BIOSTAT 650 SEC 002 HW5

Jaehoon Kim

2024-10-16

Problem 1

(a)

Let the full model be:

$$
\begin{aligned}
Y_{dep} =& \beta_0 + \beta_{age}X_{age} + \beta_{sex}X_{sex} + \beta_{RE}X_{RE} + \beta_{edu}X_{edu} + \beta_{hty}X_{hty} + \beta_{hch}X_{hch} + \beta_{AF}X_{AF} + \beta_{CAD}X_{CAD} + \beta_{dia}X_{dia} \\
&+ \beta_{str}X_{str} + \beta_{NIH}X_{NIH} + \beta_{com}X_{com} + \beta_{fat}X_{fat} + \epsilon
\end{aligned}
$$

$$\hat{\sigma}^2 = SSE(\beta_0, \beta_{age}, \beta_{sex}, \beta_{RE}, \beta_{edu}, \beta_{hty}, \beta_{hch}, \beta_{AF}, \beta_{CAD}, \beta_{dia}, \beta_{str}, \beta_{NIH}, \beta_{com}, \beta_{fat})/(n-15)$$

The sequence of hypothesis is

1. $H_0 : \beta_{age} = \beta_{sex} = \beta_{RE} = 0$ vs. $H_1 : \beta_{age} \neq 0$, $\beta_{sex} \neq 0$ or $\beta_{RE} \neq 0$.

$$F = \frac{SS(\beta_{age}, \beta_{sex}, \beta_{RE}|\beta_0)/3}{\hat{\sigma}^2} \sim F_{3,n-15}$$

If the null hypothesis is rejected, go to the next step; otherwise, stop.

2. $H_0 : \beta_{edu} = 0$ vs. $H_1 : \beta_{edu} \neq 0$,adjusting for age, sex and race/ethnicity.

$$F = \frac{SS(\beta_{edu}|\beta_{age}, \beta_{sex}, \beta_{RE}, \beta_0)/1}{\hat{\sigma}^2} \sim F_{1,n-15}$$

If null hypothesis is rejected, go the next step. Otherwise, stop.

3. $H_0 : \beta_{hty} = \beta_{hch} = \beta_{AF} = \beta_{CAD} = \beta_{dia} = \beta_{str} = 0$ vs. $H_1 : \beta_{hty} \neq 0, \beta_{hch} \neq 0, \beta_{AF} \neq 0, \beta_{CAD} \neq 0, \beta_{dia} \neq 0, or \beta_{str} \neq 0$,adjusting for age, sex, race/ethnicity, and education.

$$F = \frac{SS(\beta_{hty}, \beta_{hch}, \beta_{AF}, \beta_{CAD}, \beta_{dia}, \beta_{str}|\beta_{edu}, \beta_{age}, \beta_{sex}, \beta_{RE}, \beta_0)/7}{\hat{\sigma}^2} \sim F_{7,n-15}$$

If the null hypothesis is rejected, go to the next step. Otherwise, stop.

4. $H_0 : \beta_{NIH} = \beta_{com} = 0$ vs. $H_1 : \beta_{NIH} \neq 0 or \beta_{com} \neq 0$, adjusting for age, sex, race/ethnicity,education and stroke risk factors.

$$F = \frac{SS(\beta_{NIH}, \beta_{com}|\beta_{hty}, \beta_{hch}, \beta_{AF}, \beta_{CAD}, \beta_{dia}, \beta_{str}, \beta_{edu}, \beta_{age}, \beta_{sex}, \beta_{RE}, \beta_0)/2}{\hat{\sigma}^2} \sim F_{2,n-15}$$

If the null hypothesis is rejected, go to the next step. Otherwise, stop.

5.$H_0 : \beta_{fat} = 0$ vs. $H_1 : \beta_{fat} \neq 0$, adjusting for age, sex, race/ethnicity,education and stroke risk factors, stroke severity, and comorbity

$$F = \frac{SS(\beta_{fat}|\beta_{NIH}, \beta_{com}, \beta_{hty}, \beta_{hch}, \beta_{AF}, \beta_{CAD}, \beta_{dia}, \beta_{str}, \beta_{edu}, \beta_{age}, \beta_{sex}, \beta_{RE}, \beta_0)/1}{\hat{\sigma}^2} \sim F_{2,n-15}$$

If null hypothesis is rejected, go to the next step. Otherwise, stop.

