HW2

Nan Tang

4/15/2020

Problem 1

1

```
auto_model <- mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + factor(origi

full_ls <- lm(data=Auto, formula=auto_model)

summary(full_ls)$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
## factor(origin)2   2.630002360 0.5664146647   4.6432455 4.720373e-06
## factor(origin)3   2.853228228 0.5527363020   5.1620062 3.933208e-07
```

Interpretation

While values of all other variabels are held fixed, every one unit increment on cylinders will decrease mpg by 0.49 unit on average.

While values of all other variables are held fixed, one unit increment on displacement will increase mpg by 0.0199 unit on average.

While values of all other variables are held fixed, one unit increment on horsepower will decrease mpg by 0.017 unit on average.

While values of all other variables are held fixed, one unit increment on weight will decrease mpg by 0.0065 unit on average.

While values of all other variabels are held fixed, one unit increment on acceleration will increase mpg by 0.081 unit on average.

While values of all other variables are held fixed, one unit increment on year will increase mpg by 0.75 unit on average.

Categorical variable 'origin'

While values of all other variables are equal to zero, the expected mean value of mpg for US made car is -17.95.

While values of all other variabels are fixed, European cars are in average 2.63 unit higher than US cars on mpg, Japanese cars are in average 2.85 unit higher than US cars on mpg.

Hypothesis Check

At significance level of 0.05, evidences provided by p-value of t-test for each predictors are significant to reject the null hypothese of zero linear coefficients for 'displacement', 'weight', 'year', 'origin' and Intercept.

2

```
full_ls_fit <- full_ls$fitted.values

resub_MSE <- sum((full_ls_fit - Auto$mpg)^2) / length(Auto$mpg)

print(resub_MSE)
```

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

3

```
jap_df <- data.frame(cylinders=3, displacement=100, horsepower=85, weight=3000, acceleration=20, year=8

predict(full_ls, jap_df)
```

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

4

```
summary(full_ls)$coefficient[c(8, 9),]
```

```
##                  Estimate Std. Error  t value     Pr(>|t|)
## factor(origin)2 2.630002  0.5664147 4.643246 4.720373e-06
## factor(origin)3 2.853228  0.5527363 5.162006 3.933208e-07
```

While holding all other variables fixed, mpg of Japanese car is averagely 2.85 higher than mpg of American car.

5

```
full_ls$coefficients[[4]] * 10
```

```
## [1] -0.1818346
```

While holding all other variables fixed, 10-unit increment on horsepower will averagely decrease mpg by 0.18 unit.

Problem 2

1

```
auto_origin <- factor(Auto$origin, levels=c('3', '1', '2'))

ori_ls <- lm(Auto$mpg ~ auto_origin)
summary(ori_ls)$coefficients
```

```
##                Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)   30.450633  0.7196327  42.314129 1.214739e-147
## auto_origin1 -10.417164  0.8275617 -12.587779  1.023502e-30
## auto_origin2  -2.847692  1.0580718  -2.691398  7.422377e-03
```

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.

where β_0 is average mpg for Japanese cars; β_1 is difference in average mpg between US cars and Japanese cars; β_2 is difference in average mpg between European cars and Japanese cars.

Estimators: $\hat{\beta_0} = 30.451, \hat{\beta_1} = -10.417, \hat{\beta_2} = -2.848$

In this case, if the auto is made in Japan, $x_{i1} = x_{i2} = 0$, $y_i = \beta_0 + \epsilon = 30.451 + \epsilon$.

If the auto is made in US, $x_{i1} = 1, x_{i2} = 0$, $y_i = \beta_0 + \beta_1 + \epsilon = 20.033 + \epsilon$.

If the auto is made in Europe, $x_{i1} = 0, x_{i2} = 1$, $y_i = \beta_0 + \beta_2 + \epsilon = 27.603 + \epsilon$.

Predicted mpg on average for Japanese car is 30.451, for US car is 20.033, for European car is 27.603.

2

```
auto_origin <- factor(Auto$origin, levels=c('1', '2', '3'))

ori_ls <- lm(Auto$mpg ~ auto_origin)
summary(ori_ls)$coefficients
```

```
##                Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)   20.033469  0.4086405 49.024678 1.383741e-168
## auto_origin2   7.569472  0.8767164  8.633889  1.543152e-16
## auto_origin3  10.417164  0.8275617 12.587779  1.023502e-30
```

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.

where β_0 is average mpg for American cars; β_1 is difference in average mpg between European cars and American cars; β_2 is difference in average mpg between Japanese cars and American cars.

