Department of Statistics
University of Wisconsin, Madison
PhD Qualifying Exam Part II
Wednesday, August 29, 2012
1:00-4:00pm, Room 133 SMI

- There are a total of FOUR (4) problems in this exam. Please do a total of TWO (2) problems.
- Each problem must be done in a separate exam book.
- Please turn in TWO (2) exam books.
- Please write your code name and **NOT** your real name on each exam book.

1. Denote by $\mathcal{B}$ the Borel sigma field on the real line R, and let $(\Omega, \mathcal{F})$ be a measurable space. Define a mapping $Q(t, A)$ for $t \in R$ and $A \in \mathcal{F}$ such that $Q(t, \cdot)$ is a probability measure on $(\Omega, \mathcal{F})$ for each $t \in R$, and $Q(\cdot, A)$ is a Borel function for each $A \in \mathcal{F}$.

   (a) Assume that $f(\omega)$ is a measurable function on $(\Omega, \mathcal{F})$. For each $t \in R$, define the integral of $f(\omega)$ with respect to probability measure $Q(t, \cdot)$ as follows,

   $$H(t) = \int_{\Omega} f(\omega) Q(t, d\omega).$$

   Show that $H(t)$ is a measurable function on $(R, \mathcal{B})$.

   (b) Assume that μ is a probability measure on $(R, \mathcal{B})$. For each $A \in \mathcal{F}$, define

   $$P^{\mu}(A) = \int_{R} Q(t, A) \mu(dt).$$

   Show that $P^{\mu}(\cdot)$ is a probability measure on $(\Omega, \mathcal{F})$.

   (c) Denote by Π a probability measure on $(R, \mathcal{B})$, P a probability measure on $(\Omega, \mathcal{F})$, and T a random variable on $(\Omega, \mathcal{F}, P)$. Assume that Π, P and T satisfy $\Pi = P \circ T^{-1}$, and for $A \in \mathcal{F}$,

   $$P(A) = \int_{R} Q(t, A) \Pi(dt),$$

   where $P \circ T^{-1}$ denotes the induced probability measure by T. Prove or disprove the following statement:

   $$\text{For any } A \in \mathcal{F}, \qquad P(A|T) = Q(T, A) \text{ almost surely.}$$

2. Suppose that $X, X_1, X_2, ...X_n$ are i.i.d. with mean μ, variance σ^2, and $E|X|^4 < \infty$. Define the sample mean $\bar{X}_n$ and sample variance S_n^2 using the following

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

This problem studies the asymptotic behavior of $\bar{X}_n$ and S_n^2 with some applications to statistical theory. We now define *excess kurtosis*, which will be denoted by γ. The quantity $\mu_4 = E(X-\mu)^4$ is the fourth central moment of X. The ratio $\frac{\mu_4}{\sigma^4}$ is the *kurtosis* of X and $\gamma \equiv \frac{\mu_4}{\sigma^4} - 3$ is the *excess kurtosis* of X.

In this problem $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. $P(\mu)$ denotes a Poisson distribution with mean μ.

(a) Prove that $\gamma \geq -2$. Construct a random variable when equality holds. Find γ for $N(\mu, \sigma^2)$ and $P(\mu)$.

(b) Find the limiting distribution of $\frac{\sqrt{n}}{\sqrt{2\sigma^2}}\left(S_n^2 - \sigma^2\right)$. Express the asymptotic variance as a function of γ.

(c) Let σ_0^2 be a specified value. Suppose that you want to test $H_0 : \sigma^2 = \sigma_0^2$ versus $H_A : \sigma^2 > \sigma_0^2$. Recall that the α-level test based on normal theory rejects when

$$\frac{n-1}{\sigma_0^2} S_n^2 \geq \chi^2_{n-1,\alpha},$$

where $\chi^2_{n-1,\alpha}$ denotes the upper α tail for chi-square distribution with $n-1$ degrees of freedom. Find the limiting probability of a Type I error as a function γ.

- For the remainder of the problem let $X \sim P(\mu)$.

(d) Show that the sequence $\left\{\sqrt{n}\left|\bar{X}_n - \mu\right|\right\}$ is uniformly integrable. Prove that the following limit exists

$$\lim_{n\to\infty} E\left(\sqrt{n}\left|\bar{X}_n - \mu\right|\right).$$

Express the limit in terms of μ.

