Stats101A, Spring 2023 - Homework 7

Luke Villanueva - 206039397

05/19/23

```
library(tidyverse)
df <- read.delim(paste0(getwd(),"/waistweightheight.txt"))
glimpse(df)
```

```
## Rows: 507
## Columns: 7
## $ Waistcm <dbl> 71.5, 79.0, 83.2, 77.8, 80.0, 82.5, 82.0, 76.8, 68.5, 77.5, 81~
## $ wtKg    <dbl> 65.6, 71.8, 80.7, 72.6, 78.8, 74.8, 86.4, 78.4, 62.0, 81.6, 76~
## $ HTCm    <dbl> 174.0, 175.3, 193.5, 186.5, 187.2, 181.5, 184.0, 184.5, 175.0,~
## $ gen     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,~
## $ Waist   <dbl> 28.15, 31.10, 32.76, 30.63, 31.50, 32.48, 32.28, 30.24, 26.97,~
## $ Height  <dbl> 68.50, 69.02, 76.18, 73.43, 73.70, 71.46, 72.44, 72.64, 68.90,~
## $ Weight  <dbl> 144.65, 158.32, 177.94, 160.08, 173.75, 164.93, 190.51, 172.87~
```

Problem A

a.

```
m1 <- lm(Weight ~ Waist + Height, data = df)
```

i.

```
# SSReg
ssreg <- sum(anova(m1)[-3,2])
ssreg
```

```
## [1] 387916.6
```

```
# RSS
rss <- anova(m1)[3,2]
rss
```

```
## [1] 50259.23
```

```
# SYY
ssreg + rss
```

```
## [1] 438175.8
```

SSReg is 387916.6

RSS is 50259.23

SYY is 438175.8

ii.

```
summary(m1)
```

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

R-squared is 0.8853

Adjusted R-squared is 0.8848

iii.

```
summary(m1)
```

```
##
## Call:
## lm(formula = Weight ~ Waist + Height, data = df)
##
## 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 slope of the variable "Height" is estimated to be 2.4884. This means that provided that all observations have the same "Waist", the observations have an average change in Weight of 2.4884 units as Height increases by 1 unit.

b.

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

```
## Rows: 507
## Columns: 8
## $ Waistcm   <dbl> 71.5, 79.0, 83.2, 77.8, 80.0, 82.5, 82.0, 76.8, 68.5, 77.5, ~
## $ wtKg      <dbl> 65.6, 71.8, 80.7, 72.6, 78.8, 74.8, 86.4, 78.4, 62.0, 81.6, ~
## $ HTCm      <dbl> 174.0, 175.3, 193.5, 186.5, 187.2, 181.5, 184.0, 184.5, 175.~
## $ gen       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ Waist     <dbl> 28.15, 31.10, 32.76, 30.63, 31.50, 32.48, 32.28, 30.24, 26.9~
## $ Height    <dbl> 68.50, 69.02, 76.18, 73.43, 73.70, 71.46, 72.44, 72.64, 68.9~
## $ Weight    <dbl> 144.65, 158.32, 177.94, 160.08, 173.75, 164.93, 190.51, 172.~
## $ worthless <dbl> 0.966061669, -2.173410541, 4.566335483, 8.966940460, 4.98302~
```

i.

```
m2 <- lm(Weight ~ Waist + Height + worthless, new.df)

ssreg2 <- sum(anova(m2)[-4,2])
ssreg2
```

```
## [1] 387928.3
```

```
rss2 <- anova(m2)[4,2]
rss2
```

```
## [1] 50247.49
```

```
ssreg2 + rss2
```

```
## [1] 438175.8
```

SSReg is 387928.3

RSS is 50247.49

SYY is 438175.8

ii.

SSReg and RSS have changed from part a. SSReg changed because it adds on the sum squares of the worthless variable. RSS changed because the model is altered due to the worthless variable, so the residuals' sum squares are altered. SYY stays the same because the actual data of Weight hasn't changed, so the variance of Weight stays the same.

iii.

```
summary(m2)
```

```
##
## 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
```

R-squared has not changed while the adjusted R-squared dropped by 0.0002 when the worthless variable is added. This implies that the newly added variable "worthless" is most likely not useful to describe the behavior of Weight.

c.

i.

```
m3 <- lm(Weight ~ worthless + Waist + Height, new.df)

# ssreg
ssreg3 <- sum(anova(m3)[-4,2])
ssreg3
```

```
## [1] 387928.3
```

```
# rss
rss3 <- anova(m3)[4,2]
rss3
```

```
## [1] 50247.49
```

```
# syy
ssreg3 + rss3
```

```
## [1] 438175.8
```

SSReg is 387928.3

RSS is 50247.49

SYY is 438175.8

ii.

All variables have stayed the same. This is due to the fact the same variables were used in this model. The only difference with this model and the last is the order of partial F tests. SSReg and RSS do not depend on the order of the variables, and so, this implies that SYY does not change either because of the order of the variables.

iii.

