STAT 8260 Exam 2 – Tuesday, April 10
SHOW ALL WORK

Name: Answer Key

1. Suppose we take a random sample of $n = 6$ adult father-son pairs and we measure the height of each man. Suppose that 3 of the sons in this sample were first-born children in their family and the other 3 happened to be second-born. Let (y_{ij}, x_{ij}) by the (son's height, father height) pair of measurements corresponding to the jth son of the ith birth order, $i = 1, 2$, $j = 1, 2, 3$. Consider the following five models for such data.

$$
\begin{aligned}
(i) \quad & y_{ij} = \alpha_i + e_{ij} \\
(ii) \quad & y_{ij} = \alpha + \beta x_{ij} + e_{ij} \\
(iii) \quad & y_{ij} = \alpha_i + \beta_i x_{ij} + e_{ij} \\
(iv) \quad & y_{ij} = \mu + \alpha I(i = 2) + \beta x_{ij} + e_{ij} \\
(v) \quad & y_{ij} = \mu + \alpha I(i = 2) + \beta x_{ij} I(i = 1) + \gamma x_{ij} I(i = 2) + e_{ij}
\end{aligned}
$$

where $I(\cdot)$ denotes an indicator variable taking the value 1 when the condition inside the parentheses is true and the value 0 otherwise. All models have the same assumptions on the e_{ij}s: $e_{11}, \ldots e_{23} \overset{iid}{\sim} N(0, \sigma^2)$.

a. **(8 pts.)** Which, if any, of these models are equivalent.

(iii) and (v) are equivalent, the others are all different.

You could write down the model matrices for all 5 models and check which ones have the same column space, but I intended this to be much easier than that. Model (i) specifies 2 means, α_1 and α_2 for the 2 groups, respectively. Model (ii) specifies a single regression line (common intercept α and slope β) for all subjects. Model (iii) allows two regression lines, $\alpha_1 + \beta_1 X$ for group 1, and $\alpha_2 + \beta_2 X$ for group 2. Model (iv) specifies 2 different intercepts: μ and $\mu + \alpha$ for groups 1 and 2, respectively, but a single common slope β. Finally, Model (v), like model (iii) specifies distinct intercepts, μ and $\mu + \alpha$, for the 2 groups, and distinct slopes, β and γ, for the 2 groups.

b. **(9 pts.)** Write down model (iii) in vector/matrix notation.

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} 1 & 0 & x_{11} & 0 \\ 1 & 0 & x_{12} & 0 \\ 1 & 0 & x_{13} & 0 \\ 0 & 1 & 0 & x_{21} \\ 0 & 1 & 0 & x_{22} \\ 0 & 1 & 0 & x_{23} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{23} \end{pmatrix}$$

$$\underline{y} = \underline{X}\,\underline{\beta} + \underline{e}$$

c. **(9 pts.)** Under the maintained hypothesis that model (v) holds, express the hypothesis that H_0 : {model (ii) holds}, in the form of the general linear hypothesis on the parameters of model (v).

$$\underline{C}\underline{\beta} = \underline{t} \quad \text{where} \quad C = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}, \beta = \begin{pmatrix} \mu \\ \alpha \\ \beta \\ \gamma \end{pmatrix}$$

$$\text{and } \underline{t} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

d. **(9 pts.)** Based upon the results in the table below, compute an appropriate test statistic (give me its numeric value for these data) and give its reference distribution for testing the hypothesis in part (c).

Model	MSE	R^2
(ii)	9.9268	.0432
(v)	3.2262	.8445

(Several way to compute F)

$$F = \frac{(R^2_{FM} - R^2_{RM})/h}{(1 - R^2_{FM})/(n-k-1)} = \frac{(.8445 - .0432)/2}{(1 - .8445)/(6-3-1)} = 5.15$$

$\sim F(h, n-k-1) = F(2, 2)$ under H_0

2. Consider the regression model

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{pmatrix} = \beta \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \end{pmatrix}$$
$$= \beta \mathbf{j}_4 + \mathbf{e}$$

where $\begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} \sim N\left(\mathbf{0}, \sigma_1^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ independent of $\begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix} \sim N(\mathbf{0}, \sigma_2^2 \mathbf{I})$ where $\sigma_2^2 = 2\sigma_1^2$ and $\rho = .5$.

a. **(13 pts.)** Derive a simple, non-matrix formula for $\hat{\beta}_{\text{BLUE}}$, the best linear unbiased estimator of β in this model.

