Biost/Stat 533
Spring 2008

Homework Key #4

1. (a) Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from a distribution with mean θ and finite variance σ^2. Find a BLUE of θ (and justify that it is, in fact, the Best Unbiased Linear Estimate).
(b) Explain the statement in lecture notes 8 that $RSS/(n-p)$ is a generalization of the sample variance.

Solution:(a) The linear model here is $\mathbf{Y} = \mathbf{1_n}\theta + \epsilon$ where $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$, $\mathbf{1_n} \in \mathbb{R}^n$ is the vector whose elements are all 1, and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)^T$ with $E[\epsilon] = 0$ and $Var(\epsilon) = \sigma^2\mathbf{I}$. The least squares estimate $\hat{\theta}$ of θ is

$$\hat{\theta} \;=\; (\mathbf{1_n}^T\mathbf{1_n})^{-1}\mathbf{1_n}^T\mathbf{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \equiv \overline{Y}.$$

Let $a = 1$. Because $\mathbf{1_n}$ is of full rank, $\hat{\theta} = a\hat{\theta}$ is the BLUE of $\theta = a\theta$ by applying the Corrolary in the page 4 of the lecture note 8.

Another solution for (a): Let $\eta = \mathbf{1_n}\theta$ (note that this η plays the same role as $\theta \equiv X\beta$ in the page 2 of the lecture note 8). Also let $\mathbf{c} = \frac{1}{n}\mathbf{1_n}$. Then $\hat{\eta} = \mathbf{1_n}\overline{Y} = (\overline{Y}, \ldots, \overline{Y})$ and $\mathbf{c}^T\eta = \theta$. Thus applying the Theorem in the page 2 of the lecture note 8, the estimate $\mathbf{c}^T\hat{\eta} = \overline{Y}$ is the BLUE of $\mathbf{c}^T\eta = \theta$.
(b) In part (a) the elementary statistics gives the sample variance $\frac{\Sigma(Y_i - \overline{Y})^2}{n-1}$. This is exactly the same as $RSS/(n-1)$ computed in the context of linear model theory. Recall that this sample variance is the unbiased estimate of σ^2. Furthermore note that under the normality assumption of Y_i, this is independent of the mean of data, which is the least squares estimate of the parameter in the linear model context and RSS/σ^2 is distributed as χ^2_{n-1}. Similar properties for $RSS/(n-p)$ in other linear models hold and hence this is the generalization of the sample variance.

We will complete the calculations for the simple one-way ANOVA model. By considering both $r = p$ and $r < p$, one can understand more clearly the properties of $\hat{\beta}$ and $\hat{\theta}$.
2. To begin, consider the full rank model version of the model:

$$\mathbf{Y}^{2J\times 1} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1J} \\ Y_{21} \\ \vdots \\ Y_{2J} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \\ 1 & 0 \\ 0 & 1 \\ \vdots & \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1J} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2J} \end{pmatrix} = \mathbf{X}^{2J\times 2}\beta^{2\times 1} + \epsilon^{2J\times 1}$$

(a) Interpret the model parameters.
(b) Compute $\hat{\beta}$. Is it unique (yes/no)? Explain.
(c) Compute $\mathbf{X}\hat{\beta} \equiv \hat{\theta}$. Is it unique (yes/no)? Explain.
(d) Compute the hat matrix (always unique) $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Solution:(a) The parameter μ_1 is the mean of the group 1 and the parameter μ_2 is the mean of the group 2.
(b) The estimate $\hat{\beta}$ is unique since the model has full rank and computed as follows.

$$\begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} J & 0 \\ 0 & J \end{pmatrix}^{-1} \begin{pmatrix} \Sigma Y_{1j} \\ \Sigma Y_{2j} \end{pmatrix} = \begin{pmatrix} \overline{Y}_1 \\ \overline{Y}_2 \end{pmatrix}$$

where $\overline{Y}_i$ is the sample mean of the group i.
(c) We have

$$\hat{\theta} = \mathbf{X}\hat{\beta} = \begin{pmatrix} \overline{Y}_1 \\ \vdots \\ \overline{Y}_1 \\ \overline{Y}_2 \\ \vdots \\ \overline{Y}_2 \end{pmatrix}$$

Because the hat matrix $\mathbf{P}$ is unique and $\mathbf{PY} = \mathbf{X}\hat{\beta}$ for any least squares estimate $\hat{\beta}$ of β, $\mathbf{X}\hat{\beta}$ is unique.
(d) We have

$$\begin{aligned} P &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X} \\ &= \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{pmatrix} J & 0 \\ 0 & J \end{pmatrix}^{-1} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix} \\ &= \frac{1}{J} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}. \end{aligned}$$