```
summary(m3)
```

```
##
## 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
```

R-squared and adjusted R-squared have not changed from part b's model. This is because the order does not matter to calculate R-squared.

d.

Adjusted R-squared is a more reliable guide to see if the new variable is useful to add because it takes into account the degrees of freedom as well. This allows the adj. R-squared to change positively only if the RSS is heavily impacted by the new variable. In other words, RSS needs to go down drastically by the new variable for adj. R-squared to go up significantly.

e.

We cannot look purely at the pattern of SSReg because the RSS could also be increasing dramatically due to the new variables. And so, this makes it so that even if SSReg is increasing, the increasing RSS implies that the new model does not necessarily explain more of the variation. Partial tests are useful guidelines on whether add a new variable or not because they test for each added variable whether the parameters would be equal to 0 or not. In short, the tests look at if the parameters are statistically significant enough to have a value that's not 0, or are they insignificant so that if they were 0, they'd barely affect the model.

Problem B.

```r
cars <- read.csv(paste0(getwd(),"/cars04.csv"))
glimpse(cars)
```

```
## Rows: 234
## Columns: 13
## $ Vehicle.Name         <chr> "Chevrolet Aveo 4dr", "Chevrolet Aveo LS 4dr hatc~
## $ Hybrid               <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ SuggestedRetailPrice <int> 11690, 12585, 14610, 14810, 16385, 13670, 15040, ~
## $ DealerCost           <int> 10965, 11802, 13697, 13884, 15357, 12849, 14086, ~
## $ EngineSize           <dbl> 1.6, 1.6, 2.2, 2.2, 2.2, 2.0, 2.0, 2.0, 2.0, 2.0,~
## $ Cylinders            <int> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
## $ Horsepower           <int> 103, 103, 140, 140, 140, 132, 132, 130, 110, 130,~
## $ CityMPG              <int> 28, 28, 26, 26, 26, 29, 29, 26, 27, 26, 26, 32, 3~
## $ HighwayMPG           <int> 34, 34, 37, 37, 37, 36, 36, 33, 36, 33, 33, 38, 4~
## $ Weight               <int> 2370, 2348, 2617, 2676, 2617, 2581, 2626, 2612, 2~
## $ WheelBase            <int> 98, 98, 104, 104, 104, 105, 105, 103, 103, 103, 1~
```

```
## $ Length                <int> 167, 153, 183, 183, 183, 174, 174, 168, 168, 168,~
## $ Width                 <int> 66, 66, 69, 68, 69, 67, 67, 67, 67, 67, 67, 67, 6~
```

```
# model
cars.m <- lm(SuggestedRetailPrice ~ . -Vehicle.Name -Hybrid, cars)
```

a.

```
testm <- lm(SuggestedRetailPrice ~ . -Hybrid, cars)
testm
```

The resulting linear model is not comprehensible because it combines a categorical predictor alongside a numerical predictor. And so, the model tries to default the fitting to each unique Vehicle.Name and provide an intercept and slope for each one. And so, this model does not help describe a relationship between Sugg. Retail Price and the attributes of the car.

b.

$$Y_{SuggestedRetailPrice} = B_0X_{DealerCost} + B_1x_{EngineSize} + B_2X_{Cylinders} + B_3X_{Horsepower} + B_4X_{CityMPG}$$
$$+B_5X_{HighwayMPG} + B_6X_{Weight} + B_7X_{WheelBase} + B_8X_{Length}$$
$$+B_9X_{Width} + Intercept$$
$$\Longrightarrow$$
$$Y_{SuggestedRetailPrice} = 1.0542X_{DealerCost} - 32.2472x_{EngineSize} + 228.3295X_{Cylinders} + 2.3621X_{Horsepower} - 16.7424X_{CityMI}$$
$$+46.7575X_{HighwayMPG} + 0.6992X_{Weight} + 27.0534X_{WheelBase} - 7.3202X_{Length}$$
$$+ - 84.7085X_{Width} + 349.9763$$

c.

```
summary(cars.m)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ . - Vehicle.Name - Hybrid,
##     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
## 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 Cylinders,

slope: 228.32952

t-stat: 3.171

p-val: 0.001730

From the t-stat and the p-val, we can conclude that Cylinders is a good predictor variable to describe the behavior of Sugg.Retail Price because it is unlikely to retrieve the estimated slope again through more samplings. Therefore, it is likely Cylinders has a correlation to Sugg.Retail Price.

d.

fval = tval^2

```
sqrt(anova(cars.m)[3,4])
```

```
## [1] 3.099739
```

So, t-stat of Cylinders is 3.099639

e.

```
summary(cars.m)
```

```
##
## Call:
## lm(formula = SuggestedRetailPrice ~ . - Vehicle.Name - Hybrid,
##     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
## 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
```

F-Stat is $2.073 * 10^4$. Since the F-stat is very big, this means at least one of the variables has a statistically significant slope that is not 0.

f.

```
# model without fuel consumption
cars.mNoFuel <- lm(SuggestedRetailPrice ~ . -Vehicle.Name -Hybrid -CityMPG -HighwayMPG, cars)

# anova test
anova(cars.mNoFuel, cars.m)
```

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

Since the F-stat is 3.8994 with a p-val of 0.02165, this means that under a 5% significance level, the impact of CityMPG and HighwayMPG is statistically significant enough to alter Sugg. Retail Price. This implies that the model including those two variables would explain more of the data's variation rather than the model excluding them.