(c) What are your model and assumptions for analyzing the data to be collected?

(d) Show the outline of your computations and the ANOVA for analyzing the data.

(e) If you are able to make and test twice the number of CPs in Part b, show your new design and the ANOVA.

6. An experiment has been designed to study the effect of three lubricating oils on fuel economy in diesel engines, which is measured after the engine has been running for 15 minutes. Five trucks were available for this study and the following data were collected:

Oil	Truck	Fuel Consumption
1	1	0.5
1	2	0.634
1	3	0.487
1	4	0.329
1	5	0.512
2	1	0.535
2	2	0.675
2	3	0.52
2	4	0.435
2	5	0.54
3	1	0.513
3	2	0.595
3	3	0.488
3	4	0.4
3	5	0.51

Oil sum of sq = 0.006706, Truck sum of sq = 0.092100, Total sum of sq = 0.103028

(a) What design did the experimenter use?

(b) Identify your model, show your ANOVA, and analyze the data collected.

(c) What are your conclusions?

(d) Did the experimenter use an efficient design? Explain why or why not.

(e) If Oil #2 is currently used by the trucking company, construct a meaningful set of orthogonal contrasts for Oil types

(f) How do the sum of squares for these contrasts in Part e relate to relate to the Oil sum of squares given above?