Tutorial 10

Steve Hof

07/04/2021

df = read.csv(file = "randle\_stats.csv")  
dfa = read.csv(file = "randle\_advanced.csv")

calc = function(x) {  
 threes = x$X3P \* 3  
 twos = x$FG \* 2  
 ft = x$FT   
 rbd = x$TRB \* 1.2  
 blk = x$BLK \* 2  
 stl = x$STL \* 2  
 to = x$TOV \* -1  
 ass = x$AST \* 1.5  
 return(threes + twos + ft + rbd + blk + stl + to + ass)  
}

# clean df data  
df$MP = gsub(":", ".", df$MP)  
df$X = gsub("@", "away", df$X)  
df$X[!nzchar(df$X)] = "home"  
  
df = df %>% select(-c(Date, Age, Tm, Rk, G, X.1))  
cols.num = c("GS", "MP", "FG", "FGA", "FG.", "X3P", "X3PA", "X3P.",  
 "FT", "FTA", "FT.", "ORB", "DRB", "TRB", "AST", "STL",  
 "BLK", "TOV", "PF", "PTS", "GmSc", "X...")  
df[cols.num] = sapply(df[cols.num], as.numeric)  
df$fpts = calc(df)  
  
#clean dfa data  
dfa$MP = gsub(":", ".", dfa$MP)  
dfa$X = gsub("@", "away", dfa$X)  
dfa$X[!nzchar(dfa$X)] = "home"  
  
dfa = dfa %>% select(-c(Date, Age, Tm, Rk, G, X.1, Opp, GS))  
cols.numa = c("MP", "TS.", "eFG.", "ORB.", "DRB.", "TRB.",  
 "AST.", "STL.", "BLK.", "TOV.", "USG.", "ORtg",  
 "DRtg", "GmSc", "BPM")  
dfa[cols.numa] = sapply(dfa[cols.numa], as.numeric)  
cols.fac = c("X")  
dfa[cols.fac] = sapply(dfa[cols.fac], as.factor)  
  
# calculate fantasy points and add as column in dfa  
dfa$fpts = df$fpts

fit1 = lm(fpts ~ ., data = dfa)  
summary(fit1)

##   
## Call:  
## lm(formula = fpts ~ ., data = dfa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.7142 -1.0879 0.1556 1.3356 4.8254   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.16994 16.53301 -1.039 0.306576   
## Xhome -0.03347 0.83880 -0.040 0.968407   
## MP 0.65845 0.17682 3.724 0.000731 \*\*\*  
## TS. -77.75592 17.56823 -4.426 9.9e-05 \*\*\*  
## eFG. 26.28377 9.21657 2.852 0.007444 \*\*   
## ORB. 0.06653 0.24084 0.276 0.784093   
## DRB. 0.28729 0.21334 1.347 0.187279   
## TRB. -0.19658 0.42440 -0.463 0.646271   
## AST. -0.03459 0.04829 -0.716 0.478890   
## STL. -0.03461 0.34619 -0.100 0.920969   
## BLK. -0.35627 0.33230 -1.072 0.291440   
## TOV. 0.36570 0.12840 2.848 0.007512 \*\*   
## USG. 0.61684 0.15118 4.080 0.000268 \*\*\*  
## ORtg 0.34080 0.12385 2.752 0.009557 \*\*   
## DRtg -0.08971 0.04676 -1.919 0.063706 .   
## GmSc 0.85557 0.33702 2.539 0.016030 \*   
## BPM 0.33301 0.21393 1.557 0.129105   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.483 on 33 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.9622, Adjusted R-squared: 0.9439   
## F-statistic: 52.49 on 16 and 33 DF, p-value: < 2.2e-16

(1.b) MP, TS., eFG., TOV., USG., ORtg and GmSc are the significant predictors

(1.c) The adjusted R-squared value is 0.9439

fit2 = lm(fpts ~ MP + TS. + eFG. + TOV. + USG. + ORtg + GmSc, data = dfa)  
summary(fit2)

##   
## Call:  
## lm(formula = fpts ~ MP + TS. + eFG. + TOV. + USG. + ORtg + GmSc,   
## data = dfa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.1030 -2.4183 0.0707 2.2855 5.1625   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.61815 13.77426 0.698 0.488857   
## MP 0.35762 0.16472 2.171 0.035619 \*   
## TS. -68.89221 16.07582 -4.285 0.000104 \*\*\*  
## eFG. 25.66256 9.82263 2.613 0.012414 \*   
## TOV. 0.19904 0.11032 1.804 0.078387 .   
## USG. 0.26965 0.12862 2.096 0.042105 \*   
## ORtg 0.09235 0.11894 0.776 0.441857   
## GmSc 1.68401 0.26308 6.401 1.05e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.029 on 42 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.9284, Adjusted R-squared: 0.9165   
## F-statistic: 77.8 on 7 and 42 DF, p-value: < 2.2e-16

(2.c) The adjusted R-squared is 0.9165

(2.b) Yes, our fit1 has a higher R-squared value than our fit2, so it looks like we might have accidentaly removed something that we shouldn’t have.