HW4_ISYE6414_ashish_dhiman

Ashish Dhiman | adhiman9@gatech.edu

2022-10-06

```
library(ggplot2)
setwd("~/data_projects/fall22_hw/isye6414_hw/hw4")
```

Boxplot and ANOVA

Read Data and Summary

```
head -5 ./homework04data01.csv
```

```
## "Species","Length"
## "a",7.72
## "a",7.26
## "a",9.1
## "a",3.69
```

```
data1 = read.csv("./homework04data01.csv",sep=",")
dim(data1)
```

```
## [1] 85  2
```

```
summary(data1)
```

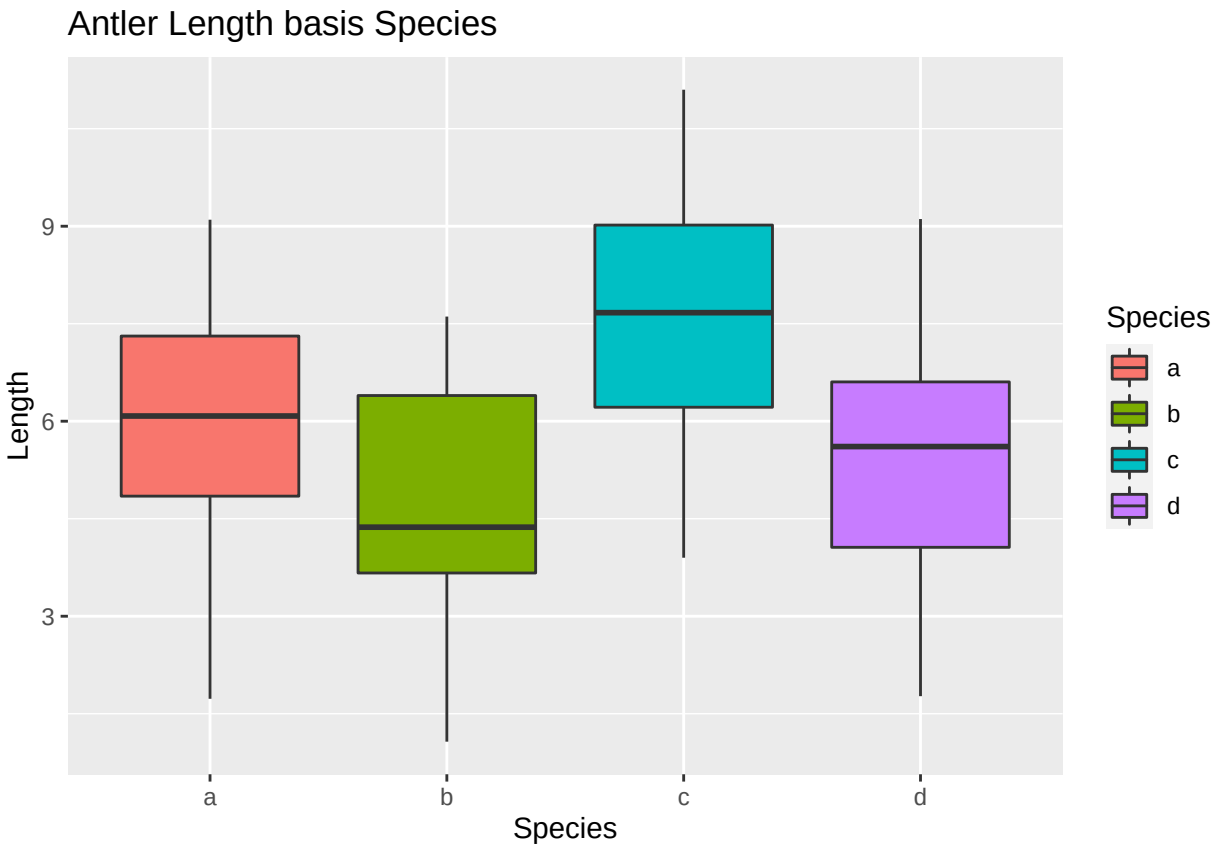
```
##    Species              Length
##  Length:85          Min.   : 1.070
##  Class :character   1st Qu.: 4.620
##  Mode  :character   Median : 6.100
##                     Mean   : 6.015
##                     3rd Qu.: 7.570
##                     Max.   :11.100
```

```
table(data1$Species)
```

```
##
##  a  b  c  d
## 20 15 22 28
```

Question 1: Box Plots

```
ggplot(data1, aes(x=Species, y=Length, fill=Species)) +
    geom_boxplot() + ggtitle("Antler Length basis Species")
```



From the box plot above we see there is a significant overlap between the ranges of antler length between the different species. Hence, without further investigation, it is very difficult to conclude if there is a difference between the antler length among species.

