Hockey

24 July 2020

Salaries and Performance in the National Hockey League (NHL)

By looking at third league modeled on the North American system we can get a better understanding of the three variables we have used to explain win perecentage: salaries, lagged win percentage, and fixed effects.

We follow the same steps as we did for both those leagues.

```
# As usual, we begin by loading the packages we will need

options(warn = -1)
library("readxl",quietly = TRUE)
library("tidyverse",quietly = TRUE)
```

```
# Now we load the data

Hockey = read_excel("NHL pay and performance.xlsx")
```

```
Hockey %>% summary()
```

```
##      Team               season          wpc            salaries       
##  Length:301         Min.   :2009   Min.   :0.2561   Min.   :29727500  
##  Class :character   1st Qu.:2011   1st Qu.:0.4268   1st Qu.:56400000  
##  Mode  :character   Median :2014   Median :0.5000   Median :64093628  
##                     Mean   :2014   Mean   :0.5000   Mean   :62101877  
##                     3rd Qu.:2016   3rd Qu.:0.5610   3rd Qu.:69553691  
##                     Max.   :2018   Max.   :0.7500   Max.   :78205257
```

```
Hockey %>% str()
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    301 obs. of  4 variables:
##  $ Team    : chr  "Anaheim Mighty Ducks" "Atlanta Flames/Thrashers" "Boston Bruins" "
##  $ season  : num  2009 2009 2009 2009 2009 ...
##  $ wpc     : num  0.512 0.427 0.646 0.5 0.561 ...
##  $ salaries: num  49838000 34262500 51535000 52225700 63100000 ...
```

We can see that we have 301 observations in total covering the seasons 2009 to 2018. This is somewhat more than we has in the NBA case, but much less than the MLB case. We can now look at the changes in total salary spending across the seasons:

```r
Sumsal <- Hockey %>%
  group_by(season)%>%
   dplyr::summarise(salaries = sum(salaries))%>%rename(allsal = salaries)
Sumsal
```

```
## # A tibble: 10 x 2
##    season     allsal
##     <dbl>      <dbl>
##  1   2009 1541615281
##  2   2010 1552508107
##  3   2011 1625195685
##  4   2012 2083240145
##  5   2013 1875078749
##  6   2014 1870670657
##  7   2015 1936588513
##  8   2016 2010863335
##  9   2017 2044256803
## 10   2018 2152647770
```

Salary inflation has not been as dramatic in the NHL as in other leagues we have looked at, but they have still increased by more than one third in a decade, which is very unlikely to be caused by improving player quality on average. As with the other leagues, the main driver of increasing salaries has been increasing team revenues and the capacity of the players to bargain for higher wages.

As before, we use left_join() to add the aggregate salaries for each season to our original dataframe:

```r
Hockey <- left_join(Hockey, Sumsal, by="season")
head(Hockey)
```

```
## # A tibble: 6 x 5
##   Team                        season   wpc salaries     allsal
##   <chr>                        <dbl> <dbl>    <dbl>      <dbl>
## 1 Anaheim Mighty Ducks          2009 0.512 49838000 1541615281
## 2 Atlanta Flames/Thrashers      2009 0.427 34262500 1541615281
## 3 Boston Bruins                 2009 0.646 51535000 1541615281
## 4 Buffalo Sabres                2009 0.5   52225700 1541615281
## 5 Calgary Flames                2009 0.561 63100000 1541615281
## 6 Carolina Hurricanes           2009 0.549 49075000 1541615281
```

```r
tail(Hockey)
```

```
## # A tibble: 6 x 5
##   Team                                      season   wpc salaries     allsal
##   <chr>                                      <dbl> <dbl>    <dbl>      <dbl>
## 1 St. Louis Eagles/Blues                      2018 0.537 65500832 2152647770
```

```
## 2 Tampa Bay Lightning                        2018 0.659 73324166 2152647770
## 3 Toronto Arenas/St. Patricks/Maple Leafs    2018 0.598 63934167 2152647770
## 4 Vancouver Canucks                          2018 0.378 72819166 2152647770
## 5 Washington Capitals                        2018 0.598 74965962 2152647770
## 6 Winnipeg Jets                              2018 0.634 68507499 2152647770
```

