DATA 605 Discussion Week 12

CUNY Spring 2021

Philip Tanofsky

17 April 2021

Prompt

Using R, build a regression model for data that interests you. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Regression Model

I chose to use some NBA team data easily accessed using the *NBAloveR* package. The goal is to determine the relationship between the Offensive Rating, Defensive Rating, and Rating Difference as compared to the team's win total in a given season. Rating Difference is a simple derived variable.

- Offensive Rating: From basketball-reference.com, for teams it is points scored per 100 possessions.
- Defensive Rating: From basketball-reference.com, for teams it is points allowed per 100 possessions.
- Rating Difference: Offensive Rating minus Defensive Rating. The larger the value, the better for a team.

Load Libraries

```
library(tidyverse)
library(ggplot2)
library(NBAloveR)
```

Load Data

As the package only allows to pull data from a single franchise, I selected five just to keep the code readable. I selected the Boston Celtics, Los Angeles Lakers, Chicago Bulls, San Antonio Spurs, and Philadelphia 76ers.

```
bos <- getTeamHistory(team_code = "bos")
lal <- getTeamHistory(team_code = "lal")
chi <- getTeamHistory(team_code = "chi")
sas <- getTeamHistory(team_code = "sas")
phi <- getTeamHistory(team_code = "phi")
```

```
nba <- rbind(bos, lal)
nba <- rbind(nba, chi)
nba <- rbind(nba, sas)
nba <- rbind(nba, phi)
```

Data Derivation

Derive the attribute Rating Difference.

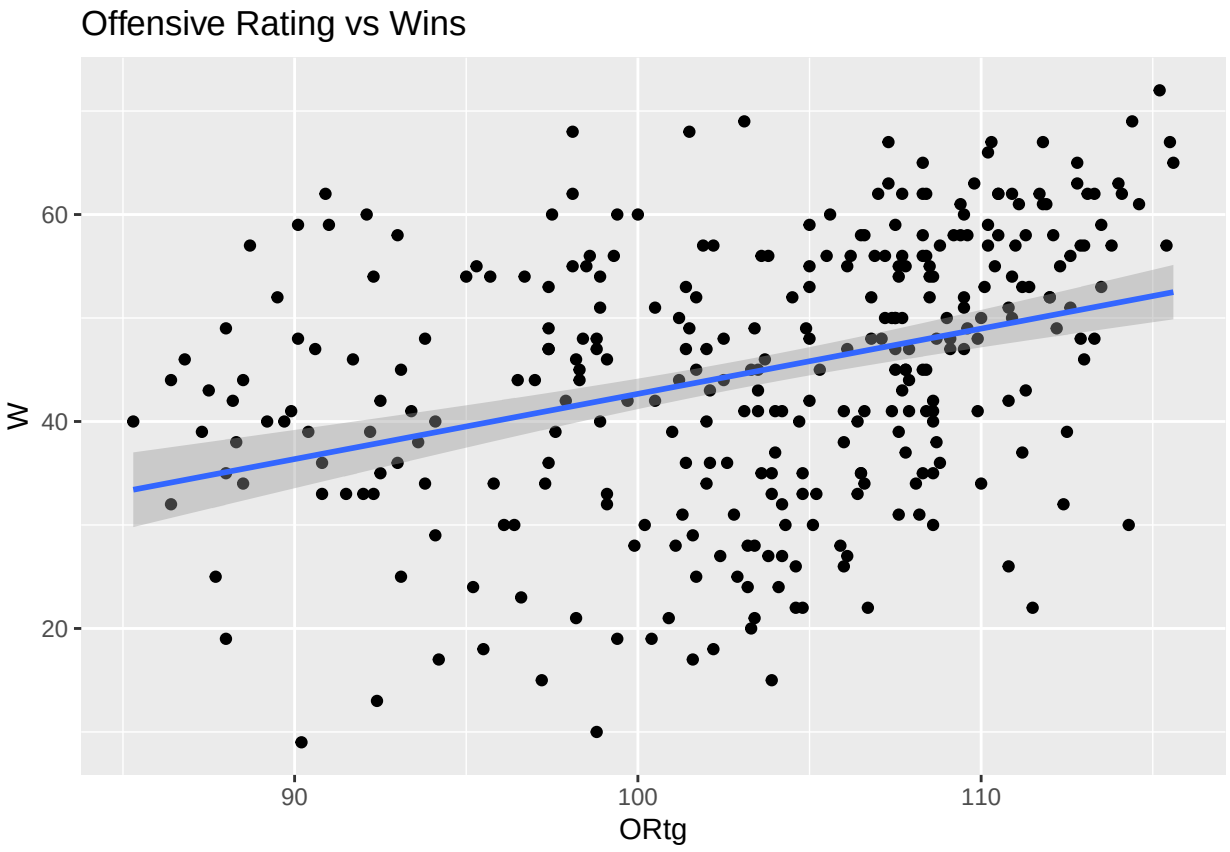
```
nba$RtgDif <- nba$ORtg - nba$DRtg

head(nba)
```

```
##    Season  Lg           Team  W  L  W/L%   Finish  SRS Pace RelativePace  ORtg
## 1 2020-21 NBA Boston Celtics 30 26 0.536 3rd of 5 2.40 97.9         -1.3 114.3
## 2 2019-20 NBA Boston Celtics 48 24 0.667 2nd of 5 5.83 99.5         -0.8 113.3
## 3 2018-19 NBA Boston Celtics 49 33 0.598 3rd of 5 3.90 99.6         -0.4 112.2
## 4 2017-18 NBA Boston Celtics 55 27 0.671 2nd of 5 3.23 96.0         -1.3 107.6
## 5 2016-17 NBA Boston Celtics 53 29 0.646 1st of 5 2.25 96.8          0.4 111.2
## 6 2015-16 NBA Boston Celtics 48 34 0.585 2nd of 5 2.84 98.5          2.7 106.8
##   RelativeORtg  DRtg RelativeDRtg               Playoffs            Coaches
## 1          2.2 112.2          0.1                        B. Stevens (30-26)
## 2          2.7 107.0         -3.6   Lost E. Conf. Finals B. Stevens (48-24)
## 3          1.8 107.8         -2.6    Lost E. Conf. Semis B. Stevens (49-33)
## 4         -1.0 103.9         -4.7   Lost E. Conf. Finals B. Stevens (55-27)
## 5          2.4 108.4         -0.4   Lost E. Conf. Finals B. Stevens (53-29)
## 6          0.4 103.6         -2.8 Lost E. Conf. 1st Rnd. B. Stevens (48-34)
##        TopWinShare RtgDif
## 1   J. Tatum (4.9)    2.1
## 2   J. Tatum (6.9)    6.3
## 3  K. Irving (9.1)    4.4
## 4  K. Irving (8.9)    3.7
## 5 I. Thomas (12.5)    2.8
## 6  I. Thomas (9.7)    3.2
```