$\underline{e} \sim N(0, \sigma_1^2 V)$ where $V = \begin{pmatrix} 1 & .5 & 0 & 0 \\ .5 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{pmatrix}$

$\hat{\beta}_{BLUE}$ is the GLS estimator $(X^T V^{-1} X)^{-1} X^T V^{-1} y$ where $X = \underline{j}_4$

$V^{-1} = \begin{pmatrix} \frac{4}{3}\begin{pmatrix} 1 & -.5 \\ -.5 & 1 \end{pmatrix} & \underline{0} \\ \underline{0} & \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \end{pmatrix}$ $\quad j^T V^{-1} = \begin{pmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$ (sum of columns of V^{-1})

$(j^T V^{-1} j)^{-1} = \frac{3}{7}$

$(j^T V^{-1} j)^{-1} j^T V^{-1} y = \frac{3}{7}\left(\frac{2}{3} \quad \frac{2}{3} \quad \frac{1}{2} \quad \frac{1}{2}\right) y = \left(\frac{2}{7}, \frac{2}{7}, \frac{3}{14}, \frac{3}{14}\right) \underline{y}$

$= \frac{2}{7} y_{1\cdot} + \frac{3}{14} y_{2\cdot}$

b. **(13 pts.)** Assuming the model is correct, compute the ratio

$$\frac{\text{var}(\bar{y}_{..})}{\text{var}(\hat{\beta}_{\text{BLUE}})},$$

where $\bar{y}_{..} = \frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2} y_{ij}$ is the sample mean.

$\bar{y}_{..}$ is the OLS estimator, with variance

$$\text{Var}(\bar{y}_{..}) = (X^TX)^{-1}X^TVX(X^TX)^{-1} = (\underline{j}^T\underline{j})^{-1}\underline{j}^T\sigma_1^2V\underline{j}(\underline{j}^T\underline{j})^{-1}$$

$$= \sigma_1^2\frac{1}{16}\underline{j}^TV\underline{j} \quad , \quad \underline{j}^TV\underline{j} = \left(\tfrac{3}{2}, \tfrac{3}{2}, 2, 2\right)\underline{j} = 7 \qquad (\underline{j}^T\underline{j} = n = 4)$$

$$= \sigma_1^2\, 7/16$$

$$\text{Var}(\hat{\beta}_{BLUE}) = \sigma_1^2(X^TV^{-1}X)^{-1} = \sigma_1^2(\underline{j}^TV^{-1}X)^{-1} = \sigma_1^2\, 3/7$$

$$\Rightarrow \quad \frac{\text{Var}(\bar{y}_{..})}{\text{Var}(\hat{\beta}_{BLUE})} = \frac{\sigma_1^2\frac{7}{16}}{\sigma_1^2\frac{3}{7}} = \frac{49}{48}$$

notice > 1

3. Consider the classical linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{X}$ is $n \times (k+1)$ with $\text{rank}(\mathbf{X}) = k+1 < n$. Suppose this model is fit to data $\mathbf{y}$ yielding the ordinary least squares regression parameter estimate $\hat{\boldsymbol{\beta}}$ and MSE, s^2. We wish to form a prediction interval for an additional observation y_0 conforming to this model. That is, $y_0 = \mathbf{x}_0^T\boldsymbol{\beta} + e_0$ where $\mathbf{x}_0$ is the a vector of explanatory variables corresponding to y_0 ($\mathbf{x}_0^T$ can be considered as an additional row of $\mathbf{X}$) and $e_0 \sim N(0, \sigma^2)$ independent of $\mathbf{e}$.

(13 pts.) For $\hat{y}_0 = \mathbf{x}_0^T\hat{\boldsymbol{\beta}}$, show that

$$\frac{y_0 - \hat{y}_0}{s\sqrt{1 + \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_0}} \sim t(n-k-1).$$

Since $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$, $\hat{y} = x_0^T\hat{\beta} \sim N(x_0^T\beta, \sigma^2 x_0^T(X^TX)^{-1}x_0)$

Since $\hat{y}_0$ is computed from $\hat{\beta}$ which is based on y a sample independent of y_0, $\text{Cov}(\hat{y}_0, y_0) = 0$. In addition $y_0 \sim N(x_0^T\beta, \sigma^2)$. So,

$y_0 - \hat{y}_0 \sim N(0, \sigma^2 + \sigma^2 x_0^T(X^TX)^{-1}x_0)$

We know that $\frac{SSE}{\sigma^2} \sim \chi^2(n-k-1)$ and $S = \sqrt{\frac{SSE}{df_E}}$

and SSE independent of $\hat{\beta}$ hence indep of $\hat{y}_0 = x_0^T\hat{\beta}$

also, SSE indep of y_0 (different samples).

