hw2

Keyi Zhong

4/24/2020

1.1

```
library(ISLR)
data <- Auto
data$origin <- factor(data$origin)
lm.fit <- lm(mpg~.,data=data[-9])
summary(lm.fit)$coefficients
```

```
##                  Estimate   Std. Error     t value     Pr(>|t|)
## (Intercept)  -17.954602067 4.6769339310  -3.8389685 1.445124e-04
## cylinders     -0.489709424 0.3212308567  -1.5244782 1.282146e-01
## displacement   0.023978644 0.0076532690   3.1331244 1.862685e-03
## horsepower    -0.018183464 0.0137085987  -1.3264276 1.854885e-01
## weight        -0.006710384 0.0006551331 -10.2427793 6.375633e-22
## acceleration   0.079103036 0.0982184978   0.8053782 4.211012e-01
## year           0.777026939 0.0517840867  15.0051297 2.332943e-40
## origin2        2.630002360 0.5664146647   4.6432455 4.720373e-06
## origin3        2.853228228 0.5527363020   5.1620062 3.933208e-07
```

One unit growth of cylinders will make 0.49 decrease in mpg, the t value is -1.526 so we should not reject the null hypothesis. One unit growth of displacement will make 0.02 increase in mpg, the t value is 2.647, so we should probabaly reject the null hypothesis if we want to make it strict. One unit growth of horsepower will make 0.017 decrease in mpg, the t value is -1.23, so we should not reject the null hypothesis. One unit growth of weight will make 0.0065 decrease in mpg, the t value is -9.9, so we should reject the null hypothesis. One unit growth of acceleration will make 0.09 increase in mpg, the t value is 0.815, so we should not reject the null hypothesis. One unit growth of year will make 0.75 increase in mpg, the t value is 14.7, so we should reject the null hypothesis. The difference in American cars and European cars is 2.63 mpg, and the t value is 4.643 so we can reject the null hypothesis. The difference in American cars and Japanese cars is 2.85 mpg, and the t value is 5.162 so we can reject the null hypothesis.

1.2

```
mean(lm.fit$residuals^2)
```

```
## [1] 10.68212
```

1.3

```
predict(lm.fit,data.frame(cylinders=3,displacement=100,horsepower=85,weight=3000,acceleration=20,year=8
```

```
##        1
## 27.89483
```

1.4

```
lm.fit$coefficients["origin3"]
```

```
##  origin3
## 2.853228
```

1.5

```
10*lm.fit$coefficients["horsepower"]
```

```
## horsepower
## -0.1818346
```

2.1

$$x_{i1} = \begin{cases} 1 & \text{if ith car is American} \\ 0 & \text{if ith car is not American} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if ith car is European} \\ 0 & \text{if ith car is not European} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith car is American} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if ith car is European} \\ \beta_0 + \epsilon_i & \text{if ith car is Japanese} \end{cases}$$

```
jap.avg <-  mean(Auto$mpg[Auto$origin==3])
us.avg <- mean(Auto$mpg[Auto$origin==1])
eu.avg <- mean(Auto$mpg[Auto$origin==2])

x1 <- rep(0,length(Auto))
x1[which(Auto$origin==1)] <- 1
x1[which(Auto$origin!=1)] <- 0

x2 <- rep(0,length(Auto))
x2[which(Auto$origin==2)] <- 1
x2[which(Auto$origin!=2)] <- 0
lm.fit2.1 <- lm(Auto$mpg~x1+x2)
summary(lm.fit2.1)$coefficients
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  30.450633  0.7196327  42.314129 1.214739e-147
## x1          -10.417164  0.8275617 -12.587779  1.023502e-30
## x2           -2.847692  1.0580718  -2.691398  7.422377e-03
```

β_0 is the aaverage mpg for all Japanese car which is 30.4506329, β_1 is the difference in average mpg between Japanese and American Car which is -10.4171635, β_2 is the difference in average mpg between Japanese and European Car which is -2.8476917. The predicted mpg for a Japanese car is 30.4506329, for an American car is 20.0334694, and for an European car is 27.6029412

