STAT 511 Exam 1 Spring 2012

1. Suppose $\boldsymbol{X}$ is an $n \times p$ design matrix. Prove that $\mathcal{C}(\boldsymbol{X}) = \mathcal{C}(\boldsymbol{P_X})$.

2. Consider a competition among 5 table tennis players labeled 1 through 5. For $1 \leq i < j \leq 5$, define y_{ij} to be the score for player i minus the score for player j when player i plays a game against player j. Suppose for $1 \leq i < j \leq 5$,

$$y_{ij} = \beta_i - \beta_j + \epsilon_{ij}, \tag{1}$$

where $\beta_1, \ldots, \beta_5$ are unknown parameters and the ϵ_{ij} terms are random errors with mean 0. Suppose four games will be played that will allow us to observe y_{12}, y_{34}, y_{25}, and y_{15}. Let

$$\boldsymbol{y} = \begin{bmatrix} y_{12} \\ y_{34} \\ y_{25} \\ y_{15} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}, \text{ and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{12} \\ \epsilon_{34} \\ \epsilon_{25} \\ \epsilon_{15} \end{bmatrix}.$$

   (a) Define a design matrix $\boldsymbol{X}$ so that model (1) may be written as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$.

   (b) Is $\beta_1 - \beta_2$ estimable? Prove that your answer is correct.

   (c) Is $\beta_1 - \beta_3$ estimable? Prove that your answer is correct.

   (d) Find a generalized inverse of $\boldsymbol{X}'\boldsymbol{X}$.

   (e) Write down a general expression for the normal equations.

   (f) Find a solution to the normal equations in this particular problem involving table tennis players.

   (g) Find the Ordinary Least Squares (OLS) estimator of $\beta_1 - \beta_5$.

   (h) What must we assume about $\boldsymbol{\epsilon}$ in order for the OLS estimator of $\beta_1 - \beta_5$ to be unbiased?

   (i) What must we assume about $\boldsymbol{\epsilon}$ in order for the OLS estimator of $\beta_1 - \beta_5$ to have the smallest variance among all linear unbiased estimators?

   (j) Give a linear unbiased estimator of $\beta_1 - \beta_5$ that is not the OLS estimator.

3. Suppose $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ for some unknown $\sigma^2 > 0$. Let $\hat{\boldsymbol{y}} = \boldsymbol{P_X y}$.

   (a) Determine the distribution of

$$\begin{bmatrix} \hat{\boldsymbol{y}} \\ \boldsymbol{y} - \hat{\boldsymbol{y}} \end{bmatrix}.$$

   (b) Determine the distribution of $\hat{\boldsymbol{y}}'\hat{\boldsymbol{y}}$.

4. 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. Furthermore, suppose

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

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

(a) Provide the BLUE of μ_1.

(b) Provide the BLUE of μ_2.

(c) Determine the standard error of the BLUE of μ_2.

(d) Conduct a test of $H_0 : \mu_1 = \mu_2$. Provide a test statistic, the distribution of that test statistic (be very precise), a p-value, and a conclusion.

(e) Provide an F-statistic for testing $H_0 : \mu_3 = \mu_4$.

(f) 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 test statistic, its degrees of freedom, a p-value, and a conclusion.

(g) Provide a matrix $\boldsymbol{C}$ and a vector $\boldsymbol{d}$ so that the null hypothesis of the test in part (f) may be written as $H_0 : \boldsymbol{C\beta} = \boldsymbol{d}$, where $\boldsymbol{\beta} = (\mu_1, \ldots, \mu_5)'$.

(h) Fill in the missing entries in the ANOVA table produced by the R command `anova(o3)`. (This is the last R command in the provided code.)

```
> 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
```

There are actually two data points
here that are plotted on top of each other.