Prelim Exam **Linear Models** **Spring 2017**

May 2017

Preliminary Examination: LINEAR MODELS

Answer all questions and show all work.

1. Consider a completely randomized experiment in which a total of 10 rats were randomly assigned to 5 treatment groups with 2 rats in each treatment group. Suppose the different treatments correspond to different doses of a drug in milliliters per gram of body weight as indicated in the following table.

Treatment	1	2	3	4	5
Dose of Drug (mL/g)	0	2	4	8	16

Suppose for $i = 1, \ldots, 5$, and $j = 1, 2$, Y_{ij} denotes the weight at the end of the study of the jth rat from the i treatment group. Further more, suppose

$$Y_{ij} = \mu_i + \epsilon_{ij},$$

where $\mu_1, \ldots, \mu_5$ are unknown parameters and the ϵ_{ij} terms are $i.i.d.$ $N(0, \sigma^2)$ for some unknown parameter $\sigma^2 > 0$. Use the R code and partial output provided with this exam to give numerical answer the following questions.

a. Provide the BLUE of μ_2.

b. Determine the standard error of the BLUE of μ_2.

c. Conduct a test for $H_0 : \mu_1 = \mu_2$ vs. $H_a : \mu_1 \neq \mu_2$. Provide a test statistic, the distribution of the test statistic under H_0 and H_a, respectively.

d. Does a simple linear regression model with body weight as a response and dose as a *quantitative* explanatory variable fit these data adequately? Provide a matrix $\mathbf{A}$ and a vector $\mathbf{c}$ so that the null hypothesis of the test can be written as $H_0 : \mathbf{A}'\boldsymbol{\beta} = \mathbf{c}$, where $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_5)'$. Also provide a test statistic, its distribution under H_0.

```
> plot(d,y)  #See plot on the back of this page.

> dose=as.factor(d)
> o1=lm(y~dose)
> summary(o1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  351.000      6.576  53.372 4.37e-08 ***
dose2        -10.000      9.301  -1.075 0.331406
dose4         -6.000      9.301  -0.645 0.547277
dose8        -17.000      9.301  -1.828 0.127119
dose16       -70.500      9.301  -7.580 0.000634 ***

> anova(o1)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
dose         6505.6
Residuals     432.5

> is.numeric(d)
[1] TRUE

> o2=lm(y~d)
> anova(o2)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
d            5899.6
Residuals    1038.5

> o3=lm(y~d+dose)
> anova(o3)

Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
d                                   0.0004245 ***
dose                                0.1907591
Residuals
```

2. Let $Y_0, Y_1, \ldots Y_n$ follow a general linear model,

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta_i; i = 0, 1, \ldots, n$$

where the vector $\boldsymbol{\beta}$ is p-dimensional; $\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_n$ are known covariates, and $\delta_0, \delta_1, \ldots, \delta_n$ are mean-zero error terms. Write $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)'$, $\mathbf{X} \equiv (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$, and $\boldsymbol{\delta} \equiv (\delta_1, \ldots, \delta_n)'$. Notice that Y_0 is NOT part of $\mathbf{Y}$ and plays a special role in this question.

Define

$$\Sigma_{YY} = var(\mathbf{Y}), \; \boldsymbol{\sigma}_{Y_0} = cov(\mathbf{Y}, Y_0), \sigma_{00} = var(Y_0).$$

The parameters $\boldsymbol{\beta}$ are unknown, but assume that Σ_{YY}, $\boldsymbol{\sigma}_{Y_0}$, and σ_{00} are known. Suppose that $\mathbf{Y}$ is observed with measurement error and the data are:

$$\mathbf{Z} = \mathbf{Y} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ is independent of $\mathbf{Y}$, and $\epsilon_1, \ldots, \epsilon_n$ are *i.i.d.* random variables with mean zero and known variance σ_e^2. Notice that there is NO observation Z_0. We wish to make inference on Y_0 and $\boldsymbol{\beta}$ based on the data $\mathbf{Z}$ using only first- and second-moment assumptions. (In the case of Y_0 we call the inference prediction, and in the case of $\boldsymbol{\beta}$ we call the inference estimation.)

a. Derive the expression for $\Sigma_{ZZ} \equiv var(\mathbf{Z})$ in terms of the parameters defined above.

b. We first consider inference on $\boldsymbol{\beta}$ (i.e., estimation).

b-i. Give the Generalized Least Squares (GLS) estimator $\hat{\boldsymbol{\beta}}_{GLS}$. [Assume that all necessary matrix inverses exist.]

