HUL315

ASSIGNMENT 3

Answer 1.

The alcohol prohibition is a good move in India. Many factors can be studied while focusing on it as its also a great revenue source.

Answer 2.

Part a-

Using R Studio to perform pooled OLS estimate to trmgpa as the dependent variable and spring, sat, hsperc, female, black, white, frstsem, tothrs, crsgpa, and season as explanatory variable.

Using the following code-

```
# Loading Libraries

library(readxl)

# Creating Dataframe

columns <-
c("term","sat","tothrs","cumgpa","season","frstsem","crsgpa","verbmath","trmgpa","hssize","hsrank","id","sprin
g","female","black","white","ctrmgpa","ctothrs","ccrsgpa","ccrspop","cseason","hsperc","football")

D <- read_excel("C:/Users/ROHAN SHARMA/Downloads/GPA3.xls",col_names=FALSE)

colnames(D) <- columns

# Part a

#Defing functions

function1 <- trmgpa~spring+sat+hsperc+female+black+white+frstsem+tothrs+crsgpa+season

# Extracting summary

model1=lm(formula=function1,data=D)

summary(model1)
```

This gives the following summary-

```
> summary(model1)

Call:
lm(formula = function1, data = D)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.84899 -0.33132  0.01915  0.38002  1.57924 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.7528474  0.3479049  -5.038 5.94e-07 ***
spring      -0.0580066  0.0480368  -1.208    0.228    
sat          0.0016984  0.0001494  11.367  < 2e-16 ***
hsperc      -0.0086610  0.0010363  -8.358 3.28e-16 ***
female       0.3504013  0.0518524   6.758 2.89e-11 ***
black       -0.2541495  0.1229216  -2.068    0.039 *  
white       -0.0233146  0.1173954  -0.199    0.843    
frstsem     -0.0346585  0.0760345  -0.456    0.649    
tothrs      -0.0003389  0.0007267  -0.466    0.641    
crsgpa       1.0478655  0.1041144  10.065  < 2e-16 ***
season      -0.0272904  0.0490460  -0.556    0.578    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5519 on 721 degrees of freedom
Multiple R-squared:  0.4776,	Adjusted R-squared:  0.4703 
F-statistic: 65.91 on 10 and 721 DF,  p-value: < 2.2e-16
```

This gives the equation-

$$
\begin{aligned}
trmgpa = -1.75 - 0.058 spring + 0.0017 sat - 0.0087 hsperc \\
+ 0.350\, female - 0.254 black - 0.023 white - 0.035\, frstsem \\
- 0.0003 tothrs + 1.048 crsgps - 0.027 season
\end{aligned}
$$

And $R^2 = 0.4776$, $\bar{R}^2 = 0.4703$ and $n = 732$

This clearly shows that the **coefficient of season** is -0.027, which indicates that the GPA changes by -0.027 which change in season, keeping the other variables constant.

Now finding the statistical significance of this coefficient-

The t-statistic for season is-

$$t = -0.556$$

The p-vale for season is-

$$p - value = 0.578$$

Now as $p - value > t$, this shows that the coefficient of season is statistically insignificant.

Part b-

The ability levels are not included, and it's correlated with the variable season. Now as in the fall only football is played, here the variable season is negatively correlated with the negative term.

Using the code-

```
# Loading Libraries


library(readxl)


# Creating Dataframe


columns <-
c("term","sat","tothrs","cumgpa","season","frstsem","crsgpa","verbmath","trmgpa","hssize","hsrank","id","spring","fe
male","black","white","ctrmgpa","ctothrs","ccrsgpa","ccrspop","cseason","hsperc","football")

D <- read_excel("C:/Users/ROHAN SHARMA/Downloads/GPA3.xls",col_names=FALSE)

colnames(D) <- columns


# Part a

#Defing functions


function1 <- trmgpa~spring+sat+hsperc+female+black+white+frstsem+tothrs+crsgpa+season


#Part b

#For the fall semester

D2 = subset(D,term == 1)

model2 = lm(formula = function1,data = D2)

summary(model2)
```

Gives the summary –

```
> summary(model2)

Call:
lm(formula = function1, data = D2)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.66315 -0.34265  0.03271  0.40701  1.58943 

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.569e+00  5.258e-01  -2.984  0.00304 ** 
spring              NA         NA      NA       NA    
sat          1.637e-03  2.161e-04   7.576 3.11e-13 ***
hsperc      -8.523e-03  1.494e-03  -5.706 2.44e-08 ***
female       3.242e-01  7.204e-02   4.501 9.19e-06 ***
black       -2.399e-01  1.769e-01  -1.356  0.17603    
white       -4.794e-02  1.686e-01  -0.284  0.77625    
frstsem     -2.017e-02  9.602e-02  -0.210  0.83376    
tothrs      -2.057e-05  1.214e-03  -0.017  0.98650    
crsgpa       1.028e+00  1.540e-01   6.677 9.37e-11 ***
season      -1.156e-01  8.392e-02  -1.378  0.16906    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5611 on 356 degrees of freedom
Multiple R-squared:  0.447,	Adjusted R-squared:  0.433 
F-statistic: 31.97 on 9 and 356 DF,  p-value: < 2.2e-16
```

