Blog 1: Metrics of Simple Linear Regression in R

Michael O'Donnell

September 27, 2020

The purpose of this blog is simply to define the metrics given to you in R from a Simple Linear Regression Model

To look at a linear model in R, let's use help

```
help(lm)
```

```
## starting httpd help server ... done
```

Now, start by loading a dataset This dataset contains regular season data for all NBA teams from 2014-2018

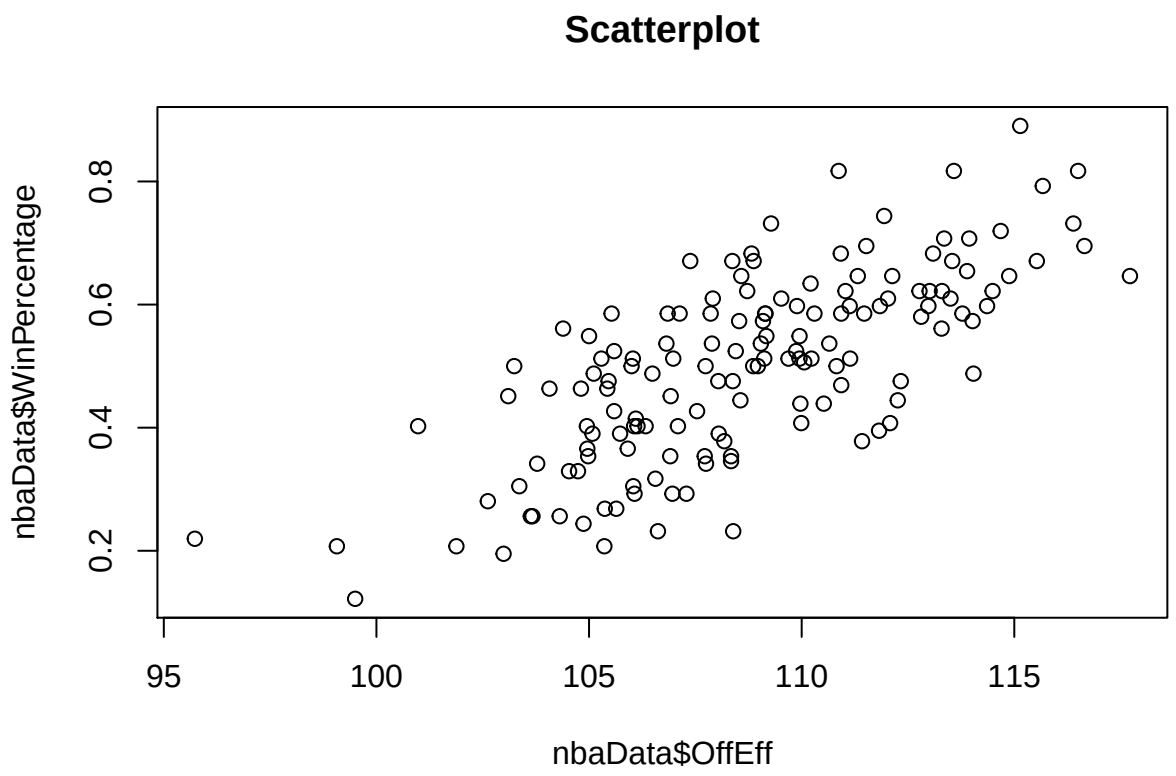
```
nbaData <- read.csv("data/nba_data.csv")
colnames(nbaData)[1] <- "Team"

head(nbaData, 3)
```

```
##             Team Season SeasonType Win Loss MatchCount WinPercentage
## 1  Atlanta Hawks   2018        REG  28   53         81     0.3456790
## 2 Boston Celtics   2018        REG  49   33         82     0.5975610
## 3  Brooklyn Nets   2018        REG  42   40         82     0.5121951
##      Pts OppPts   Pace OffEff DefEff EFgPercentage OppEFgPercentage
## 1 112.93 119.21 103.46 108.34 114.73         0.521            0.541
## 2 112.39 107.95  98.97 112.98 108.22         0.534            0.514
## 3 112.24 112.32 100.30 110.23 110.23         0.520            0.512
##   TsPercentage OppTsPercentage RebRate EffPts OppEffPts FastBreakPts
## 1        0.555           0.580   50.07 125.25    138.43        15.26
## 2        0.567           0.550   49.25 132.42    119.59        16.24
## 3        0.556           0.548   50.18 122.98    127.00        11.62
##   OppFBPts PointsInPaint OppPointsInPaint PointsOffTO OppPointsOffTO
## 1    16.51         51.19            49.36       21.14          16.88
## 2    13.17         44.78            45.93       14.82          18.12
## 3    11.83         48.76            51.20       17.35          15.38
##   SecondChancePTS OppSecondChancePTS PersonalFoulsPG OppPersonalFoulsPG
## 1           14.11              14.51          23.519             22.124
## 2           12.48              13.52          21.500             22.037
## 3           13.82              14.40          20.354             19.537
##   ShootingFoulsPG ShootingFoulsDrawnPG LessThnEightFeedUsage
## 1          14.889               12.642                 43.55
## 2          12.268               13.415                 43.45
## 3          12.134               10.549                 36.19
##   EightToSixteenFeedUsage SixteenToTwentyFourFeetUsage
## 1                   11.46                         4.80
## 2                   11.46                         4.89
## 3                   14.82                        10.90
##   TwentyFourPlusFeetUsage AvgShotDistance OppAvgShotDistance
## 1                   39.91           13.06              13.34
```

```
## 2                   39.96           13.18              12.89
## 3                   38.00           14.00              13.49
##   AvgMadeShotDistance OppMadeAvgShotDis
## 1               10.34             10.75
## 2               10.70             10.45
## 3               11.64             10.85
```