3. Now consider a non-full rank version of the model:

$$\mathbf{Y}^{2J\times 1} = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1J} \\ Y_{21} \\ \vdots \\ Y_{2J} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & & \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & & \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1J} \\ \epsilon_{21} \\ \vdots \\ \epsilon_{2J} \end{pmatrix} = \mathbf{X}^{2J\times 3}\beta^{3\times 1} + \epsilon^{2J\times 1}$$

(a) Interpret the model parameters.
(b) Recall that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y}$. There are infinite number of generalized inverses for $\mathbf{X}^T\mathbf{X}$. Here are two of them:

$$(\mathbf{X}^T\mathbf{X})_1^{-} = J^{-1}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad (\mathbf{X}^T\mathbf{X})_2^{-} = J^{-1}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Use $(\mathbf{X}^T\mathbf{X})_1^{-}$ and $(\mathbf{X}^T\mathbf{X})_2^{-}$ to compute $\hat{\beta}$. Is $\hat{\beta}$ unique (yes/no)? Explain.
(c) Compute $\mathbf{X}\hat{\beta} \equiv \hat{\theta}$ for both case in (a). Is it unique (yes/no)? Compare your results with Problem 2(c).
(d) Compute the hat matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ and compare your results with Problem 2(d).

Solution: (a) The parameter μ is the overall mean and α_i is the difference between the mean of the group i and the overall mean.
(b) Since we have

$$\mathbf{X}^T\mathbf{Y} = \begin{pmatrix} \Sigma Y_{1j} + \Sigma Y_{2j} \\ \Sigma Y_{1j} \\ \Sigma Y_{2j} \end{pmatrix},$$

estimates of β are

$$\begin{aligned} \hat{\beta}_1 &\equiv (\mathbf{X}^T\mathbf{X})_1^{-}\mathbf{X}^T\mathbf{Y} \\ &= \frac{1}{J}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} \Sigma Y_{1j} + \Sigma Y_{2j} \\ \Sigma Y_{1j} \\ \Sigma Y_{2j} \end{pmatrix} \\ &= \frac{1}{J}\begin{pmatrix} \Sigma Y_{2j} \\ \Sigma Y_{1j} - \Sigma Y_{2j} \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} \overline{Y_{2\cdot}} \\ \overline{Y_{1\cdot}} - \overline{Y_{2\cdot}} \\ 0 \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned}
\hat{\beta}_2 &\equiv (\mathbf{X}^T\mathbf{X})_2^{-}\mathbf{X}^T\mathbf{Y} \\
&= \frac{1}{J}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} \Sigma Y_{1j} - \Sigma Y_{2j} \\ \Sigma Y_{1j} \\ \Sigma Y_{2j} \end{pmatrix} \\
&= \frac{1}{J}\begin{pmatrix} 0 \\ \Sigma Y_{1j} \\ \Sigma Y_{2j} \end{pmatrix} \\
&= \begin{pmatrix} 0 \\ \overline{Y}_{1\cdot} \\ \overline{Y}_{2\cdot} \end{pmatrix}.
\end{aligned}$$

Clearly this computation shows that $\hat{\beta}$ is not unique.

(c) We have

$$\hat{\theta}_1 \equiv \mathbf{X}\hat{\beta}_1 = \begin{pmatrix} \overline{Y}_{1\cdot} \\ \vdots \\ \overline{Y}_{1\cdot} \\ \overline{Y}_{2\cdot} \\ \vdots \\ \overline{Y}_{2\cdot} \end{pmatrix} \quad \text{and} \quad \hat{\theta}_2 \equiv \mathbf{X}\hat{\beta}_2 = \begin{pmatrix} \overline{Y}_{1\cdot} \\ \vdots \\ \overline{Y}_{1\cdot} \\ \overline{Y}_{2\cdot} \\ \vdots \\ \overline{Y}_{2\cdot} \end{pmatrix}.$$

Using the same reasoning in Problem 2(c), $\hat{\theta}$ is unique as is more clear in part (d). Moreover this $\hat{\theta}$ is the same as $\hat{\theta}$ in Problem 2(c).

