STAT 666 Final

Maggie Buffum

June 11, 2019

Problem 1

(10 points) Suppose we fit the following model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij},\ i = 1, 2,\ j = 1, 2,$$

where $\epsilon_{ij} \sim \text{iid} N(0, \sigma^2)$, σ^2 is unknown, and $\sum_{i=1}^{2} \tau_i = 0$. Develop a suitable sequence of nested hypotheses, given the relevant sum of squares. Complete the ANOVA table.

Consider the following sequence of nested hypotheses,

$$\begin{aligned} S_{H_0} &= \{Y_{ij} = \mu + \tau_i + \epsilon_{ij},\ \sum_{i=1}^{2} \tau_i = 0\} \\ \supset S_{H_1} &= \{Y_{ij} = \mu + \epsilon_{ij}\} \\ \supset S_{H_2} &= \{Y_{ij} = \epsilon_{ij}\} \end{aligned}$$

We need to find the fitted values under the different hypotheses to evaluate the sequential sums of squares in the ANOVA table. Let's first estimate μ using H_1:

$$\begin{aligned} \frac{d}{d\mu} S_{H_1} &= \frac{d}{d\mu} \sum_{i=1}^{2} \sum_{j=1}^{2} (Y_{ij} - \mu)^2 \\ &= 2 \sum_{i=1}^{2} \sum_{j=1}^{2} (Y_{ij} - \mu)(-1) = \text{ (set) } 0 \\ &\implies \sum_{i=1}^{2} \sum_{j=1}^{2} Y_{ij} - \sum_{i=1}^{2} \sum_{j=1}^{2} \mu = 0 \\ &\implies \sum_{i=1}^{2} \sum_{j=1}^{2} \mu = \sum_{i=1}^{2} \sum_{j=1}^{2} Y_{ij} \\ &\implies 4\mu = 4\bar{Y}_{\cdot\cdot} \\ &\implies \mu = \bar{Y}_{\cdot\cdot} \end{aligned}$$

We can estimate τ_i using H_0:

$$\begin{aligned}
\frac{d}{d\tau_i} S_{H_1} &= \frac{d}{d\tau_i} \sum_{j=1}^{2} (Y_{ij} - \mu - \tau_i)^2 \\
&= 2 \sum_{j=1}^{2} (Y_{ij} - \mu - \tau_i)(-1) = \text{ (set) } 0 \\
&\implies \sum_{j=1}^{2} Y_{ij} - \sum_{j=1}^{2} \mu - \sum_{j=1}^{2} \tau_i = 0 \\
&\implies 2\hat{\tau}_i = 2\bar{Y}_{i\cdot} - 2\hat{\mu} \\
&\implies \hat{\tau}_i = \bar{Y}_{i\cdot} - \hat{\mu} \\
&\implies \hat{\tau}_i = \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}
\end{aligned}$$

Now we can evaluate the fitted values for each hypothesis in the sequence:

$$\begin{aligned}
\hat{y}_{H_0} &= \{\hat{\mu} + \hat{\tau}_i\} = \{\bar{Y}_{\cdot\cdot} + \bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\} = \{\bar{Y}_{i\cdot}\} \\
\hat{y}_{H_1} &= \{\hat{\mu}\} = \{\bar{Y}_{\cdot\cdot}\}
\end{aligned}$$

Finally, we can estimate the sum of squares attributed to each parameter. $R(H_1|H_2)$ gives us the sum of squares attributed to μ:

$$R(H_1|H_2) = (\hat{y}_{H_1} - \hat{y}_{H_2})'(\hat{y}_{H_1} - \hat{y}_{H_2}) = (\bar{Y}_{\cdot\cdot})'(\bar{Y}_{\cdot\cdot}) = \sum_{i=1}^{2}\sum_{j=1}^{2} \bar{Y}_{\cdot\cdot}^2 = 4\bar{Y}_{\cdot\cdot}^2$$

