Homework3

Tianqi Wang

9/11/2018

Question 2

Eventhough it is true that adding predictor variables to a regression model can never reduce R^2, but higher R^2 does not necessarrily mean a better model, putting values has little influence on the model would just cause overfitting. A good model should use few variables to generate more information.

Question 3

Even though adjusted R^2 does penalized by the number of independent variables, but after this “penalization”, the value would not mean the portion that the variation of dependent variable could explained by the independent variables anymore. i.e the adjusted R^2 could be negative.

Question 4

```
Kelley <- read.csv("KelleyBlueBookData.csv")
Kelley$Cylinder <- as.factor(Kelley$Cylinder)
Kelley$Type <- as.factor(Kelley$Type)
Mymodel <- lm(Price ~ Mileage + Type + Cylinder + Liter + Cruise +
    Sound + Leather, data = Kelley)
summary(lm(Price ~ Mileage + Type + Cylinder + Liter + Cruise +
    Sound + Leather, data = Kelley))
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Type + Cylinder + Liter + Cruise +
##     Sound + Leather, data = Kelley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13959.3  -3197.7   -547.9   2504.3  17603.0
##
## Coefficients:
##                  Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)    29433.5616   1608.3572  18.300 < 0.0000000000000002 ***
## Mileage           -0.1871      0.0220  -8.505 < 0.0000000000000002 ***
## TypeCoupe     -18570.1180    874.1339 -21.244 < 0.0000000000000002 ***
## TypeHatchback -18310.1556   1085.1462 -16.873 < 0.0000000000000002 ***
## TypeSedan     -15468.4303    799.9942 -19.336 < 0.0000000000000002 ***
## TypeWagon      -9452.1225   1000.0199  -9.452 < 0.0000000000000002 ***
## Cylinder6       1360.1311   1075.4525   1.265             0.206349
## Cylinder8      14164.7491   1959.0043   7.231     0.00000000000113 ***
## Liter           1115.8414    621.2359   1.796             0.072849 .
## Cruise          4650.7921    473.6264   9.820 < 0.0000000000000002 ***
## Sound             14.7921    404.3379   0.037             0.970826
## Leather         1677.9449    433.2245   3.873             0.000116 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5097 on 792 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7341
## F-statistic: 202.6 on 11 and 792 DF,  p-value: < 0.00000000000000022
```

(a)

The coefficient of $\beta_{leather}$ is 1677.9499, since the p-value is 0.000116 < 0.001, we can reject the null hypothesis: $H_0 : B_{leather} = 0$

(b)

The coefficient of $\beta_{leather}$ is 1677.9499, which means cars with leather chairs would price higher 1677.9449 on average, else being equal.

(c)

```
Mymodel2 <- lm(Price ~ Mileage + Type + Cylinder + Cruise + Leather,
    data = Kelley)
anova(Mymodel, Mymodel2)
```

```
## Analysis of Variance Table
##
## Model 1: Price ~ Mileage + Type + Cylinder + Liter + Cruise + Sound +
##     Leather
## Model 2: Price ~ Mileage + Type + Cylinder + Cruise + Leather
##   Res.Df         RSS Df Sum of Sq      F Pr(>F)
## 1    792 20573527636
## 2    794 20657786793 -2 -84259157 1.6218 0.1982
```

From the p-value of the F-test of restricted model and unrestricted model, which is 0.1982 > 0.1, we can not reject the null hypothesis $H_0 : \beta_{Liter} = \beta_{Sound} = 0$, hence we could drop these two variables without signicantly decreasing R^2.

(d)

Although there are three exclusive categories, but the category 4-cylinder would be 1-"6-cylinder"-"8-cylinder" which would be 1 if the observation is not in the other two categories.

