Predicting Games Played

Kyle Joecken

Question

Can we use data from previous seasons to predict how much time a particular NHL skater will get on the ice (TOI) in various situations (even strength, short-handed, power play) in the upcoming 2014-2015 season?

Data

As before, we load our database of season-wide NHL data into a data frame called `skaterstats`. We'll also grab our games played predictions and store it in a data frame called `skatpred15`.

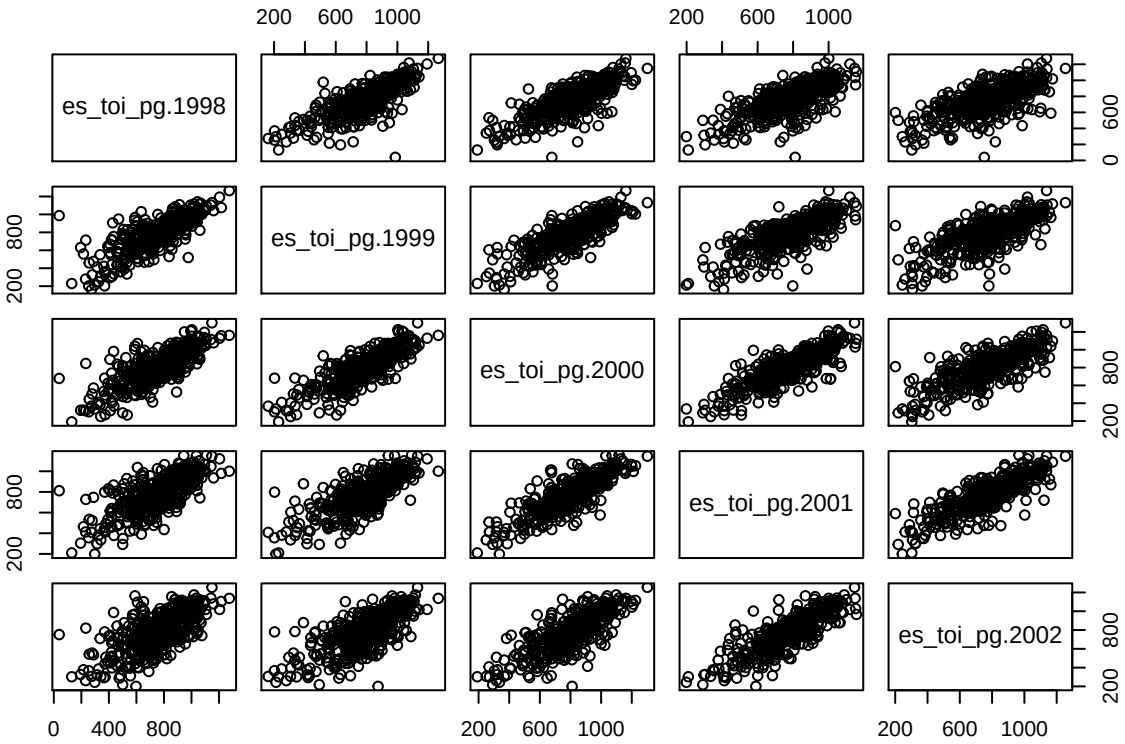
Features

We first shave off all but the useful variables, as we did before. We should break up the data into three separate chunks, one for each of the three situations.

```
skattoi <- skaterstats[, c(1, 2, 6, 45:47)]
skattoi$es_toi_pg <- skattoi$es_toi / skattoi$games_played
skattoi$sh_toi_pg <- skattoi$sh_toi / skattoi$games_played
skattoi$pp_toi_pg <- skattoi$pp_toi / skattoi$games_played
skattoipg <- skattoi[, c(1, 2, 7:9)]
estoipg <- reshape(skattoipg[, c(1:3)], timevar = "season",
                   idvar = "nhl_num", direction = "wide")
shtoipg <- reshape(skattoipg[, c(1, 2, 4)], timevar = "season",
                   idvar = "nhl_num", direction = "wide")
pptoipg <- reshape(skattoipg[, c(1, 2, 5)], timevar = "season",
                   idvar = "nhl_num", direction = "wide")
```