$R(H_0|H_1)$ gives us the sum of squares attributed to τ_i:

$$\begin{aligned}
R(H_0|H_1) &= (\hat{y}_{H_0} - \hat{y}_{H_1})'(\hat{y}_{H_0} - \hat{y}_{H_1}) = (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})'(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}) = \sum_{i=1}^{2}\sum_{j=1}^{2} (\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} (\bar{Y}_{i\cdot}^2 + \bar{Y}_{\cdot\cdot}^2 - 2\bar{Y}_{i\cdot}\bar{Y}_{\cdot\cdot}) = \sum_{i=1}^{2}\sum_{j=1}^{2} \bar{Y}_{i\cdot}^2 + \sum_{i=1}^{2}\sum_{j=1}^{2} \bar{Y}_{\cdot\cdot}^2 - 2\sum_{i=1}^{2}\sum_{j=1}^{2} \bar{Y}_{i\cdot}\bar{Y}_{\cdot\cdot} \\
&= 2\sum_{i=1}^{2} \bar{Y}_{i\cdot}^2 + 4\bar{Y}_{\cdot\cdot}^2 - 8\bar{Y}_{\cdot\cdot}^2 \\
&= 2\sum_{i=1}^{2} \bar{Y}_{i\cdot}^2 - 4\bar{Y}_{\cdot\cdot}^2
\end{aligned}$$

We evaluate the sum of squares due to error as

$$\begin{aligned}
SSE = SST - R(H_0) &= \sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2 - (\hat{y}_{H_0})'(\hat{y}_{H_0}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2 - (\bar{Y}_{i\cdot})'(\bar{Y}_{i\cdot}) = \sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2 - \sum_{i=1}^{2} \bar{Y}_{i\cdot}^2
\end{aligned}$$

The resulting ANOVA table is provided below.

Source	SS	Df	MSE	F
τ_i	$2\sum_{i=1}^{2} \bar{Y}_{i\cdot}^2 - 4\bar{Y}_{\cdot\cdot}^2$	1	$2\sum_{i=1}^{2} \bar{Y}_{i\cdot}^2 - 4\bar{Y}_{\cdot\cdot}^2$	$\frac{4\sum_{i=1}^{2} \bar{Y}_{i\cdot}^2 - 8\bar{Y}_{\cdot\cdot}^2}{\sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2 - \sum_{i=1}^{2} \bar{Y}_{i\cdot}^2}$
Error	$\sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2 - \sum_{i=1}^{2} \bar{Y}_{i\cdot}^2$	2	$(\sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2 - \sum_{i=1}^{2} \bar{Y}_{i\cdot}^2)/2$	
Total	$\sum_{i=1}^{2}\sum_{j=1}^{2} Y_{ij}^2$	3		

Problem 2

(25 points) We would like to see how a level of light intensity (X) is related to a measure of plant growth (Y). We fit a regression model through the origin:

X	Y
3.5	1.6
4.6	1.8
4.6	2.2

As the data have tiles values in X, we apply a lack of fit text at the level of significance $\alpha = 0.05$.

(a) State the null hypothesis.

The model fits the data relatively well; there is no lack of fit. The expected mean squares for lack of fit equals the mean squares for pure error.

(b) State the alternative hypothesis.

The model fits the data poorly; there is a lack of fit. The expected mean squares for lack of fit is significantly larger than the mean squares for pure error than would be normal under the F-distribution.

(c) Compute SSE, SSPE, and SSLF.

```
X <- matrix(data=XY$X)
Y <- matrix(data=XY$Y)

P <- X %*% solve(t(X)%*%X) %*% t(X)
Y_hat <- P %*% Y

Xe <- matrix(data=c(
  1,0,0,
  0,1,1
),nrow = 3,byrow = F)

Pe <- Xe %*% solve(t(Xe)%*%Xe) %*% t(Xe)
Y_bar <- Pe %*% Y


SSE  <- t(Y) %*% (diag(3) - P) %*% Y
SSLF <- t(Y) %*% (Pe - P) %*% Y
SSPE <- t(Y) %*% (diag(3) - Pe) %*% Y

tab <- data.frame(
```

```
  SSE=SSE,SSLF=SSLF,SSPE=SSPE
)
kable(tab)
```

SSE	SSLF	SSPE
0.0847499	0.0047499	0.08

(d) Derive the F-test statistic to test your null hypothesis.

