HW 7

Solutions

Introduction

In this assignment, we'll return to our hockey data set. Specifically, we'll use the cleaned file which contains variables regarding how each player did in the season *following* the one listed in each row. Here's that code combined into a newer frame, `nhl.data1`.

```
library(RCurl)
library(mosaic)
url <- getURL("https://raw.githubusercontent.com/statsbylopez/StatsSports/master/NHL.csv")
nhl.data <- read.csv(text = url)
nhl.data <- filter(nhl.data, TOI > 500, Position!="D", Position!="DL", Position!="DR")
nrow(nhl.data)
```

```
## [1] 3006
```

```
nhl.data <- na.omit(nhl.data)
nhl.data <- mutate(nhl.data, Shots_Sixty = Shots/TOI*60)
nhl.data1 <- nhl.data %>%
  arrange(Name, Season) %>%
  group_by(Name) %>%
  mutate(f.Goals = lead(Goals),
         f.Assists = lead(Goals),
         f.PDO = lead(PDO),
         f.CFRel_Percent = lead(CFRel_Percent),
         f.CF_Percent = lead(CF_Percent),
         Sh_Percent = Goals/Shots,
         f.Sh_Percent = lead(Sh_Percent),
         f.PDO = lead(PDO))
head(nhl.data1)
```

```
## Source: local data frame [6 x 27]
## Groups: Name [4]
##
##             Name Position    Team Games   Season   Age Salary Goals
##           (fctr)   (fctr)  (fctr) (int)    (int) (int)  (dbl) (int)
## 1    Aaron.Voros        L     NYR    58 20082009    27  1.200     5
## 2    Adam.Burish        C     CHI    83 20082009    25  0.713     9
## 3    Adam.Burish        C     DAL    63 20102011    27  1.000     7
## 4    Adam.Burish        C     DAL    65 20112012    28  1.300     6
## 5 Adam.Cracknell        C EDM/VAN    50 20152016    30  0.575     5
## 6      Adam.Hall       RC     T.B    74 20082009    28  0.600     4
## Variables not shown: Assists (int), Goals_Sixty (dbl), Assists_Sixty
##   (dbl), CF_Percent (dbl), PDO (dbl), CFRel_Percent (dbl), Corsi (int),
##   CorsiFor (int), CorsiAgainst (int), Shots (int), TOI (dbl), Shots_Sixty
##   (dbl), f.Goals (int), f.Assists (int), f.PDO (dbl), f.CFRel_Percent
##   (dbl), f.CF_Percent (dbl), Sh_Percent (dbl), f.Sh_Percent (dbl)
```

The data set `nhl.data1` contains the the players' production in the season following (abbreviated with an `f.`).

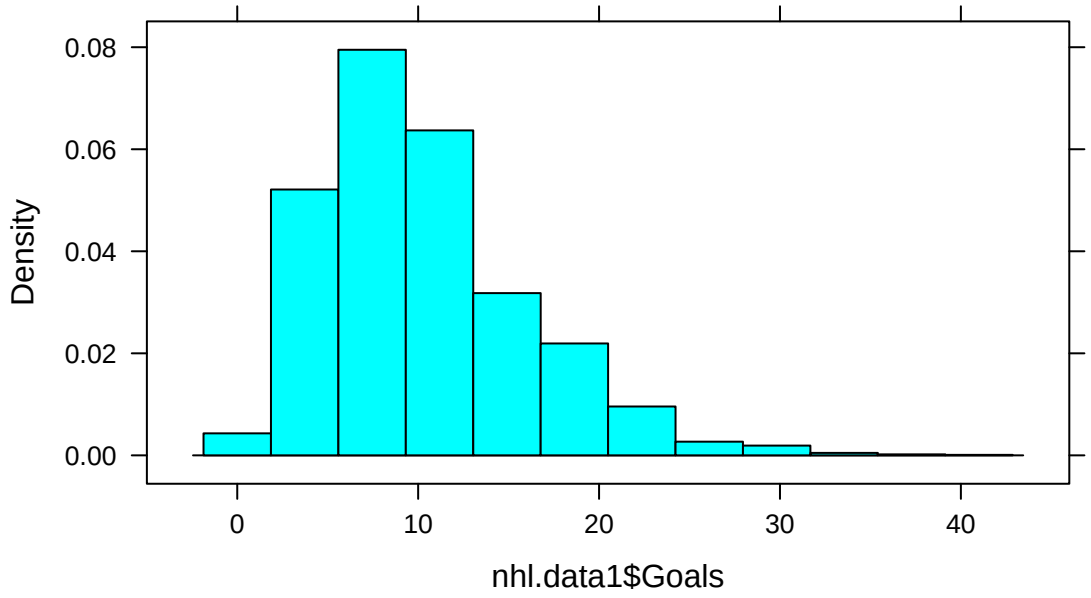
Links to future goals scored

1. A hockey coach is thinking about predicting a player's future goals using a regression model. Use evidence from our most recent lab to convince him it may be preferred to predict a players future relative Corsi percentage (`f.CFRel_Percent`).