Estimators: $\hat{\beta_0} = 20.033, \hat{\beta_1} = 7.569, \hat{\beta_2} = 10.417$

In this case, if the auto is made in America, $x_{i1} = x_{i2} = 0$, $y_i = \beta_0 + \epsilon = 20.033 + \epsilon$.

If the auto is made in Europe, $x_{i1} = 1, x_{i2} = 0$, $y_i = \beta_0 + \beta_1 + \epsilon = 27.063 + \epsilon$.

If the auto is made in Japan, $x_{i1} = 0, x_{i2} = 1$, $y_i = \beta_0 + \beta_2 + \epsilon = 30.451 + \epsilon$.

Predicted mpg on average for Japanese car is 30.451, for US car is 20.033, for European car is 27.603.

3

```
ori_US <- rep(-1, nrow(Auto))
ori_US[which(Auto$origin == 1)] <- 1

ori_EU <- rep(-1, nrow(Auto))
ori_EU[which(Auto$origin == 2)] <- 1

ori_ls <- lm(Auto$mpg ~ ori_US + ori_EU)
summary(ori_ls)$coefficients
```

```
##              Estimate Std. Error    t value      Pr(>|t|)
## (Intercept) 23.818205  0.4383582  54.335028 9.487991e-184
## ori_US      -5.208582  0.4137808 -12.587779  1.023502e-30
## ori_EU      -1.423846  0.5290359  -2.691398  7.422377e-03
```

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.

Estimators: $\hat{\beta}_0 = 23.82, \hat{\beta}_1 = -5.21, \hat{\beta}_2 = -1.42$

In this case, if the auto is made in America, $x_{i1} = 1, x_{i2} = -1$, $y_i = \beta_0 + \beta_1 - \beta_2 + \epsilon = 20.033 + \epsilon$.

If the auto is made in Europe, $x_{i1} = -1, x_{i2} = 1$, $y_i = \beta_0 - \beta_1 + \beta_2 + \epsilon = 27.063 + \epsilon$.

If the auto is made in Japan, $x_{i1} = -1, x_{i2} = 11$, $y_i = \beta_0 - \beta_1 - \beta_2 + \epsilon = 30.451 + \epsilon$.

Predicted mpg on average for Japanese car is 30.451, for US car is 20.033, for European car is 27.603.

4

```
new_auto <- Auto
new_auto$origin[which(Auto$origin == 3)] <- 0

ori_ls <- lm(data=new_auto, formula=mpg ~ origin)
summary(ori_ls)$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
```

$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

Estimators: $\hat{\beta}_0 = 25.239, \hat{\beta}_1 = -1.845$

In this case, if the auto is made in Japan, $x_i = 0$, $y_i = \beta_0 + \epsilon = 25.239 + \epsilon$.

If the auto is made in America, $x_i = 1$, $y_i = \beta_0 + \beta_1 + \epsilon = 23.394 + \epsilon$.

If the auto is made in Europe, $x_i = 2$, $y_i = \beta_0 + 2\beta_2 + \epsilon = 21.549 + \epsilon$.

Predicted mpg on average for Japanese car is 25.239, for US car is 23.394, for European car is 21.549.

5

Although we use differnt values to represent levels of categorical variables, the predications came out from those different models are exactly same. Predicted mpg on average for Japanese car is 25.239, for US car is 23.394, for European car is 21.549. In this case, we can see representation of values in categorical variables (value of X1, X2) or number of variables will not affect prediction.

Problem 3

1

The expected value of weight for individual who is 64 inches tall is 142.1

2

Let X^* denotes height in unit of feet.

$\beta_1^* = \frac{\sum(x_i^* - \bar{x}^*)(y_i - \bar{y})}{\sum(x_i^* - \bar{x}^*)^2} = \frac{\sum(\frac{1}{12}x_i - \frac{1}{12}\bar{x})(y_i - \bar{y})}{\sum(\frac{1}{12}x_i - \frac{1}{12}\bar{x}^*)^2} = 12\beta_1 = 57.6$

$\beta_0^* = \bar{y} - \frac{1}{12}\bar{x}\beta_1^* = \beta_0 = -165.1$

The expected value of weight for individual who is 5.333 feet tall is 142.1

3

The least square model can be represented as:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 \frac{X_1}{X_2} + \epsilon$

$= \beta_0 + (\beta_1 + \frac{\beta_2}{12})X_1 + \epsilon$

Least square solution for coefficients in this problem and previous one on X_1, Y should be equivalient. Therefore, implicit expression for β_1, β_2 is $\beta_1 + \beta_2/12 = 4.8$, $\beta_0 = -165.1$.

4

Resubstitution error for these three model should be exactly same.