Next, let's have a look at all players that played in all of the first 5 years of our data and use `pairs(` to look at correlation between years visually.

```
estoi9802 <- estoipg[!is.na(estoipg$es_toi_pg.1998) &
                             !is.na(estoipg$es_toi_pg.1999) &
                             !is.na(estoipg$es_toi_pg.2000) &
                             !is.na(estoipg$es_toi_pg.2001) &
                             !is.na(estoipg$es_toi_pg.2002), 2:6]
pairs(estoi9802)
```



That's pretty. Next, we fit linear models going further and further back, then run `ANOVA()` to see how far back we can go while still feasibly gaining explained variance.

```
fit1 <- lm(es_toi_pg.2002 ~ es_toi_pg.2001, data = estoi9802)
fit2 <- lm(es_toi_pg.2002 ~ es_toi_pg.2001 + es_toi_pg.2000, data = estoi9802)
fit3 <- lm(es_toi_pg.2002 ~ es_toi_pg.2001 + es_toi_pg.2000 + es_toi_pg.1999,
           data = estoi9802)
fit4 <- lm(es_toi_pg.2002 ~ es_toi_pg.2001 + es_toi_pg.2000 + es_toi_pg.1999
            + es_toi_pg.1998, data = estoi9802)
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: es_toi_pg.2002 ~ es_toi_pg.2001
## Model 2: es_toi_pg.2002 ~ es_toi_pg.2001 + es_toi_pg.2000
## Model 3: es_toi_pg.2002 ~ es_toi_pg.2001 + es_toi_pg.2000 + es_toi_pg.1999
## Model 4: es_toi_pg.2002 ~ es_toi_pg.2001 + es_toi_pg.2000 + es_toi_pg.1999 +
##     es_toi_pg.1998
##   Res.Df     RSS Df Sum of Sq     F  Pr(>F)
## 1    433 5298679
## 2    432 4961749  1    336930 29.50 9.4e-08 ***
## 3    431 4921038  1     40712  3.56    0.06 .
## 4    430 4910537  1     10501  0.92    0.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As in the games played analysis, we see that little to nothing is gained by adding the third and fourth seasons previous as regressors, and that the coefficients aren't statistically significant anyway. As before, we'll simply use the previous two seasons as regressors.

The analysis for power play and short-handed time is similar; for power play TOI, there is statistically nothing gained by considering the season three years previous. For short-handed TOI, the season three years previous is statistically significant but not practically significant; for simplicity and legibility, we stay consistent by using the previous two seasons to predict the next. The code is included in the `.Rmd` file but hidden for brevity.

Algorithm

As we have decided upon the same basic model as we did for games played, we will proceed in much the same manner for TOI.

```
##   baseline season n-1 season n-2
##    74.8824     0.6954     0.2081
```

We repeat the above analysis (but continue to hide the code for brevity) for power play and short-handed TOI.