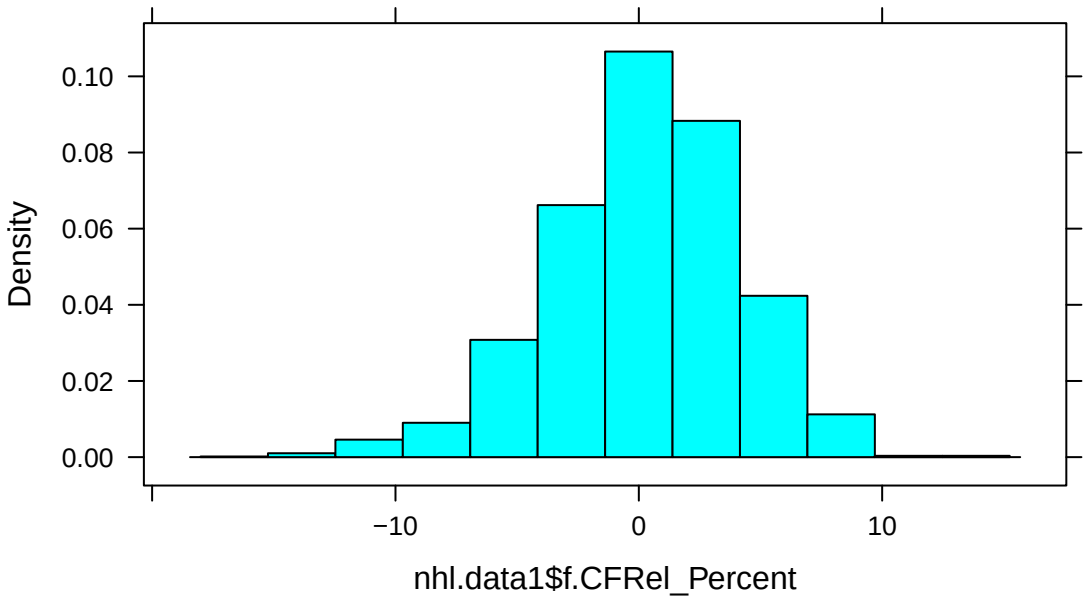
Answer: There are two reasons to possibly prefer relative Corsi over goals alone. First Corsi is more repeatable from one year to the next, suggesting our predictions would be more accurate. Second, Corsi accounts for things a player does on the ice apart from his goals.

2. There's an additional benefit to using `f.CFRel_Percent` as an outcome, (instead of `f.Goals`). Using univariate statistics (histograms, etc), identify why this is.

```
histogram(nhl.data1$Goals)
```




```
histogram(nhl.data1$f.CFRel_Percent)
```



Goal scoring outcomes are strongly skewed right, which may make it troublesome to use as the outcome in a typical regression model.

3. Using the AIC criterion and the explanatory variables `Goals`, `Assists`, `CF_Percent`, `PDO`, `CFRel_Percent`, `Shots`, `Salary`, and `Age`, derive which linear regression fit is optimal with the `f.CFRel_Percent` outcome. Reminder: it is generally not a good idea to use variables that are strongly correlated in the same regression model.

```
fit <- lm(f.CFRel_Percent ~ CFRel_Percent + Goals + Assists + Age, data = nhl.data1)
AIC(fit)
```