Then since $t = \dfrac{(y_0 - \hat{y}_0) / [\sigma\sqrt{1 + x_0^T(X^TX)^{-1}x_0}]}{\sqrt{\frac{SSE/\sigma^2}{df_E}}} = \dfrac{num}{den}$

when $num \sim N(0,1)$ ← independent

$den = \sqrt{\chi^2/df}$ ←

$\Rightarrow t \sim t(df_E)$

$= t(n-k-1)$

4. Consider the classical linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where $\mathbf{X}$ is $n \times p$ with $\text{rank}(\mathbf{X}) = p < n$. Define $\tilde{\sigma}^2_\ell = SSE/(n-\ell)$, an estimator of σ^2. Recall the mean square error of an estimator T of a parameter θ is $\text{E}\{(T-\theta)^2\}$ but can also be represented as variance + bias^2.

a. **(13 pts.)** Find the mean square error of $\tilde{\sigma}^2_\ell$, for an arbitrary value of the constant ℓ.

$$\text{Var}(\tilde{\sigma}^2_\ell) = \frac{1}{(n-\ell)^2} \text{Var}(y^T P_{C(X)^\perp} y)$$

$$= \frac{1}{(n-\ell)^2}\left[2\,\text{tr}\left(P_{C(X)^\perp}\sigma^2 I P_{C(X)^\perp}\sigma^2 I\right) + 4(X\beta)^T P_{C(X)^\perp}\sigma^2 I \underbrace{P_{C(X)^\perp} X\beta}_{=0}\right]$$

$$= \frac{1}{(n-\ell)^2} 2\sigma^4 \underbrace{\text{tr}\left(P_{C(X)^\perp}\right)}_{=n-p} = \frac{n-p}{(n-\ell)^2} 2\sigma^4$$

$$\text{Bias}(\tilde{\sigma}^2_\ell) = \text{Bias}\left(\frac{n-p}{n-\ell}\underbrace{\frac{SSE}{n-p}}_{=s^2}\right) = E\left(\frac{n-p}{n-\ell} s^2\right) - \sigma^2$$

$$= \frac{n-p}{n-\ell}\underbrace{E(s^2)}_{=\sigma^2} - \sigma^2 = \sigma^2\left(\frac{n-p}{n-\ell} - 1\right)$$

$$\text{MSE}(\tilde{\sigma}^2_\ell) = \frac{n-p}{(n-\ell)^2} 2\sigma^4 + \sigma^4\left(\frac{n-p}{n-\ell} - 1\right)^2$$

b. **(13 pts.)** Find the value of ℓ that yields the estimator $\tilde{\sigma}^2_\ell$ with the smallest mean square error.

From part (a) $MSE(\tilde{\sigma}^2_\ell) = \frac{4\sigma^4(n-p)}{(n-\ell)^3} + \frac{2\sigma^4(n-p)}{(n-\ell)^2}\left(\frac{n-p}{n-\ell} - 1\right)^2 \equiv M$

$$\frac{\partial M}{\partial \ell} = -2(n-p)(n-\ell)^{-3}(-1)2\sigma^4 + 2\sigma^4\left(\frac{n-p}{n-\ell} - 1\right)\left(-(n-p)(n-\ell)^{-2}(-1)\right)$$

$$= \frac{4\sigma^4(n-p)}{(n-\ell)^3} + \frac{2\sigma^4(n-p)}{(n-\ell)^2}\left(\frac{n-p}{n-\ell} - 1\right) \overset{\text{set}}{=} 0$$

(By multiplying through by $\frac{(n-\ell)^3}{2\sigma^4(n-p)}$)

$$\Rightarrow 2 + (n-\ell)\left(\frac{n-p}{n-\ell} - 1\right) = 0$$

$$\Rightarrow 2 + (n-p-n+\ell) = 0 \Rightarrow \ell = p-2$$

$\Rightarrow \frac{SSE}{n-p+2}$ is the minimum MSE estimator of σ^2