The box plot however, does give us an indication of possible differences:

1. Mean of c is larger than every one else
2. Mean of a and d are very similar

Question 2: ANOVA to test difference in mean with $\alpha = 0.05$

```
anova_v0 = aov(Length ~ Species, data=data1)
model.tables(anova_v0, type = "means")
```

```
## Tables of means
## Grand mean
##
## 6.014706
```

```
##
##  Species
##          a      b      c      d
##      6.063  4.697  7.594  5.445
## rep 20.000 15.000 22.000 28.000
```

Hypothesis test for equal means:

$$H_0 : \mu_a = \mu_b = \mu_c = \mu_d$$

$$H_a : \text{Some means are different}$$

```
summary(anova_v0)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Species      3   90.0  30.001   8.121 8.58e-05 ***
## Residuals   81  299.2   3.694
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above table we see that the p-value is 8.58e-5. This is fairly smaller than 5%, thus we have sufficient statistical evidence to reject Null Hypothesis, i.e. All species have similar means

Question 3: Identify SSE, SSTr, MSE, MSTr

From above table:

- SSE = 299.2
- SSTr = 90
- MSE = SSE/(N-k) = 299.2/81 = 3.694 (from table)

```
299.2/81
```

```
## [1] 3.693827
```

- MSTr = SSTr/k-1 = 90.0/3 = 30.001 (from table)

```
90.0/3
```

```
## [1] 30
```

Question 4: Pairwise comparison of means

```
TukeyHSD(anova_v0)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Length ~ Species, data = data1)
##
## $Species
##           diff         lwr        upr     p adj
## b-a -1.3656667 -3.08777369  0.3564404 0.1683081
## c-a  1.5306364 -0.02706805  3.0883408 0.0559481
## d-a -0.6176429 -2.09373459  0.8584489 0.6920003
## c-b  2.8963030  1.20807683  4.5845292 0.0001303
## d-b  0.7480238 -0.86520632  2.3612539 0.6183736
## d-c -2.1482792 -3.58469904 -0.7118594 0.0010291
```

The above table gives the 95% CI for the 6 pairwise difference of means above.

As expected from the box plot, c-b and d-c both have low p-values. Similarly, p-value for c-a is about 5%, hinting there is a slight difference in means, espcially for alpha <6%.

While d-b has lowest absolute diff and the highest p-value.

Question 5: Multiple Linear regression

y: monthly labor hours required

x1: monthly X-ray exposures

x2: monthly occupied bed days

x3: average length of patients' stays (in days).

```
data2 = read.csv("./homework04Hospital.csv",sep=",")
dim(data2)
```

```
## [1] 16  4
```

```
summary(data2)
```

```
##       Xray          BedDays            Length          Hours
##  Min.   : 2048   Min.   :  472.9   Min.   : 3.900   Min.   :  566.5
##  1st Qu.: 5765   1st Qu.: 1359.3   1st Qu.: 4.810   1st Qu.: 1609.4
##  Median : 9616   Median : 2279.7   Median : 5.560   Median : 2233.1
##  Mean   :17036   Mean   : 4280.4   Mean   : 5.824   Mean   : 4643.1
##  3rd Qu.:16684   3rd Qu.: 3879.5   3rd Qu.: 6.223   3rd Qu.: 3812.7
##  Max.   :86533   Max.   :15524.0   Max.   :10.780   Max.   :18854.5
```