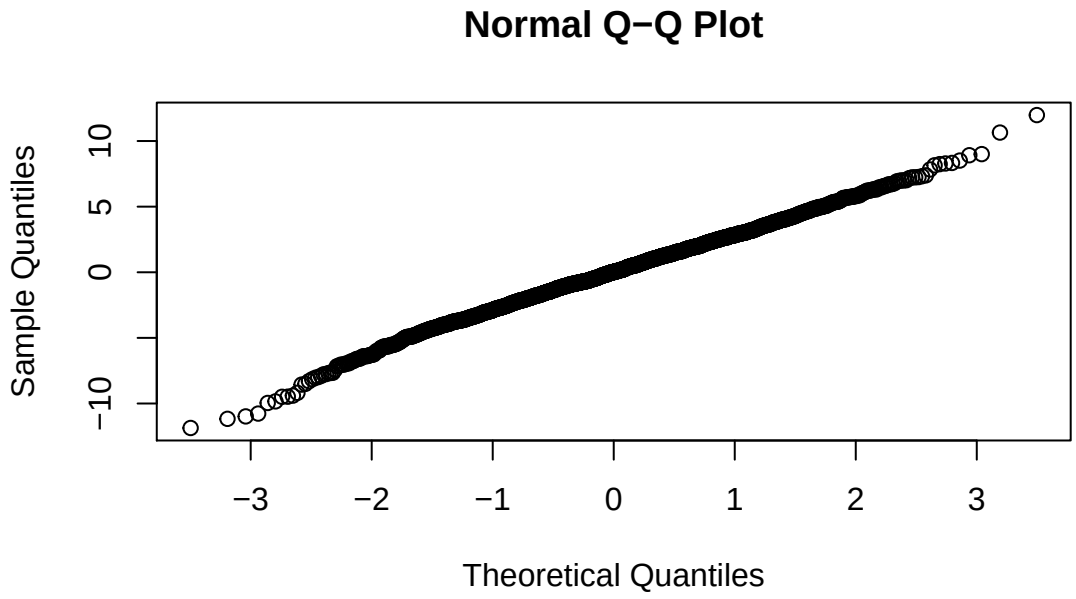
```
## [1] 10628.2
```

4. Which variables more relevant to a player's performance in a given season? Do any of the associations surprise you?

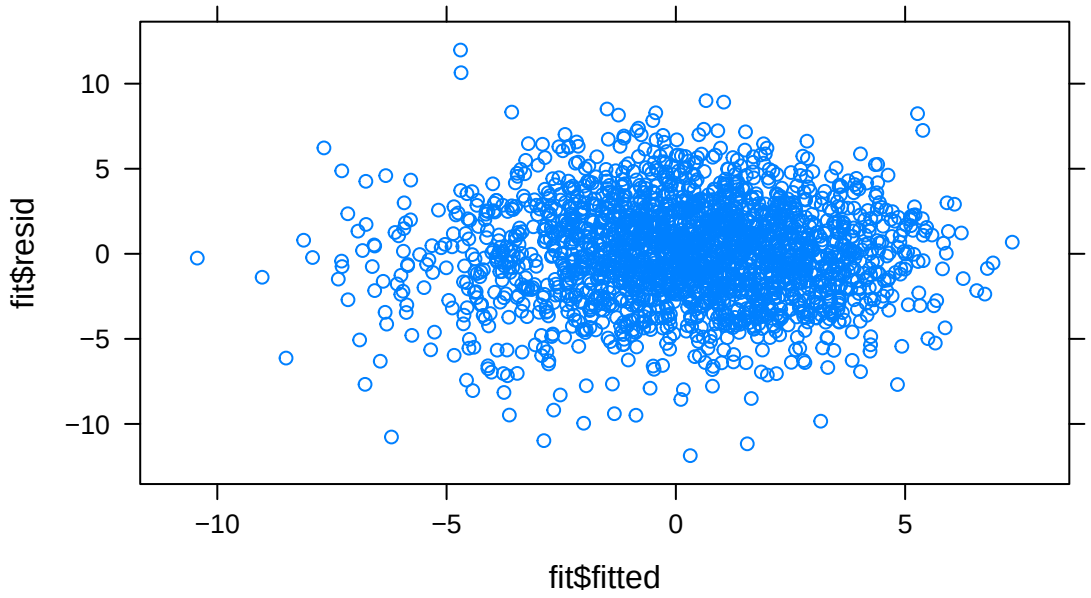
Answers will vary: `PDO` tends to not be relevand, while `Age` surprisingly does. As a player gets older, given his metrics from a season before, `f.CFRel_Percent` performance tends to drop. It's also interesting that assists seem more strongly linked to `f.CFRel_Percent` than goals, at least after accounting for `CFRel_Percent` from the prior season.

5. Check the assumptions for the linear regression model that you chose.

```
qqnorm(fit$resid)
```




```
xyplot(fit$resid~fit$fitted)
```



There's a bit of skewness in the residuals (moreso than the normal distribution), but by and large, the assumptions are reasonable. There appears to be an independence between the residuals and fitted values.

Project

6. Write a three paragraph description of your project.