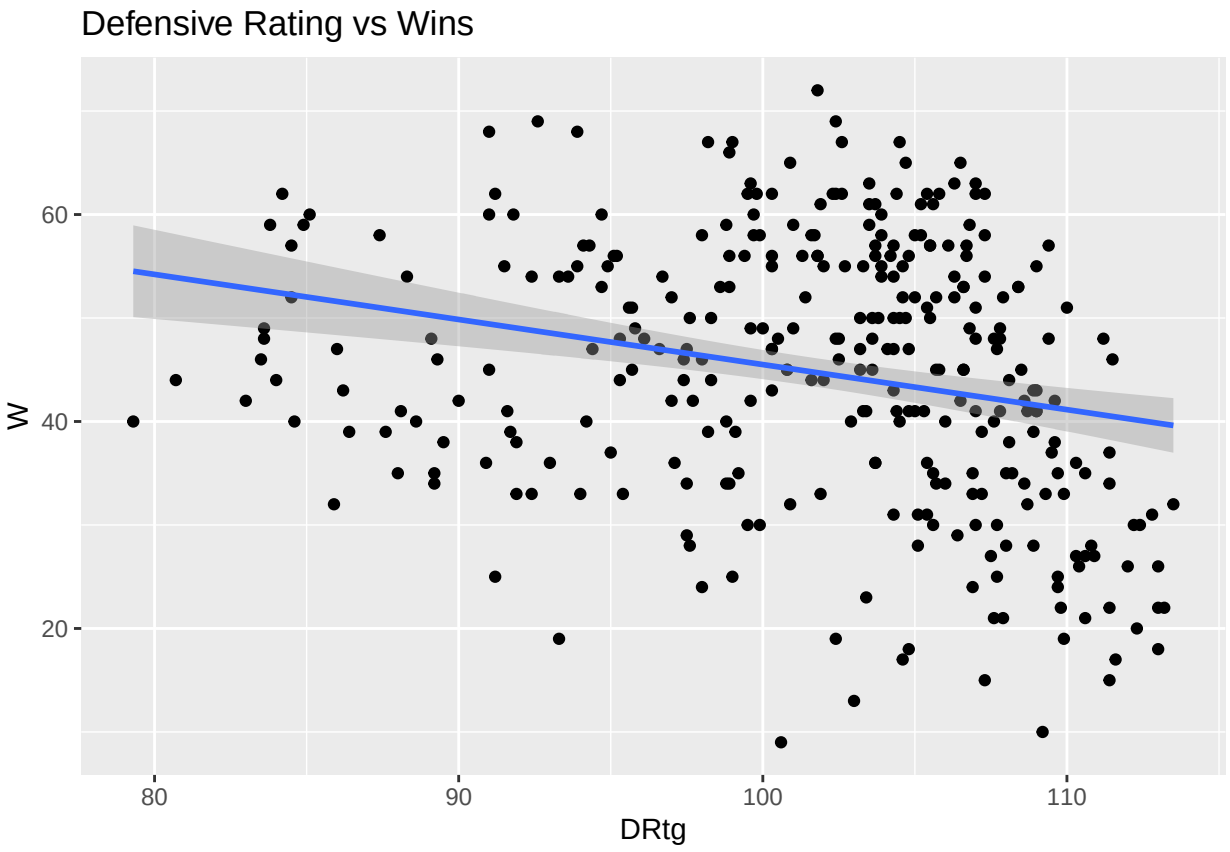
EDA Plots

Plot each rating attribute against team wins.

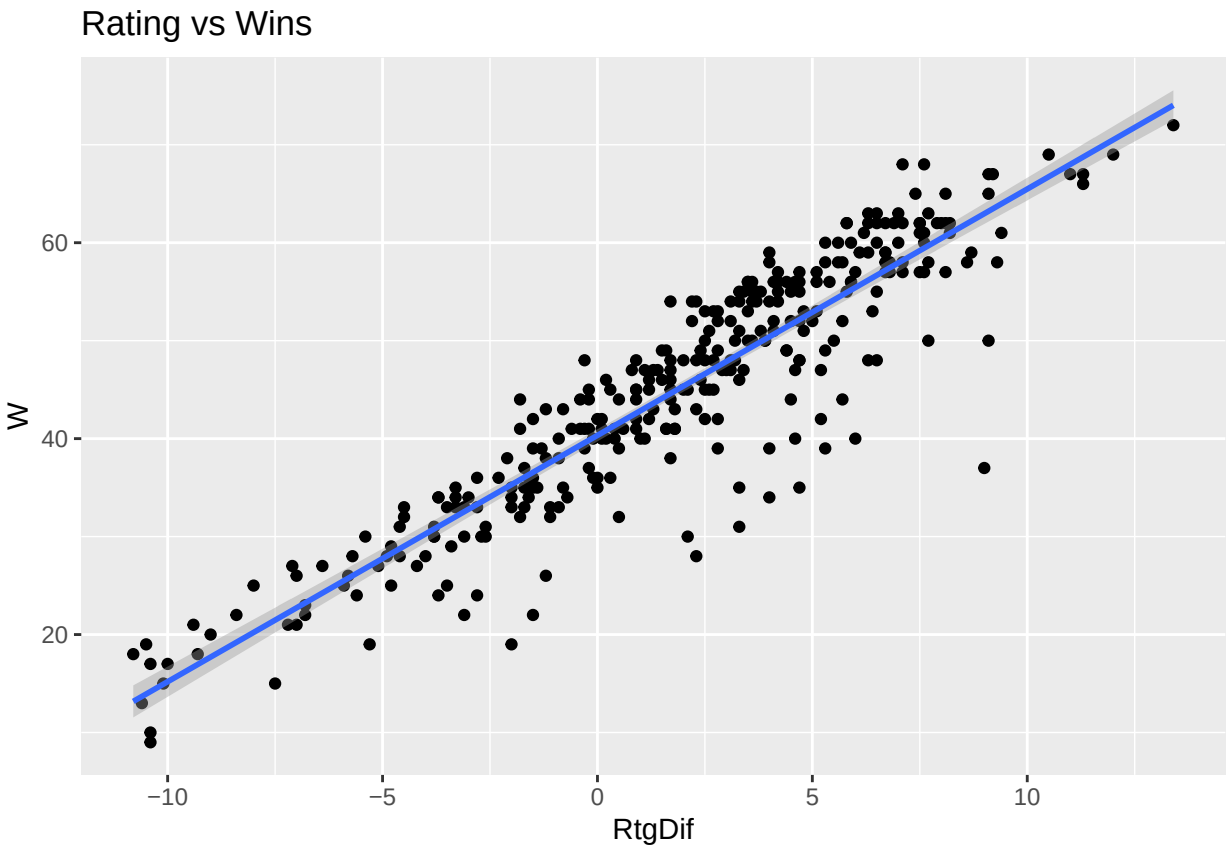
```
nba %>%
  ggplot(aes(x=ORtg, y=W)) +
  geom_point() +
  labs(title = 'Offensive Rating vs Wins') + geom_smooth(method='lm', formula= y~x)
```




```
nba %>%
  ggplot(aes(x=DRtg, y=W)) +
  geom_point() +
  labs(title = 'Defensive Rating vs Wins') + geom_smooth(method='lm', formula= y~x)
```




```
nba %>%
  ggplot(aes(x=RtgDif, y=W)) +
  geom_point() +
  labs(title = 'Rating vs Wins') + geom_smooth(method='lm', formula= y~x)
```