```
lr_model = lm(Hours ~ ., data = data2)
summary(lr_model)
```

```
##
## Call:
## lm(formula = Hours ~ ., data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -677.23 -270.19   60.93  228.32  517.70
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1946.80204  504.18193   3.861  0.00226 **
## Xray           0.03858    0.01304   2.958  0.01197 *
## BedDays        1.03939    0.06756  15.386 2.91e-09 ***
## Length      -413.75780   98.59828  -4.196  0.00124 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 387.2 on 12 degrees of freedom
## Multiple R-squared:  0.9961, Adjusted R-squared:  0.9952
## F-statistic:  1028 on 3 and 12 DF,  p-value: 9.919e-15
```

From the above table we can have beta of the different predictors

- $\beta_{xray} = 0.03858$
- $\beta_{bed\_days} = 1.03939$
- $\beta_{length} = -413.75780$

Question 6: Interpretation of each coefficient

- $\beta_{xray} = 0.03858$: This implies that with each unit monthly exposure of xray the expected monthly labor hours required increase by 0.03858, given rest of the predictors remain constant.
- $\beta_{bed\_days} = 1.03939$. Similarly with each unit increase in monthly occupied bed days the expected monthly labor hours required increase by 1.03939, given rest of the predictors remain constant.
- $\beta_{length} = -413.75780$ With each unit increase in average length of patients' stay (in days), the expected monthly labor hours required decrease by 413.7580, given rest of the predictors remain constant. The negative effect seems a little counter intuitive here, but might because with greater length of stay in hospital, the labor requirement is more spread out, but we need more information on data collection to identify this.

```
test = lm(Hours ~ Length, data = data2)
summary(test)
```

```
##
## Call:
## lm(formula = Hours ~ Length, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6068.1 -2395.8 -1177.8   180.5 13193.4
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6636.4     4613.4  -1.439    0.172
## Length        1936.6      765.2   2.531    0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4769 on 14 degrees of freedom
## Multiple R-squared:  0.3139, Adjusted R-squared:  0.2649
## F-statistic: 6.406 on 1 and 14 DF,  p-value: 0.02399
```

Thus in absence of other predictors, Length is postiively effecting y. Also note that the intercept value is negative here, and p-value has also changed.

Question 7: Hypothesis Test on BedDays

$$H_0 : \beta_{bed\_days} = 0 \quad H_a : \beta_{bed\_days} \neq 0$$

test statistic = 15.386

Critical t value, from qt function

```
print ("Upper tale t")
```

```
## [1] "Upper tale t"
```

```
qt(0.975, df = 12)
```

```
## [1] 2.178813
```

```
print ("Lower tale t")
```

```
## [1] "Lower tale t"
```

```
qt(0.025, df = 12)
```

```
## [1] -2.178813
```

Thus are |critical t value| is 2.178813

Now given out test statistic is significantly larger than critical upper tail t value, we have sufficient evidence to reject the null hypothesis and conclude that BedDays variable is a significant predictor, given all other predictor variables in the model.

Question 8: Hypothesis Test on Xray

$$H_0 : \beta_{xray} = 0 \quad H_a : \beta_{xray} \neq 0$$

For the above we have p-value as $0.01197 = 1.197\%$ (this is two tailed probability here)

For $\alpha = 5\%, p < \alpha$ Hence we can reject the null hypothesis in this case, and have sufficient statistical evidence to conclude that **Xray variable is a significant predictor**, given all other predictor variables in the model.

However if our confidence level changes, i.e. for $\alpha = 1\%, p > \alpha$ Hence we can not reject the null hypothesis in this case, and conclude that **Xray variable is not a significant predictor**, given all other predictor variables in the model.

Question 8: Hypothesis Test on Length

$$H_0 : \beta_{length} = 0 \quad H_a : \beta_{length} \neq 0$$

For the above we have p-value as $0.00124 = 0.124\%$ (this is two tailed probability here)

For $\alpha = 5\%, p < \alpha$ Hence we can reject the null hypothesis in this case, and have sufficient statistical evidence to conclude that **length variable is a significant predictor**, given all other predictor variables in the model.

Even for more stricter confidence level, i.e. for $\alpha = 1\%, p < \alpha$ Hence we can reject the null hypothesis again in this case, and conclude that **length variable is a significant predictor**, given all other predictor variables in the model.