705604096_stats101a_hw7

Jade Gregory

2023-05-17

Question 1

Part A

```
wwh <- read.delim('waistweightheight.txt')
```

a)

```
weight.lm <- lm(Weight ~ Waist + Height, data = wwh)
weight.lm
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height, data = wwh)
##
## Coefficients:
## (Intercept)        Waist       Height
##    -165.533        4.960        2.488
```

Weight = 2.488 * Height + 4.96 * Waist - 165.533

i.

```
anova(weight.lm)
```

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq F value    Pr(>F)
## Waist       1 358074  358074 3590.77 < 2.2e-16 ***
## Height      1  29843   29843  299.26 < 2.2e-16 ***
## Residuals 504  50259     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

RSS = 50259 SSReg = 358074 + 29843 = 387917 SYY = RSS + SSReg = 438176

ii.

```
summary(weight.lm)
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height, data = wwh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.760  -6.405  -0.420   5.656  45.474
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -165.5332     8.2517  -20.06   <2e-16 ***
## Waist          4.9605     0.1229   40.37   <2e-16 ***
## Height         2.4884     0.1438   17.30   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.986 on 504 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8848
## F-statistic:  1945 on 2 and 504 DF,  p-value: < 2.2e-16
```

The multiple r-squared value is 0.8853 and the adjusted r-squared value is 0.8848.

iii. The slope for height represents that among people with the same waist size, those who are 1 inch taller are an average of 2.488 pounds heavier.

b)

```
set.seed(23)
new.df <- transform(wwh,worthless = rnorm(dim(wwh)[1],0,5))
```

i.

```
weight.lm2 <- update(weight.lm, .~. + new.df$worthless)
anova(weight.lm2)
```

```
## Analysis of Variance Table
##
## Response: Weight
##                    Df Sum Sq Mean Sq   F value Pr(>F)
## Waist               1 358074  358074 3584.4800 <2e-16 ***
## Height              1  29843   29843  298.7400 <2e-16 ***
## new.df$worthless    1     12      12    0.1176 0.7318
## Residuals         503  50247     100
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(weight.lm2)
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height + new.df$worthless, data = wwh)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.981  -6.384  -0.350   5.800  45.435
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -165.54777    8.25903 -20.044   <2e-16 ***
## Waist               4.95999    0.12300  40.325   <2e-16 ***
## Height              2.48874    0.14397  17.286   <2e-16 ***
## new.df$worthless    0.02992    0.08724   0.343    0.732
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.995 on 503 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8846
## F-statistic:  1294 on 3 and 503 DF,  p-value: < 2.2e-16
```

RSS = 50247 SSReg = 358074 + 29843 + 12 = 387929 SYY = RSS + SSReg = 438176

ii. RSS decreases by 12 units, and SSReg increases by 12 while SYY stays the same value. 12 is the value of the new.df$worthless sum squared, which was added into this model.

iii. The r-squared value is the same with a value of 0.8853, and the adjusted r-squared value has decreased by 0.0002 from its value in part (a) with a new value of 0.8846.

c)

```
weight.lm3 <- lm(Weight ~ worthless + Waist + Height, data = new.df)
anova(weight.lm3)
```

```
## Analysis of Variance Table
##
## Response: Weight
##            Df Sum Sq Mean Sq   F value Pr(>F)
## worthless   1     58      58    0.5828 0.4456
## Waist       1 358020  358020 3583.9463 <2e-16 ***
## Height      1  29850   29850  298.8086 <2e-16 ***
## Residuals 503  50247     100
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(weight.lm3)
```

```
##
## Call:
## lm(formula = Weight ~ worthless + Waist + Height, data = new.df)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -32.981  -6.384  -0.350   5.800  45.435
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -165.54777    8.25903 -20.044   <2e-16 ***
## worthless      0.02992    0.08724   0.343    0.732
## Waist          4.95999    0.12300  40.325   <2e-16 ***
## Height         2.48874    0.14397  17.286   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 9.995 on 503 degrees of freedom
## Multiple R-squared:  0.8853, Adjusted R-squared:  0.8846
## F-statistic:  1294 on 3 and 503 DF,  p-value: < 2.2e-16
```

RSS = 50247 SSReg = 58 + 358020 + 29850 = 387928 SYY = RSS + SSReg = 438175

The RSS is the same of 50247, while the SSReg value is one unit less than in (b). Therefore, the SYY value is one unit less than the value in (b). Both the r-squared value and adjjusted r-squared values are the same as in part (b).

d) I believe the adjusted r-squared value is more reliable to tell whether a new variable should be added to our model or not because the adjusted r-squared value takes into account the addition of new variables whereas the r-squared value will always increase with the addition of new variables, despite the significance of those variables. This can lead to a misrepresentation of how good a specific regression model is for our data. The adjusted r-squared value will actually decrease in value if it decides that a new variable does not contribute to the fit of a regression, by taking into account the degrees of freedom taken by each variable. Therefore, we can confidently say that the adjusted r-squared value is a better measure of whether or not to add a new variable to a model.

e) Usually, the SSReg will increase as we add new variables to a model, regardless of if the additional variable improves the model or not. This is why we cannot just look at SSReg to decide whether to add a new variable. Partial tests are useful for telling us whether we should add a new variable or not because they are able to check the significance of each variable while using the full model. We are able to split the analysis into smaller parts to get a more detailed picture of the model. Partial F-tests are particularly useful because they assess the significance of variables by providing f statistics and p-values that we are able to analyze the statistical significance of. If a variable is deemed statistically significant, then we are able to add that variable to the model as it would increase the productivity of the model.