b-ii. Consider any linear estimator,

$$\hat{\boldsymbol{\beta}}(\mathbf{A}) \equiv \mathbf{A}\mathbf{Z},$$

where $\mathbf{A}$ is a $p \times n$ matrix. Now define the $p \times p$ matrix,

$$Q(\mathbf{A}) \equiv E\{(\boldsymbol{\beta} - \mathbf{A}\mathbf{Z})(\boldsymbol{\beta} - \mathbf{A}\mathbf{Z})'\}.$$

We wish to minimize $\mathbf{k}'Q(\mathbf{A})\mathbf{k}$ with respect to $\mathbf{A}$ for any given $\mathbf{k}$, where $\mathbf{A}$ is restricted to satisfy $E(\hat{\boldsymbol{\beta}}(\mathbf{A})) = \boldsymbol{\beta}$. Show that $\hat{\boldsymbol{\beta}}_{GLS}$ is a solution to this optimization problem. [Hint: Minimize $\mathbf{k}'Q(\mathbf{A})\mathbf{k}$ with respect to $\mathbf{a} \equiv \mathbf{A}'\mathbf{k}$.]

c. We now consider inference on Y_0 (i.e., prediction)

c-i. Assume $\boldsymbol{\beta}$ is known in this part (this assumption will be relaxed in the next part). Define the loss function,

$$L(Y_0, \hat{Y}_0) = (\hat{Y}_0 - Y_0)^2.$$

The optimal linear predictor is obtained by minimizing $E(L(Y_0, \hat{Y}_0))$; you may assume that it is given by,

$$\hat{Y}_0(\boldsymbol{\beta}) = \mathbf{x}_0'\boldsymbol{\beta} + \boldsymbol{\sigma}_{Y_0}'\Sigma_{ZZ}^{-1}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}).$$

Give an expression for the mean squared prediction error,

$$M_1(\boldsymbol{\beta}) \equiv E(\hat{Y}_0(\boldsymbol{\beta}) - Y_0)^2.$$

How does $M_1(\boldsymbol{\beta})$ vary with $\boldsymbol{\beta}$?

c-ii. In reality, $\boldsymbol{\beta}$ is unknown. One way to deal with this is to "plug in" an estimator, $\hat{\boldsymbol{\beta}}(\mathbf{A}) \equiv \mathbf{AZ}$, to yield the estimated predictor, $\hat{Y}_0(\hat{\boldsymbol{\beta}}(\mathbf{A}))$, and the estimated mean squared prediction error, $M(\hat{\boldsymbol{\beta}}(\mathbf{A}))$. Show that $\hat{Y}_0(\hat{\boldsymbol{\beta}}(\mathbf{A}))$ is of the form,

$$\mathrm{b}(\mathbf{A})'\mathbf{Z},$$

and give a formula for $\mathrm{b}(\mathbf{A})$. Also give a formula for $M_1(\hat{\boldsymbol{\beta}}(\mathbf{A}))$.

d. Now, $M_1(\hat{\boldsymbol{\beta}}(\mathbf{A}))$ does not account for the variability in $\hat{\boldsymbol{\beta}}(\mathbf{A})$. From (c)-ii, recall that $\hat{Y}_0(\hat{\boldsymbol{\beta}}(\mathbf{A})) = \mathrm{b}(\mathbf{A})'\mathbf{Z}$, where $\hat{\boldsymbol{\beta}}(\mathbf{A}) = \mathbf{AZ}$. It is not hard to show (but you do NOT have to show it here) that any linear unbiased predictor can be written as $\hat{Y}_0(\hat{\boldsymbol{\beta}}(\mathbf{A}))$ where $\hat{\boldsymbol{\beta}}(\mathbf{A})$ is an unbiased estimator of $\boldsymbol{\beta}$. You may assume that (*no need to prove*)

$$\begin{aligned} M_2(\mathbf{A}) &\equiv E(\mathrm{b}(\mathbf{A})'\mathbf{Z} - Y_0)^2 \\ &= \sigma_{00} - \boldsymbol{\sigma}_{Y0}'\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\sigma}_{Y0} + (\mathrm{x}_0 - \mathbf{X}'\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\sigma}_{Y0})'Q(\mathbf{A})(\mathrm{x}_0 - \mathbf{X}'\boldsymbol{\Sigma}_{ZZ}^{-1}\boldsymbol{\sigma}_{Y0}). \end{aligned}$$

d-i. Use this to find the Best Linear Unbiased Predictor (BLUP), Y_0^*, of Y_0.

d-ii. Derive the expression for $M^* \equiv E[(Y_0^* - Y_0)^2]$. Compare the two expressions, $M_1(\hat{\boldsymbol{\beta}}_{GLS})$ and M^*.