2.2

$$x_{i1} = \begin{cases} 1 & \text{if ith car is Japanese} \\ 0 & \text{if ith car is not Japanese} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if ith car is European} \\ 0 & \text{if ith car is not European} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if ith car is Japanese} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if ith car is European} \\ \beta_0 + \epsilon_i \quad \text{if ith car is American} \end{cases}$$

```r
jap.avg <-  mean(Auto$mpg[Auto$origin==3])
us.avg <- mean(Auto$mpg[Auto$origin==1])
eu.avg <- mean(Auto$mpg[Auto$origin==2])

x1 <- rep(0,length(Auto))
x1[which(Auto$origin==3)] <- 1
x1[which(Auto$origin!=3)] <- 0

x2 <- rep(0,length(Auto))
x2[which(Auto$origin==2)] <- 1
x2[which(Auto$origin!=2)] <- 0
lm.fit2.2 <- lm(Auto$mpg~x1+x2)
summary(lm.fit2.2)$coefficients
```

```
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 20.033469  0.4086405 49.024678 1.383741e-168
## x1          10.417164  0.8275617 12.587779  1.023502e-30
## x2           7.569472  0.8767164  8.633889  1.543152e-16
```

β_0 is the average mpg for all American car which is 20.0334694, β_1 is the difference in average mpg between American and Japanese Car which is 10.4171635, β_2 is the difference in average mpg between American and European Car which is 7.5694718. The predicted mpg for a Japanese car is 30.4506329, for an American car is 20.0334694, and for an European car is 27.6029412

2.3

$$x_{i1} = \begin{cases} 1 & \text{if ith car is Japanese} \\ -1 & \text{if ith car is not Japanese} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if ith car is European} \\ -1 & \text{if ith car is not European} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 - \beta_2 + \epsilon_i & \text{if ith car is Japanese} \\ \beta_0 - \beta_1 + \beta_2 + \epsilon_i & \text{if ith car is European} \\ \beta_0 - \beta_1 - \beta_2 + \epsilon_i & \text{if ith car is American} \end{cases}$$

```
x1 <- rep(0,length(Auto))
x1[which(Auto$origin==3)] <- 1
x1[which(Auto$origin!=3)] <- -1

x2 <- rep(0,length(Auto))
x2[which(Auto$origin==2)] <- 1
x2[which(Auto$origin!=2)] <- -1
lm.fit2.3 <- lm(Auto$mpg~x1+x2)
summary(lm.fit2.3)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 29.026787  0.5290359 54.867331 3.309713e-185
## x1           5.208582  0.4137808 12.587779  1.023502e-30
## x2           3.784736  0.4383582  8.633889  1.543152e-16
```

```
beta0 <- lm.fit2.3$coefficients[1]
beta1 <- lm.fit2.3$coefficients[2]
beta2 <- lm.fit2.3$coefficients[3]
```

The predicted mpg for a Japanese car is 30.4506329, for an American car is 20.0334694, and for an European car is 27.6029412

2.4

```
origin <- Auto$origin
mpg <- Auto$mpg
origin <- replace(origin,which(origin==3),0)
lm.fit2.4 <- lm(mpg~origin)
summary(lm.fit2.4)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 25.239473  0.7332078 34.423357 2.865396e-120
## origin      -1.845337  0.6384470 -2.890353  4.063214e-03
```

$$x_i = \begin{cases} 0 & \text{if ith car is Japanese} \\ 1 & \text{if ith car is American} \\ 2 & \text{if ith car is European} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \epsilon_i & \text{if ith car is Japanese} \\ \beta_0 + \beta_1 + \epsilon_i & \text{if ith car is European} \\ \beta_0 + 2\beta_1 + \epsilon_i & \text{if ith car is American} \end{cases}$$

β_0 is -17.9546021 and β_1 is -0.4897094. The predicted mpg for a Japanese car is 25.2394727, for an American car is , and for an European car is

2.5

```r
e1 <- mean(lm.fit2.1$residuals^2)
e2 <- mean(lm.fit2.2$residuals^2)
e3 <- mean(lm.fit2.3$residuals^2)
e4 <- mean(lm.fit2.4$residuals^2)
```

The first two results have the same predictions for Japanese, American and European cars average mpg, so that their training squared errors also should be the same but with different coefficient and interpretations and interceptions except the last two which is 40.5987307. The training squared errors for the third is 40.5987307 and for the fourth is 59.4884444. The third fit the best.

3.1

```r
-165.1+4.8*64
```

```
## [1] 142.1
```

3.2

The coefficient estimate β_0^* is -165.1 and β_1^* is 57.6. 142.1.

3.3

Because X_1 and X_2 are highly dependent, $X_1 = X_2 * 12$, either β_1 or β_2 can be zero. As long as $\beta_1 + \beta_2/12 = 4.8$

3.4

The training mean squared erroes in three models should be the same because change unit means there is a linear relationship between two variables which will not affect the model.