STATS 752
Written Report for Final Project

Amandeep Sidhu (400076920)

07 April, 2023

Contents

Introduction

Throughout NHL history, it has been thought that the more a team would spend, the better their playoff result would be. However, the best counter-example to this phenomena would be the Toronto Maple Leafs who typically spend in the upper bracket of teams but fail to make it past the first round. Though the salary cap remains ‘flat’, teams abuse the rules by using the LTIR to spend above the cap (the best example of this was the Tampa Bay Lightning who spent well above the cap in their back to back championships). In this final project, we have created a data set that contains the final cap hits for all 32 NHL teams across the years 2019-2022 and also collected their respective season results. The season results consist of 5 levels: missed playoffs, 1st round exit, 2nd round exit, 3rd round exit and a Stanley cup final appearance. The source used to obtain this information was a publicly available resource for the teams cap hits across these years (Wuest 2015). Moreover, the final cap hit was also split into 5 levels (denoted as salary tiers) to ensure that a factor variable could be attained. In the methodology section, we will observe the first few rows of data set and the data entries will be explained (i.e. what do the factors in the cap hit and finish column mean). The final data set that will be used contains 125 observations with 23 (out of 25) non-empty cells that contain the counts of the team in each cell. Refer to the appendix to observe the code on how the data was collected from this website. That will be the data set that we will perform ANOVA on in the later sections of this report.

The problem at hand is as follows: determine whether there is an effect of an NHL teams spending and their respective playoff success. This will be attained by using various hypothesis tests (that are appropriate) and try to draw some conclusions based on the metrics observed. The tests will be determined as we need to investigate (using techniques from lecture) whether or not interaction occurs in the data.

Methodology

Let us observe the first few rows of the final dataset that will be used to compute the ANOVA table.

```
head(nhl_df_finish[, c(1,2,3,5,6)])
```

```
##                   TEAM FINAL CAP HIT  LTIR USED salary_tier finish
## 1   Detroit Red Wings      86422598 $6,922,598           A      1
## 2 Washington Capitals      79623040   $143,128           C      2
## 3 Pittsburgh Penguins      79536752   $331,741           C      2
## 4     Edmonton Oilers      80007877   $864,200           B      1
## 5     San Jose Sharks      79128253         $0           C      4
## 6        Dallas Stars      81113278 $2,135,504           B      3
```

As we see, the salary tier levels have been split into 5 levels: A (which corresponds to a cap hit above 84 million dollars), B (which corresponds to a cap hit between 80 million and 84 million dollars), C (which corresponds to a cap hit between 76 million and 80 million dollars), D (which corresponds to a cap hit between 73 million and 76 million dollars) and finally E (which corresponds to a cap hit below 73 million dollars).

In addition, the 'finish' variable also has 5 levels defined: 1 (which corresponds to missing the playoffs that year), 2 (which corresponds to a 1st round finish), 3 (which corresponds to a 2nd round finish), 4 (which corresponds to a 3rd round finish) and 5 (which corresponds to making the Stanley cup final).

Now, let us observe the two-way crossed classification table with these factor levels defined. The labels at the top will correspond to the salary tier and the labels on the right will correspond to the finish of each team. The 5 levels for each factor were defined as discussed above.

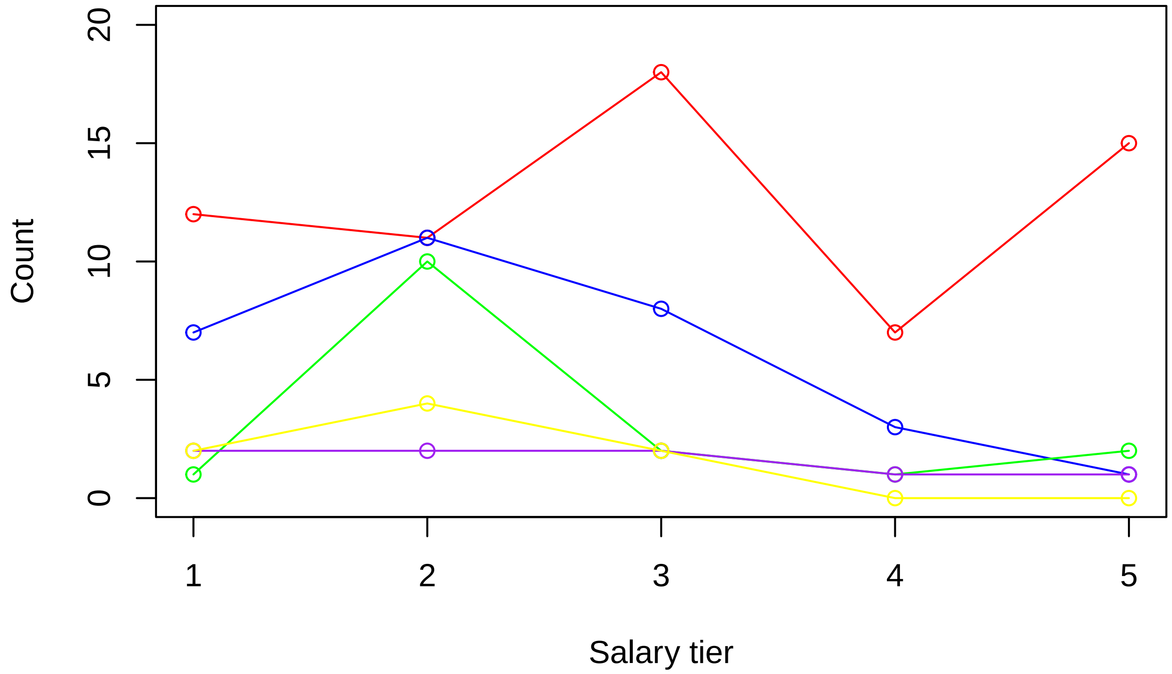
```r
table1
```

```
##
##      A  B  C  D  E
##   1 12 11 18  7 15
##   2  7 11  8  3  1
##   3  1 10  2  1  2
##   4  2  2  2  1  1
##   5  2  4  2  0  0
```