(e) Let $g(x) = x^\beta$ for $x \geq 0$ and $0 < \beta < \infty$. Find the limiting distribution of $\sqrt{n}\left(g\left(\bar{X}_n\right) - g(\mu)\right)$. Express the asymptotic variance as a function of μ and β.

3. An experiment was conducted to study the effect of factors that might affect the sweetness of wine measured in grams of sugar per liter of juice. Two experimental factors were of interest: the fertilization of the plants, and type of pruning method used. The fertilizer treatments consisted of: (1) amendment with nitrogen only (NO); or (2) amendment with nitrogen and phosphorus (NP). The pruning methods are defined in terms of the number of buds left. In this experiment pruning was done to leave either 50, 45, 40, or 30 buds. Pruning to 50 buds is least aggressive; pruning to 30 is most aggressive.

   The experiment was conducted in three vineyards at a large winery. We will simply label these as the East, West, and Central vineyard. Within each vineyard, two rows of vines were chosen at random, and each one was randomly assigned to receive one of the fertilizer treatments. Within each of the rows, four trunks were chosen, and each of the trunks was randomly assigned to receive on of the four pruning methods. Finally, at harvest time, two grapes were sampled from each trunk (call them Grape 1 and Grape 2) and the sweetness of each grape was determined.

   The data are shown below. The columns are, in order, the sweetness, the vineyard, the fertilizer, the amount of pruning, and the grape sampled for each trunk.

```
45 E NO 50 1
53 E NO 50 2
51 E NO 45 1
61 E NO 45 2
53 E NO 40 1
66 E NO 40 2
54 E NO 30 1
66 E NO 30 2
76 E NP 50 1
86 E NP 50 2
71 E NP 45 1
81 E NP 45 2
71 E NP 40 1
82 E NP 40 2
62 E NP 30 1
73 E NP 30 2
31 C NO 50 1
43 C NO 50 2
27 C NO 45 1
40 C NO 45 2
35 C NO 40 1
44 C NO 40 2
49 C NO 30 1
57 C NO 30 2
61 C NP 50 1
73 C NP 50 2
61 C NP 45 1
71 C NP 45 2
73 C NP 40 1
82 C NP 40 2
```

```
81 C NP 30 1
88 C NP 30 2
66 W NO 50 1
76 W NO 50 2
53 W NO 45 1
66 W NO 45 2
79 W NO 40 1
96 W NO 40 2
83 W NO 30 1
98 W NO 30 2
79 W NP 50 1
88 W NP 50 2
75 W NP 45 1
83 W NP 45 2
86 W NP 40 1
100 W NP 40 2
101 W NP 30 1
113 W NP 30 2
```

These data were also analyzed in SAS as follows:

```
options ls=80;

data a;
    infile "grapes.txt" ;
    input sweet yard$ fert$ prune grape;

proc glm;
    class yard fert prune grape;
    model sweet = yard|fert|prune|grape;

    lsmeans yard fert prune grape;
```

Here is some edited output.

```
                        The GLM Procedure

                     Class Level Information

              Class          Levels    Values

              yard                3    C E W
              fert                2    NO NP
              prune               4    30 40 45 50
              grape               2    1 2
```

```
Dependent Variable: sweet

                                     Sum of
Source                      DF      Squares    Mean Square   F Value    Pr > F

Model                       47  17422.81250      370.69814         .         .
Error                        0      0.00000              .
Corrected Total             47  17422.81250

Source                      DF    Type I SS    Mean Square   F Value    Pr > F

yard                         2  5924.625000    2962.312500         .         .
fert                         1  5742.187500    5742.187500         .         .
yard*fert                    2   805.875000     402.937500         .         .
prune                        3  1772.729167     590.909722         .         .
yard*prune                   6  1215.208333     202.534722         .         .
fert*prune                   3    82.729167      27.576389         .         .
yard*fert*prune              6   369.958333      61.659722         .         .
grape                        1  1441.020833    1441.020833         .         .
yard*grape                   2    10.791667       5.395833         .         .
fert*grape                   1     6.020833       6.020833         .         .
yard*fert*grape              2     4.041667       2.020833         .         .
prune*grape                  3     6.562500       2.187500         .         .
yard*prune*grape             6    35.375000       5.895833         .         .
fert*prune*grape             3     3.562500       1.187500         .         .
yard*fert*prun*grape         6     2.125000       0.354167         .         .


                        Least Squares Means

                        yard    sweet LSMEAN

                        C         57.2500000
                        E         65.6875000
                        W         83.8750000

                        fert    sweet LSMEAN

                        NO        58.0000000
                        NP        79.8750000

                       prune    sweet LSMEAN

                       30         77.0833333
                       40         72.2500000
                       45         61.6666667
                       50         64.7500000
```

```
grape     sweet LSMEAN

1           63.4583333
2           74.4166667
```