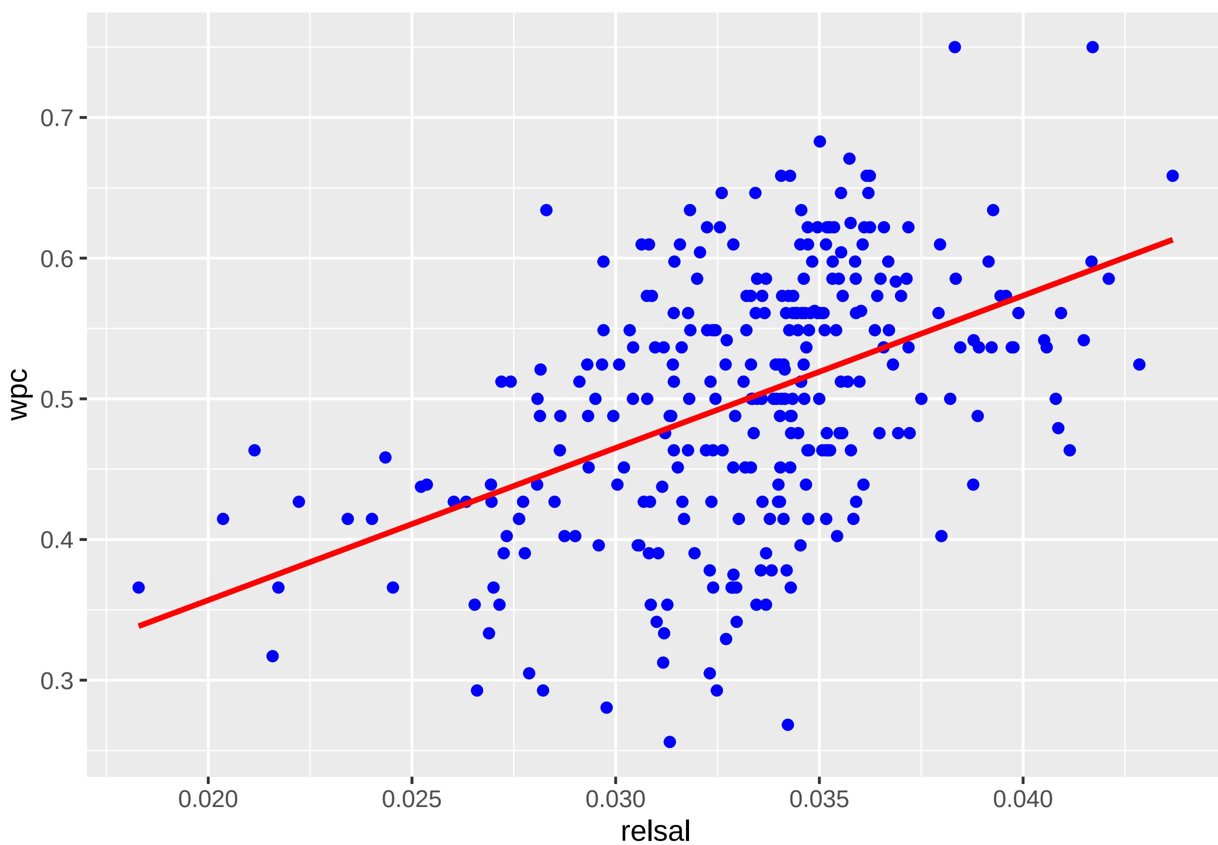
We can now create a variable which we call 'relsal', which measures the share of a team's salary spend in the total spending of all teams in that season:

```r
# Here we create the variable 'relsal' for the Hockey

Hockey[,'relsal']= Hockey[,'salaries']/Hockey[,'allsal']
```

Before running a regression, it makes sense to look at the relationship between salaries and win percentage on a chart. To do this we use ggplot(). Since our argument is that higher relative salaries mean better players which in turns leads to more wins, we put relsal on the x axis and wpc on the y axis.

```r
ggplot(data = Hockey,aes(x = relsal,y = wpc )) + geom_point(color='blue')+
  geom_smooth(method = "lm", se = FALSE,color = "red")
```