The F-statistic is the ratio of MS Lack of Fit and MS Pure Error, or

$$F(H_0) = \frac{SSLF/(m-p)}{SSPE/(N-m)} \sim F_{m-p,N-m}$$

where p is the rank of X, m is the number of unique levels of X, and N is the total number of observations. Therefore,

```
p <- Rank(X)
m <- Rank(Xe)
N <- NROW(X)

MSLF <- SSLF/(m-p)
MSPE <- SSPE/(N-m)

F <- MSLF/MSPE
p_value <- pf(q = F,df1 = m-p,df2 = N-m)

tab <- data.frame(
  Source = c("Residual","Lack of Fit","Pure Error"),
  Df = c(N-p,m-p,N-m),
  SS = c(SSE,SSLF,SSPE),
  MS = c("",round(MSLF,7),round(MSPE,7)),
  F  = c("",round(F,7),""),
  p_value = c("",round(p_value,7),"")
)
kable(tab)
```

Source	Df	SS	MS	F	p_value
Residual	2	0.0847499			
Lack of Fit	1	0.0047499	0.0047499	0.0593733	0.1521577
Pure Error	1	0.0800000	0.08		

(e) Draw the geometry of lack of fit and pure-error based on this dataset. Does your geometry support the result from (d)?

Given the following, we can represent the lack of fit and pure error geometrically (shown in the following

figure):

```
I <- diag(3)
# (I-P)y:
(I-P)%*%Y
```

```
##              [,1]
## [1,]  0.06069269
## [2,] -0.22308961
## [3,]  0.17691039
```

```
# (I-Pe)y:
(I-Pe)%*%Y
```

```
##      [,1]
## [1,]  0.0
## [2,] -0.2
## [3,]  0.2
```

```
# (Pe-P)y:
(Pe-P)%*%Y
```

```
##              [,1]
## [1,]  0.06069269
## [2,] -0.02308961
## [3,] -0.02308961
```

```
# PeY:
Pe%*%Y
```

```
##      [,1]
## [1,]  1.6
## [2,]  2.0
## [3,]  2.0
```

```
# Py:
P%*%Y
```

```
##          [,1]
## [1,] 1.539307
## [2,] 2.023090
## [3,] 2.023090
```

giving us the following geometric form of lack of fit and pure error:

Problem 3

(10 points) Consider the regression lines $Y_{l,i} = \beta_{l,0} + \beta_{l,1} X_{l,i} + \epsilon_{l,i}$, where $l = 1, 2$ and $\epsilon_{l,i} \sim \text{iid} N(0, \sigma^2)$. Given observations $(x_{l,i}, y_{l,i})$, $i = 1, \ldots, n_l$, $l = 1, 2$, carry out the following test:

$$H_0 : \beta_{1,1} = \beta_{2,1} \text{ vs. } H_0 : \beta_{1,1} \neq \beta_{2,1}$$