```
##   baseline season n-1 season n-2
##     9.1610     0.6886     0.1882
```

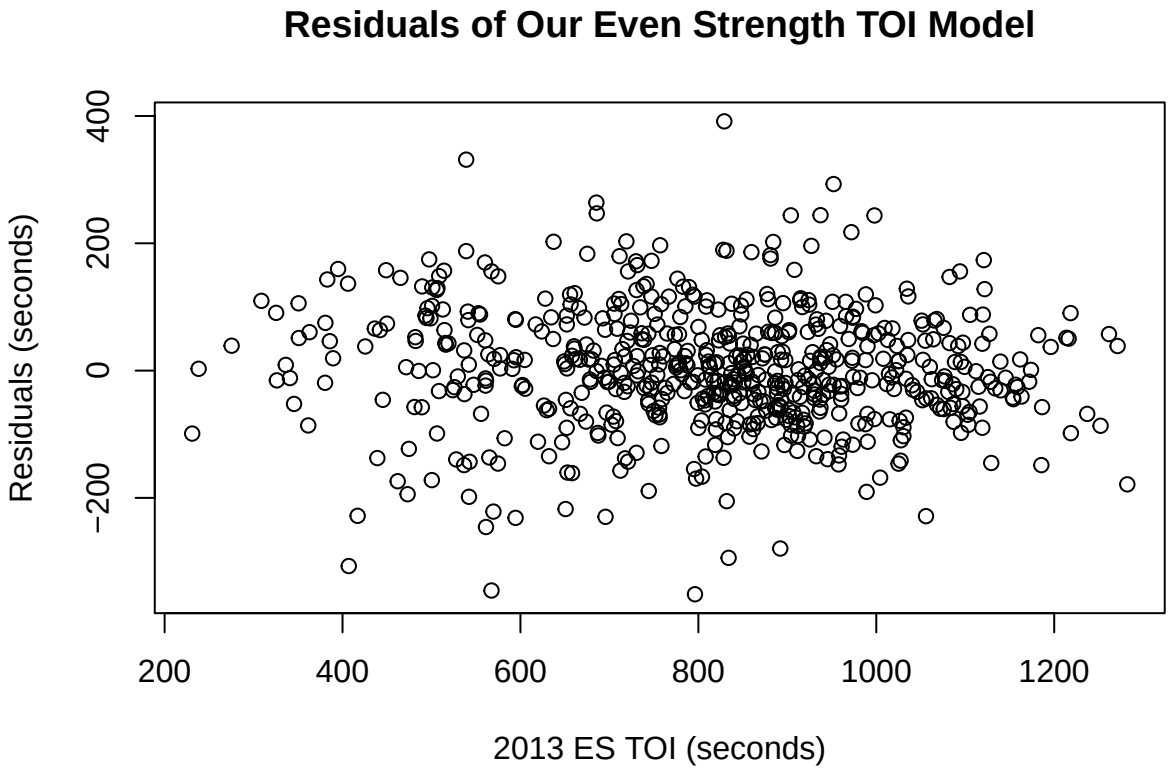
```
##   baseline season n-1 season n-2
##    18.0077     0.6383     0.1628
```

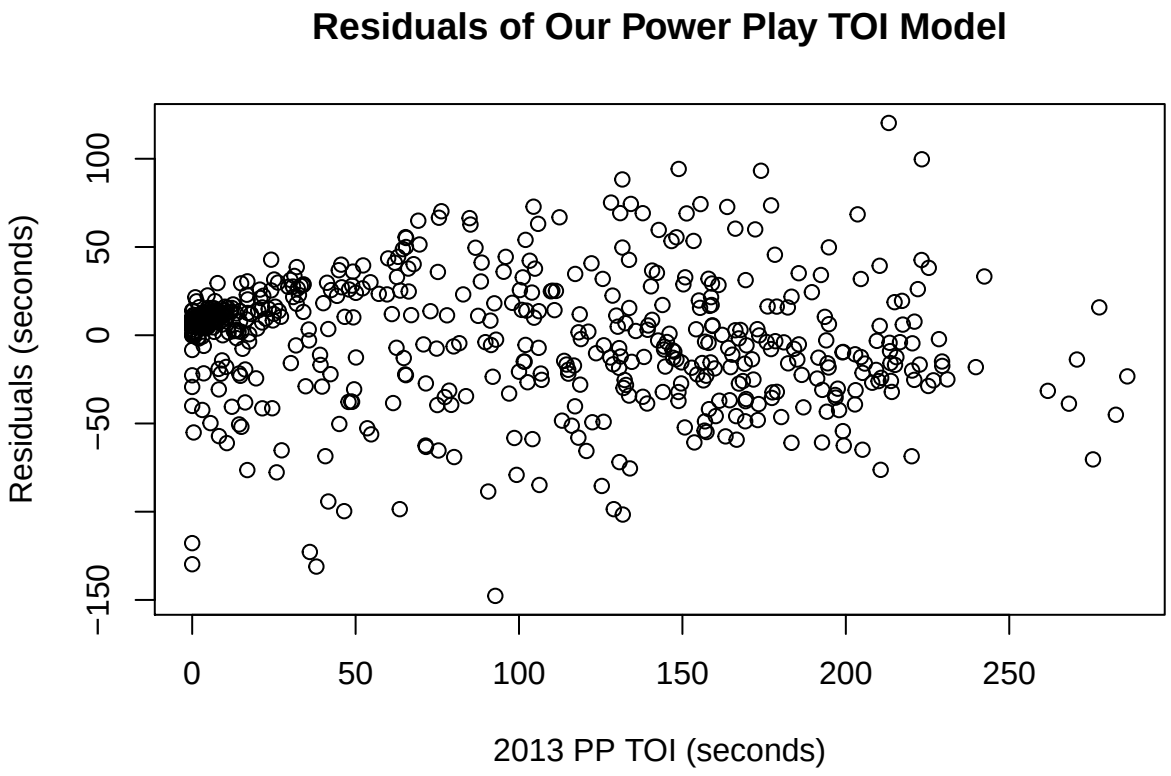
We let x, y and z represent situational TOI in seasons n, $n-1$ and $n-2$, respectively. Choosing some legible coefficients, we now have the following models:

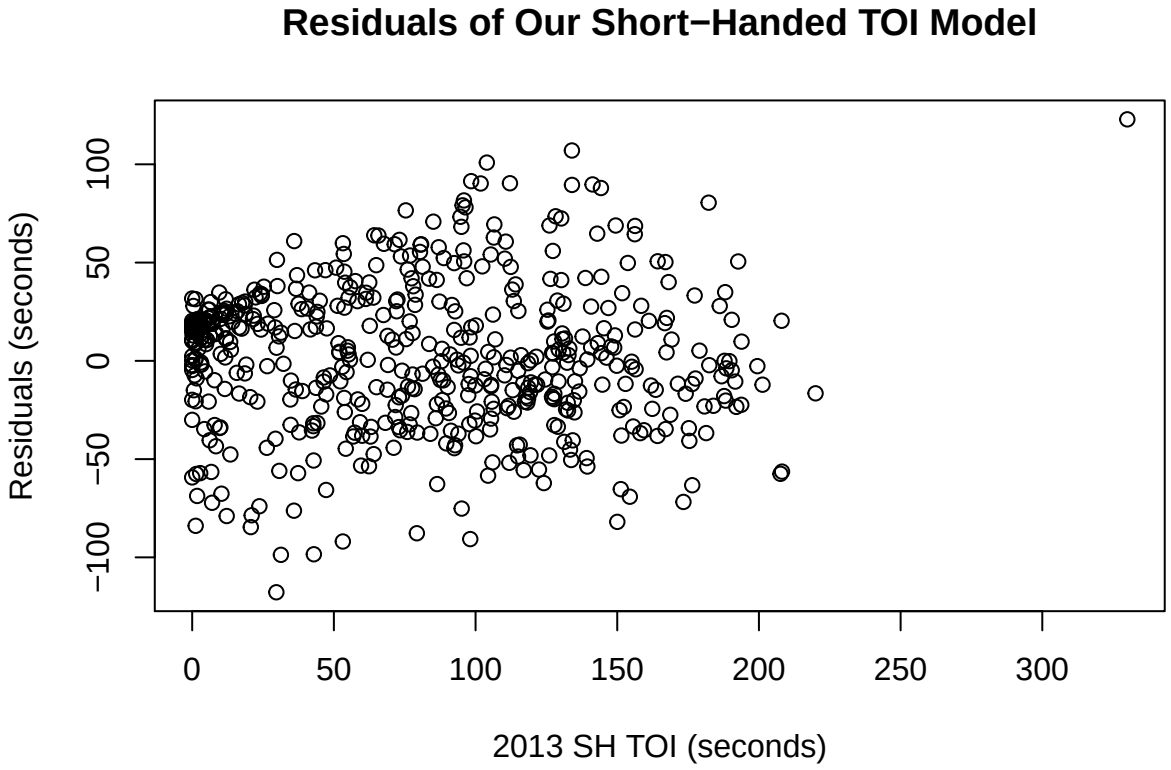
$$x_{es} \sim 75 + \frac{7}{10} y_{es} + \frac{1}{5} z_{es}$$
$$x_{pp} \sim 9 + \frac{13}{20} y_{pp} + \frac{1}{5} z_{pp}$$
$$x_{sh} \sim 18 + \frac{13}{20} y_{sh} + \frac{3}{20} z_{sh}$$

Evaluation

Let's try to use these models to predict TOI for 2014.


Residuals of Our Even Strength TOI Model
Residuals (seconds)
400
200
0
−200
200
400
600
800
1000
1200
2013 ES TOI (seconds)


Residuals of Our Power Play TOI Model
Residuals (seconds)
100
50
0
−50
−150
0
50
100
150
200
250
2013 PP TOI (seconds)



Looks like Darren Helm himself one crazy game in 2013.