Self Test

Re-run the regplot with smaller dots so that there is no overlap.

```
wpcsal1_lm = lm(formula = 'wpc ~ relsal', data = Hockey)
wpcsal1_lm %>% summary()
```

```
##
## Call:
## lm(formula = "wpc ~ relsal", data = Hockey)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.242579 -0.053059  0.004707  0.058350  0.194760
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14039    0.03943   3.560 0.000431 ***
## relsal      10.82427    1.17868   9.183  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0808 on 299 degrees of freedom
## Multiple R-squared:   0.22,  Adjusted R-squared:  0.2174
## F-statistic: 84.33 on 1 and 299 DF,  p-value: < 2.2e-16
```

For the NHL we find that the coefficient of relsal is quite similar to that of the NBA. In fact, the regression looks quite similar in terms of the intercept (this is the value of win percentage if relsal were zero) and the R-squared.

Self Test

Create a subset of the data which includes only the 2018 season and run the regression of wpc on relsal. How do the results compare to the results above?

Let's see how the addition of the lagged dependent variable changes our relsal estimate.

```
Hockey <-  Hockey %>% arrange(Team,season)
head(Hockey)
```

```
## # A tibble: 6 x 6
##   Team                 season   wpc salaries     allsal relsal
##   <chr>                 <dbl> <dbl>    <dbl>      <dbl>  <dbl>
## 1 Anaheim Mighty Ducks   2009 0.512 49838000 1541615281 0.0323
## 2 Anaheim Mighty Ducks   2010 0.476 55207500 1552508107 0.0356
## 3 Anaheim Mighty Ducks   2011 0.573 53977500 1625195685 0.0332
## 4 Anaheim Mighty Ducks   2012 0.415 74655277 2083240145 0.0358
## 5 Anaheim Mighty Ducks   2013 0.625 67064166 1875078749 0.0358
```

```
## 6 Anaheim Mighty Ducks   2014 0.659 64127746 1870670657 0.0343
```

```r
tail(Hockey)
```

```
## # A tibble: 6 x 6
##   Team           season   wpc salaries     allsal relsal
##   <chr>           <dbl> <dbl>    <dbl>      <dbl>  <dbl>
## 1 Winnipeg Jets   2013 0.5   57705244 1875078749 0.0308
## 2 Winnipeg Jets   2014 0.451 63682588 1870670657 0.0340
## 3 Winnipeg Jets   2015 0.524 63322574 1936588513 0.0327
## 4 Winnipeg Jets   2016 0.427 52338400 2010863335 0.0260
## 5 Winnipeg Jets   2017 0.488 64113622 2044256803 0.0314
## 6 Winnipeg Jets   2018 0.634 68507499 2152647770 0.0318
```

```r
Hockey <- Hockey %>%
       group_by(Team)%>%
       mutate(wpc_lag = dplyr::lag(wpc))%>%
       ungroup()
head(Hockey)
```

```
## # A tibble: 6 x 7
##   Team                 season   wpc salaries     allsal relsal wpc_lag
##   <chr>                 <dbl> <dbl>    <dbl>      <dbl>  <dbl>   <dbl>
## 1 Anaheim Mighty Ducks   2009 0.512 49838000 1541615281 0.0323  NA
## 2 Anaheim Mighty Ducks   2010 0.476 55207500 1552508107 0.0356   0.512
## 3 Anaheim Mighty Ducks   2011 0.573 53977500 1625195685 0.0332   0.476
## 4 Anaheim Mighty Ducks   2012 0.415 74655277 2083240145 0.0358   0.573
## 5 Anaheim Mighty Ducks   2013 0.625 67064166 1875078749 0.0358   0.415
## 6 Anaheim Mighty Ducks   2014 0.659 64127746 1870670657 0.0343   0.625
```

```r
tail(Hockey)
```

```
## # A tibble: 6 x 7
##   Team           season   wpc salaries     allsal relsal wpc_lag
##   <chr>           <dbl> <dbl>    <dbl>      <dbl>  <dbl>   <dbl>
## 1 Winnipeg Jets   2013 0.5   57705244 1875078749 0.0308   0.451
## 2 Winnipeg Jets   2014 0.451 63682588 1870670657 0.0340   0.5
## 3 Winnipeg Jets   2015 0.524 63322574 1936588513 0.0327   0.451
## 4 Winnipeg Jets   2016 0.427 52338400 2010863335 0.0260   0.524
## 5 Winnipeg Jets   2017 0.488 64113622 2044256803 0.0314   0.427
## 6 Winnipeg Jets   2018 0.634 68507499 2152647770 0.0318   0.488
```