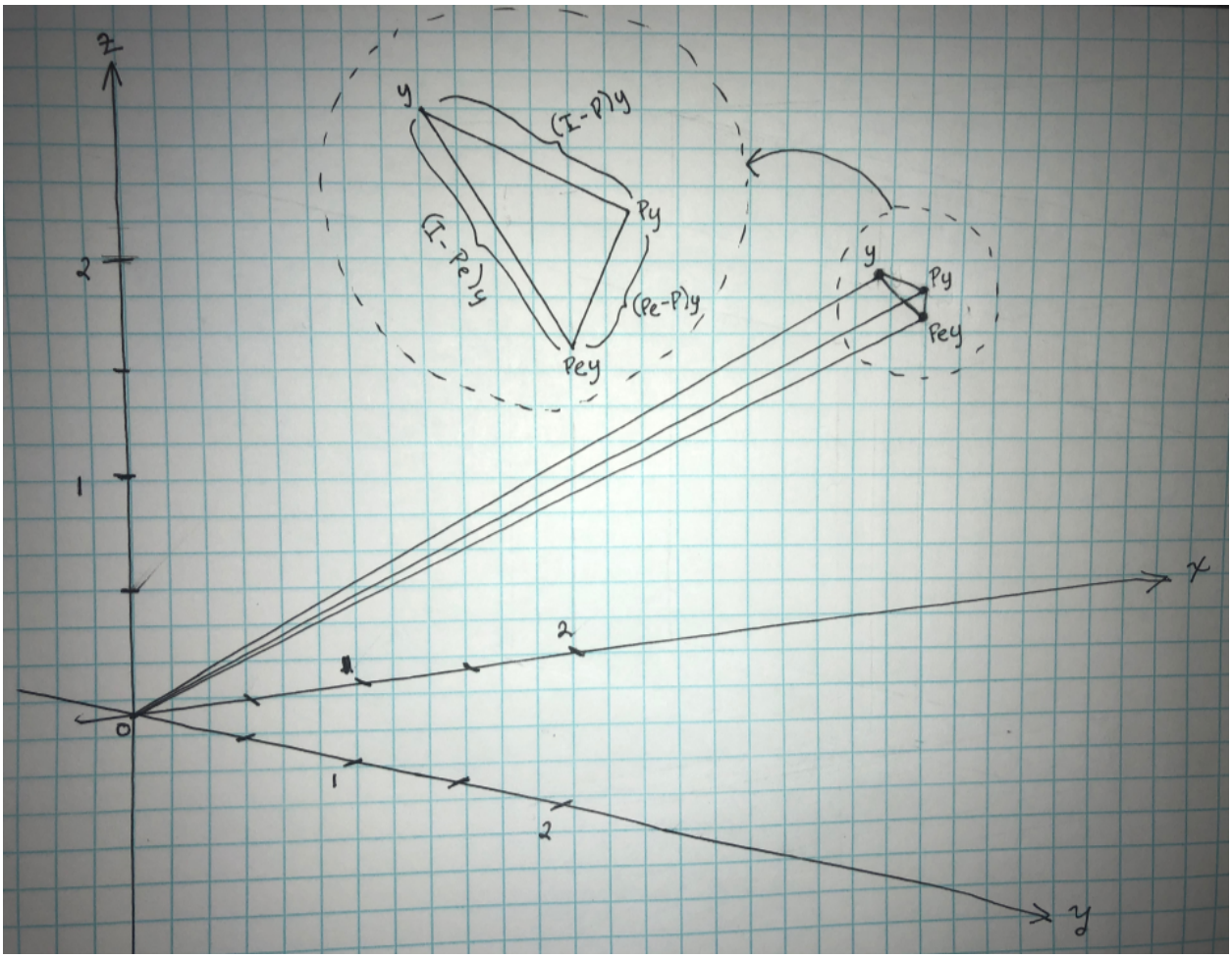

Figure 1: Geometric Representation of Lack of Fit and Pure Error

The two given models can be formatted as follows:

$$\begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,n_1} \\ y_{2,1} \\ \vdots \\ y_{2,n_2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_{1,1} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{1,n_1} & 0 \\ 0 & 1 & 0 & x_{2,1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{1,n_2} \end{bmatrix} \begin{bmatrix} \beta_{1,0} \\ \beta_{2,0} \\ \beta_{1,1} \\ \beta_{2,1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,1} \\ \vdots \\ \epsilon_{1,n_1} \\ \epsilon_{2,1} \\ \vdots \\ \epsilon_{2,n_2} \end{bmatrix},$$