(d) We have

$$\begin{aligned}
\mathbf{P}_1 &\equiv \mathbf{X}(\mathbf{X}^T\mathbf{X})_1^{-}\mathbf{X}^T \\
&= \frac{1}{J}\begin{pmatrix} 1 & 1 & 0 \\ & \vdots & \vdots \\ \vdots & 1 & 0 \\ & 0 & 1 \\ & \vdots & \vdots \\ 1 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & 0 \\ 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}
\end{aligned}$$

$$
= \frac{1}{J} \begin{pmatrix} 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & \vdots \\ 1 & -1 & \vdots \\ \vdots & \vdots & \vdots \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & \cdots & \cdots & \cdots & \cdots & 1 \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}
$$

$$
= \frac{1}{J} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}
$$

and

$$
\begin{aligned}
\mathbf{P}_2 &\equiv \mathbf{X}(\mathbf{X}^T\mathbf{X})_2^- \mathbf{X}^T \\
&= \frac{1}{J}\mathbf{X} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{X}^T
\end{aligned}
$$

$$
= \frac{1}{J} \begin{pmatrix} 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ \vdots & 1 & 0 \\ \vdots & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \mathbf{X}^T
$$

$$
= \frac{1}{J} \begin{pmatrix} 1 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ 1 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 1 \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 1 & \cdots & 1 \end{pmatrix}
$$

Hence, we have

$$
\mathbf{P}_1 = \mathbf{P}_2 = \mathbf{P} \text{ from 2(d).}
$$

4. Let $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, (i = 1, \cdots, n)$ where $E[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2 \mathbf{I}$. Prove that the least squares estimates of β_0 and β_1 are uncorrelated if and only if $\overline{x} = 0$.

Solution: Let

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

We can assume that the design matrix $\mathbf{X}$ in this problem is of full rank because the question implicitly assumes so by saying "the" least squares estimate. Practically, if the design matrix is not of full rank, then we have the same value of x_i for each i and fitting to a linear model is not reasonable.

Thus, we have

$$\begin{aligned} \text{var}(\hat{\beta}) &= \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \\ &= \sigma^2 \begin{pmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{pmatrix}^{-1} \\ &= \frac{\sigma^2}{n \Sigma x_i^2 - (\Sigma x_i)^2} \begin{pmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{pmatrix} \end{aligned}$$

Hence, $\hat{\beta}_0$ and $\hat{\beta}_1$ are uncorrelated if and only if

$$\frac{-\Sigma x_i}{n \Sigma x_i^2 - (\Sigma x_i)^2} = 0 \;\Leftrightarrow\; \Sigma x_i = 0 \;\Leftrightarrow \overline{x} = 0$$

as desired.

5. (a) Let

$$Y_i = \beta_0 + \beta_1 (x_{i1} - \overline{x}_1) + \beta_2 (x_{i2} - \overline{x}_2) + \epsilon_i \quad i = 1, 2, \cdots, n$$

where $\overline{x}_j = \Sigma_{i=1}^n x_{ij}/n, E[\epsilon] = 0$ and $Var[\epsilon] = \sigma^2 \mathbf{I}$. If $\hat{\beta}_1$ is the least squares estimate of β_1 then show that

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma_{i=1}^n (x_{i1} - \overline{x}_1)^2 (1 - r_{12}^2)}$$

where r_{12} is the correlation coefficient of pairs (x_{i1}, x_{i2}).

(b) Comment on the impact of the highly correlated predictors of x_1 and x_2 in linear model.

Solution: (a) From the design matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} - \overline{x}_1 & x_{n2} - \overline{x}_2 \end{pmatrix}$$

we compute

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 & 0 \\ 0 & \Sigma(x_{i1}-\overline{x}_1)^2 & \Sigma(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2) \\ 0 & \Sigma(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2) & \Sigma(x_{i2}-\overline{x}_2)^2 \end{pmatrix}.$$

Notice that $\mathbf{X}'\mathbf{X}$ is block diagonal so that the corresponding diagonal block in the inverse is the inverse of the diagonal block. Let A be the lower diagonal block of $\mathbf{X}'\mathbf{X}$. This is only a 2×2 matrix so the first element of its inverse is given by