(e)

Firstly I used the assigned values to do the prediction.

```
new <- data.frame(Mileage = 15000, Type = as.factor("Convertible"),
    Cylinder = as.factor(6), Liter = 3, Cruise = 1, Sound = 1,
    Leather = 1)
predict(Mymodel, new)
```

```
##        1
## 37678.35
```

I might hesitate to be confidend in such prediction, since when I use `xtabs` to check the frequency for different combinations of `Cylinder` and `Type`, I found there is no observation that is convertible with 6 cylinders.

```
xtabs(~Type + Cylinder, data = Kelley)
```

```
##              Cylinder
## Type            4   6   8
##   Convertible  30   0  20
##   Coupe        80  40  20
##   Hatchback    30  30   0
##   Sedan       190 240  60
##   Wagon        64   0   0
```

(f)

```r
new <- data.frame(Mileage = 15000, Type = as.factor("Convertible"),
    Cylinder = as.factor(4), Liter = 2, Cruise = 1, Sound = 1,
    Leather = 1)
predict(Mymodel, new)
```

```
##        1
## 35202.38
```

(g)

```r
predict(Mymodel, new, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 35202.38 33630.31 36774.45
```

The confidence interval of the predicted value at $\alpha = 0.05$ significance level means there are 0.95 chance that the mean of truly value would be contained in this confidence interval.

(h)

```r
predict(Mymodel, new, interval = "prediction", level = 0.9)
```

```
##        fit     lwr      upr
## 1 35202.38 26706.2 43698.56
```

The prediction inverval of the predictede value at $\alpha = 0.1$ significance level means there are 0.9 chance that the truly value would be contained in this interval.

(i)

```r
summary(Mymodel)
```

```
##
## Call:
## lm(formula = Price ~ Mileage + Type + Cylinder + Liter + Cruise +
##     Sound + Leather, data = Kelley)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13959.3  -3197.7   -547.9   2504.3  17603.0
##
## Coefficients:
##                  Estimate  Std. Error t value             Pr(>|t|)
## (Intercept)    29433.5616   1608.3572  18.300 < 0.0000000000000002 ***
## Mileage           -0.1871      0.0220  -8.505 < 0.0000000000000002 ***
## TypeCoupe     -18570.1180    874.1339 -21.244 < 0.0000000000000002 ***
## TypeHatchback -18310.1556   1085.1462 -16.873 < 0.0000000000000002 ***
## TypeSedan     -15468.4303    799.9942 -19.336 < 0.0000000000000002 ***
## TypeWagon      -9452.1225   1000.0199  -9.452 < 0.0000000000000002 ***
## Cylinder6       1360.1311   1075.4525   1.265             0.206349
## Cylinder8      14164.7491   1959.0043   7.231     0.00000000000113 ***
## Liter           1115.8414    621.2359   1.796             0.072849 .
## Cruise          4650.7921    473.6264   9.820 < 0.0000000000000002 ***
## Sound             14.7921    404.3379   0.037             0.970826
## Leather         1677.9449    433.2245   3.873             0.000116 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5097 on 792 degrees of freedom
## Multiple R-squared:  0.7378, Adjusted R-squared:  0.7341
## F-statistic: 202.6 on 11 and 792 DF,  p-value: < 0.00000000000000022
```

So R^2 is 0.7377879, it means 73.8% variability of Price could be explained by the variabilities of all independent variables in this model. R_a^2 is 0.734146, on the basis of these two pieces of information alone, I can not say there is a really strong evidence of overfitting, since there are no much difference between R^2 and R_a^2.