Part B

```
cars <- read.csv('cars04.csv')
```

```
cars.new <- cars[,c(-1, -2)]
cars.new.lm <- lm(SuggestedRetailPrice ~ ., data = cars.new)
cars.new.lm
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ ., data = cars.new)
```

```
##
## Coefficients:
## (Intercept)   DealerCost   EngineSize    Cylinders   Horsepower      CityMPG
##    349.9763       1.0542     -32.2472     228.3295       2.3621     -16.7424
##  HighwayMPG       Weight    WheelBase       Length        Width
##     46.7575       0.6992      27.0534      -7.3202     -84.7085
```

a) We do not use Vehicle.Name because it is description of the car that is non-numeric or categorical and therefore does not belong in computation. Rstudio would not have the capacity to account for the inferred meaning of the vehicle’s name.

b) SuggestedRetailPrice = 349.9763 + 1.0542 * DealerCost - 32.2472 * EngineSize + 228.3295 * Cylinders + 2.3621 * Horsepower - 16.7424 * CityMPG + 46.7575 * HighwayMPG + 0.6992 * Weight + 27.0534 * WheelBase - 7.3202 * Length - 84.7085 * Width

c)

```
summary(cars.new.lm)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ ., data = cars.new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1403.85  -276.86   -55.03   257.55  2584.11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  349.97628 1461.40052   0.239 0.810953
## DealerCost     1.05418    0.00564 186.923  < 2e-16 ***
## EngineSize   -32.24720  123.05642  -0.262 0.793523
## Cylinders    228.32952   71.99492   3.171 0.001730 **
## Horsepower     2.36212    1.42851   1.654 0.099624 .
## CityMPG      -16.74239   21.46286  -0.780 0.436181
## HighwayMPG    46.75754   24.17910   1.934 0.054403 .
## Weight         0.69920    0.20751   3.370 0.000887 ***
## WheelBase     27.05345   16.36168   1.653 0.099644 .
## Length        -7.32019    7.12296  -1.028 0.305209
## Width        -84.70850   30.21238  -2.804 0.005496 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 532.3 on 223 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 2.073e+04 on 10 and 223 DF,  p-value: < 2.2e-16
```

For the Cylinders variable, the estimated slope is 228.32952 with a t statistic of 3.171 and a p-value of 0.001730. From the t statistic and the p-value, using a significance level of 0.05, we can conclude that there is a relationship between Cylinders and SuggestedRetailPrice variables. With the p-value of 0.001730 being less than 0.05, we reject our null hypothesis that states there is no significant association between the Cylinders and SuggestedRetailPrice variables.

d)

```
anova(cars.new.lm)
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##             Df     Sum Sq    Mean Sq    F value    Pr(>F)
## DealerCost   1 5.8714e+10 5.8714e+10 2.0724e+05 < 2.2e-16 ***
## EngineSize   1 7.7453e+06 7.7453e+06 2.7338e+01 3.925e-07 ***
## Cylinders    1 2.7222e+06 2.7222e+06 9.6084e+00  0.002186 **
## Horsepower   1 7.0394e+05 7.0394e+05 2.4847e+00  0.116377
## CityMPG      1 2.1856e+05 2.1856e+05 7.7150e-01  0.380714
## HighwayMPG   1 2.1052e+05 2.1052e+05 7.4310e-01  0.389601
## Weight       1 1.2563e+06 1.2563e+06 4.4344e+00  0.036341 *
## WheelBase    1 3.9621e+04 3.9621e+04 1.3990e-01  0.708785
## Length       1 1.6483e+06 1.6483e+06 5.8179e+00  0.016673 *
## Width        1 2.2271e+06 2.2271e+06 7.8611e+00  0.005496 **
## Residuals  223 6.3178e+07 2.8331e+05
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
sqrt(9.6084)
```

```
## [1] 3.099742
```

The t statistic is 3.099742. We take the square root of the f statistic to get this value. Sqrt(9.6084) = 3.099742.

