Department of Statistics

University of Wisconsin, Madison

PhD Qualifying Exam Option B

August 28, 2018

12:30-4:30pm, Room 133 SMI

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

1. Let $X_1, \ldots, X_n$ be independent, identically distributed copies of X, a random variable with the Log-Normal(μ, σ) distribution. Recall that the probability density function (pdf) of the Log-Normal(μ, σ) distribution is given by

$$f(x; \mu, \sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(x) - \mu)^2}{2\sigma^2}\right\}, \quad x > 0, \quad -\infty < \mu < \infty, \quad 0 < \sigma < \infty.$$

Assume that μ is unknown and $\sigma = 1$. The parameter of interest in this problem is the population mean, $\theta = \mathbb{E}_\mu(X)$. Solve the following problems:

(a) Give an expression for θ as a function of μ.

(b) Find $\widehat{\theta}$, the maximum likelihood estimator (MLE) of θ.

(c) Find $\widetilde{\theta}$, the uniformly minimum variance unbiased estimator (UMVUE) of θ.

(d) In terms of mean squared error (MSE), which estimator has better performance: $\widehat{\theta}$ or $\widetilde{\theta}$?

(e) Show that $\widehat{\theta}$ and $\widetilde{\theta}$ are both asymptotically normal with a common asymptotic variance, and hence the two estimators have equivalent performance asymptotically. Calculate the common asymptotic variance.

(f) **Definition.** Given two sequences of estimators $\{V_n\}$ and $\{W_n\}$ of θ which satisfy $\sqrt{n}(V_n - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma_V^2)$ and $\sqrt{n}(W_n - \theta) \xrightarrow{d} \mathrm{N}(0, \sigma_W^2)$, the *asymptotic relative efficiency* (ARE) of V_n to W_n is given by σ_V^2/σ_W^2.

Calculate the asymptotic relative efficiency (ARE) of $\widetilde{\theta}$ to $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, the sample mean.

HINT: The moment generating function of a standard normal random variable $Z \sim \mathrm{N}(0,1)$ is given by $M_Z(t) = e^{t^2/2}$.

2. This question consists of two separate parts.

(a) (Part 1) Let X be an inverse Gaussian random variable, following the probability density function,

$$f(x;\lambda,\,\mu) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left\{\frac{-\lambda\,(x-\mu)^2}{2\mu^2 x}\right\}, \quad \text{for } x \in (0,\infty),$$

with parameters $\lambda \in (0,\infty)$ and $\mu \in (0,\infty)$.

i. Derive $\mathrm{E}(X)$ and $\mathrm{var}(X)$.
HINT: Consider the exponential family distribution.

(b) (Part 2) Let $\boldsymbol{X} = (X_1,\ldots,X_p)^T \sim N(\boldsymbol{\mu},\Sigma)$ be a multivariate Gaussian random vector, following the probability density function,

$$f(\boldsymbol{x};\boldsymbol{\mu},\Sigma) \;=\; \frac{1}{\sqrt{(2\pi)^p}\{\det(\Sigma)\}^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\},$$

with the mean vector $\boldsymbol{\mu}$ and a positive definite covariance matrix Σ. Assume $\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n\}$ are i.i.d. samples of $\boldsymbol{X}$, with $n > p$.

i. Consider the parameter space $\Theta_0 = \{(\boldsymbol{\mu},\Sigma) : \boldsymbol{\mu} \in \mathbb{R}^{p\times 1},\, \Sigma = \mathrm{I}_{p\times p}\}$, where $\mathrm{I}_{p\times p}$ denotes an identity matrix of dimension $p \times p$. Find the maximum likelihood estimator $\widehat{\boldsymbol{\mu}}_{\mathrm{MLE};0}$ of $\boldsymbol{\mu}$ in Θ_0.