We now run our regression again, but adding wpc_lag into the regression equation:

```r
wpcsal2_lm = lm(formula = 'wpc ~ wpc_lag + relsal', data = Hockey)
wpcsal2_lm %>% summary()
```

```
##
```

```
## Call:
## lm(formula = "wpc ~ wpc_lag + relsal", data = Hockey)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.232626 -0.055605  0.006524  0.054086  0.225338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.07164    0.04327   1.656    0.099 .
## wpc_lag      0.29664    0.05699   5.205 3.88e-07 ***
## relsal       8.42007    1.35197   6.228 1.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07797 on 266 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.2854, Adjusted R-squared:  0.2801
## F-statistic: 53.13 on 2 and 266 DF,  p-value: < 2.2e-16
```

Once again adding the lagged dependent variable is justified both in terms of the statistical significance of the variable and the addition to R-squared (whether adjusted or not). However, the impact on our main variable of interest, relsal, is relatively small. Its value has fallen from 10.8 to 8.4, which while it does suggest that there was some omitted variable bias, it is not as great as in the NBA case, while in the MLB case the coefficient of relsal was much smaller to begin with.

Let's now see what changes if we include fixed effects:

```r
wpcsal3_lm <- lm(wpc ~ wpc_lag + relsal + factor(Team),
                    data = Hockey)

wpcsal3_lm %>% summary()
```

```
##
## Call:
## lm(formula = wpc ~ wpc_lag + relsal + factor(Team), data = Hockey)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.215744 -0.044546  0.001814  0.042355  0.205660
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                              0.252704   0.063501
## wpc_lag                                                  0.054225   0.064166
```

```
## relsal                                                 8.070876   1.532997
## factor(Team)Atlanta Flames/Thrashers                  -0.060812   0.059269
## factor(Team)Boston Bruins                             -0.013688   0.034750
## factor(Team)Buffalo Sabres                            -0.133202   0.035599
## factor(Team)Calgary Flames                            -0.082288   0.034958
## factor(Team)Carolina Hurricanes                       -0.088485   0.035998
## factor(Team)Chicago Blackhawks                         0.009418   0.034757
## factor(Team)Colorado Rockies/Avalanche                -0.070699   0.035609
## factor(Team)Columbus Blue Jackets                     -0.060764   0.035190
## factor(Team)Dallas Stars                              -0.039750   0.035056
## factor(Team)Detroit Cougars/Red Wings                 -0.063663   0.034717
## factor(Team)Edmonton Oilers                           -0.144130   0.036344
## factor(Team)Florida Panthers                          -0.085445   0.035976
## factor(Team)Los Angeles Kings                         -0.021707   0.034723
## factor(Team)Minnesota North Stars/Wild                -0.050658   0.034843
## factor(Team)Montreal Canadiens                        -0.057953   0.034757
## factor(Team)Nashville Predators                       -0.010692   0.034982
## factor(Team)New Jersey Devils                         -0.080442   0.035036
## factor(Team)New York Islanders                        -0.055420   0.035745
## factor(Team)New York Rangers                          -0.034675   0.034910
## factor(Team)Ottawa Eagles/Senators                    -0.070489   0.034964
## factor(Team)Philadelphia Flyers                       -0.072542   