And in the spring season, the ability levels and season have a positive correlation. This is done using the code

```
# Loading Libraries

library(readxl)

# Creating Dataframe

columns <-
c("term","sat","tothrs","cumgpa","season","frstsem","crsgpa","verbmath","trmgpa","hssize","hsrank","id","spring","female","black","white","ctrmgpa","ctothrs","ccrsgpa","ccrspop","cseason","hsperc","football")

D <- read_excel("C:/Users/ROHAN SHARMA/Downloads/GPA3.xls",col_names=FALSE)

colnames(D) <- columns

# Part a

#Defing functions

function1 <- trmgpa~spring+sat+hsperc+female+black+white+frstsem+tothrs+crsgpa+season

#For the spring semester

D3 = subset(D,term == 2)

model3 = lm(formula = function1,data = D3)

summary(model3)
```

Gives the summary-

```
> summary(model3)

Call:
lm(formula = function1, data = D3)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.86902 -0.31699  0.00582  0.32627  1.39303 

Coefficients: (2 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.9553415  0.4708465  -4.153 4.11e-05 ***
spring              NA         NA      NA       NA    
sat          0.0017585  0.0002088   8.424 9.02e-16 ***
hsperc      -0.0086450  0.0014540  -5.946 6.56e-09 ***
female       0.3686613  0.0770826   4.783 2.53e-06 ***
black       -0.2568414  0.1721247  -1.492    0.137    
white       -0.0051944  0.1650027  -0.031    0.975    
frstsem             NA         NA      NA       NA    
tothrs      -0.0005233  0.0009332  -0.561    0.575    
crsgpa       1.0721855  0.1443017   7.430 8.08e-13 ***
season       0.0008929  0.0647980   0.014    0.989    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5464 on 357 degrees of freedom
Multiple R-squared:  0.5101,    Adjusted R-squared:  0.4991 
F-statistic: 46.46 on 8 and 357 DF,  p-value: < 2.2e-16
```

Now due to this not included variable ability levels, which is correlated with the variable season, the model suffers a problem of biased estimators of pooled OLS.

Part c-

On the glance, the variable like sat, hsperc, female, black, white can be dropped from the model, as they are not affected by the variable semester.

Using the following code to find out the summary-

```
# Loading Libraries


library(readxl)


# Creating Dataframe


columns <-
c("term","sat","tothrs","cumgpa","season","frstsem","crsgpa","verbmath","trmgpa","hssize","hsrank","id","spring","female","black","white","ctrmgpa","ctothrs","ccrsgpa","ccrspop","cseason","hsperc","football")

D <- read_excel("C:/Users/ROHAN SHARMA/Downloads/GPA3.xls",col_names=FALSE)

colnames(D) <- columns


#Part c

#Defining functions


deltatrmgpa <- diff(D$trmgpa)

deltafrstsem <- diff(D$frstsem)

deltatothrs <- diff(D$tothrs)

deltacrsgpa <- diff(D$crsgpa)

deltaseason <- diff(D$season)


deltatrmgpa <- c(deltatrmgpa[seq(length(deltafrstsem))%%2 == 1])

deltafrstsem <- c(deltafrstsem[seq(length(deltafrstsem))%%2 == 1])

deltatothrs <- c(deltatothrs[seq(length(deltafrstsem))%%2 == 1])

deltacrsgpa <- c(deltacrsgpa[seq(length(deltafrstsem))%%2 == 1])

deltaseason <- c(deltaseason[seq(length(deltafrstsem))%%2 == 1])


function2 <- deltatrmgpa~deltatothrs+deltacrsgpa+deltafrstsem+deltaseason

D4 = data.frame(deltatrmgpa,deltafrstsem,deltatothrs,deltacrsgpa,deltaseason)


#Extracting summary


model4 = lm(formula = function2,data = D4)

summary(model4)
```

The summary is-

```
> summary(model4)

Call:
lm(formula = function2, data = D4)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.46328 -0.33017  0.01223  0.36800  2.04326 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.23656    0.20597  -1.149    0.252    
deltatothrs   0.01215    0.01439   0.845    0.399    
deltacrsgpa   1.13635    0.11881   9.564   <2e-16 ***
deltafrstsem  0.01915    0.06927   0.276    0.782    
deltaseason  -0.06450    0.04252  -1.517    0.130    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5784 on 361 degrees of freedom
Multiple R-squared:  0.208,	Adjusted R-squared:  0.1992 
F-statistic:  23.7 on 4 and 361 DF,  p-value: < 2.2e-16
```

$$\Delta trmgpa = -0.237 + 0.019\Delta tothrs + 1.136\Delta crsgpa - 0.065\Delta season$$

This gives $R^2 = 0.208$ and $n = 366$

Now the coefficient of $\Delta season$ is -0.065, and the t-statistic is $t = -1.517$, and the p-value is $p - value = 0.130$.

Part d-

One of the major factors excluded here the academic load. Now as we know that no of courses per semester may vary college to college, it should be considered. This is directly connected, as can be seen in our college to, athletes with less number of course load may score higher GPA.