ii. Consider the parameter space $\Theta = \{(\boldsymbol{\mu},\Sigma) : \boldsymbol{\mu} \in \mathbb{R}^{p\times 1},\, \Sigma \in \mathbb{R}^{p\times p} \text{ is positive definite}\}$. Find the maximum likelihood estimator $\widehat{\boldsymbol{\mu}}_{\mathrm{MLE}}$ of $\boldsymbol{\mu}$ in Θ.
HINT 1: You may use without proving that the maximum likelihood estimator $\widehat{\Sigma}_{\mathrm{MLE}}$ of Σ in Θ is the sample covariance matrix S_n.
HINT 2: For a positive definite matrix A, the partial derivative of $\boldsymbol{x}^T A\boldsymbol{x}$ with respect to $\boldsymbol{x}$ is $\frac{\partial \boldsymbol{x}^T A\boldsymbol{x}}{\partial \boldsymbol{x}} = 2A\boldsymbol{x}$.

iii. Consider the hypothesis

$$H_0 : \Sigma = \mathrm{I}_{p\times p} \quad \text{versus} \quad H_1 : \Sigma \neq \mathrm{I}_{p\times p}.$$

For the maximum likelihood ratio test statistic, defined by $\Lambda_n = \frac{\max_{\theta\in\Theta} f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n;\theta)}{\max_{\theta\in\Theta_0} f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n;\theta)}$, discuss the limit distribution of $2\log(\Lambda_n)$ under H_0 as $n \to \infty$.

3. Consider a linear model written in the form

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \ldots + \mathbf{X}_k\beta_k + \epsilon \tag{1}$$

where $\mathbf{Y}$ is $n \times 1$, $\mathbf{X}_i$ is $n \times p_i$ for $1 \leq i \leq k$, β_i is $p_i \times 1$, and ϵ is an $n \times 1$ vector of errors that has a normal distribution with mean 0 and covariance matrix $\sigma^2 I_{n\times n}$.[1] The total number of linear regression coefficients is $p = \sum_{i=1}^{k} p_i < n$. We suppose also that $\mathbf{X}_i^\top \mathbf{X}_i$ is invertible for each $i = 1, \ldots, k$, and

$$\mathbf{X}_i^\top \mathbf{X}_j = 0 \text{ for } i \neq j. \tag{2}$$

Let

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix} \tag{3}$$

be the vector of least squares estimates under the full model (1).

(a) Show that $\widehat{\beta}_i = (\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\mathbf{X}_i^\top \mathbf{Y}$, in other words, the estimate of β_i under the full model (1) is the same as under the model $\mathbf{Y} = \mathbf{X}_i\beta_i + \epsilon$.

(b) Define the residual vector $\mathbf{e} = \mathbf{Y} - \sum_{i=1}^{k} \mathbf{X}_i\widehat{\beta}_i$. Show that

$$\mathbf{Y}^\top \mathbf{Y} = \sum_{i=1}^{k} \widehat{\beta}_i^\top \mathbf{X}_i^\top \mathbf{X}_i \widehat{\beta}_i + \mathbf{e}^\top \mathbf{e}. \tag{4}$$

Explain how to interpret (4) as an "analysis of variance decomposition" of the total sum of squares $\mathbf{Y}^\top \mathbf{Y}$ into k sums of squares due to regression (one corresponding to each of the submatrices $\mathbf{X}_1, \ldots, \mathbf{X}_k$) and the sum of squares due to error. What are the corresponding degrees of freedom?

(c) Suppose we want to test the hypothesis $H_0 : \beta_1 = \ldots = \beta_l = 0$ for some l between 1 and k, against the alternative hypothesis H_1 that not all of $\beta_1, \ldots, \beta_l$ are 0. Consider the following test statistic

$$\frac{(\sum_{i=1}^{l} \widehat{\beta}_i^\top \mathbf{X}_i^\top \mathbf{X}_i \widehat{\beta}_i)/(\sum_{i=1}^{l} p_i)}{(\mathbf{e}^\top \mathbf{e})/(n-p)} \tag{5}$$

and find its distribution when H_0 is true.

