Department of Statistics
University of Wisconsin-Madison
PhD Qualifying Exam Part II
September 6, 2007
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, **NOT** your real name, on each exam book.

1. *Some definitions and facts:* Let $(\Omega, \mathcal{F}, P)$ be a probability space and suppose that $\mathcal{D}$ is a sub-$\sigma-$algebra of $\mathcal{F}$. For an integrable random variable X, the *conditional expectation* of X given $\mathcal{D}$, $E[X|\mathcal{D}]$, is the unique (up to changes on events of probability zero) $\mathcal{D}$-measurable random variable such that

$$\int_D X dP = \int_D E[X|\mathcal{D}] dP$$

for all $D \in \mathcal{D}$.

Problem: Assume that X is a random variable defined on $(\Omega, \mathcal{F}, P)$ satisfying $E[X^2] < \infty$ and $\{\mathcal{D}_n\}$ is a sequence of sub-σ-algebras in $\mathcal{F}$ (not necessarily increasing).

(a) Suppose $\mathcal{D}_1 \subset \mathcal{D}_2$ and that $E[E[X|\mathcal{D}_1]^2] = E[E[X|\mathcal{D}_2]^2]$. Show that $E[X|\mathcal{D}_1] = E[X|\mathcal{D}_2]$ a.s.

(b) Define $X_n = E[X|\mathcal{D}_n]$. Suppose $\lim_{n\to\infty} E[X_n^2] = E[X^2]$. Show that

$$\lim_{n\to\infty} E[(X - X_n)^2] = 0.$$

(c) Suppose $\{\mathcal{D}_n\}$ is increasing, that is, $\mathcal{D}_n \subset \mathcal{D}_{n+1}$ for $n = 1, 2, \ldots$, and $X_n = E[X|\mathcal{D}_n]$. Prove that there exists a random variable Y such that $\lim_{n\to\infty} E[(Y - X_n)^2] = 0$ and that $X = Y$ if and only if the assumption of part (b) holds. (This result is a version of the martingale convergence theorem, but you may not use the martingale convergence theorem in your solution.)

2. Let $X_1, X_2, \ldots, X_n$ be independent random variables with

$$P(X_k = x) = \begin{cases} k^{-2}/2, & x = -k \\ (1 - k^{-2})/2, & x = -1/2 \\ (1 - k^{-2})/2, & x = 1/2 \\ k^{-2}/2, & x = k. \end{cases}$$

Define $\mu_k = E(X_k)$, $\sigma_k^2 = \text{var}(X_k)$ and $s_n^2 = \sum_{k=1}^n \sigma_k^2$.

(a) The Feller condition states:

$$\text{For any } \epsilon > 0, \quad \max_{1 \le k \le n} P(|(X_k - \mu_k)/s_n| > \epsilon) \to 0 \quad \text{as } n \to \infty.$$

Is the Feller condition satisfied in this problem?

(b) Is the Lindeberg condition satisfied in this problem?

(c) Consider the random sequence $\{Y_k\}$, where

$$Y_k = \begin{cases} X_k, & \text{if } |X_k| = 1/2, \\ 0, & \text{otherwise.} \end{cases}$$

Derive the asymptotic distribution of $\sum_k Y_k$.

(d) Show that $\sum_{k=1}^n (X_k - Y_k) = O_P(1)$.

(e) Derive the asymptotic distribution of $\sum_k X_k$.

(f) Are there any contradictions between the conclusions of parts (a), (b) and (e)?

3. Data from a sample of 388 cars were obtained on the following variables: horse power (`hp`), number of cylinders (`cyl`), miles per gallon rating (`mpg`), engine displacement (`disp`), weight (`weight`), and acceleration (`acc`; time to reach 60mph from rest). An engineer wants to study the influence of the variables on `hp`. Figure 1 shows plots of `hp` versus each of the other five variables.