3. In *regularized* regression on p predictors, we seek an estimator $\hat{\boldsymbol{\beta}}_R(\boldsymbol{\Lambda})$ in the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \;\; E(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}, \;\; var(\boldsymbol{\varepsilon}|\mathbf{X}) = \sigma^2\mathbf{I}_n,$$

that minimizes,

$$SS_{\boldsymbol{\Lambda}}(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}'\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\boldsymbol{\beta}.$$

a. Show that the least-squares estimator for $\boldsymbol{\beta}$ in the augmented model (with artificial observations)

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \boldsymbol{\Lambda} \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \tilde{\boldsymbol{\varepsilon}} \end{bmatrix}$$

is $\hat{\boldsymbol{\beta}}_R(\boldsymbol{\Lambda}) \equiv \underset{\boldsymbol{\beta}}{argmin}\, SS_{\boldsymbol{\Lambda}}(\boldsymbol{\beta})$ and give the expression for it.

For parts (b)-(d), assume that $\mathbf{X} \equiv [\mathrm{x}_1, \cdots, \mathrm{x}_p]$ is *orthogonal* and that we now focus on one regularizing parameter λ and set $\boldsymbol{\Lambda} = \sqrt{\lambda}\,\mathbf{I}_p$. Denote $\hat{\boldsymbol{\beta}}_R(\lambda) \equiv \hat{\boldsymbol{\beta}}_R(\boldsymbol{\Lambda})$.

b. If $\hat{\beta}$ is the least-squares estimate for the model, show that

$$(\hat{\boldsymbol{\beta}}_R(\boldsymbol{\Lambda}))_i = s_i(\lambda)\hat{\beta}_i,$$

for $i = 1, \ldots, p$ and provide the expression for $s_i(\lambda)$, called the i-th *shrinkage factor.*

c. Define the hat matrix $\mathbf{H}(\lambda)$ with respect to $\hat{\beta}_R(\lambda)$ and show that $\mathbf{H}(\lambda)$ is symmetric but *not*, in general, idempotent.

d. Find the *effective degrees of freedom*, defined as the trace of $\mathbf{H}(\lambda)$,

$$df(\lambda) \equiv tr(\mathbf{H}(\lambda)).$$

Verify that $df(0) = p$ (no regularization), and describe the behavior of $df(\lambda)$ as $\lambda \to \infty$. [Hint: Recall that $tr(AB) = tr(BA)$.]

4. Consider the following model:

$$y_{ij} = \mu + \alpha_i + e_{ij}\ , i = 1, \ldots, a\ , j = 1, \ldots, n$$

where μ is the population mean of the response, α_i denotes the effect of the ith randomly selected treatment and is assumed to be distributed $i.i.d.$ $N(0, \sigma_\alpha^2)$, and e_{ij} denotes the random error and is distributed $i.i.d.$ $N(0, \sigma^2)$. It is also assumed that α_i and e_{ij} are independent random variables. Let $\mathbf{y} = (y_{11}, \ldots, y_{an})^T$.
Let $SST = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \overline{y}_{..})^2$, $SSA = n \sum_{i=1}^a (y_{i.} - \overline{y}_{..})^2$ and $SSE = \sum_{i=1}^a \sum_{j=1}^n (y_{ij} - \overline{y}_{i.})^2$.

a. Show that the log likelihood function of $(\mu, \sigma_\alpha^2, \sigma^2)$ is

$$-\frac{N}{2}\log(2\pi) - \frac{N-a}{2}\log\sigma^2 - \frac{a}{2}\log(\sigma^2 + n\sigma_\alpha^2) - \frac{SSE}{2\sigma^2} - \frac{SSA}{2(\sigma^2 + n\sigma_\alpha^2)} - \frac{N(\overline{y}_{..} - \mu)^2}{2(\sigma^2 + n\sigma_\alpha^2)},$$

where $N = an$, $\mu \in \mathbf{R}$, $\sigma_\alpha^2 \geq 0$ and $\sigma^2 \geq 0$.

b. Derive the REML (Restricted Maximum Likelihood) estimators of $(\sigma_\alpha^2, \sigma^2)$.