[1] $I_{n\times n}$ represents the $n \times n$ identity matrix.

(d) In a clinical trial of a new drug designed to reduce blood pressure, 30 men are sampled and their age in years, weight in kg. and reduction in blood pressure (B.P.) are recorded. Assume B.P. for the ith patient is y_i, age is x_{i1} and weight is x_{i2}. Consider the following relevant summary statistics: $\sum_{i=1}^n y_i = 293.9$, $\sum_{i=1}^n (y_i - \bar{y})^2 = 1846.7$, $\sum_{i=1}^n y_i(x_{i1} - \bar{x}_{\cdot 1}) = 1137.5$, $\sum_{i=1}^n y_i(x_{i2} - \bar{x}_{\cdot 2}) = 659.4$, $\sum_{i=1}^n (x_{i1} - \bar{x}_{\cdot 1})^2 = 2309.2$, $\sum_{i=1}^n (x_{i2} - \bar{x}_{\cdot 2})^2 = 9800.4$, $\sum_{i=1}^{30} (x_{i1} - \bar{x}_{\cdot 1})(x_{i2} - \bar{x}_{\cdot 2}) = 204.4$. Here $n = 30$, $\bar{x}_{\cdot j} = n^{-1} \sum_{i=1}^n x_{ij}$, $j = 1, 2$, and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$.

Write the model in the form (1) with $\mathbf{X}_1 = (1, \ldots, 1)^\top$ and

$$\mathbf{X}_2 = \begin{pmatrix} x_{11} - \bar{x}_{\cdot 1} & x_{12} - \bar{x}_{\cdot 2} \\ x_{21} - \bar{x}_{\cdot 1} & x_{22} - \bar{x}_{\cdot 2} \\ \vdots & \vdots \\ x_{n1} - \bar{x}_{\cdot 1} & x_{n2} - \bar{x}_{\cdot 2} \end{pmatrix}. \tag{6}$$

That is, $p_1 = 1$, $p_2 = 2$, and $k = 2$.

(i) Calculate $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

(ii) Let $\beta_2 = (\beta_{21}, \beta_{22})^\top$. Test the two-sided hypotheses $H_{01} : \beta_{21} = 0$ and $H_{02} : \beta_{22} = 0$. What do you conclude?

(iii) The following diagnostics are computed. Let $\mathbf{H} = [h_{ij}]_{i,j=1}^{30}$ be the hat matrix. The 30 diagonal values h_{ii} of $\mathbf{H}$ are

```
 [1] 0.11988922 0.05834409 0.05441139 0.04329834 0.04426311 0.23335081
 [7] 0.05412480 0.07538321 0.15539322 0.05653703 0.22390930 0.28184081
[13] 0.07378709 0.09070571 0.05286586 0.09623293 0.04718068 0.05292954
[19] 0.08476036 0.04018092 0.06126449 0.17467493 0.12319317 0.05940402
[25] 0.04473446 0.11528190 0.13849663 0.20402401 0.07108001 0.06845797
```

The values of the externally studentized residuals are

```
 0.46875921  0.79424610 -1.54423164 -0.15305068  0.19914681  1.66431746
-1.16755649 -0.04586624 -0.46479608  0.36124336 -0.12249977 -0.35573657
-0.25456885  0.29313257  1.28308925 -0.54395312  1.06428922  0.26464545
-1.20258139 -0.76206507 -0.13879668 -0.24710292 -2.77067342  0.65613355
 0.27234328  3.14658055  0.92988566 -0.31857324 -1.45304215  0.31074510
```

The values of DFFITS, defined as externally studentized residuals times $\sqrt{h_{ii}/(1 - h_{ii})}$, are

```
 0.17301000  0.19770033 -0.37043008 -0.03255987  0.04285734  0.91821099
-0.27929242 -0.01309633 -0.19936605  0.08843088 -0.06579837 -0.22285382
-0.07185220  0.09258255  0.30313688 -0.17749866  0.23682986  0.06256369
-0.36596829 -0.15592196 -0.03545781 -0.11367910 -1.03854827  0.16489162
 0.05893533  1.13583950  0.37283843 -0.16128721 -0.40194146  0.08423936
```

Comment on these values from the point of view of determining which observations are (1) point of high leverage (outlying X observations), (2) outlying Y observations, (3) influential.

