Statistical Modelling Techniques Homework-6

Berke Kaygısız

```
data(iris) # I used it to call the data over R.
```

PROBLEM Does sepal length have a statistically significant effect on petal length?

```
colSums(is.na(iris)) # I used it to check for missing observations in each column.
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width      Species
##            0            0            0            0            0
```

```
sum(is.na(iris)) # I used it to check the total number of missing observations.
```

```
## [1] 0
```

```
str(iris) # I used it to study the structure of the data set.
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(iris) # I used it to review the first few observations.
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
summary(iris) # I used it to review summary statistics.
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
```

```
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

I checked the missing values in the data set, there are no missing values in the data set. There are 150 observations and 5 variables in the Dataset. The types of variables are as follows; "Sepal.Length": A variable consisting of numeric values. "Sepal.Width": A variable consisting of numeric values. "Petal.Length": A variable consisting of numeric values. "Petal.Width": A variable consisting of numeric values. "Species": A factor variable with three different level. With summary statistics, I obtained min, max, quartiles, median, mean values for each variable.

```
regmodel <- lm(Petal.Length ~ Sepal.Length, data = iris) # I used it to create the regression model.


summary(regmodel) # I used it to view the model summary.
```

```
##
## Call:
## lm(formula = Petal.Length ~ Sepal.Length, data = iris)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.47747 -0.59072 -0.00668  0.60484  2.49512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7.10144    0.50666  -14.02   <2e-16 ***
## Sepal.Length  1.85843    0.08586   21.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8678 on 148 degrees of freedom
## Multiple R-squared:   0.76,  Adjusted R-squared:  0.7583
## F-statistic: 468.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

I set up the regression model this way because I wanted to investigate the effect of the Sepal.Length variable on Petal.length. According to the results of regression analysis; According to the r-squared value, x explain y 76% of the time. The model makes sense because the f-statistic value is 468.6 and the p-value is 2.2e-16, which is a very small value compared to 0.05. It intersects the Y coordinate at -7.10144. When Sepal.Length increases by one unit, Petal.Length will increase by 1.85843 units.

```
anova_result <- aov(Sepal.Length ~ Species, data = iris) # I used it to do the ANOVA analysis.


summary(anova_result) # I used it to view the ANOVA results.
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
```

```
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the F statistic is 119.3 and the p-value is <2e-16, at least one group mean between species is different from each other. According to the results of ANOVA analysis, there are statistically significant differences in sepal length between different species.

```
data(mtcars) # I used it to call the data over R.
```

PROBLEM Does car weight have a statistically significant effect on car performance?

```
colSums(is.na(mtcars)) # I used it to check for missing observations in each column.
```

```
##  mpg  cyl disp   hp drat   wt qsec   vs   am gear carb
##    0    0    0    0    0    0    0    0    0    0    0
```

```
sum(is.na(mtcars)) # I used it to check the total number of missing observations.
```

```
## [1] 0
```

```
str(mtcars) # I used it to study the structure of the data set.
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
head(mtcars) # I used it to review the first few observations.
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
summary(mtcars) # I used it to review summary statistics.
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

I checked the missing values in the data set, there are no missing values in the data set. There are 32 observations and 11 variables in the Dataset. Variables and their types are as follows; mpg: Performance of vehicles in miles per gallon (numerical) cyl: Number of cylinders (numeric) disp: Engine cylinder volume (numerical) hp: Horsepower (numeric) drat: Rear axle ratio (numeric) wt: Weight (numeric) qsec: 1/4 mile time (numeric) etc: Engine (V) or straight engine (binary) am: Gear type (automatic or manual) (binary) gear: Number of gears (numeric) carb: Carburetor type (numeric) With summary statistics, I obtained min, max, quartiles, median, mean values for each variable.

```r
regression_model <- lm(mpg ~ wt, data = mtcars) # I used it to create the regression model.


summary(regression_model) # I used it to view the model summary.
```

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.5432 -2.3647 -0.1252  1.4096  6.8727 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
## wt           -5.3445     0.5591  -9.559 1.29e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.046 on 30 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7446 
## F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
```

That's how I set up the regression model because I wanted to investigate the effect of car weight on car performance. According to the results of regression analysis; Since R-squared is 0.7528, X explains Y by 75.28%. The model is significant because the f-statistic value is 91.38 and the p-value is 1.294e-10, less than 0.05. It intersects the Y coordinate at point 37.2851. There is a negative correlation between car weight and car performance, so when car weight increases by one unit, car performance will decrease by 5.3445 units.

```
ancova_model <- lm(mpg ~ wt + as.factor(cyl), data = mtcars) # I used it to do the ANCOVA analysis.


summary(ancova_model) # I used it to view the ANCOVA results.
```

```
##
## Call:
## lm(formula = mpg ~ wt + as.factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.9908     1.8878  18.006  < 2e-16 ***
## wt               -3.2056     0.7539  -4.252 0.000213 ***
## as.factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
## as.factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374, Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

When we look at the CYL variable, 6-cylinder cars have 4.2556 units lower performance than 8-cylinder cars. When we look at the significance levels, we can understand that all variables are statistically significant on the model.