0.035053
## factor(Team)Phoenix Coyotes                           -0.068486   0.035986
## factor(Team)Pittsburgh Pirates/Penguins                0.026207   0.034921
## factor(Team)San Jose Sharks                           -0.003769   0.034665
## factor(Team)St. Louis Eagles/Blues                     0.028620   0.034878
## factor(Team)Tampa Bay Lightning                       -0.021701   0.034966
## factor(Team)Toronto Arenas/St. Patricks/Maple Leafs -0.085818   0.035595
## factor(Team)Vancouver Canucks                         -0.067311   0.034891
## factor(Team)Washington Capitals                        0.009975   0.034824
## factor(Team)Winnipeg Jets                             -0.025499   0.039277
##                                                       t value Pr(>|t|)
## (Intercept)                                             3.980 9.19e-05 ***
## wpc_lag                                                 0.845  0.39893
## relsal                                                  5.265 3.16e-07 ***
## factor(Team)Atlanta Flames/Thrashers                   -1.026  0.30592
## factor(Team)Boston Bruins                              -0.394  0.69401
## factor(Team)Buffalo Sabres                             -3.742  0.00023 ***
## factor(Team)Calgary Flames                             -2.354  0.01940 *
## factor(Team)Carolina Hurricanes                        -2.458  0.01469 *
## factor(Team)Chicago Blackhawks                          0.271  0.78665
## factor(Team)Colorado Rockies/Avalanche                 -1.985  0.04825 *
## factor(Team)Columbus Blue Jackets                      -1.727  0.08553 .
## factor(Team)Dallas Stars                               -1.134  0.25799
## factor(Team)Detroit Cougars/Red Wings                  -1.834  0.06795 .
```

```
## factor(Team)Edmonton Oilers                            -3.966 9.70e-05 ***
## factor(Team)Florida Panthers                           -2.375  0.01835 *
## factor(Team)Los Angeles Kings                          -0.625  0.53248
## factor(Team)Minnesota North Stars/Wild                 -1.454  0.14730
## factor(Team)Montreal Canadiens                         -1.667  0.09677 .
## factor(Team)Nashville Predators                        -0.306  0.76014
## factor(Team)New Jersey Devils                          -2.296  0.02256 *
## factor(Team)New York Islanders                         -1.550  0.12238
## factor(Team)New York Rangers                           -0.993  0.32160
## factor(Team)Ottawa Eagles/Senators                     -2.016  0.04493 *
## factor(Team)Philadelphia Flyers                        -2.069  0.03959 *
## factor(Team)Phoenix Coyotes                            -1.903  0.05824 .
## factor(Team)Pittsburgh Pirates/Penguins                 0.750  0.45372
## factor(Team)San Jose Sharks                            -0.109  0.91350
## factor(Team)St. Louis Eagles/Blues                      0.821  0.41271
## factor(Team)Tampa Bay Lightning                        -0.621  0.53544
## factor(Team)Toronto Arenas/St. Patricks/Maple Leafs   -2.411  0.01668 *
## factor(Team)Vancouver Canucks                          -1.929  0.05491 .
## factor(Team)Washington Capitals                         0.286  0.77480
## factor(Team)Winnipeg Jets                              -0.649  0.51684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0735 on 236 degrees of freedom
##   (32 observations deleted due to missingness)
## Multiple R-squared:  0.4366, Adjusted R-squared:  0.3602
## F-statistic: 5.715 on 32 and 236 DF,  p-value: 4.548e-16
```