4. A flower shop owner who grows her own flowers carried out an experiment to evaluate the impact of a vitamin solution on two different types of fiddle-leaf fig floor plants. The vitamin solution works by injection to the leaf. She grew 18 fiddle-leaf fig floor plants in 18 different pots, 6 with genotype G1 and 12 with genotype G2. The pots were arranged on a low bench in her greenhouse with a completely randomized design. On each fiddle-leaf fig floor plant, one leaf was selected for injection with the vitamin solution and another one was randomly selected for a pseudo control solution. At the end of a two week period, she used a device to record scores for the color of each leaf, with high values indicating healthy leaves and low values unhealthy leaves.

   Let y_{ijk} denote the score for genotype i ($i = 1$ for genotype G1, and $i = 2$ for genotype G2), solution j ($j = 1$ for control and $j = 2$ for the vitamin solution), and fiddle-leaf fig plant k ($k = 1, \cdots, 6$ for genotype G1 and $k = 7, \cdots, 18$ for genotype G2).

   Consider the following R output to answer parts **(a)-(d)**. The R output may contain more parts than you need to utilize.

```
> dat
      Plant Genotype Control   Trt
1         1       G1   97.09 88.72
2         2       G1   90.40 79.59
3         3       G1   85.94 75.98
4         4       G1   92.83 82.29
5         5       G1   91.90 83.45
6         6       G1   78.59 75.97
7         7       G2   82.46 77.03
8         8       G2   77.97 87.10
9         9       G2   89.34 80.48
10       10       G2   94.57 84.08
11       11       G2   83.93 86.66
12       12       G2   87.42 77.19
13       13       G2   71.75 68.59
14       14       G2   77.80 80.79
15       15       G2   77.29 81.43
16       16       G2   72.06 74.53
17       17       G2   75.78 81.71
18       18       G2   83.47 73.08

y <- as.vector(t(cbind(dat$Control,dat$Trt)))
genotype <- factor(rep(1:2, c(12,24)))
```

```
tGroup <- factor(rep(1:2, 18))
plant <- factor(rep(dat$Plant, each = 2))

y
 [1] 97.09 88.72 90.40 79.59 85.94 75.98
 [7] 92.83 82.29 91.90 83.45 78.59 75.97
[13] 82.46 77.03 77.97 87.10 89.34 80.48
[19] 94.57 84.08 83.93 86.66 87.42 77.19
[25] 71.75 68.59 77.80 80.79 77.29 81.43
[31] 72.06 74.53 75.78 81.71 83.47 73.08

genotype
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2
[21] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Levels: 1 2

tGroup
 [1] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
[21] 1 2 1 2 1 2 1 2 1 2 1 2 1 2 1 2
Levels: 1 2

> plant
 [1] 1  1  2  2  3  3  4  4  5  5  6  6  7
[14] 7  8  8  9  9  10 10 11 11 12 12 13 13
[27] 14 14 15 15 16 16 17 17 18 18
18 Levels: 1 2 3 4 5 6 7 8 9 10 11 ... 18

## Fit m1
m1 <- lmer(y ~ genotype*tGroup + (1|plant), REML = FALSE)

summary(m1)

Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: y ~ genotype * tGroup + (1 | plant)

     AIC      BIC   logLik deviance df.resid
   235.4    244.9   -111.7    223.4       30

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.36535 -0.65981 -0.06035  0.62692  1.80776
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 plant    (Intercept) 16.50    4.062
 Residual             16.88    4.108
Number of obs: 36, groups:  plant, 18

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         89.458      2.359   37.93
genotype2           -8.305      2.889   -2.88
tGroup2             -8.458      2.372   -3.57
genotype2:tGroup2    6.694      2.905    2.30
```

(a) Write down the model fitted in the m1 object in mathematical notation. State the underlying assumptions explicitly and specify the estimated values of the parameters in this model.

