S530 HW 3

Erick Castillo

10/24/2021

Problem 1

This problem uses the data set infmort. The specified model in this case log(mortality)= $\beta_0 + \beta_1 log(income)_{1i} + \beta_2(region)_{2i} + \epsilon_i$.

A. State the null and alternative hypothesis for the global F-test for this model. Perform the test and summarize the results.

The test uses the following hypothesis: $H_0 : \beta_1 = \beta_2 = 0$ $H_A : \beta_i \neq 0$ for some $i \in \{1, 2\}$

The test is performed in the following chunk:

```
log.mort <- log(infmort$mortality)
full.mod1 <- lm(log.mort~log(income)+region, data = infmort)
redu.mod1 <- lm(log.mort~1, data = infmort)

anova(redu.mod1, full.mod1)
```

```
## Analysis of Variance Table
##
## Model 1: log.mort ~ 1
## Model 2: log.mort ~ log(income) + region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    100 93.769
## 2     96 35.980  4    57.789 38.547 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above ANOVA output, we can conclude that there is strong evidence to suggest that at least one of the slope parameters is significant in the model. This means that I would reject H_0.

B. Explain the practical meaning of the hypothesis $H_0 : \beta_2 = 0$ in the context of the above model.

The null hypothesis $H_0 : \beta_2 = 0$ poses the question of whether the slope of the region variable in the above model is significant. The alternative hypothesis would be $H_1 : \beta_2 \neq 0$.

C. Perform the test introduced in part B of this problem.

The following chunk performs the test to see if the slope of the region variable is significant:

```
redu.mod2 <- lm(log.mort~log(income), data = infmort)
anova(redu.mod2, full.mod1)
```

```
## Analysis of Variance Table
##
## Model 1: log.mort ~ log(income)
## Model 2: log.mort ~ log(income) + region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     99 46.685
## 2     96 35.980  3    10.705 9.5211 1.449e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above ANOVA output we can conclude that there is strong evidence to suggest that slope of the region variable is significant in the model. That is, I would reject $H_0 : \beta_2 = 0$.

Problem 2

This problem uses the sat data set.

A. Using total as the response variable and expend and takers as response variables, test the hypothesis that $\beta_{expend} = \beta_{takers} = 0$. Do any of the two predictors have an effect on total?

The following chunk of code runs a Global F-Test on the two predictor variables above:

```
full.mod2 <- lm(total~expend+takers, data = sat)
redu.mod3 <- lm(total~1, data = sat)

anova(redu.mod3, full.mod2)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ 1
## Model 2: total ~ expend + takers
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     49 274308
## 2     47  49520  2    224788 106.67 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above ANOVA output indicates that there is strong evidence that at least one of slopes of the aforementioned predictors is not zero. In other words, I would reject $H_0 : \beta_{expend} = \beta_{takers} = 0$.

To see if the predictors have an effect on total see the following output:

```
summary(full.mod2)
```

```
##
## Call:
## lm(formula = total ~ expend + takers, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.400 -22.884   1.968  19.142  68.755
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 993.8317    21.8332  45.519  < 2e-16 ***
## expend       12.2865     4.2243   2.909  0.00553 **
## takers       -2.8509     0.2151 -13.253  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.46 on 47 degrees of freedom
## Multiple R-squared:  0.8195, Adjusted R-squared:  0.8118
## F-statistic: 106.7 on 2 and 47 DF,  p-value: < 2.2e-16
```

Observing the p-values in the above model, it is clear that expend and takers are both useful predictors in the model.

B. Now add ratio to the model. Test the hypothesis that $\beta_{ratio} = 0$. Compare this model to the previous one using an F-test. Show that the F-test and t-test here are equivalent.

The following chunk of code adds ratio to the model and tests the above hypothesis:

```
fuller.mod <- update(full.mod2, .~. + ratio)
anova(full.mod2, fuller.mod)
```

```
## Analysis of Variance Table
##
## Model 1: total ~ expend + takers
## Model 2: total ~ expend + takers + ratio
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     47 49520
## 2     46 48627  1    892.74 0.8445 0.3629
```

Notice that the above test indicates that the slope of the ratio variable appears to be insignificant, that is, we would fail to reject $H_0 : \beta_{ratio} = 0$. Notice that the p-value of the above output is 0.3629, this is the p-value for the F-test of the variable ratio.

The following chunk of code is a summary of the model with ratio.