The least squares estimates of the βs are obtained by minimizing the function $S(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$. Let's find $\hat{\beta}_{l,0}$ first:

$$\begin{aligned}
\frac{dS(\beta)}{d\beta_{l,0}} = \frac{d}{d\beta_{l,0}} \left(\sum_{i=1}^{n_l} (y_{l,i} - \beta_{l,0} - \beta_{l,1} x_{l,i})^2 \right) &= \text{ (set) } 0 \\
\implies 2 \sum_{i=1}^{n_l} (y_{l,i} - \beta_{l,0} - \beta_{l,1} x_{l,i})(-1) &= 0 \\
\implies \sum_{i=1}^{n_l} \beta_{l,0} = \sum_{i=1}^{n_l} y_{l,i} - \beta_{l,1} \sum_{i=1}^{n_l} x_{l,i} & \\
\implies n_l \beta_{l,0} = n_l \bar{y}_{l,.} - n_l \beta_{l,1} \bar{x}_{l,.} & \\
\implies \hat{\beta}_{l,0} = \bar{y}_{l,.} - \hat{\beta}_{l,1} \bar{x}_{l,.} &
\end{aligned}$$

and now $\hat{\beta}_{l,1}$:

$$\begin{aligned}
\frac{dS(\beta)}{d\beta_{l,1}} = \frac{d}{d\beta_{l,1}} \left(\sum_{i=1}^{n_l} (y_{l,i} - \beta_{l,0} - \beta_{l,1} x_{l,i})^2 \right) &= \text{ (set) } 0 \\
\implies 2 \sum_{i=1}^{n_l} (y_{l,i} - \beta_{l,0} - \beta_{l,1} x_{l,i})(-x_{l,i}) &= 0 \\
\implies \sum_{i=1}^{n_l} y_{l,i} x_{l,i} - \beta_{l,0} \sum_{i=1}^{n_l} x_{l,i} - \beta_{l,1} \sum_{i=1}^{n_l} x_{l,i}^2 &= 0 \\
\implies \hat{\beta}_{l,1} \sum_{i=1}^{n_l} x_{l,i}^2 = \sum_{i=1}^{n_l} y_{l,i} x_{l,i} - n_l \hat{\beta}_{l,0} \bar{x}_{l,.} & \\
\implies \hat{\beta}_{l,1} \sum_{i=1}^{n_l} x_{l,i}^2 = \sum_{i=1}^{n_l} y_{l,i} x_{l,i} - n_l (\bar{y}_{l,.} - \hat{\beta}_{l,1} \bar{x}_{l,.}) \bar{x}_{l,.} & \\
\implies \hat{\beta}_{l,1} \sum_{i=1}^{n_l} x_{l,i}^2 - n_l \hat{\beta}_{l,1} \bar{x}_{l,.}^2 = \sum_{i=1}^{n_l} y_{l,i} x_{l,i} - n_l \bar{y}_{l,.} \bar{x}_{l,.} & \\
\implies \hat{\beta}_{l,1} \left(\sum_{i=1}^{n_l} x_{l,i}^2 - n_l \bar{x}_{l,.}^2 \right) = \sum_{i=1}^{n_l} y_{l,i} x_{l,i} - n_l \bar{y}_{l,.} \bar{x}_{l,.} & \\
\implies \hat{\beta}_{l,1} = \frac{\sum_{i=1}^{n_l} y_{l,i} x_{l,i} - n_l \bar{y}_{l,.} \bar{x}_{l,.}}{\sum_{i=1}^{n_l} x_{l,i}^2 - n_l \bar{x}_{l,.}^2} & \\
\implies \hat{\beta}_{l,1} = \frac{\sum_{i=1}^{n_l} (y_{l,i} - \bar{y}_{l,.})(x_{l,i} - \bar{x}_{l,.})}{\sum_{i=1}^{n_l} (x_{l,i} - \bar{x}_{l,.})^2} &
\end{aligned}$$