(j)

```
anova(Mymodel)
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df      Sum Sq     Mean Sq  F value                Pr(>F)
## Mileage     1  1605590375  1605590375  61.8089   0.00000000000001236 ***
## Type        4 24553392857  6138348214 236.3023 < 0.00000000000000022 ***
## Cylinder    2 28681267906 14340633953 552.0581 < 0.00000000000000022 ***
## Liter       1   233414398   233414398   8.9855             0.0028063 **
## Cruise      1  2408426579  2408426579  92.7150 < 0.00000000000000022 ***
## Sound       1    16078757    16078757   0.6190             0.4316660
## Leather     1   389684357   389684357  15.0013             0.0001163 ***
## Residuals 792 20573527636    25976676
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic I calculated from the `anova` table is

```
(sum(anova(Mymodel)["Sum Sq"][-length(anova(Mymodel)["Sum Sq"][,
    1]), 1])/sum(anova(Mymodel)["Df"][-length(anova(Mymodel)["Df"][,
    1]), 1]))/anova(Mymodel)["Mean Sq"][length(anova(Mymodel)["Mean Sq"][,
    1]), 1]
```

```
## [1] 202.5868
```

Which is same as the result in `summary` : 202.5868218, since the F-statistic is large and the p-value is very small and less than 0.01, I can reject the null hypothesis, at 0.01 significance level.

Question 5

(a)

```
Brand <- read.csv("BrandPreference.csv")
(Mymodel <- summary(lm(BrandLiking ~ MoistureContent + Sweetness,
    data = Brand)))
```

```
##
## Call:
## lm(formula = BrandLiking ~ MoistureContent + Sweetness, data = Brand)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.400 -1.762  0.025  1.587  4.200
##
## Coefficients:
##                  Estimate Std. Error t value      Pr(>|t|)
## (Intercept)       37.6500     2.9961  12.566 0.00000001200 ***
## MoistureContent    4.4250     0.3011  14.695 0.00000000178 ***
## Sweetness          4.3750     0.6733   6.498 0.00002011047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.693 on 13 degrees of freedom
## Multiple R-squared:  0.9521, Adjusted R-squared:  0.9447
## F-statistic: 129.1 on 2 and 13 DF,  p-value: 0.000000002658
```

(b)

```
len <- nrow(Brand)
X <- cbind(rep(1, len), Brand$MoistureContent, Brand$Sweetness)
Y <- Brand$BrandLiking
solve(t(X) %*% X) %*% t(X) %*% Y
```

```
##        [,1]
## [1,] 37.650
## [2,]  4.425
## [3,]  4.375
```

The result is same as what I got in `summary` .

(c)

```
diag(solve(t(X) %*% X))^0.5 * Mymodel$sigma
```

```
## [1] 2.9961032 0.3011197 0.6733241
```

The result is same as what I got in `summary`

(d)

HY is :

```
X %*% solve(t(X) %*% X) %*% t(X) %*% Y
```

```
##        [,1]
##  [1,] 64.10
##  [2,] 72.85
##  [3,] 64.10
##  [4,] 72.85
##  [5,] 72.95
##  [6,] 81.70
##  [7,] 72.95
##  [8,] 81.70
##  [9,] 81.80
## [10,] 90.55
## [11,] 81.80
## [12,] 90.55
## [13,] 90.65
## [14,] 99.40
## [15,] 90.65
## [16,] 99.40
```

$\hat{\beta}_0 + \hat{\beta}_1 * MoistureContent + \hat{\beta}_2 * Sweetness$ is :

```
X %*% matrix(Mymodel$coefficients[, 1], nrow = 3, ncol = 1)
```

```
##        [,1]
##  [1,] 64.10
##  [2,] 72.85
##  [3,] 64.10
##  [4,] 72.85
##  [5,] 72.95
##  [6,] 81.70
##  [7,] 72.95
##  [8,] 81.70
##  [9,] 81.80
## [10,] 90.55
## [11,] 81.80
## [12,] 90.55
## [13,] 90.65
## [14,] 99.40
## [15,] 90.65
## [16,] 99.40
```

This two methods we get the same result.

Question 6

$$
\begin{aligned}
H \times H = X(X^TX)^{-1}X^T \times X(X^TX)^{-1}X^T &= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\
&= X(X^TX)^{-1}(X^TX(X^TX)^{-1})X^T \\
&= X(X^TX)^{-1}X^T
\end{aligned}
$$

$$= H$$

Question 7

The restricted residual sum of squares would be bigger since fewer variables are included in the model to capture the variability of the dependent variable.

Question 8

