MS Exam Spring 2018

Problem 1

A company has developed three specialized training workshops for their employees. Each of the 12 employees is randomly assigned to one of the three workshops. The company would like to develop a model that could be used to predict each employee's performance score (a number from 0 to 100) based on their attendance at the workshops. The previous year's performance score is also available for use as a predictor. The following table shows and ordered pair for each employee, consisting of the current performance score and the previous year's score.

```
##         [,1] [,2] [,3]
##   [1,] "A"  "70" "58"
##   [2,] "A"  "68" "60"
##   [3,] "A"  "70" "62"
##   [4,] "A"  "72" "65"
##   [5,] "B"  "75" "60"
##   [6,] "B"  "72" "60"
##   [7,] "B"  "74" "62"
##   [8,] "B"  "71" "60"
##   [9,] "C"  "72" "66"
## [10,] "C"  "73" "61"
## [11,] "C"  "72" "62"
## [12,] "C"  "73" "65"
```

(a) Write an appropriate model for the situation described above, allowing for different slopes and different intercepts for the three workshops. Hint: there are 12 observations, and you will need to use at least one indicator variable.

Clues:

- Use the previous year's performance score as a predictor, i.e., we need to include it in the design matrix as a continuous variable.

- Different slopes and different intercepts for the three workshops: means that we need to include separate indicators for each workshop, but also interact them with the predictors to get separate slopes.

Model:
$$y_i = \beta_0 + \beta_1 x_i + I_B(\gamma_0 + \gamma_1 x_i) + I_C(\delta_0 + \delta_1 x_i) + \epsilon_i$$

where

$$I_B = \begin{cases} 1 & \text{for } 5 \leq i \leq 8 \\ 0 & \text{otherwise} \end{cases}$$

$$I_C = \begin{cases} 1 & \text{for } 9 \leq i \leq 12 \\ 0 & \text{otherwise} \end{cases}$$

(b) Write the matrix form of the appropriate model. Show the contents and dimensions of all matrices.

$$\mathbf{Y} = \mathbf{Xb} + \epsilon$$

$$
\begin{bmatrix} 70 \\ 68 \\ 70 \\ 72 \\ 75 \\ 72 \\ 74 \\ 71 \\ 72 \\ 73 \\ 72 \\ 73 \end{bmatrix}_{12\times1}
=
\begin{bmatrix}
1 & 0 & 0 & 58 & 0 & 0 \\
1 & 0 & 0 & 60 & 0 & 0 \\
1 & 0 & 0 & 62 & 0 & 0 \\
1 & 0 & 0 & 65 & 0 & 0 \\
1 & 1 & 0 & 60 & 60 & 0 \\
1 & 1 & 0 & 60 & 60 & 0 \\
1 & 1 & 0 & 62 & 62 & 0 \\
1 & 1 & 0 & 60 & 60 & 0 \\
1 & 0 & 1 & 66 & 0 & 60 \\
1 & 0 & 1 & 61 & 0 & 60 \\
1 & 0 & 1 & 62 & 0 & 62 \\
1 & 0 & 1 & 65 & 0 & 60
\end{bmatrix}_{12\times6}
\begin{bmatrix} \beta_0 \\ \gamma_0 \\ \delta_0 \\ \beta_1 \\ \gamma_1 \\ \delta_1 \end{bmatrix}_{6\times1}
+ \epsilon_{12\times1}
$$

(c) Suppose that you wish to test for equality of the three slopes. Write the matrix form of the reduced model. What will be the numerator and denominator degrees of freedom for the additional sum of squares F test?

To test the equality of the three slopes, we want to see whether γ_1 and δ_1 are jointly significant. So we will drop them from the model and run an additional sum of squares F-test. The reduced model is

$$
\begin{bmatrix} 70 \\ 68 \\ 70 \\ 72 \\ 75 \\ 72 \\ 74 \\ 71 \\ 72 \\ 73 \\ 72 \\ 73 \end{bmatrix}_{12\times1}
=
\begin{bmatrix}
1 & 0 & 0 & 58 \\
1 & 0 & 0 & 60 \\
1 & 0 & 0 & 62 \\
1 & 0 & 0 & 65 \\
1 & 1 & 0 & 60 \\
1 & 1 & 0 & 60 \\
1 & 1 & 0 & 62 \\
1 & 1 & 0 & 60 \\
1 & 0 & 1 & 66 \\
1 & 0 & 1 & 61 \\
1 & 0 & 1 & 62 \\
1 & 0 & 1 & 65
\end{bmatrix}_{12\times4}
\begin{bmatrix} \beta_0 \\ \gamma_0 \\ \delta_0 \\ \beta_1 \end{bmatrix}_{4\times1}
+ \epsilon_{12\times1}
$$