Having estimated the parameters, we can evaluate the SSE as

$$
\begin{aligned}
SSE &= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(y_{l,i}-\hat{\beta}_{l,0}-\hat{\beta}_{l,1}x_{l,i})^2\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(y_{l,i}-(\bar{y}_{l,.}-\hat{\beta}_{l,1}\bar{x}_{l,.})-\hat{\beta}_{l,1}x_{l,i})^2\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.}+\hat{\beta}_{l,1}\bar{x}_{l,.}-\hat{\beta}_{l,1}x_{l,i})^2\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}((y_{l,i}-\bar{y}_{l,.})-\hat{\beta}_{l,1}(x_{l,i}-\bar{x}_{l,.}))^2\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}((y_{l,i}-\bar{y}_{l,.})^2+\hat{\beta}_{l,1}^2(x_{l,i}-\bar{x}_{l,.})^2-2\hat{\beta}_{l,1}(y_{l,i}-\bar{y}_{l,.})(x_{l,i}-\bar{x}_{l,.}))\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.})^2+\hat{\beta}_{l,1}^2\sum_{i=1}^{n_l}(x_{l,i}-\bar{x}_{l,.})^2-2\hat{\beta}_{l,1}\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.})(x_{l,i}-\bar{x}_{l,.})\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.})^2+\hat{\beta}_{l,1}^2\sum_{i=1}^{n_l}(x_{l,i}-\bar{x}_{l,.})^2-2\hat{\beta}_{l,1}^2\sum_{i=1}^{n_l}(x_{l,i}-\bar{x}_{l,.})^2\right] \\
&= \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.})^2-\hat{\beta}_{l,1}^2\sum_{i=1}^{n_l}(x_{l,i}-\bar{x}_{l,.})^2\right] \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.})^2-\sum_{l=1}^{2}\hat{\beta}_{l,1}^2\sum_{i=1}^{n_l}(x_{l,i}-\bar{x}_{l,.})^2
\end{aligned}
$$

We are interested in testing the hypothesis that

$$H:\beta_{1,1}=\beta_{2,1}=\beta$$

Under the null hypothesis, we can estimate $\beta_{l,0}$ and β_1 by minimizing the function $S(\beta)=\sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-\beta_{l,0}-\beta_1 x_{l,i})^2$. Let's first evaluate $\hat{\beta}_{l,0,H}$:

$$
\begin{aligned}
\frac{dS(\beta)}{d\beta_{l,0}} &= \frac{d}{d\beta_{l,0}}\left[\sum_{i=1}^{n_l}(y_{l,i}-\beta_{l,0}-\beta_1 x_{l,i})^2\right] \\
&= 2\sum_{i=1}^{n_l}(y_{l,i}-\beta_{l,0}-\beta_1 x_{l,i})(-1) = \text{ (set) } 0 \\
&= \sum_{i=1}^{n_l}\beta_{l,0}=\sum_{i=1}^{n_l}(y_{l,i}-\beta_1 x_{l,i}) \\
&= n_l\hat{\beta}_{l,0}=n_l\bar{y}_{l,.}-n_l\hat{\beta}_H\bar{x}_{l,.} \\
&= \hat{\beta}_{l,0}=\bar{y}_{l,.}-\hat{\beta}_H\bar{x}_{l,.}
\end{aligned}
$$

Next we evaluate $\hat{\beta}_H$:

$$\begin{aligned}
\frac{dS(\beta)}{d\beta_1} &= \frac{d}{d\beta_1}\left[\sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-\beta_{l,0}-\beta_1 x_{l,i})^2\right] \\
&= 2\sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-\beta_{l,0}-\beta_1 x_{l,i})(-x_{l,i}) = \text{ (set) } 0 \\
&\implies \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-\hat{\beta}_{l,0,H}-\hat{\beta}_H x_{l,i})x_{l,i} = 0 \\
&\implies \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-(\bar{y}_{l,.}-\hat{\beta}_H\bar{x}_{l,.})-\hat{\beta}_H x_{l,i})x_{l,i} = 0 \\
&\implies \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i}-\bar{y}_{l,.}+\hat{\beta}_H\bar{x}_{l,.}-\hat{\beta}_H x_{l,i})x_{l,i} = 0 \\
&\implies \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}(x_{l,i}(y_{l,i}-\bar{y}_{l,.})+\hat{\beta}_H\bar{x}_{l,.}x_{l,i}-\hat{\beta}_H x_{l,i}^2)\right] = 0 \\
&\implies \sum_{l=1}^{2}\left[\sum_{i=1}^{n_l}x_{l,i}(y_{l,i}-\bar{y}_{l,.})+\hat{\beta}_H\sum_{i=1}^{n_l}(\bar{x}_{l,.}x_{l,i}-x_{l,i}^2)\right] = 0 \\
&\implies \sum_{l=1}^{2}\sum_{i=1}^{n_l}x_{l,i}(y_{l,i}-\bar{y}_{l,.})-\hat{\beta}_H\sum_{l=1}^{2}\sum_{i=1}^{n_l}x_{l,i}(x_{l,i}-\bar{x}_{l,.}) = 0 \\
&\implies \hat{\beta}_H = \frac{\sum_{l=1}^{2}\sum_{i=1}^{n_l}x_{l,i}(y_{l,i}-\bar{y}_{l,.})}{\sum_{l=1}^{2}\sum_{i=1}^{n_l}x_{l,i}(x_{l,i}-\bar{x}_{l,.})}
\end{aligned}$$