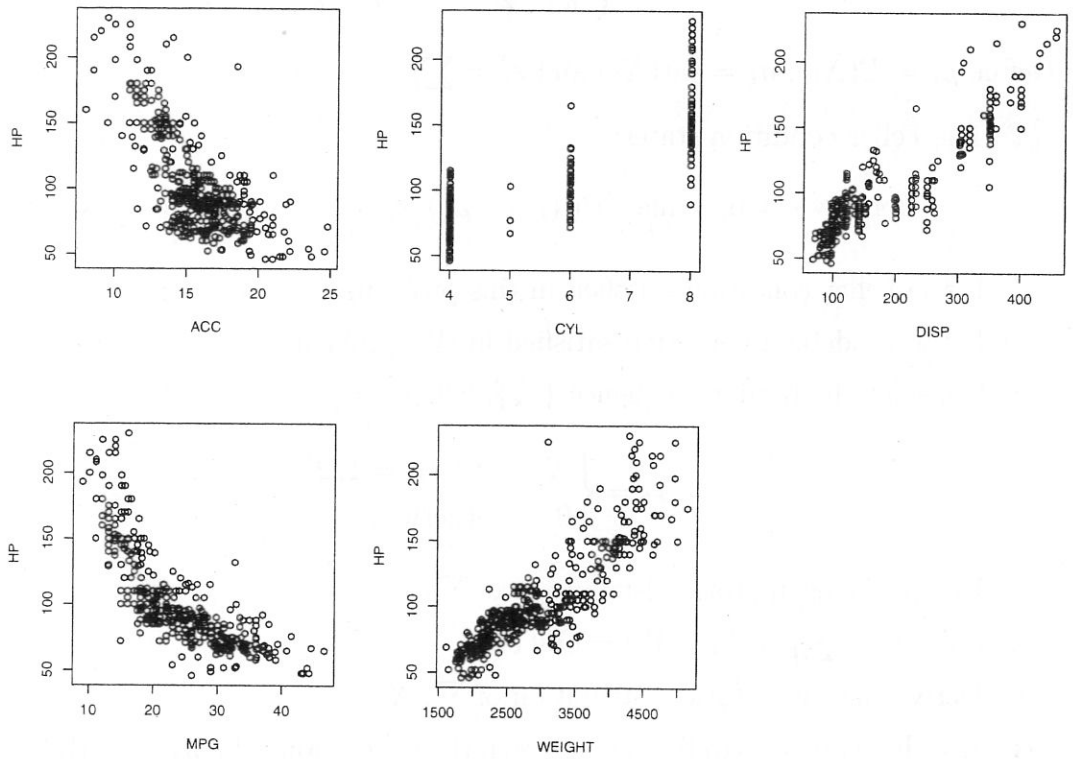

Figure 1: Plots of `hp` versus predictor variables

The engineer first fitted two models to the data:

$$\texttt{hp} = \beta_0 + \beta_1\texttt{acc} + \beta_2\texttt{cyl} + \beta_3\texttt{disp} + \beta_4\texttt{mpg} + \beta_5\texttt{weight} + \varepsilon \quad (1)$$

$$\texttt{hp} = \gamma_0 + \gamma_2\texttt{cyl} + \varepsilon \quad (2)$$

Model (1) is a multiple linear regression using all five predictor variables and model (2) is a simple linear regression of `hp` on `cyl`. The results from the `R` package are:

```
Call: lm(formula = hp ~ cyl + disp + mpg + weight + acc)
Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.666309   9.250348  12.828  < 2e-16 ***
cyl          -2.515602   1.253062  -2.008   0.0454 *
disp          0.142016   0.026798   5.300 1.97e-07 ***
mpg          -0.399036   0.156467  -2.550   0.0112 *
weight        0.018050   0.002458   7.342 1.28e-12 ***
acc          -4.663439   0.303906 -15.345  < 2e-16 ***
---
Residual standard error: 12.91 on 382 degrees of freedom
Multiple R-Squared: 0.8901,     Adjusted R-squared: 0.8886
F-statistic: 618.7 on 5 and 382 DF,  p-value: < 2.2e-16

Call: lm(formula = hp ~ cyl)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.1114     3.5128  -0.601    0.548
cyl          19.3972     0.6107  31.764   <2e-16 ***
---
Residual standard error: 20.37 on 386 degrees of freedom
Multiple R-Squared: 0.7233,     Adjusted R-squared: 0.7226
F-statistic:  1009 on 1 and 386 DF,  p-value: < 2.2e-16
```

(a) Explain why the sign of the coefficient of `cyl` changes between models (1) and (2)? Describe clearly the effect of `cyl` on `hp`. You may refer to the plots in Figure 2 if necessary.