For Offensive Rating, a slight positive linear trend is visible, yet the scatterplot indicates wide variability in the relationship. Similar for Defensive Rating, except the linear relationship is negative. As for Rating Difference, the scatterplot shows a strong positive relationship between the rating difference and the team wins.

```
corr <- cor(nba$RtgDif, nba$W, use = "complete.obs")
corr
```

```
## [1] 0.9151873
```

The correlation between Rating Difference and team wins is above 90%, indicating a high correlation between the two variables.

```
model_ortg <- lm(W ~ ORtg, data=nba)
summary(model_ortg)
```

```
##
## Call:
## lm(formula = W ~ ORtg, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -31.914  -8.165   1.713   9.013  26.528
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -20.34924    9.76658  -2.084    0.038 *
## ORtg          0.63019    0.09424   6.687 1.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.14 on 320 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.1226, Adjusted R-squared:  0.1199
## F-statistic: 44.71 on 1 and 320 DF,  p-value: 1.017e-10
```

```
model_drtg <- lm(W ~ DRtg, data=nba)
summary(model_drtg)
```

```
##
## Call:
## lm(formula = W ~ DRtg, data = nba)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -36.239  -9.704   0.741  10.228  27.284
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 89.06285    9.78860   9.099  < 2e-16 ***
## DRtg        -0.43562    0.09609  -4.533 8.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.56 on 320 degrees of freedom
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.06035,    Adjusted R-squared:  0.05741
## F-statistic: 20.55 on 1 and 320 DF,  p-value: 8.214e-06
```

```
model_rtgdif <- lm(W ~ RtgDif, data=nba)
summary(model_rtgdif)
```

```
##
## Call:
## lm(formula = W ~ RtgDif, data = nba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.9646  -2.2960   0.7784   3.3895   9.8133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.33265    0.31117  129.62   <2e-16 ***
## RtgDif       2.51466    0.06191   40.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.223 on 320 degrees of freedom
```

```
##   (7 observations deleted due to missingness)
## Multiple R-squared:  0.8376, Adjusted R-squared:  0.8371
## F-statistic:  1650 on 1 and 320 DF,  p-value: < 2.2e-16
```

After generating the linear regression models for each rating variable against the team wins, a clear relationship is identified between the rating difference and the team wins with an R-squared value of greater than 83%. I do believe for the derived rating difference value, the linear model is appropriate. Alone, the Offensive Rating, nor the Defensive Rating, indicates whether a team is good or not. A high-scoring team may also be very bad at defensive. By taking the difference, teams with a good offensive and a good defensive would expected to be considered "good", and thus win more games. Also, by taking the difference, the pace of play will not impact the model quite as much. A high scoring team in the 1990s may not be the same as a high-scoring team in the 2010s. By taking the difference, the style and pace of games for a certain era is largely masked.