STAT 741 **Solutions to Assignment 1**

1. Let $\mathbf{Y}_k$, $k = 1, \ldots, K$ be independent random vectors of dimension n_k satisfying linear models $\mathbf{Y}_k = \mathbf{X}_k\boldsymbol{\beta} + \boldsymbol{\varepsilon}_k$, where $\boldsymbol{\varepsilon}_k$ has mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$. Note that the parameter vector $\boldsymbol{\beta}$ is the same for each k. Prove that any set $\widehat{\boldsymbol{\beta}}$ of least squares estimates from the complete set of $n = \sum_k n_k$ observations satisfies

$$\left(\sum_k \mathbf{X}_k^T\mathbf{X}_k\right)\widehat{\boldsymbol{\beta}} = \sum_k \mathbf{X}_k^T\mathbf{X}_k\widehat{\boldsymbol{\beta}}_k,$$

where $\widehat{\boldsymbol{\beta}}_k$ is any set of least squares estimates based on $\mathbf{Y}_k$ and $\mathbf{X}_k$ alone.

[*Hint:* Partition the matrix $\mathbf{X}$ and the observation vector $\mathbf{Y}$ of the combined data set into those of the K component data sets.]

Solution: We can write

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_K \end{bmatrix}.$$

Then

$$\mathbf{X}^T\mathbf{X} = [\mathbf{X}_1^T \ \cdots \ \mathbf{X}_K^T] \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_K \end{bmatrix} = \sum_{k=1}^K \mathbf{X}_k^T\mathbf{X}_k,$$

and similarly $\mathbf{X}^T\mathbf{Y} = \sum_{k=1}^K \mathbf{X}_k^T\mathbf{Y}_k$. The least squares estimates satisfy the normal equations

$$\mathbf{X}^T\mathbf{X}\widehat{\boldsymbol{\beta}} = \left(\sum_{k=1}^K \mathbf{X}_k^T\mathbf{X}_k\right)\widehat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{Y} = \sum_{k=1}^K \mathbf{X}_k^T\mathbf{Y}_k$$

and

$$\mathbf{X}_k^T\mathbf{X}_k\widehat{\boldsymbol{\beta}}_k = \mathbf{X}_k^T\mathbf{Y}_k \qquad k = 1, \ldots, K.$$

Sum the latter system of equations to obtain the result.

2. Suppose one has data (x_{ij}, Y_{ij}), $i = 1, 2$, $j = 1, \ldots, n_i$, and one fits parallel linear regression lines

$$Y_{ij} = \alpha_i + \beta(x_{ij} - \bar{x}_{i\cdot}) + \varepsilon_{ij}$$

under the assumption that the ε_{ij} are i.i.d. $N(0, \sigma^2)$.

(a) Set up and solve the least squares equations for the parameters.

(b) Show how to test the null hypothesis that the two lines are identical. Give the distribution of your test statistic under the null hypothesis.

Solution: (a) Let $\mathcal{S} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} [Y_{ij} - \alpha_i - \beta(x_{ij} - \bar{x}_{i\cdot})]^2$. The normal equations are

$$\begin{aligned} \frac{\partial \mathcal{S}}{\partial \alpha_i} &= 0 = -2 \sum_{j=1}^{n_i} (Y_{ij} - \alpha_i - \beta(x_{ij} - \bar{x}_{i\cdot})) = -2 \sum_{j=1}^{n_i} (Y_{ij} - \alpha_i), \qquad i = 1, 2, \\ \frac{\partial \mathcal{S}}{\partial \beta} &= 0 = -2 \sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})(Y_{ij} - \alpha_i - \beta(x_{ij} - \bar{x}_{i\cdot})). \end{aligned}$$

Therefore, $\hat{\alpha}_i = \bar{Y}_{i\cdot}$, $i = 1, 2$, and

$$\hat{\beta} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})(Y_{ij} - \bar{Y}_{i\cdot})}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2} = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot}) Y_{ij}}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2}.$$

The unknown variance is estimated by

$$s^2 = \frac{1}{n_1 + n_2 - 3} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\cdot} - \hat{\beta}(x_{ij} - \bar{x}_{i\cdot}))^2.$$

(b) If the lines are identical, the problem reduces to simple linear regression. The least squares estimates are $\hat{\alpha}_0 = \bar{Y}_{\cdot\cdot}$ and

