Homework 7

Jun Ryu, UID: 605574052

2023-05-19

```
knitr::opts_chunk$set(echo = TRUE)

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.3.0      v stringr 1.5.0
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

Question 1

```
waist <- read.table("waistweightheight.txt", header = T)
head(waist)
```

```
##   Waistcm wtKg  HTCm gen Waist Height Weight
## 1    71.5 65.6 174.0   1 28.15  68.50 144.65
## 2    79.0 71.8 175.3   1 31.10  69.02 158.32
## 3    83.2 80.7 193.5   1 32.76  76.18 177.94
## 4    77.8 72.6 186.5   1 30.63  73.43 160.08
## 5    80.0 78.8 187.2   1 31.50  73.70 173.75
## 6    82.5 74.8 181.5   1 32.48  71.46 164.93
```

a)

```
model <- lm(Weight~Waist+Height, data = waist)
```

```
summary(model)
```

i)

```
##
## Call:
## lm(formula = Weight ~ Waist + Height, data = waist)
##
## 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
```

```
anova(model)
```

```
## 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
```

Looking at the ANOVA table, SSReg is $358074 + 29843$, which is 387917. RSS is 50259, and SYY is the sum of SSReg and RSS, which is $387917 + 50259$ or 438176.

ii) From the summary table, the R-squared value is 0.8853 and the adjusted R-squared value is 0.8848.

iii) From the summary table, the slope for the height is 2.4884, which indicates that among people of the same waist size, people whose height is 1 inch taller are on average 2.4884 pounds heavier.

b)

```
set.seed(23)
new.df <- transform(waist, worthless = rnorm(dim(waist)[1],0,5))
model2 <- lm(Weight~Waist+Height+worthless, data = new.df)
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height + worthless, 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 ***
## Waist          4.95999    0.12300  40.325   <2e-16 ***
## Height         2.48874    0.14397  17.286   <2e-16 ***
## 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
```

```
anova(model2)
```

```
## 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 ***
## worthless   1     12      12    0.1176 0.7318
## Residuals 503  50247     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

i) Looking at the ANOVA table, SSReg is $358074 + 29843 + 12$, which is 387929. RSS is 50247, and SYY is the sum of SSReg and RSS, which is $387929 + 50247$ or 438176.

ii) The SYY has stayed the same, but SSReg has gone up due to the addition of a new variable, while the RSS has gone down.

iii) The R-squared value has remained the same, but the adjusted R-squared value has gone down from the previous model.

c)

```
model3 <- lm(Weight~worthless+Waist+Height, data = new.df)
summary(model3)
```

```
##
## 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
```

```
anova(model3)
```

```
## 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
```

From part b), we first notice that both the R-squared and the adjusted R-squared values remain the same. Also, our RSS value has remained the same at 50247, but our SSReg value has changed by 1 giving us a value of 387928 instead.

d)

Adjusted R-squared value is a better indicator at telling us whether we should add a new variable because the R-squared value will always go up if you add a new variable, whereas the adjusted R-squared value can go either direction and indicate how useful the added variable is based on which direction it goes.

e)

Looking at SSReg is not as effective because adding a new variable will most likely make the value go up regardless of its significance. Instead, partial tests are better since looking at values like the p-value tells us more concretely whether the new variable is statistically significant or not (e.g. less than 0.05).

Question 2

```
cars <- read.csv("cars04.csv")
head(cars)
```

```
##                  Vehicle.Name Hybrid SuggestedRetailPrice DealerCost EngineSize
## 1          Chevrolet Aveo 4dr      0                11690      10965        1.6
## 2 Chevrolet Aveo LS 4dr hatch      0                12585      11802        1.6
## 3      Chevrolet Cavalier 2dr      0                14610      13697        2.2
## 4      Chevrolet Cavalier 4dr      0                14810      13884        2.2
## 5   Chevrolet Cavalier LS 2dr      0                16385      15357        2.2
## 6           Dodge Neon SE 4dr      0                13670      12849        2.0
##   Cylinders Horsepower CityMPG HighwayMPG Weight WheelBase Length Width
## 1         4        103      28         34   2370        98    167    66
## 2         4        103      28         34   2348        98    153    66
## 3         4        140      26         37   2617       104    183    69
## 4         4        140      26         37   2676       104    183    68
## 5         4        140      26         37   2617       104    183    69
## 6         4        132      29         36   2581       105    174    67
```

```
model_car <- lm(SuggestedRetailPrice ~ DealerCost+EngineSize+Horsepower+CityMPG+
                  HighwayMPG+Weight+WheelBase+Length+Width+Cylinders, data = cars)
```

a)

If we were to include the column Vehicle.Name, since the column entries are not of a numeric type, it will categorize each vehicle name as its own variable, and that is certainly not what we want.

b)

```
summary(model_car)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ DealerCost + EngineSize +
##     Horsepower + CityMPG + HighwayMPG + Weight + WheelBase +
##     Length + Width + Cylinders, data = cars)
##
## 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
## 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 **
## Cylinders    228.32952   71.99492   3.171 0.001730 **
## ---
```

```
## 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 equation of the fitted model is SuggestedRetailPrice = 349.97628 + 1.05418*DealerCost - 32.2472*EngineSize + 2.36212*Horsepower - 16.74239*CityMPG + 46.75754*HighwayMPG + 0.6992*Weight + 27.05345*WheelBase - 7.32019*Length - 84.70850*Width + 228.32592*Cylinders

c)

For the Cylinders variable, the estimated slope is 228.32592. The t-statistic is 3.171 and the p-value is 0.00173. From these last two values, we can conclude that Cylinders is a useful predictor (by itself) in predicting the SuggestedRetailPrice.

d)

```
anova(model_car)
```

```
## 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 ***
## Horsepower   1 1.0860e+06 1.0860e+06 3.8331e+00  0.051496 .
## CityMPG      1 1.9693e+05 1.9693e+05 6.9510e-01  0.405327
## HighwayMPG   1 5.4432e+04 5.4432e+04 1.9210e-01  0.661576
## Weight       1 1.3086e+06 1.3086e+06 4.6190e+00  0.032697 *
## WheelBase    1 6.4650e+04 6.4650e+04 2.2820e-01  0.633335
## Length       1 1.9825e+06 1.9825e+06 6.9977e+00  0.008742 **
## Width        1 1.4838e+06 1.4838e+06 5.2374e+00  0.023043 *
## Cylinders    1 2.8496e+06 2.8496e+06 1.0058e+01  0.001730 **
## Residuals  223 6.3178e+07 2.8331e+05
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

You can get the t-statistics by simply taking the square root of the F-values shown in the ANOVA table (since Cylinders is placed last meaning that all other variables are controlled for both values).

e)

Using the summary command, we obtain the F-value for Cylinders by simply squaring the value. The F-value represents whether Cylinders is statistically significant given that all other variables are controlled and included in the model.

f)

```r
model_reduced <- lm(SuggestedRetailPrice ~ DealerCost+EngineSize+Horsepower+Weight+
                      WheelBase+Length+Width+Cylinders, data = cars)
anova(model_car, model_reduced)
```

```
## Analysis of Variance Table
##
## Model 1: SuggestedRetailPrice ~ DealerCost + EngineSize + Horsepower +
##     CityMPG + HighwayMPG + Weight + WheelBase + Length + Width +
##     Cylinders
## Model 2: SuggestedRetailPrice ~ DealerCost + EngineSize + Horsepower +
##     Weight + WheelBase + Length + Width + Cylinders
##   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
```

Since our p-value is less than 0.05 (0.02165), we can reject the null hypothesis that fuel consumption has no affect on the suggested retail price.