(b)

Let the following vector notations of beta coeffecients be defined as such:

(i) $\boldsymbol{\beta}_{demo} = (\beta_{age}, \beta_{sex}, \beta_{RE})^T$
(ii) $\boldsymbol{\beta}_{Edu} = (\beta_{edu})^T$
(iii) $\boldsymbol{\beta}_{Risk} = (\beta_{hty}, \beta_{hch}, \beta_{AF}, \beta_{CAD}, \beta_{dia}, \beta_{str})^T$
(iv) $\boldsymbol{\beta}_{Stroke} = (\beta_{NIH}, \beta_{com})^T$
(v) $\boldsymbol{\beta}_{Fatalism} = (\beta_{fat})^T$

Let A, B denote the SS numerator and SSE denominator for the F test respectively.

Step	Var. Tested	$SS(num.)$	$SS(Denom.)$	$TestStatistics$
$I.$	$\boldsymbol{\beta}_{demo}$	$SS(\boldsymbol{\beta}_{demo}\|\beta_0)$	$SSE(\boldsymbol{\beta}_{Fatalism}, \boldsymbol{\beta}_{Stroke}, \boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$F = \frac{A/3}{B/597}$
$II.$	$\boldsymbol{\beta}_{Edu}$	$SS(\boldsymbol{\beta}_{Edu}\|\boldsymbol{\beta}_{demo}, \beta_0)$	$SSE(\boldsymbol{\beta}_{Fatalism}, \boldsymbol{\beta}_{Stroke}, \boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$F = \frac{A/1}{B/597}$
$III.$	$\boldsymbol{\beta}_{Risk}$	$SS(\boldsymbol{\beta}_{Risk}\|\boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$SSE(\boldsymbol{\beta}_{Fatalism}, \boldsymbol{\beta}_{Stroke}, \boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$F = \frac{A/7}{B/597}$
$IV.$	$\boldsymbol{\beta}_{Stroke}$	$SS(\boldsymbol{\beta}_{Stroke}\|\boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$SSE(\boldsymbol{\beta}_{Fatalism}, \boldsymbol{\beta}_{Stroke}, \boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$F = \frac{A/2}{B/597}$
$V.$	$\boldsymbol{\beta}_{Fatalism}$	$SS(\boldsymbol{\beta}_{Fatalism}\|\boldsymbol{\beta}_{Stroke}, \boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$SSE(\boldsymbol{\beta}_{Fatalism}, \boldsymbol{\beta}_{Stroke}, \boldsymbol{\beta}_{Risk}, \boldsymbol{\beta}_{Edu}, \boldsymbol{\beta}_{demo}, \beta_0)$	$F = \frac{A1}{B/597}$