Here we find ten fixed effects that are statistically significant. The fixed effects add considerably to the R-squared of the regression, and only marginally reduce the value of relsal. However, the most striking impact of the fixed effects is to reduce the value of of the lagged dependent variable to the point where it is statistically insignificant. This is in contrast to what we found in all three of the other other leagues. Because it is statistically insignificant in this version, and since we want to keep the fixed effects, we can drop the lagged dependent variable, which we do in this regression:

```r
wpcsal4_lm <- lm(wpc ~ relsal + factor(Team),
                    data = Hockey)

wpcsal4_lm %>% summary()
```

```
##
## Call:
## lm(formula = wpc ~ relsal + factor(Team), data = Hockey)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.207166 -0.045519  0.000437  0.043597  0.210607
##
## Coefficients:
##                                                       Estimate Std. Error
## (Intercept)                                          0.2562708  0.0522363
## relsal                                               8.7600186  1.3796018
## factor(Team)Atlanta Flames/Thrashers                -0.0472432  0.0497402
## factor(Team)Boston Bruins                           -0.0005806  0.0326583
## factor(Team)Buffalo Sabres                          -0.1277661  0.0326180
## factor(Team)Calgary Flames                          -0.0801552  0.0325899
## factor(Team)Carolina Hurricanes                     -0.0783579  0.0331344
## factor(Team)Chicago Blackhawks                       0.0137214  0.0326038
## factor(Team)Colorado Rockies/Avalanche              -0.0806411  0.0327971
## factor(Team)Columbus Blue Jackets                   -0.0601293  0.0326860
## factor(Team)Dallas Stars                            -0.0482120  0.0326736
## factor(Team)Detroit Cougars/Red Wings               -0.0517945  0.0325915
## factor(Team)Edmonton Oilers                         -0.1436206  0.0327036
## factor(Team)Florida Panthers                        -0.0826116  0.0329827
## factor(Team)Las Vegas Golden Knights                 0.0832319  0.0764354
## factor(Team)Los Angeles Kings                       -0.0232854  0.0325904
## factor(Team)Minnesota North Stars/Wild              -0.0527307  0.0325993
## factor(Team)Montreal Canadiens                      -0.0607271  0.0326336
## factor(Team)Nashville Predators                     -0.0096055  0.0328027
## factor(Team)New Jersey Devils                       -0.0680370  0.0325851
## factor(Team)New York Islanders                      -0.0624305  0.0333040
## factor(Team)New York Rangers                        -0.0407972  0.0329456
## factor(Team)Ottawa Eagles/Senators                  -0.0790807  0.0325838
## factor(Team)Philadelphia Flyers                     -0.0728543  0.0328197
## factor(Team)Phoenix Coyotes                         -0.0646767  0.0334408
## factor(Team)Pittsburgh Pirates/Penguins              0.0242823  0.0327659
## factor(Team)San Jose Sharks                          0.0070410  0.0325944
## factor(Team)St. Louis Eagles/Blues                   0.0289372  0.0327680
## factor(Team)Tampa Bay Lightning                     -0.0407336  0.0326338
## factor(Team)Toronto Arenas/St. Patricks/Maple Leafs -0.0913668  0.0326703
## factor(Team)Vancouver Canucks                       -0.0616002  0.0327150
## factor(Team)Washington Capitals                      0.0141051  0.0327453
## factor(Team)Winnipeg Jets                           -0.0345894  0.0360771
##                                                     t value Pr(>|t|)
## (Intercept)                                           4.906 1.62e-06 ***
## relsal                                                6.350 9.19e-10 ***
## factor(Team)Atlanta Flames/Thrashers                 -0.950 0.343070
## factor(Team)Boston Bruins                            -0.018 0.985830
## factor(Team)Buffalo Sabres                           -3.917 0.000114 ***
## factor(Team)Calgary Flames                           -2.460 0.014545 *
```

```
## factor(Team)Carolina Hurricanes                            -2.365 0.018750 *
## factor(Team)Chicago Blackhawks                              0.421 0.674199
## factor(Team)Colorado Rockies/Avalanche                     -2.459 0.014573 *
## factor(Team)Columbus Blue Jackets                          -1.840 0.066932 .
## factor(Team)Dallas Stars                                   -1.476 0.141235
## factor(Team)Detroit Cougars/Red Wings                      -1.589 0.113193
## factor(Team)Edmonton Oilers                                -4.392 1.62e-05 ***
## factor(Team)Florida Panthers                               -2.505 0.012850 *
## factor(Team)Las Vegas Golden Knights                        1.089 0.277168
## factor(Team)Los Angeles Kings                              -0.714 0.475549
## factor(Team)Minnesota North Stars/Wild                     -1.618 0.106938
## factor(Team)Montreal Canadiens                             -1.861 0.063857 .
## factor(Team)Nashville Predators                            -0.293 0.769881
## factor(Team)New Jersey Devils                              -2.088 0.037744 *
## factor(Team)New York Islanders                             -1.875 0.061940 .
## factor(Team)New York Rangers                               -1.238 0.216681
## factor(Team)Ottawa Eagles/Senators                         -2.427 0.015884 *
## factor(Team)Philadelphia Flyers                            -2.220 0.027267 *
## factor(Team)Phoenix Coyotes                                -1.934 0.054158 .
## factor(Team)Pittsburgh Pirates/Penguins                     0.741 0.459292
## factor(Team)San Jose Sharks                                 0.216 0.829136
## factor(Team)St. Louis Eagles/Blues                          0.883 0.377977
## factor(Team)Tampa Bay Lightning                            -1.248 0.213046
## factor(Team)Toronto Arenas/St. Patricks/Maple Leafs  -2.797 0.005538 **
## factor(Team)Vancouver Canucks                              -1.883 0.060792 .
## factor(Team)Washington Capitals                             0.431 0.666995
## factor(Team)Winnipeg Jets                                  -0.959 0.338543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07284 on 268 degrees of freedom
## Multiple R-squared:  0.4318, Adjusted R-squared:  0.364
## F-statistic: 6.366 on 32 and 268 DF,  p-value: < 2.2e-16
```