For this analysis, we model the relationship between Offensive Efficiency and WinPercentage Y: WinPercentage X: OffEff Let's start with a scatterplot

```
plot(nbaData$OffEff, nbaData$WinPercentage, main = "Scatterplot")
```



Let's quickly look at the correlation between the variables

```
cor(nbaData$OffEff, nbaData$WinPercentage)
```

```
## [1] 0.7309047
```

Build simple linear regression model (first variable in Y, second is x)

```
model <- lm(WinPercentage ~ OffEff, nbaData)
model
```

```
##
```

```
## Call:
## lm(formula = WinPercentage ~ OffEff, data = nbaData)
##
## Coefficients:
## (Intercept)       OffEff
##    -2.63460      0.02884
```

Now, lets see a summary of the model

```
summary(model)
```

```
##
## Call:
## lm(formula = WinPercentage ~ OffEff, data = nbaData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25923 -0.06440 -0.00524  0.07322  0.25462
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.634599   0.240738  -10.94   <2e-16 ***
## OffEff       0.028836   0.002213   13.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1037 on 148 degrees of freedom
## Multiple R-squared:  0.5342, Adjusted R-squared:  0.5311
## F-statistic: 169.7 on 1 and 148 DF,  p-value: < 2.2e-16
```

From the summary above, let's define all the variables

Residuals: the difference between the observed variables and the model variables. Here, you want symmetry and a median value around 0. We have that in this case! To take this further, you can plot the residuals to see if they're normally distributed

Coefficients Estimate: these represent the model's intercept and slope terms. The intersection of Estimate and (Intercept) is the model's intercept. The intersection of Estimate and OffEff is the model's slope.

Coefficients Std. Error: these measure the average amount the coeffiecient estimates vary from the observed variables. Ideally, we wan these low to show that the model is not much different from the observed data.

Coefficient t value: how many standard deviations our coefficient estimate is from 0. The further this is from 0 the better chance there exists a relationship between the variables.

Coefficient Pr(>|t|): the probability of observing any value >= t. A small p-value indicates unlikely relationship between predictor (OffEff) and response (WinPercentage). Typically, a p-value less than 5% is pretty good. You can see the ***'s line up to the sifniciance codes. We have a very small p-value in this example, which will allow us to reject the null hypothesis and conclude a relationship between OffEff and WinPercentage.

Residual Standard Error: measure of quality of linear regression fit, how close on average is the abline to the data points? On average, our data points are within 0.1037 of our regression line. With 148 degrees of freedom, or data points that went into the estimation (data points - variables, 150 data points - 2 variables).

Multiple R-squared: proportion of variance, how well is the model fitting the data? We have 53% is our model, so rougly 53% of the variance of WinPercentage can be explained by the OffEff.

Adjusted R-squared: this metric controls against the increase in variables. So in Multiple Linear Regression with more than 1 variable, this metric is more important to compare with the multiple r-squared.

F-statistic: different than t-tests, an f-test can measure multiple variables at once (the full model) to indicate a relationship among our predictors and response. Summary, the variance explained by the parameters in your model. First look at the p-value, if it is less than your significance value then you have sufficient evidence that your model fits the model better than without predictor variables. Hence, regject the null hypothesis.

Now for fun, lets see the abline with the scatterplot

```
plot(nbaData$OffEff, nbaData$WinPercentage)
abline(model, col = 2)
```