```
summary(fuller.mod)
```

```
##
## Call:
## lm(formula = total ~ expend + takers + ratio, data = sat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.284 -21.130   1.414  16.709  66.073
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1035.4739    50.3155  20.580   <2e-16 ***
## expend        11.0140     4.4521   2.474   0.0171 *
## takers        -2.8491     0.2155 -13.222   <2e-16 ***
## ratio         -2.0282     2.2071  -0.919   0.3629
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32.51 on 46 degrees of freedom
## Multiple R-squared:  0.8227, Adjusted R-squared:  0.8112
## F-statistic: 71.16 on 3 and 46 DF,  p-value: < 2.2e-16
```

Now, looking at the p-value of the variable ratio, we can see that it is 0.3629. This is the p-value generated using a t-test. That is, the F-test and t-test p-values are equivelant, meaning that in both cases we would fail to reject $H_0 : \beta_{ratio} = 0$.

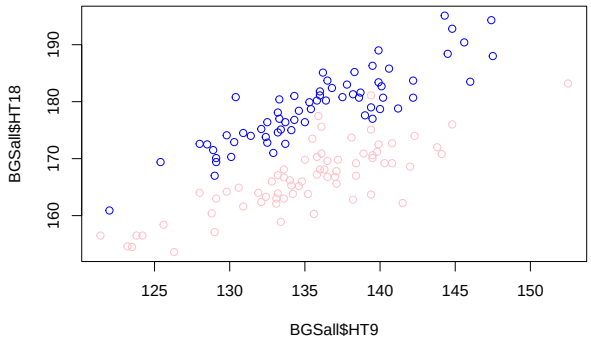
Problem 3

This problem uses the BGSall data set in the alr4 package. We will be considering the regression of HT18 on HT9 and the grouping factor Sex.

A. Draw the scatterplot of HT18 vs HT9, use different colors to indicate the different sexes in the dataset. Comment on an appropriate model for the data.

```
BGSall$Sex <- factor(BGSall$Sex) # 0 is for males, 1 for females

plot(BGSall$HT9, BGSall$HT18, col = c('blue','pink')[BGSall$Sex])
```



It is clear from the above plot that the observed males tend to be taller than the observed females. In this case, I would probably need to implement a dummy/indicator variable into the model to account for the difference in the heights.

B. We fit the model, and test the significance of the indicator variable:

```
mod1 <- lm(HT18~HT9+Sex, data = BGSall)
summary(mod1)
```

```
##
## Call:
## lm(formula = HT18 ~ HT9 + Sex, data = BGSall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -10.4694  -2.0952  -0.0136   1.7101  10.4467
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  48.51731    7.33385   6.616 8.27e-10 ***
## HT9           0.96006    0.05388  17.819  < 2e-16 ***
## Sex1        -11.69584    0.59036 -19.811  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.432 on 133 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8494
## F-statistic: 381.7 on 2 and 133 DF,  p-value: < 2.2e-16
```

We can see from the above output that the dummy variable is significant in the model, as the t-value's distance is very far from 0.

C. Obtain a 95% confidence interval for the difference between males and females.

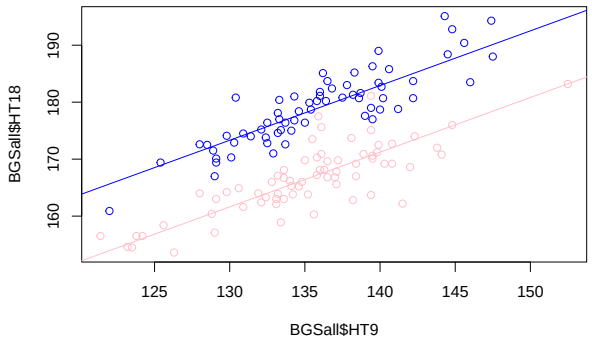
```
confint(mod1)
```

```
##                   2.5 %     97.5 %
## (Intercept)  34.0112360  63.023384
## HT9           0.8534845   1.066628
## Sex1        -12.8635477 -10.528134
```

Recall that females are coded as 1 in this dataset. Using the above output, we can be 95% confident that women in this dataset are between 10.53cm and 12.86cm shorter than men.

D. Add the parallel regression line to the scatterplot generated in part A of this problem:

```
plot(BGSall$HT9, BGSall$HT18, col = c('blue','pink')[BGSall$Sex])
abline(a = 48.517, b = 0.96, col = 'blue') # regression line for males
abline(a = 36.821, b = 0.96, col = 'pink') # regression line for females
```

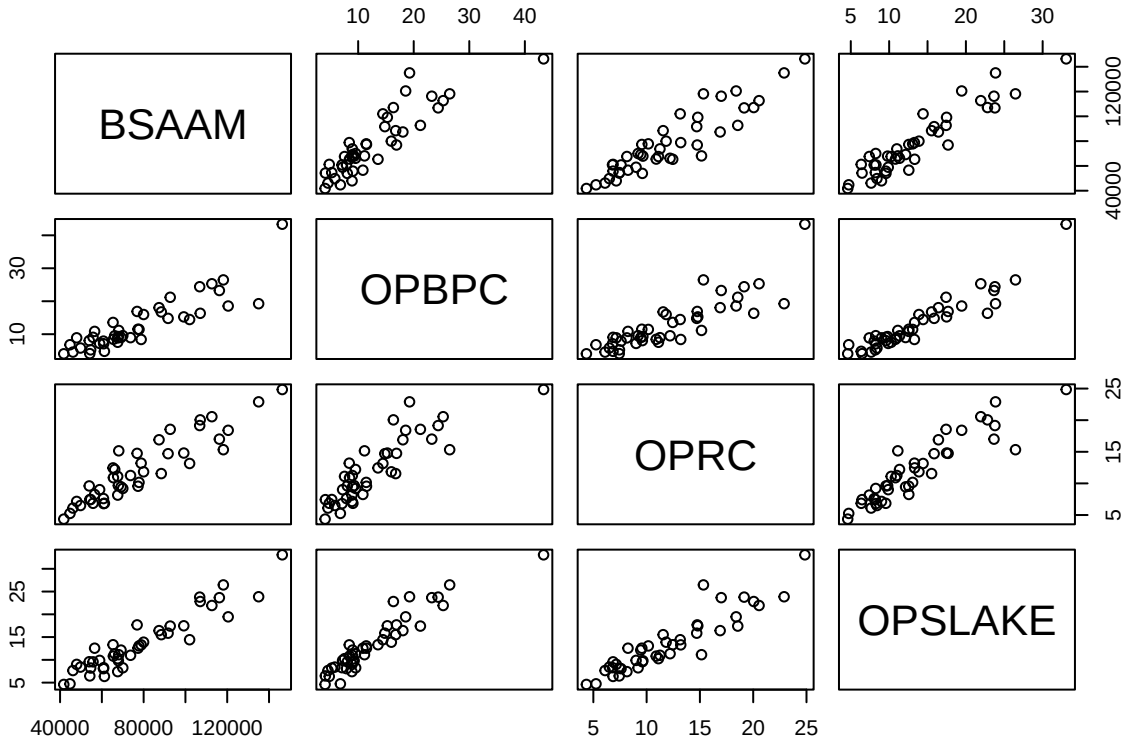


Problem 4

This problem uses the water data set in alr4. Use BSAAM as the response and OPBPC, OPRC, and OPSLAKE as predictors.

A. Examine the scatterplot matrix. Explain what the correlation matrix should look like, then compute the correlation matrix to verify.

```
water1 <- water[, c('BSAAM','OPBPC','OPRC','OPSLAKE')]
pairs(water1)
```



It appears that the correlation matrix should have high values for all the relationships that are present. That is, $r \geq 0.5$ for all entries. The following code should verify this:

```
corrplot(cor(water1), method = 'number')
```

And it does.

B. Get the regression summary of BSAAM on the three regressors with OPBPC, OPRC, and OPSLAKE included sequentially.

```
mod2 <- lm(BSAAM~OPBPC+OPRC+OPSLAKE, data=water)
summary(mod2)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4   4741.9  19921.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
## OPBPC          40.61     502.40   0.081  0.93599
## OPRC         1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

The $\Pr(>|t|)$ column is the p-value of the corresponding variable, while the other variables that are present in the model. For example, OPBPC would have a high p-value given OPRC and OPSLAKE are included in the model.

C. Find SSR(OPSLAKE|OPRC, OPBPC) and SSE(OPBPC, OPRC) using anova().

```
mod2.reduced <- lm(BSAAM~OPBPC+OPRC, data=water)
anova(mod2.reduced, mod2)
```

```
## Analysis of Variance Table
##
## Model 1: BSAAM ~ OPBPC + OPRC
## Model 2: BSAAM ~ OPBPC + OPRC + OPSLAKE
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1     40 3331163233
## 2     39 2689509185  1 641654049 9.3045 0.004097 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the above output, we can find that SSE(OPBPC, OPRC)$= 3,331,163,233$.

Notice that SSE(OPBPC, OPRC)$-$SSE(OPBPC, OPRC, OPSLAKE)$=$SSR(OPSLAKE|OPBPC, OPRC). We can then calculate that SSR(OPSLAKE|OPBPC, OPRC)$= 641,654,048$