```
Body <- read.csv("BodyFatPercentage.csv")
model1 <- lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + NECK + CHEST +
    HIP + THIGH, data = Body)
model2 <- lm(BODYFAT ~ AGE + NECK + CHEST + THIGH, data = Body)
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: BODYFAT ~ AGE + WEIGHT + HEIGHT + NECK + CHEST + HIP + THIGH
## Model 2: BODYFAT ~ AGE + NECK + CHEST + THIGH
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    244 6159.1
## 2    247 6395.0 -3   -235.97 3.1161 0.02684 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So the F-statistic of restrcited model and unrestricted model is:

$$\frac{(6395.0-6159.1)/3}{6159.1/244} = 3.11$$

Which is the same as we can directly observe from the anova table.

Question 9

- 7 dummy variables we need to use.
- We just need to transform the variable into factor and use `lm` to run the regression and R would create the dummy variables automatically.
- It would cause multicollinearity since the eighth dummy variable is the linear combination of previous 7 ones.

Question 10

(a)

```
FEV <- read.csv("FEV.csv")
(model10 <- summary(lm(FEV ~ AGE + SEX + SMOKER + AGE * SEX +
    SEX * SMOKER + SMOKER * AGE, data = FEV)))
```

```
##
## Call:
## lm(formula = FEV ~ AGE + SEX + SMOKER + AGE * SEX + SEX * SMOKER +
##     SMOKER * AGE, data = FEV)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.70166 -0.31145 -0.01758  0.29933  1.81958
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.69028    0.10775   6.406    0.000000000287297 ***
## AGE          0.18033    0.01104  16.332 < 0.0000000000000002 ***
## SEX         -0.76220    0.14633  -5.209    0.000000255937855 ***
## SMOKER       2.14912    0.37905   5.670    0.000000021556621 ***
## AGE:SEX      0.10936    0.01474   7.419    0.000000000000373 ***
## SEX:SMOKER   0.01048    0.14886   0.070                0.944
## AGE:SMOKER  -0.17079    0.02838  -6.017    0.000000002966795 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5058 on 647 degrees of freedom
## Multiple R-squared:  0.6628, Adjusted R-squared:  0.6597
## F-statistic:   212 on 6 and 647 DF,  p-value: < 0.00000000000000022
```

(b)

If we move from a subject that is a female and a non-smoker to a subject that is male and a smoker, then the (SEX,SMOKER) variable would change from (0,0) to (1,1), and the forced expiratory volume would change `-0.762+2.149-0.010+(0.109-0.171)*AGE = 1.377-0.062*AGE` , on average, else being equal, where `AGE` is the based on the age of the observation.

(c)

The coefficient of AGE is 0.18033, and the coefficient of interaction terms of AGE:SEX,AGE:SMOKER are 0.10936 and -0.17079 respectively. Which means:

- For female non-smoker, each age increase accompany with FEV increase 0.18033, on average, else being equal.
- For female smoker, each age increase accompany with FEV increase 0.18033-0.17079=0.00954, on average, else being equal.
- For male non-smoker, each age increase accompany with FEV increase 0.18033+0.01048=0.19081, on average, else being equal.
- For male smoker, each age increase accompany with FEV increase 0.18033+0.01048-0.17079=0.01002, on average, else being equal.

For all types of samples, the coefficient are all positive, that means all FEV would increase as age grows. If the data set was full of individuals aged 40-65 years old, this interpret would be a bit comfused.

(d)

The coeffient of the interation term is -0.17, which means for smokers, the slope of Price agains Age would be 0.17 less, that is 0.18-0.17 = 0.01. Which means for each one unit Age increase, the increase of forced expiratory volume would be 0.17 less for smokers, compared with non-smokers.

(e)

I think the total years for smoking would better explain how FEV decreases over time for smokers, since the smokers distribution in different ages would be different.