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\cdot\cdot}) Y_{ij}}{\sum_{i=1}^{2} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{\cdot\cdot})^2}.$$

One finds

$$\|\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}_0\|^2 = \sum_{i=1}^{2} \sum_{j=1}^{n_i} [\hat{\alpha}_i + \hat{\beta}(x_{ij} - x_{i\cdot}) - \hat{\alpha}_0 + \hat{\beta}_0(x_{ij} - x_{\cdot\cdot})]^2.$$

The test statistic is $\mathcal{F} = \|\hat{\boldsymbol{\eta}} - \hat{\boldsymbol{\eta}}_0\|^2 / s^2$, which has an F distribution with $(1, n_1 + n_2 - 3)$ degrees of freedom under the null hypothesis.

3. Let $\mathbf{Y}$ be a random vector of dimension n with mean vector $\boldsymbol{\eta}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Let $Q(\mathbf{Y}) = \mathbf{Y}^T\mathbf{B}\mathbf{Y}$ be a quadratic form in Y. Writing $\mathbf{e} = \mathbf{Y} - \boldsymbol{\eta}$, prove that $E[Q(\mathbf{Y})] = Q(\boldsymbol{\eta}) + E[Q(\mathbf{e})]$ and that $E[Q(\mathbf{e})] = \text{trace}(\mathbf{B}\boldsymbol{\Sigma})$.

Solution:

$$Q(\mathbf{Y}) = \mathbf{Y}^T\mathbf{B}\mathbf{Y} = (\boldsymbol{\eta} + \mathbf{e})^T\mathbf{B}(\boldsymbol{\eta} + \mathbf{e}) = \boldsymbol{\eta}^T\mathbf{B}\boldsymbol{\eta} + 2\boldsymbol{\eta}^T\mathbf{B}\mathbf{e} + \mathbf{e}^T\mathbf{B}\mathbf{e}$$

Therefore

$$E[Q(\mathbf{Y})] = Q(\boldsymbol{\eta}) + E[Q(\mathbf{e})]$$

because $E[\mathbf{e}] = 0$. Moreover, $Q(\mathbf{e}) = \mathbf{e}^T\mathbf{B}\mathbf{e}$ is a scalar and equals its own trace. Therefore

$$E[Q(\mathbf{e})] = E[\text{trace}(\mathbf{e}^T\mathbf{B}\mathbf{e})] = E[\text{trace}(\mathbf{B}\mathbf{e}\mathbf{e}^T)] = \text{trace}[\mathbf{B}E(\mathbf{e}\mathbf{e}^T)] = \text{trace}(\mathbf{B}\Sigma).$$

4. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Let the matrix $\mathbf{X}$ be partitioned as $[\mathbf{X}_1 \mid \mathbf{X}_2]$ and let the parameter vector $\boldsymbol{\beta}$ be partitioned conformably, so that $\mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$.

(a) Suppose that $\boldsymbol{\beta}_2 = \mathbf{0}$ but one fits the full model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Show that the resulting least squares estimator $\widehat{\boldsymbol{\beta}}$ is unbiased.

(b) Suppose instead that $\boldsymbol{\beta}_2 \neq \mathbf{0}$ and the reduced model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ is estimated. Show that $\widehat{\boldsymbol{\beta}}_1$, the least squares estimator under this misspecified model, is biased unless $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$.

Solution: Assume $\mathbf{X}$ has full rank. Then $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ has expected value

$$E[\widehat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}.$$

Even if $\boldsymbol{\beta}_2 = \mathbf{0}$, $\widehat{\boldsymbol{\beta}}$ is still unbiased. If one falsely assumes $\boldsymbol{\beta}_2 = \mathbf{0}$ and fits the reduced model, then $\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{Y}$ is computed and

$$\begin{aligned} E[\widehat{\boldsymbol{\beta}}_1] = (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}\boldsymbol{\beta} &= (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{X}_2\boldsymbol{\beta}_2. \end{aligned}$$

Since $\mathbf{X}_1^T\mathbf{X}_1$ has full rank, $\widehat{\boldsymbol{\beta}}_1$ is unbiased for all $\boldsymbol{\beta}$ if and only if $\mathbf{X}_1^T\mathbf{X}_2 = \mathbf{0}$.