We can test the additional regression sum of squares from include separate slopes by testing the hypothesis $H_0 : \beta_1 = \gamma_1 = \delta_1$ using the F-statistic

$$
F = \frac{\frac{SSE_{Red} - SSE_{Full}}{DFE_{Red} - DFE_{Full}}}{\frac{SSE_{Full}}{DFE_{Full}}}
$$

Total degrees of freedom in the full model is $n - 1 = 12 - 1 = 11$, and degrees of freedom for error in the full model is $n - p = 12 - 6 = 6$. Total degrees of freedom in the reduced model is $n - 1 = 12 - 1 = 11$, and degrees of freedom for error in the reduced model is $n - p = 12 - 4 = 8$. Therefore, the numerator degrees of freedom is $8 - 6 = 2$, and the denominator degrees of freesom is 6.

Problem 2

A company that produces textiles is trying to determine if the final quality is determined by production site (A), machine operator (B), and thread type (C). The company operates 3 production sites, and all 3 will participate in the experiment. At each of the 3 sites, 5 machine operators will be randomly selected. The company uses two different types of thread. Each of the operators will produce 3 samples of cloth using each type of thread, yielding a total of $3 \times 5 \times 2 \times 3 = 90$ observations.

(a) State which effects are fixed at which effects are random.

Factors:

- (A) Production Site - 3 sites total, all used in the experiment (fixed)
- (B) Machine Operator - 5 operators randomly selected (random) - nested within site
- (C) Thread Type - 2 types of thread, both in the experiment (fixed) - crossed with operator.

(b) State which effects are nested within others and which effects are crossed.

Operator is nested within site, operator is crossed with thread type, and thread type is crossed with site.

(c) Create an abbreviated ANOVA table that has two columns: one column that lists the effects in the model (including the appropriate main effects, interactions, and nested effects), and a second column that gives the degrees of freedom for each item in the first column.

The model is

$$y_{ijkl} = \mu + \alpha_i + \beta_{i(j)} + \gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{i(j)k} + \epsilon_{(ijk)l}$$

We can create an effects table:

Model Term	Factor	DF	Check DF	Trick
α_i	A	$a-1$	$= a-1$	df_A
$\beta_{i(j)}$	$B(A)$	$a(b-1)$	$= ab-a$	$a(df_B*)$
γ_k	C	$c-1$	$= c-1$	df_C
$(\alpha\gamma)_{ik}$	AC	$(a-1)(c-1)$	$= ac-a-c+1$	$(df_A)(df_C)$
$(\beta\gamma)_{i(j)k}$	$B(A)C$	$a(b-1)(c-1)$	$= abc-ab-ac+a$	$(df_B)(df_C)$
$\epsilon_{(ijk)l}$	$Error$	$abc(n-1)$	$= abcn-abc$	$abc(df_E*)$
$--$	$Total$	$abcn-1$	$= abcn-1$	df_T

An abbreviated ANOVA table for this design is

```
##    Factor DF
## 1      A   2
## 2   B(A) 12
## 3      C   1
## 4     AC   2
## 5  B(A)C 12
## 6  Error 60
## 7  Total 89
```

Problem 3

The multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \epsilon_i$$

was fit to a data set of 75 observations. The regression SS's (SSR) were partitioned sequentially into the

following:

$$SSR(X_1) = 108$$
$$SSR(X_2|X_1) = 163$$
$$SSR(X_3|X_1X_2) = 29$$
$$SSR(X_4|X_1X_2X_3) = 41$$
$$SSR(X_5|X_1X_2X_3X_4) = 26$$

The model $Y_i + \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \beta_5 X_{i5} + \epsilon_i$ was also fit to the same data and the following ANOVA was calculated:

Source	SS
Regression	214
Residual Error	489
Total	703

Answer the following from the above information:

(a) Calculate the F-statistic for testing the hypothesis (H_0) that X_3, X_4, X_5 have no significant effect on the response Y.