As we can see, there are more than 15 non-empty cells (only 2 in our case). Next, we must determine whether there may be interaction present in our data set. Using the technique presented in class, where we plot the lines of at different levels to see if any cross each other, we can try to determine whether there is any interaction.

```r
#Create a line graph containing the different finish

#get the counts per subgroup of salary tier and finish
library(dplyr)

grouped <- nhl_df_finish %>% group_by(salary_tier, finish)
counts <- grouped %>% summarize(count = n())
counts <- counts %>% rename(num_occurrences = count)
counts <- counts %>% arrange(salary_tier, finish)
count_df <- as.data.frame(counts)

#need to add two rows,

r1 <- c("D", 5, 0)
r2 <- c("E", 5, 0)
```

```
count_df <- rbind(count_df, r1, r2)
count_df <- count_df %>% arrange(salary_tier, finish)


#plot the data


#extract the vector of counts
v1 <- count_df[count_df['finish'] == '1', 3]
v2 <- count_df[count_df['finish'] == '2', 3]
v3 <- count_df[count_df['finish'] == '3', 3]
v4 <- count_df[count_df['finish'] == '4', 3]
v5 <- count_df[count_df['finish'] == '5', 3]

#plot
plot(v1, type = "o", col = "red", xlab = "Salary tier", ylab = "Count",
     ylim = c(0,20))
lines(v2, type = "o", col = "blue")
lines(v3, type = "o", col = "green")
lines(v4, type = "o", col = "purple")
lines(v5, type = "o", col = "yellow")
```



Note, on the x-axis, A = 1, B = 2, C = 3, D = 4 and E = 5 as R is just using those as the index in the plot function. So, as we can clearly see, the lines cross one another and we can conclude that there is interaction between the factors in this data set. Thus, the full ANOVA table will be the one with γ.

So, the full ANOVA table here is as follows (Searle 1997).

Source of Variation	df	Sum of Squares	Mean Squared	F
$\alpha\|\mu$	a -1	$SS(\alpha\|\mu)$	$MS(\alpha\|\mu)$	$F(\alpha\|\mu)$
$\beta\|\mu,\alpha$	b-1	$SS(\beta\|\mu,\alpha)$	$MS(\beta\|\mu,\alpha)$	$F(\beta\|\mu,\alpha)$
$\gamma\|\mu,\alpha,\beta$	s - a - b + 1	$SS(\gamma\|\mu,\alpha,\beta)$	$MS(\gamma\|\mu,\alpha,\beta)$	$F(\gamma\|\mu,\alpha,\beta)$
error	n - s	SSE	MSE	
total	n- 1			

$$SS(\alpha|\mu) = \sum_{i=1}^{a} \frac{Y_{i..}^2}{n_{i.}} - \frac{Y_{...}^2}{n_{..}}$$

$$SS(\beta|\mu, \alpha) = \beta^{\vec{0}}_{b-1}\vec{r}$$

$$SS(\gamma|\mu, \alpha, \beta) = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{Y_{ij.}}{n_{ij}} - \sum_{i=1}^{a} \frac{Y_{i..}^2}{n_{i.}} - \beta^{\vec{0}}_{b-1}\vec{r}$$

$$MS(\alpha|\mu) = \frac{SS(\alpha|\mu)}{a-1}$$

$$MS(\beta|\mu, \alpha) = \frac{SS(\beta|\mu, \alpha)}{b-1}$$

$$MS(\gamma|\mu, \alpha, \beta) = \frac{SS(\gamma|\mu, \alpha, \beta)}{s-a-b+1}$$

$$MSE = \frac{SSE}{n-s}$$

$$F(\alpha|\mu) = \frac{MS(\alpha|\mu)}{MSE}$$

$$F(\beta|\mu, \alpha) = \frac{MS(\beta|\mu, \alpha)}{MSE}$$

$$F(\gamma|\mu, \alpha, \beta) = \frac{MS(\gamma|\mu, \alpha, \beta)}{MSE}$$

When computing the table, we will see the following results:

Source of Variation	df	Sum of Squares	Mean Squared	F
$\alpha\|\mu$	4	379.7855	94.94638	9494638
$\beta\|\mu, \alpha$	4	77.86667	19.46667	1946667
$\gamma\|\mu, \alpha, \beta$	14	118	8.428571	842857.1
error	2	3e-05	1.5e-05	
total	24	575.6522		

This will be used for our testing in the results section (i.e. drawing conclusion using the F statistics calculated). To observe how the statistics were calculated, refer to the appendix section. Note, the aov() function will not work as theres a setting that cannot read in zero cells (which is present in our data). Some observations include a very small SSE which may suggest overfitting the model but the overall data is explained well in terms of variance accounted for. Remember, large F values suggest that the variance of the response variable is explained very well by the model itself.

Results of Testing

Test 1: Perform One Test that is appropriate

$H_0 : \alpha_1 = ... = \alpha_i$ for $a = 1, ..., a$ levels, which is equivalent to testing the following $\alpha_i + \frac{1}{n_i}\sum_{j=1}^{b} n_{ij}\beta_j$ for $i = 1, ..., a$.

H_a : at least one equality fails

Assume significance level of $\alpha = 0.05$.

Our test statistic is $F(\alpha|\mu) = \frac{MS(\alpha|\mu)}{MSE}$

From the full ANOVA table above, we observe that our F value is 9494638 which is much larger than $F_{0.05}(4, 4) = 6.388$. As a result, we reject our null hypothesis and conclude that one of the equality's fails. This is a very interesting observation as we can conclude that the amount of money spent (in terms of final cap hit) does have some sort of a significant impact on the result of the teams season (in terms of playoff success). As a result, teams looking to improve their play should look to spend more money and play within the salary cap rules (like Tampa Bay lightning did with their LTIR usage).

Test 2: Perform a test using the test statistic as instructed

Using (110) of Searle, we can arrive at the following two hypothesis'

H_0 : there is an additive relationship between the two variables/predictors (i.e. the interaction term should be 0)

H_a : there is a significant impact of an interaction term (i.e. should be present in the model)

Assume a significance level of $\alpha = 0.05$.

Our test statistic is $F(\gamma|\mu, \alpha, \beta) = \frac{MS(\gamma|\mu,\alpha,\beta)}{MSE}$.

From the full ANOVA table above, we observe that our F value is 842857.1 which is much larger than $F_{0.05}(14, 2) = 19.424$. As a result, we reject our null hypothesis and conclude that the interaction term is significant. This is a very interesting observation as we can observe the change in the finish versus salary tier factors. This means, if we were to observe the marginals and fix one of the factors, we see that an increase in cap hit spent would suggest more success for the overall team. Likewise, if we were to fix the cap spent along that specific marginal, we see teams tend to do better with the more money they spend. As a result, teams should spend more money to do better overall.

Conclusion

Overall, we see that there is a significant relationship between the two factors collected by our study. As a result, NHL teams should look to abuse the cap hit rules in place and try to spend money over the salary cap as teams who have done this have shown great success in playoff runs. However, if teams wish to perform poorly in an attempt to draft better, they should look to trade larger contracts and players in order to perform worse and potentially recoup a talented prospect.

References

Searle, Ray. 1997. *Linear Models.*

Wuest, Matthew. 2015. *Capfriendly.* https://www.capfriendly.com/.