Result table below

```r
mySASData = read.sas7bdat("C:/Users/aquil/Desktop/STAT 650/completedata.sas7bdat")

var = c(
"Depression",
"Age",
"Sex",
"R_E",
"Education",
"Htn",
"HiChol",
"Afib",
"Cad",
"Db",
"CurrentSmoker",
"HxStroke",
"NIHScore",
"Comorbidity1",
"Fatalism"
)
newData = mySASData[,var]
n = nrow(newData)
full_model = paste0(var[-1], collapse = ' + ')
full_model = paste0("Depression",' ~ ', full_model)
design = model.matrix(as.formula(full_model), data = newData)
full.lm = lm(formula = as.formula(full_model),data = newData)
#Analysis of Variance Table
aov = anova(full.lm)
aov
```

```
## Analysis of Variance Table
##
## Response: Depression
```

```
##                 Df  Sum Sq Mean Sq F value    Pr(>F)    
## Age              1   640.1  640.10 26.2341 4.087e-07 ***
## Sex              1    87.6   87.59  3.5899  0.058616 .  
## R_E              1     9.3    9.35  0.3831  0.536168    
## Education        1   108.0  107.95  4.4244  0.035846 *  
## Htn              1   561.3  561.34 23.0062 2.042e-06 ***
## HiChol           1    16.5   16.47  0.6750  0.411656    
## Afib             1    22.7   22.66  0.9286  0.335606    
## Cad              1     4.2    4.22  0.1729  0.677737    
## Db               1   186.5  186.47  7.6424  0.005877 ** 
## CurrentSmoker    1    41.6   41.59  1.7046  0.192194    
## HxStroke         1    48.0   48.01  1.9678  0.161200    
## NIHScore         1    42.6   42.55  1.7440  0.187136    
## Comorbidity1     1    20.2   20.16  0.8264  0.363687    
## Fatalism         1   812.9  812.85 33.3143 1.261e-08 ***
## Residuals      597 14566.5   24.40                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
SSY = sum(aov$`Sum Sq`)
SSQ = aov$`Sum Sq`
MSE = aov$`Mean Sq`[15]
#Sum of Squares for each step:
ss1 = sum(SSQ[1:3])
print(ss1)
```

```
## [1] 737.0381
```

```r
ss2 = sum(SSQ[4])
print(ss2)
```

```
## [1] 107.9535
```

```r
ss3 = sum(SSQ[5:11])
print(ss3)
```

```
## [1] 880.7591
```

```r
ss4 = sum(SSQ[12:13])
print(ss4)
```

```
## [1] 62.71662
```

```r
ss5 = sum(SSQ[14])
print(ss5)
```

```
## [1] 812.8529
```

```r
#Test Statistics for each step:
fstat1 = ss1/3/MSE
pval1 = 1-pf(q = fstat1, df1 = 3, df2 = n-15)
print(c(fstat1, pval1))
```

```
## [1] 1.006902e+01 1.764708e-06
```

```r
fstat2 = ss2/1/MSE
pval2 = 1-pf(q = fstat2, df1 = 1, df2 = n-15)
print(c(fstat2, pval2))
```

```
## [1] 4.42441053 0.03584616
```

```r
fstat3 = ss3/7/MSE
pval3 = 1-pf(q = fstat3, df1 = 7, df2 = n-15)
print(c(fstat3, pval3))
```

```
## [1] 5.156769e+00 1.028334e-05
```

```r
fstat4 = ss4/2/MSE
pval4 = 1-pf(q = fstat4, df1 = 2, df2 = n-15)
print(c(fstat4, pval4))
```

```
## [1] 1.2852016 0.2773589
```

```r
fstat5 = ss5/1/MSE
pval5 = 1-pf(q = fstat5, df1 = 1, df2 = n-15)
print(c(fstat5, pval5))
```

```
## [1] 3.331429e+01 1.260721e-08
```

Step	Tested Var.	SS(Num.)	SS(Denom.)	Test Stat.	Dist.	p-value	Decision	Stopping Rule	Decision
I	β_{Demo}	737.0381	14566.5	1.006902e+01	$F_{3,597}$	1.764708e-06	`Reject`	`Do not stop`	`Collect`
II	β_{Edu}	107.9535	14566.5	4.42441053	$F_{1,597}$	0.03584616	`Reject`	`Do not stop`	`Collect`
III	β_{Risk}	880.7591	14566.5	5.156769e+00	$F_{7,597}$	1.028334e-05	`Reject`	`Do not stop`	`Collect`
IV	β_{Stroke}	62.71662	14566.5	1.2852016	$F_{2,597}$	0.2773589	`Fail to Reject`	`Stop`	`Not Collect`
V	$\beta_{Fatalism}$	812.8529	14566.5	3.331429e+01	$F_{1,597}$	1.260721e-08	`Reject`	`NA`	`Not Collect`

If we use the stopping rule as shown above, We would not be able to test whether fatalism is associated with depression after adjusting for all other covariates previously tested, even though we are able to see that the p-value for the test would have been very small had we been able to test it.