Now we can use these estimates to evaluate the sum of squared errors under the null hypothesis, SSE_H:

$$\begin{aligned}
SSE_H &= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \hat{\beta}_{l,0,H} - \hat{\beta}_H x_{l,i})^2 \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - (\bar{y}_{l,.} - \hat{\beta}_H \bar{x}_{l,.}) - \hat{\beta}_H x_{l,i})^2 \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}((y_{l,i} - \bar{y}_{l,.}) - \hat{\beta}_H (x_{l,i} - \bar{x}_{l,.}))^2 \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}((y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 (x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H (y_{l,i} - \bar{y}_{l,.})(x_{l,i} - \bar{x}_{l,.})) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})(x_{l,i} - \bar{x}_{l,.}) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H \sum_{l=1}^{2}\left(\sum_{i=1}^{n_l} y_{l,i}x_{l,i} - n_l \bar{y}_{l,.}\bar{x}_{l,.}\right) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H \sum_{l=1}^{2}\left(\sum_{i=1}^{n_l} y_{l,i}x_{l,i} - \sum_{i=1}^{n_l} \bar{y}_{l,.}x_{l,i}\right) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H \sum_{l=1}^{2}\sum_{i=1}^{n_l} x_{l,i}(y_{l,i} - \bar{y}_{l,.}) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i}^2 - \bar{x}_{l,.}x_{l,i}) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - 2\hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 - \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2
\end{aligned}$$

Recall that $(Q/s)/(SSE/(N-r)) \sim F_{s,N-r}$ (Result 7,2,1). One way to define Q is $Q = SSE_H - SSE$ (equation 7.2.12). Let's solve for Q:

$$\begin{aligned}
Q = SSE_H - SSE &= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 - \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - \left(\sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 - \sum_{l=1}^{2}\hat{\beta}_{l,1}^2 \sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2\right) \\
&= \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 - \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - \sum_{l=1}^{2}\sum_{i=1}^{n_l}(y_{l,i} - \bar{y}_{l,.})^2 + \sum_{l=1}^{2}\hat{\beta}_{l,1}^2 \sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 \\
&= \sum_{l=1}^{2}\hat{\beta}_{l,1}^2 \sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2 - \hat{\beta}_H^2 \sum_{l=1}^{2}\sum_{i=1}^{n_l}(x_{l,i} - \bar{x}_{l,.})^2
\end{aligned}$$

Finally, s is the rank of $\mathbf{C}'$ in the hypothesis test, that is, equal to the number of estimable hypothesis we wish to test simultaneously. Clearly $s = 1$. The rank of our original design matrix is r, and so $r = 4$. Therefore, the test statistic testing the hypothesis of equal slopes is

$$F(H) = \frac{SSE_H - SSE}{SSE/(n_1 + n_2 - 4)} \sim F1, n_1 + n_2 - 4.$$

Problem 4

(10 points) Consider a linear regression model with serially correlated errors

$$Y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \epsilon_t,\ t = 1, \ldots, N,$$

where the subscript t is used to indicate time and errors terms from different time periods are correlated. Prove the Durbin-Watson test statistic,

$$DW = \left\{ \sum_{t=2}^{N} (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2 \right\} / \left\{ \sum_{t=1}^{N} \hat{\epsilon}_t^2 \right\},$$

the most popular test for serial correlation, is bounded between 0 and 4.