This model has a very simple interpretation: wpc = 0.256 + 8.76 x relsal + fixed effects. If we ignore fixed effects, we can identify the expected win percentage for low, average and high relative spending:

```r
print(0.256 + 8.76*0.02)
```

```
## [1] 0.4312
```

```r
print(0.256 + 8.76*0.0325)
```

```
## [1] 0.5407
```

```r
print(0.256 + 8.76*0.045)
```

```
## [1] 0.6502
```

Self Test

Based on the fixed effects regression, calculate the win percentage of:

(a) The Calgary Flames assuming the value of relsal for the team is 0.03
(b) The Edmonton Oilers assuming the value of relsal for the team is 0.04
(c) The Montreal Canadiens assuming the value of relsal for the team is 0.05

Looking at this, we can see that the numbers are slightly skewed- the performance levels are higher than one might expect at each level, and this most likely reflects the impacts of the fixed effects. It's clear that most of the fixed effects are negative, and this would bring down teams to a lower level of performance. It suggests that those teams that are able to dominate, are capable of doing so (or at least were capable of doing so during this era) because of factors other than wage spending.

Conclusion

This week we have looked at four different leagues and used salary data to assess the impact of wage spending on team performance. In every case we found it had a significant impact, but that impact varied depending on the league. The league system also mattered, as we saw when contrasting the cases of the EPL with the NBA, MLB and NHL.

We also introduced two issues which should always be considered when running regressions: omitted variable bias and heterogeneity.

Finally, we should mention one issue which arises in the context of this type of exercise. The data we study here is "observational", meaning that we collect the data based on what actually happened, during events over which we had no control. This raises the question, "How would outcomes have been different if some particular variable had had a different value?" The regression coefficients produced some answers for us, but how can we be sure that there was not some other factor which we have omitted, which was what really mattered? We can't be sure.

Scientists in laboratories typically don't have this problem - they use "experimental data" which they create in a controlled environment, so that they can control all observable factors. You can use that kind of data to measure the aerobic capacity of an athlete but, since you can't directly control the game, you can never use it to analyze game outcomes.

Some experimental scientists would go so far as to say that we can infer nothing from observational data. This is the logic of a phrase you may have encountered: "correlation is not causation". We have observed correlations in the data using regression analysis, but that does not prove that the links were causal (you could go so far as to argue that win percentage causes salaries to increase, and not vice versa).

We think that is too pessimistic a view. Observational data is certainly more challenging to work with, but it is possible for us to gain insight into the underlying relationships through careful study. It is important to be aware of the pitfalls, and it is important to focus on the logical coherence of the analysis rather than just running regressions. Another way to say this is that one should always have in mind a theory that one is trying to test, and be willing to discard that theory if the data renders it untenable. With careful thought and attention to details, it is possible to generate results which can enhance our understanding.