Appendix

Below is the code used to collect the data from the internet and store it in a usuable format on R.

```r
library(rvest)

# scrape NHL salary cap data from capfriendly.com (will do 2018-2022)

#2018-2029 data
url <- "https://www.capfriendly.com/archive/2019"

# Scrape table data
nhl_data<- url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table()

# Convert list to data frame
nhl_df_2019 <- as.data.frame(nhl_data[[1]])



#2019-2020 data
url <- "https://www.capfriendly.com/archive/2020"

# Scrape table data
nhl_data<- url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table()
```

```
# Convert list to data frame
nhl_df_2020 <- as.data.frame(nhl_data[[1]])


#2020-2021 data
# Define URL
url <- "https://www.capfriendly.com/archive/2021"

# Scrape table data
nhl_data<- url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table()

# Convert list to data frame
nhl_df_2021 <- as.data.frame(nhl_data[[1]])


#scrape 2021-2022 data
# Define URL
url <- "https://www.capfriendly.com/archive/2022"

# Scrape table data
nhl_data<- url %>%
  read_html() %>%
  html_nodes("table") %>%
  html_table()

# Convert list to data frame
nhl_df_2022 <- as.data.frame(nhl_data[[1]])
```

```
nhl_df <- rbind(nhl_df_2019, nhl_df_2020, nhl_df_2021, nhl_df_2022)

#create a categorical variable consisting of the finish of the team
# (1= miss, 2 = 1st round, 3 = 2nd round, 4 = 3rd round, 5 = SF)

finish <- c(1, 2, 2, 1, 4, 3, 5, 1, 2, 2, 1, 2, 5, 1, 1, 1, 3, 1, 2, 2, 2, 1, 1,
            1, 1, 1, 3, 1, 3, 1, 4, 1, 2, 5, 1, 3, 3, 2, 2, 4, 1, 1, 1, 3, 1, 1,
            2, 2, 5, 1, 1, 1, 1, 4, 1, 1, 2, 2, 3, 1, 1, 1, 5, 2, 4, 1, 1, 1, 2,
            2, 1, 3, 4, 2, 2, 5, 3, 3, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1,
            1, 1, 1, 5, 4, 1, 2, 2, 3, 2, 1, 3, 2, 1, 1, 5, 3, 3, 2, 2, 2, 1, 4,
            1, 1, 1, 1, 2, 1, 1, 1, 1, 1)

#lets create an arbitrary tier list of salaries present
#first convert to numeric
nhl_df$`FINAL CAP HIT`<- gsub("\\$", "", nhl_df$`FINAL CAP HIT`)
nhl_df$`FINAL CAP HIT`<- as.numeric(gsub(",", "", nhl_df$`FINAL CAP HIT`))



nhl_df$salary_tier <- ifelse(nhl_df$`FINAL CAP HIT` >= 84000000, "A",
                        ifelse(nhl_df$`FINAL CAP HIT` >= 80000000, "B",
                               ifelse(nhl_df$`FINAL CAP HIT` >= 76000000,"C",
                                      ifelse(nhl_df$`FINAL CAP HIT` >= 73000000,"D", "E"
nhl_df_finish <- cbind(nhl_df, finish)

#create a table to represent this
table1 <- table(nhl_df_finish$finish, nhl_df_finish$salary_tier)
```

Below is where the full ANOVA table values were computed.

```r
SST <- sum(12^2, 11^2, 18^2, 7^2, 15^2, 7^2, 11^2, 8^2, 3^2, 1^2, 1^2, 10^2,
           2^2, 1^2, 2^2, 2^2, 2^2, 2^2, 1^2, 1^2, 2^2, 4^2, 2^2) - (125^2/23)

#SS(alpha | mu)
SS_1 <- sum(63^2/5, 30^2/5, 16^2/5, 8^2/5, 8^2/3) - 125^2/23

#SS(beta| mu, alpha)
D_adot <- as.matrix(rbind(c(5,0,0,0,0), c(0,5,0,0,0), c(0,0,5,0,0),
                          c(0,0,0,5,0), c(0,0,0,0,3)), ncol = 5)
N <- as.matrix(rbind(c(1,1,1,1), c(1,1,1,1), c(1,1,1,1), c(1,1,1,1),
                     c(1,1,1,0)))
M <- solve(D_adot)%*%N

D_dotb <- as.matrix(rbind(c(5,0,0,0), c(0,5,0,0), c(0,0,5,0),
                          c(0,0,0,4)))
C <- D_dotb - t(N)%*%solve(D_adot)%*%N

Y_adot <- c(63, 30, 16, 8, 8)
Ydotb <- c(24, 38, 32, 12)

r <- Ydotb - t(M)%*%Y_adot
Beta_b <- solve(C)%*%r

SS_2 <- t(Beta_b)%*%r


#SS(gamma|mu, alpha, beta)
SS_3 <- sum(12^2, 11^2, 18^2, 7^2, 15^2, 7^2, 11^2, 8^2, 3^2, 1, 1, 10^2,
           2^2, 1^2, 2^2, 2^2, 2^2, 2^2, 1, 1, 2^2, 4^2, 2^2) -
```

```r
  sum(63^2/5, 30^2/5, 16^2/5, 8^2/5, 8^2/3) - SS_2


SSE <- 575.6522 - 379.7855 - 77.86667 - 118


#Find F statistics (very large could suggest overfitting)


F_1 <- 94.94638/1e-05

F_2 <- 19.46667/1e-05

F_3 <- 8.428571/1e-05
```