(b) What are the expected scores for a leaf (i) from genotype G1 injected with vitamin solution, (ii) from genotype G2 injected with control solution?

(c) What is the estimated correlation between scores of two leaves treated with the vitamin solution if (i) they are from different plants; (ii) they are from the same plant?

(d) Assume that the florist is given an estimate of the ratio of the covariance of two observations from the same plant to the error variance (i.e., ratio of the variance components of the model). Treating this ratio as fixed and known, obtain an alternative formulation in the form of $Y = \mathbf{X}\beta + \epsilon'$ to your model from part (a). Here, Y is the vector of scores (variable y in the R output), $\mathbf{X}$ is the corresponding design matrix, and β are the fixed effect parameters of model m1 denoting the sets of genotype main effect, solution effect, and the interaction effect. Explicitly state your assumptions on ϵ' and derive the best linear unbiased estimator of β.

Use the following set up for parts (e)-(g) and the R output below. The R output may contain more parts than you need to utilize.

The florist decides to consider the following treatment means model

$$y_{ijk} = \mu_{ij} + a_k + \epsilon_{ijk},$$

where μ_{11}, μ_{12}, μ_{21}, and μ_{22} are unknown parameters, a_k, $k = 1, \cdots, 18$ are independent and identically distributed (i.i.d) from $\mathcal{N}(0, \sigma_a^2)$. Furthermore, ϵ_{ijk}, $i = 1, 2$, $j = 1, 2$, and $k = 1, \cdots, 18$ are i.i.d. from $\mathcal{N}(0, \sigma_e^2)$, and are independent of ϵ_{ijk}, $\forall$ i, j, and k.

```
## Fit m2
meanY <- (dat$Control + dat$Trt)/2

m2 <- lm(meanY ~ 0 + dat$Genotype)

summary(m2)
Call:
lm(formula = meanY ~ 0 + dat$Genotype)

Residuals:
Min       1Q   Median       3Q      Max
-10.1013  -1.8787  -0.3802   2.4171   9.0537

Coefficients:
Estimate Std. Error t value Pr(>|t|)
dat$GenotypeG1   85.229      2.162   39.41   <2e-16 ***
dat$GenotypeG2   80.271      1.529   52.50   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.297 on 16 degrees of freedom
Multiple R-squared:  0.9963,Adjusted R-squared:  0.9958
F-statistic:  2155 on 2 and 16 DF,  p-value: < 2.2e-16

## Fit m3
diffY <- dat$Control - dat$Trt

m3 <- lm(diffY ~  1)

summary(m3)
Call:
lm(formula = diffY ~ 1)

Residuals:
Min      1Q  Median      3Q     Max
-13.126  -6.661   2.904   6.167   6.814
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.996      1.604   2.492   0.0233 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.803 on 17 degrees of freedom

## Fit m4
m4 <- lm(diffY ~ 0 + dat$Genotype)

summary(m4)
Call:
lm(formula = diffY ~ 0 + dat$Genotype)

Residuals:
Min       1Q   Median       3Q      Max
-10.8942  -4.6892   0.6937   3.3373   8.7258

Coefficients:
Estimate Std. Error t value Pr(>|t|)
dat$GenotypeG1    8.458      2.516   3.362  0.00397 **
dat$GenotypeG2    1.764      1.779   0.992  0.33612
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 6.163 on 16 degrees of freedom
Multiple R-squared:  0.4344,Adjusted R-squared:  0.3637
F-statistic: 6.143 on 2 and 16 DF,  p-value: 0.01048
```

(e) Test for the genotype main effect at significance level of 0.05.

(f) Test for the solution main effect at significance level of 0.05.

(g) Test for the genotype $\times$ solution interaction at significance level of 0.05.