Because both the numerator and denominator of the DW statistic are sums of squares, we already know that it must be nonnegative; that is, it is bounded below by 0. To determine the upper bound, we can rewrite the DW statistic as

$$\begin{aligned} DW &= \frac{\sum_{t=2}^{N} (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} \\ &= \frac{\sum_{t=2}^{N} (\hat{\epsilon}_t^2 + \hat{\epsilon}_{t-1}^2 - 2\hat{\epsilon}_t \hat{\epsilon}_{t-1})}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} \\ &= \frac{\sum_{t=2}^{N} \hat{\epsilon}_t^2 + \sum_{t=2}^{N} \hat{\epsilon}_{t-1}^2 - 2\sum_{t=2}^{N} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} \\ &= \frac{\sum_{t=2}^{N} \hat{\epsilon}_t^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} + \frac{\sum_{t=2}^{N} \hat{\epsilon}_{t-1}^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} - 2\frac{\sum_{t=2}^{N} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} \end{aligned}$$

The first two terms in the expression above are each bounded between 0 and 1: $\hat{\epsilon}_t \geq 0 \forall t$, and the denominators are summing over an additional positive element each such that the denominators are always at least as big as the numerators.

The third term can be positive or negative since neither error terms in the numerator are squared, and we are not guaranteed that $\hat{\epsilon}_t$ and $\hat{\epsilon}_{t-1}$ have the same sign. However, we already know that the DW statistic must be nonnegative, giving us the following inequality:

$$\frac{\sum_{t=2}^{N} \hat{\epsilon}_t^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} + \frac{\sum_{t=2}^{N} \hat{\epsilon}_{t-1}^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} \geq 2\frac{\sum_{t=2}^{N} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_{t=1}^{N} \hat{\epsilon}_t^2}$$

Since the left side the of inequality is ever at most 2, the right side of the inequality naturally must have an upper bound of 2 (i.e., the ratio must be no greater than 1) for the inequality to hold. Similarly, we have that

$$-\frac{\sum_{t=2}^{N} \hat{\epsilon}_t^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} - \frac{\sum_{t=2}^{N} \hat{\epsilon}_{t-1}^2}{\sum_{t=1}^{N} \hat{\epsilon}_t^2} \leq -2\frac{\sum_{t=2}^{N} \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum_{t=1}^{N} \hat{\epsilon}_t^2},$$

that is, the left side of the equation provides a lower bound of two, and therefore the third term in the DW statistic is bounded below by -2 (i.e., the ratio must be greater than -1) for the inequality to hold.

Given the bounds of the three terms in the DW statistic, the DW statistic can be at most 4.

Problem 5

(10 points) Compute the ridge regression model's $df(SSE_k)$, where k is the one-dimensional tuning parameter.

The degrees of freedom can be found by taking the trace the of projection matrix P_R. Using the cyclic proporty of traces, we have that

$$df_E = tr(P_R) = tr(X(X'X + kI)^{-1}X') = tr(X'X(X'X + kI)^{-1})$$

Because $X'X$ is symmetric, we can use spectral decomposition to diagonalize $X'X$ such that $D_1 = diag(\lambda_1, ..., \lambda_p)$, where λ_i are the eigenvalues of $X'X$. Moreover, we know that $(X'X + kI)^{-1}$ can be diagonalized such that $D_2 = 1/(\lambda_i + k)$. From Result 2.3.6 we know that the trace of a symmetric matrix is equal to the sum of its eigenvalues. We found that the eigenvalues of $X'X(X'X + kI)$ were $\lambda_i(\lambda_i + k)^{-1}$, and so

$$df_E = tr\left(diag\left(\frac{\lambda_i}{\lambda_i + k}\right)\right) = \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k}.$$