(a) Make a suitable plot to determine whether there is evidence of an interaction between fertilization and pruning method.

(b) Write down the ANOVA table for this experiment, indicating Source and df only.

(c) Perform a formal test to determine whether there is a main effect for fertilization.

(d) Perform a formal test to determine whether there is a main effect for pruning.

(e) Is there evidence of a linear relationship between number of buds pruned and sweetness?

(f) Suppose that the experimenters had instead decided to focus on the NP fertilizer treatment, and the use of a 40 bud pruning regime. Further, suppose they observed 8 trunks, and 3 grapes per trunk. Imagine calculating the mean sweetness from the resulting 24 grapes. Use the given SAS output to estimate the variance of this mean.

Use $\alpha = 0.05$ in all hypothesis tests.

4. This question has two problems. Answer both to get full credit.

   (a) We observe n independent samples (x_i, y_i), $i = 1, 2, \ldots, n$, such that x_i is uniformly distributed on $(0, 1)$ and $y_i = 1 + x_i + \epsilon_i$, where ϵ_i has zero mean and constant (finite) variance and is independent of x_i.

      i. Suppose you use least squares to fit the model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{1}$$

         A. If n is sufficiently large, which regression coefficients are likely to be statistically significant and which nonsignificant from zero? Justify your answer. You may use the usual 0.05 level of significance.

         B. Find the limiting values of the coefficient estimates as $n \to \infty$.

      ii. Answer the questions in part (a) if you instead fit the model

$$y_i = \beta_0 + \beta_2 x_i^2 + \beta_3 x_i^3 + \epsilon_i, \quad i = 1, 2, \ldots, n. \tag{2}$$

   (b) Concrete is made by mixing together different ingredients. To figure out how much of each ingredient to use, an engineer made 103 batches of concrete with varying amounts of seven ingredients: cement, slag, fly ash, water, superplasticizer (SP), coarse aggregate, and fine aggregate, each measured in kg per cubic meter. Slag and fly ash are cement substitutes. One measured response was slump, defined as the vertical height by which a cone of wet concrete sags. A scatterplot matrix of the data is given in Figure 1 and a multiple linear regression model fitted to slump yields the following results.

```
> summary(lm(Slump ~ Cement + Slag + Flyash + Water
                     + SP + CoarseAggr + FineAggr))

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -88.525037 203.303168  -0.435    0.664
Cement        0.010216   0.065256   0.157    0.876
Slag         -0.012966   0.090819  -0.143    0.887
Flyash        0.006176   0.066216   0.093    0.926
Water         0.258982   0.204900   1.264    0.209
SP           -0.183954   0.384827  -0.478    0.634
CoarseAggr    0.029737   0.078458   0.379    0.706
FineAggr      0.038584   0.082415   0.468    0.641

Residual standard error: 7.459 on 95 degrees of freedom
Multiple R-squared: 0.3233,Adjusted R-squared: 0.2734
F-statistic: 6.484 on 7 and 95 DF,  p-value: 2.98e-06
```

   Finding no significant variables, the engineer fitted two more models which showed that slag and water were highly significant.

```
> summary(lm(Slump ~ Slag + Water))
```