(b) The engineer also fitted a full second-order model to the data and obtained the following Type I sums of squares ANOVA table. Find the corresponding ANOVA tables for models (1) and (2).

```
             Df Sum Sq Mean Sq   F value    Pr(>F)
cyl           1 418766  418766 3903.7397 < 2.2e-16 ***
disp          1  52744   52744  491.6811 < 2.2e-16 ***
mpg           1   3571    3571   33.2888 1.690e-08 ***
weight        1   1032    1032    9.6219 0.0020716 **
acc           1  39226   39226  365.6685 < 2.2e-16 ***
I(cyl^2)      1   6282    6282   58.5595 1.758e-13 ***
I(disp^2)     1   8252    8252   76.9272 < 2.2e-16 ***
I(mpg^2)      1    722     722    6.7282 0.0098699 **
I(weight^2)   1   4047    4047   37.7235 2.121e-09 ***
```

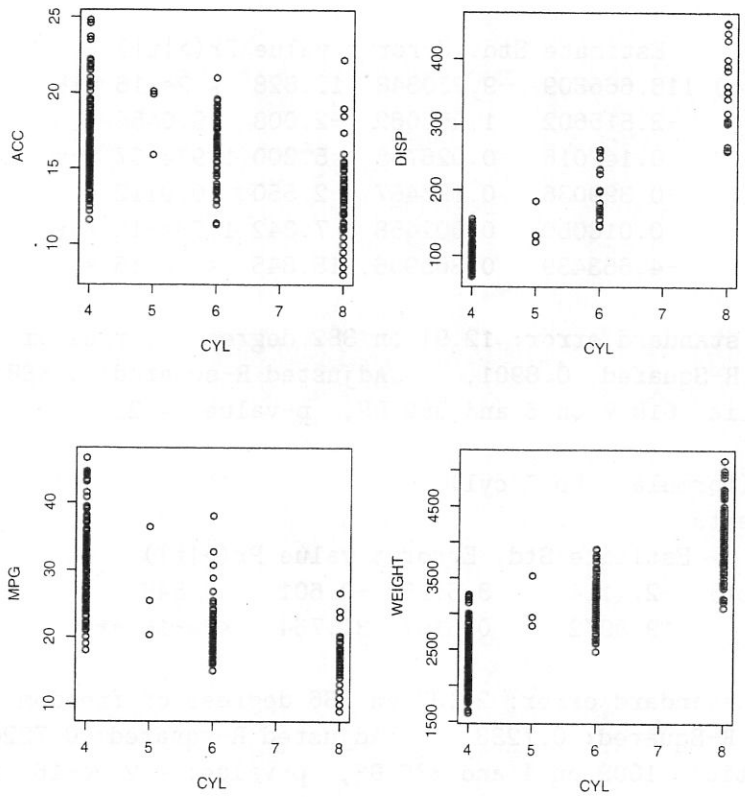

Figure 2: Plots of `acc`, `disp`, `mpg`, and `weight` versus `cyl`

```
I(acc^2)        1    436     436    4.0608 0.0446175 *
cyl:disp        1   1357    1357   12.6482 0.0004252 ***
cyl:mpg         1      7       7    0.0687 0.7933211
cyl:weight      1    151     151    1.4073 0.2362760
cyl:acc         1    346     346    3.2286 0.0731845 .
disp:mpg        1    613     613    5.7102 0.0173705 *
disp:weight     1   1316    1316   12.2646 0.0005185 ***
disp:acc        1     22      22    0.2047 0.6512234
mpg:weight      1    620     620    5.7797 0.0167070 *
mpg:acc         1     93      93    0.8653 0.3528594
weight:acc      1      5       5    0.0440 0.8339573
Residuals     367  39369     107
```

(c) Is there evidence that, as a group, the second-order terms contribute significantly to the model, given that the linear terms are present?

(d) Construct a dataset containing four observations and three variables x_1, x_2, y

such that the estimated regression coefficient of x_1 changes sign, depending on whether the model $y = \beta_0+\beta_1 x_1+\varepsilon$ or the model $y = \beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$ is fitted. Give your results in the following tabular form:

i	x_1	x_2	y
1			
2			
3			
4			

(e) Construct another dataset of the same size such that the estimated regression coefficient of x_1 is the same regardless of whether or not x_2 is in the model.