$$\begin{aligned} Var(\hat{\beta}_1) &= \sigma^2 \frac{\Sigma(x_{i2}-\overline{x}_2)^2}{|A|} \\ &= \sigma^2 \frac{\Sigma(x_{i2}-\overline{x}_2)^2}{\Sigma(x_{i1}-\overline{x}_1)^2\Sigma(x_{i2}-\overline{x}_2)^2 - (\Sigma(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} \\ &= \frac{\sigma^2}{\Sigma(x_{i1}-\overline{x}_1)^2 - \frac{(\Sigma(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2}{\Sigma(x_{i2}-\overline{x}_2)^2}} \\ &= \frac{\sigma^2}{\Sigma(x_{1i}-\overline{x}_1)^2 - \Sigma(x_{1i}-\overline{x}_1)^2 \frac{(\Sigma(x_{1i}-\overline{x}_1)(x_{2i}-\overline{x}_2))^2}{\Sigma(x_{1i}-\overline{x}_1)^2\Sigma(x_{2i}-\overline{x}_2)^2}} \\ &= \frac{\sigma^2}{\Sigma(x_{i1}-\overline{x}_1)^2(1-r_{12}^2)}. \end{aligned}$$

(b) If the predictors of x_1 and x_2 are highly correlated, r_{12}^2 is close to 1 and $\text{var}(\hat{\beta}_1)$ will be inflated, which is easily seen in the formula derived above.

6. Suppose that $\mathbf{Y} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I})$ where $\mathbf{X}$ is $n \times p$ of rank p.
(a) Find $\text{var}(S^2)$.
(b) Evaluate $E[(\mathbf{Y}^T\mathbf{A}\mathbf{Y} - \sigma^2)^2]$ for $A = \frac{1}{n-p+2}(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$.
(c) Prove that S^2 does not have minimum mean squared error among estimates of σ^2.

Solution: (a) Because $RSS/\sigma^2 = (n-p)S^2/\sigma^2 \sim \chi_{n-p}$, we have

$$\text{var}(S^2) = \text{var}\left(\frac{\sigma^2}{n-p}\frac{(n-p)S^2}{\sigma^2})\right) = \frac{\sigma^4}{(n-p)^2}\text{var}\left(\frac{(n-p)S^2}{\sigma^2})\right) = \frac{\sigma^4}{(n-p)^2}2(n-p) = \frac{2\sigma^4}{(n-p)}.$$

(b) Because $RSS = \mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}$ where $\mathbf{P}$ is the hat matrix, we have

$$\begin{aligned} \mathbf{Y}^T\mathbf{A}\mathbf{Y} &= \frac{1}{n-p+2}\mathbf{Y}^T(\mathbf{I}-\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y} \\ &= \frac{1}{n-p+2}\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y} \\ &= \frac{1}{n-p+2}RSS \\ &= \frac{\sigma^2}{n-p+2}\frac{RSS}{\sigma^2} \end{aligned}$$

so that

$$E[\mathbf{Y}^T\mathbf{A}\mathbf{Y}] = \frac{\sigma^2(n-p)}{n-p+2}.$$

Thus, we have

$$\begin{aligned}
E[(\mathbf{Y}^T\mathbf{A}\mathbf{Y}-\sigma^2)^2] &= E\left[\left(\frac{\sigma^2}{n-p+2}\frac{RSS}{\sigma^2} - \frac{n-p}{n-p+2}\sigma^2 - \frac{2}{n-p+2}\sigma^2\right)^2\right] \\
&= E[\left(\frac{\sigma^2}{n-p+2}\left(\frac{RSS}{\sigma^2}-(n-p)\right)\right)^2 \\
&\quad -2\left(\frac{\sigma^2}{n-p+2}\left(\frac{RSS}{\sigma^2}-(n-p)\right)\right)\frac{2}{n-p+2}\sigma^2 \\
&\quad +\left(\frac{2}{n-p+2}\sigma^2\right)^2] \\
&= \frac{\sigma^4}{(n-p+2)^2}\mathrm{var}\left(\frac{RSS}{\sigma^2}\right) - \frac{4\sigma^2}{(n-p+2)^2}E\left[\frac{RSS}{\sigma^2}-(n-p)\right] \\
&\quad +\frac{4}{(n-p+2)^2}\sigma^4 \\
&= \frac{\sigma^4}{(n-p+2)^2}2(n-p)+0+\frac{4\sigma^4}{(n-p+2)^2} \\
&= \frac{2\sigma^4}{n-p+2}.
\end{aligned}$$

(c) The mean squared error $MSE(S^2)$ of S^2 is greater than the mean squared error $MSE(\mathbf{Y}^T\mathbf{A}\mathbf{Y})$ of $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$ because

$$MSE(S^2) = \mathrm{var}(S^2) = \frac{2\sigma^4}{n-p} > \frac{2\sigma^4}{n-p+2} = MSE(\mathbf{Y}^T\mathbf{A}\mathbf{Y}).$$

Hence the claim follows.