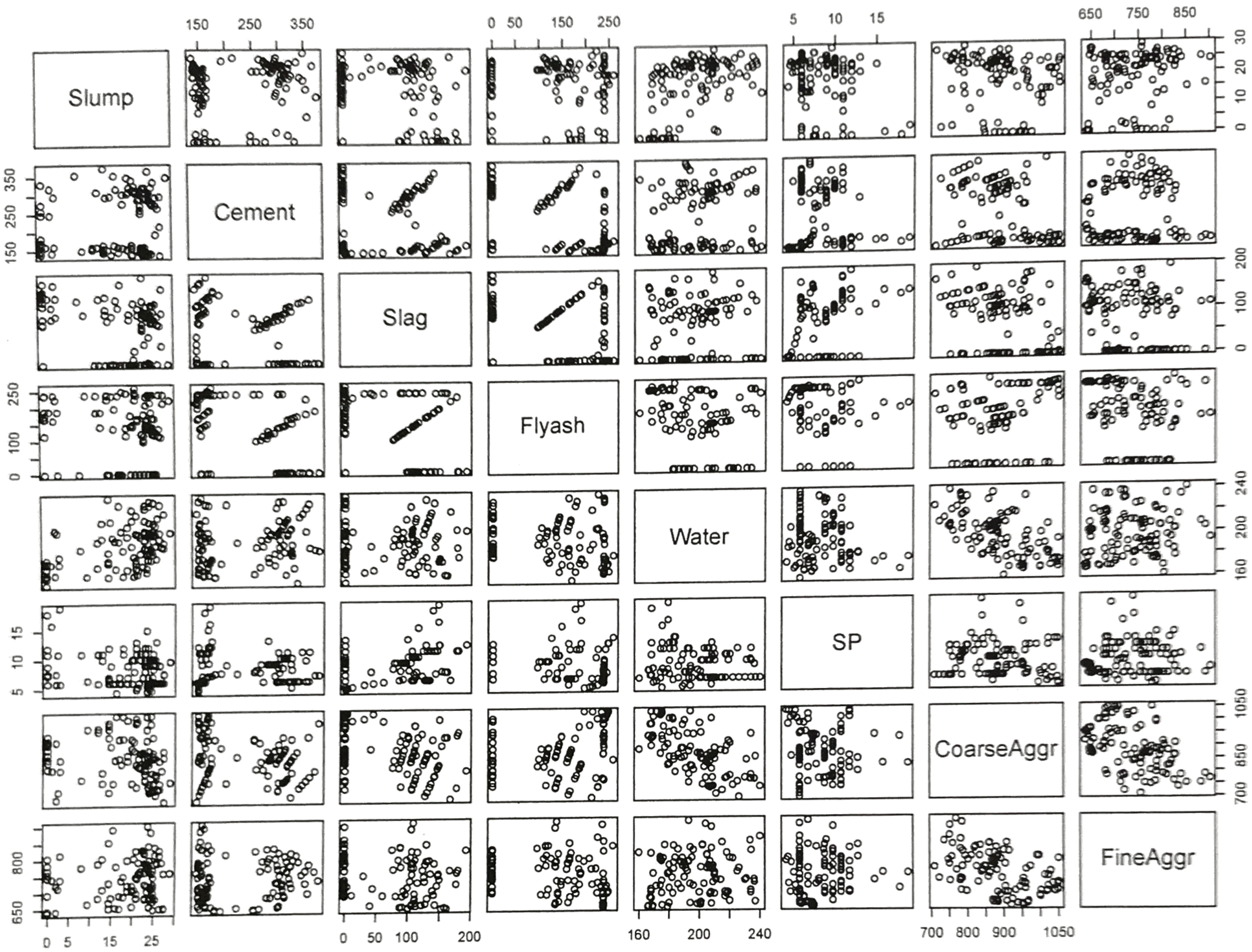

Figure 1: Scatterplot matrix of concrete data

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.09945    7.31394  -2.475  0.01502 *
Slag         -0.03933    0.01219  -3.227  0.00169 **
Water         0.19889    0.03646   5.455 3.56e-07 ***

> summary(lm(Slump ~ Slag * Water))

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.3695591  9.6825720   1.174    0.243
Slag        -0.4785987  0.1039476  -4.604 1.23e-05 ***
Water        0.0498586  0.0486295   1.025    0.308
Slag:Water   0.0022271  0.0005239   4.251 4.83e-05 ***
```

i. Are these contradictory results theoretically possible, or might the engineer have made some mistakes in his calculations?

ii. Assuming that the results are correct,

A. explain why slag and water are significant when they are the only predictors in the model but not when other predictors are present;

B. explain what the results tell you about the true relationship between slump and the other variables.

Give reasons for your answers. Mere speculations will receive no credit.