Hint: For any two vectors $u = (u_1, u_2, \ldots, u_n)$ and $v = (v_1, v_2, \ldots, v_n)$, let $\bar{u} = n^{-1}\sum_{i=1}^n u_i$, $\bar{v} = n^{-1}\sum_{i=1}^n v_i$, and $f(u, v) = n^{-1}\sum_{i=1}^n u_i v_i - \bar{u}\bar{v}$. The least squares estimates of the linear regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ are

$$\begin{aligned}
\hat{\beta}_1 &= \frac{f(x_1, y)f(x_2, x_2) - f(x_2, y)f(x_1, x_2)}{f(x_1, x_1)f(x_2, x_2) - f(x_1, x_2)^2} \\
\hat{\beta}_2 &= \frac{f(x_2, y)f(x_1, x_1) - f(x_1, y)f(x_1, x_2)}{f(x_1, x_1)f(x_2, x_2) - f(x_1, x_2)^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2.
\end{aligned}$$

4. An experiment was conducted to assess the effects of three factors on photosynthetic activity of pine tree needles. The factors were:

   A augmentation of soil around the tree with carbon (in the form of sugar). A=0 means that augmentation was not used; A=1 means that augmentation was used.

   N augmentation of soil around the tree with nitrogen (in the form of nitrous ammonia). N=0 means that augmentation was not used; N=1 means that augmentation was used.

   B branch height: whether the branch being inspected was 1.2, 1.8, or 2.4 meters above the ground. For convenience, these data have been recoded as: B = 0, 1, or 2.

   For each combination of the above factors, 4 trees were located at random in a large pine tree plantation—48 trees were used in total. So, for example, a tree might be chosen which was augmented with carbon, not augmented with nitrogen, and the branch located 2.4 m off the ground was observed. A photosynthesis meter was used to then measure photosynthetic activity of the needles on that branch.

   The data are summarized below:

A	0	0	0	0	0	0	1	1	1	1	1	1
N	0	0	0	1	1	1	0	0	0	1	1	1
B	0	1	2	0	1	2	0	1	2	0	1	2
Mean	29.30	31.90	32.80	26.00	29.1	34.60	29.10	31.50	33.4	28.90	35.90	39.90
SD	2.35	2.21	7.93	1.41	1.80	1.58	4.27	1.64	1.70	0.54	4.19	0.94

   The following (edited) results are also available to you:

```
Analysis of Variance Table
Response: y
          Sum Sq
A          75.95
N          14.34
B         379.58
A:N        75.30
A:B         7.74
N:B        69.61
A:N:B       8.53
```

(a) By making an appropriate plot, assess whether there is evidence of an interaction between augmentation by carbon and augmentation by nitrogen.

(b) By using an appropriate F test, formally assess the evidence in support of an interaction between augmentation by carbon and augmentation by nitrogen.

(c) Let μ_{ijk} represent the mean photosynthetic activity for A=i, N=j, and B=k. Construct an F test to formally test $H_0 : \mu_{010} = \mu_{011} = \mu_{012}$ and interpret, in words, the meaning of this hypothesis.

(d) The researchers are concerned that the data corresponding to the combination A=0, N=0, B=2 might contain an outlier. They believe this to be true because the standard deviation for this group appears to be large. Construct a formal test to determine whether this standard deviation is unusually large.

(e) Consider a related experiment. Again, let μ_{ijk} represent the mean photosynthetic activity for A=i, N=j, and B=k. Consider the following model for the data:

$$Y_{ijk} = \mu_{ijk} + e_{ijk}$$

where e_{ijk} is normally distributed with mean zero and variance σ^2, the correlation between e_{ijk} and $e_{ijk'}$ is $0.3^{|k-k'|}$, and otherwise e_{ijk} and $e_{i'j'k'}$ are independent.

Provide a detailed description, in words, of an experiment studying the effects of carbon augmentation, nitrogen augmentation, and branch height on photosynthetic activity, such that the experiment would likely result in such a model.

IV. Percentage Points of the F Distribution (*continued*)

$F_{0.05,\nu_1,\nu_2}$

ν_2 \ ν_1	Degrees of Freedom for the Numerator (ν_1)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Degrees of Freedom for the Denominator (ν_2)