We are testing the hypothesis

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

We need the sum of squares due to error and degrees of freedom in the reduced model. The reduced model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

We know that $SSE_{Red} = SST_{Red} - SSR_{Red}$.

$$\begin{aligned}
SSE_{Full} &= SST_{Full} - SSR_{Full} \\
&= 703 - (108 + 163 + 29 + 41 + 26) \\
&= 336
\end{aligned}$$

and

$$\begin{aligned}
SSE_{Red} &= SST_{Red} - SSR_{Red} \\
&= 703 - (163 + 108) \\
&= 432
\end{aligned}$$

The F statistic can be calculated as

```
n <- 75

SSE_Full <- 336
SSE_Red  <- 432

DFE_Full <- n - 6
DFE_Red  <- n - 3

F <- ((SSE_Red - SSE_Full)/(DFE_Red - DFE_Full))/(SSE_Full/DFE_Full)
F
```

```
## [1] 6.571429
```

(b) Calcaulate R^2 for the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$.

We can calculate the coefficient of multiple determination, R^2 (unadjusted), as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{703 - (108 + 163)}{703} = 1 - \frac{432}{703} = 1 - 0.6145092 = 0.3854908$$

(c) Calculate R^2_{adj} for the model in part (b)

We can calculate the adjusted R^2 as

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = 1 - \frac{(n-1)}{(n-p)}(1-R^2) = 1 - \frac{(75-1)}{(75-3)}(1-0.3854908) = 1 - 1.027778(0.6145092) = 0.368421$$

(d) Calculate the F_statistic for testing $H_0 : \beta_2 = \beta_4 = 0$.

The sum of squares due to error for the reduced model is given to us as 489. The degrees of freedom for error in the reduced model are $n - p = 75 - 4 = 71$. Now we can calculate the F-statistic:

```
n <- 75

SSE_Full <- 336
SSE_Red  <- 489

DFE_Full <- n - 6
DFE_Red  <- n - 4

F <- ((SSE_Red - SSE_Full)/(DFE_Red - DFE_Full))/(SSE_Full/DFE_Full)
F
```

```
## [1] 15.70982
```

Problem 4

Assume the model $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$, $i = 1, \ldots, n$ wutg tge addutuibak restructuibs tgat $\beta = 0$ and $\beta_0 = 2\beta_2$. Find the least-squares estimators of β_0, β_1, and β_2.

Substitute restrictions into the equation for y_i:

$$y_i = 2\beta_2 + \beta_2 x_i^2 + \epsilon_i$$
$$= \beta_2(2 + x_i^2) + \epsilon_i$$

Solve for $\sum_{i=1}^{n} \epsilon^2$:

$$\sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} \left(y_i - \beta_2(2 + x_i^2)\right)^2$$

Take the derivative with respect to β_2:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(y_i - \beta_2(2 + x_i^2)\right)^2$$
$$= (-2)\sum_{i=1}^{n} \left(y_i - \beta_2(2 + x_i^2)\right)(2 + x_i^2)$$

5

Set equal to 0 and solve for $\hat{\beta}_2$.

$$0 = \sum_{i=1}^{n} \left[(-2)\left(y_i - \beta_2(2 + x_i^2)\right)(2 + x_i^2)\right]$$

$$\implies 0 = (-2)\sum_{i=1}^{n}\left(y_i(2 + x_i^2) - \beta_2(2 + x_i^2)^2\right)$$

$$\implies 0 = \sum_{i=1}^{n} y_i(2 + x_i^2) - \sum_{i=1}^{n}\beta_2(2 + x_i^2)^2$$

$$\implies 0 = \sum_{i=1}^{n} y_i(2 + x_i^2) - \beta_2\sum_{i=1}^{n}(2 + x_i^2)^2$$

$$\implies \beta_2\sum_{i=1}^{n}(2 + x_i^2)^2 = \sum_{i=1}^{n} y_i(2 + x_i^2)$$

$$\implies \hat{\beta}_2 = \frac{\sum_{i=1}^{n} y_i(2 + x_i^2)}{\sum_{i=1}^{n}(2 + x_i^2)^2}$$

Now, since $\beta_0 = 2\beta_2$, we have that

$$\hat{\beta}_0 = 2\frac{\sum_{i=1}^{n} y_i(2 + x_i^2)}{\sum_{i=1}^{n}(2 + x_i^2)^2}$$

$\beta_1 = 0$, so $\hat{\beta}_1 = 0$.