e)

```
summary(cars.new.lm)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ ., data = cars.new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1403.85  -276.86   -55.03   257.55  2584.11
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   349.97628 1461.40052   0.239 0.810953
## DealerCost      1.05418    0.00564 186.923  < 2e-16 ***
## EngineSize    -32.24720  123.05642  -0.262 0.793523
## Cylinders     228.32952   71.99492   3.171 0.001730 **
## Horsepower      2.36212    1.42851   1.654 0.099624 .
## CityMPG       -16.74239   21.46286  -0.780 0.436181
## HighwayMPG     46.75754   24.17910   1.934 0.054403 .
## Weight          0.69920    0.20751   3.370 0.000887 ***
## WheelBase      27.05345   16.36168   1.653 0.099644 .
## Length         -7.32019    7.12296  -1.028 0.305209
## Width        -84.70850   30.21238  -2.804 0.005496 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 532.3 on 223 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 2.073e+04 on 10 and 223 DF,  p-value: < 2.2e-16
```

The f statistic is 2.073e+04. Because the value of the f statistic is large, we can assume that the model explains much of the total variability displayed by the data, and we can conclude from this that the model is statistically significant. Using a significance level of 0.05, we can observe that our p-value of 2.2e-16 is less than our significance level and therefore we reject our null hypothesis that states our f statistic was generated by chance.

f) Null: SuggestedRetailPrice ~ DealerCost + EngineSize + Cylinders + Horsepower + Weight + WheelBase + Length + Width Alternative: SuggestedRetailPrice ~ DealerCost + EngineSize + Cylinders + Horsepower + CityMPG + HighwayMPG + Weight + WheelBase + Length + Width

```
m.full <- lm(SuggestedRetailPrice ~ DealerCost + EngineSize + Cylinders + Horsepower + CityMPG + Highwa
m.reduced <- update(m.full, .~. -CityMPG-HighwayMPG)
m.full.anova <- anova(m.full)
m.full.anova
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)
## DealerCost    1 5.8714e+10 5.8714e+10 2.0724e+05 < 2.2e-16 ***
## EngineSize    1 7.7453e+06 7.7453e+06 2.7338e+01 3.925e-07 ***
## Cylinders     1 2.7222e+06 2.7222e+06 9.6084e+00  0.002186 **
## Horsepower    1 7.0394e+05 7.0394e+05 2.4847e+00  0.116377
## CityMPG       1 2.1856e+05 2.1856e+05 7.7150e-01  0.380714
## HighwayMPG    1 2.1052e+05 2.1052e+05 7.4310e-01  0.389601
## Weight        1 1.2563e+06 1.2563e+06 4.4344e+00  0.036341 *
## WheelBase     1 3.9621e+04 3.9621e+04 1.3990e-01  0.708785
## Length        1 1.6483e+06 1.6483e+06 5.8179e+00  0.016673 *
## Width         1 2.2271e+06 2.2271e+06 7.8611e+00  0.005496 **
## Residuals   223 6.3178e+07 2.8331e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
sum(m.full.anova$`Sum Sq`)
```

```
## [1] 58794176636
```

```
m.reduced.anova <- anova(m.reduced)
m.reduced.anova
```

```
## Analysis of Variance Table
##
## Response: SuggestedRetailPrice
##              Df     Sum Sq    Mean Sq    F value    Pr(>F)    
## DealerCost    1 5.8714e+10 5.8714e+10 2.0204e+05 < 2.2e-16 ***
## EngineSize    1 7.7453e+06 7.7453e+06 2.6651e+01 5.353e-07 ***
## Cylinders     1 2.7222e+06 2.7222e+06 9.3670e+00  0.002478 ** 
## Horsepower    1 7.0394e+05 7.0394e+05 2.4223e+00  0.121028    
## Weight        1 5.3446e+05 5.3446e+05 1.8391e+00  0.176418    
## WheelBase     1 7.3600e+02 7.3600e+02 2.5000e-03  0.959900    
## Length        1 1.4322e+06 1.4322e+06 4.9281e+00  0.027421 *  
## Width         1 1.4236e+06 1.4236e+06 4.8985e+00  0.027885 *  
## Residuals   225 6.5388e+07 2.9061e+05                           
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
sum(m.reduced.anova$`Sum Sq`)
```

```
## [1] 58794176636
```

```
sum(m.full.anova$`Sum Sq`) - sum(m.reduced.anova$`Sum Sq`)
```

```
## [1] 0
```

The change in regression is 0. Therefore our f statistic is 0.

```
1 - pf(0, 2, 223)
```

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

```
anova(m.full, m.reduced)
```

```
## Analysis of Variance Table
## 
## Model 1: SuggestedRetailPrice ~ DealerCost + EngineSize + Cylinders + 
##     Horsepower + CityMPG + HighwayMPG + Weight + WheelBase + 
##     Length + Width
## Model 2: SuggestedRetailPrice ~ DealerCost + EngineSize + Cylinders + 
##     Horsepower + Weight + WheelBase + Length + Width
##   Res.Df      RSS Df Sum of Sq      F  Pr(>F)  
## 1    223 63178392                              
## 2    225 65387880 -2  -2209488 3.8994 0.02165 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

We can determine that our f statistic has a value of 0. The p-value associated with our f statistic is 0.02165, which is less than our significance level of 0.05 and therefore we reject the null hypothesis and we can conclude that full model is the better fit to our data. The full model is the superior model to use for our data when comparing it to the reduced model.