Sports Analytics Regression

Jack Sampiere

```
library(readr)
# read data
nba_stats = read.csv('nba-stats-2013-2019.csv')
```

Part 1: An initial linear regression model

```
# split into training and test set
train = nba_stats[nba_stats$Year < 2018,]
test = nba_stats[nba_stats$Year >= 2018,]
# run regression
lm_stats_1 = lm(PTS ~ . - Team - Year - G - Playoffs - PTS, data = train)
summary(lm_stats_1)
```

```
##
## Call:
## lm(formula = PTS ~ . - Team - Year - G - Playoffs - PTS, data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.562e-13 -9.761e-14 -3.014e-14  4.631e-14  2.029e-12
##
## Coefficients: (2 not defined because of singularities)
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -4.753e-12  1.011e-12 -4.700e+00 6.29e-06 ***
## X3P          3.000e+00  8.733e-16  3.435e+15  < 2e-16 ***
## X3PA         1.024e-16  3.162e-16  3.240e-01   0.7465
## X2P          2.000e+00  3.671e-16  5.448e+15  < 2e-16 ***
## X2PA         0.000e+00  2.431e-16  0.000e+00   1.0000
## FG                  NA         NA         NA       NA
## FGA                 NA         NA         NA       NA
## FT           1.000e+00  4.327e-16  2.311e+15  < 2e-16 ***
## FTA          7.259e-16  3.582e-16  2.027e+00   0.0447 *
## ORB          2.343e-16  3.539e-16  6.620e-01   0.5090
## DRB         -1.470e-16  2.939e-16 -5.000e-01   0.6177
## AST          1.725e-16  2.405e-16  7.170e-01   0.4746
## STL         -1.550e-16  4.801e-16 -3.230e-01   0.7473
## BLK          2.267e-16  4.243e-16  5.340e-01   0.5941
## TOV          2.251e-16  3.303e-16  6.810e-01   0.4967
## PF          -4.675e-16  2.514e-16 -1.860e+00   0.0651 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.857e-13 on 136 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 2.166e+31 on 13 and 136 DF,  p-value: < 2.2e-16
```

(a) The training set R2 is 1.

(b) The nonzero coefficients are those on X3P, X2P, and FT.

(c) The result of this regression is not surprising since all points must come from either 2-point FGs, 3-point FGs, or free throws. We are thus able to perfectly predict total points with these variables in the dataset.

Part 2: A better linear regression model

```
lm_stats_2 = lm(PTS ~ X3PA + X2PA + FGA + FTA + ORB + DRB + AST + STL
                + BLK + TOV + PF, data=train)
summary(lm_stats_2)
```

```
##
## Call:
## lm(formula = PTS ~ X3PA + X2PA + FGA + FTA + ORB + DRB + AST +
##     STL + BLK + TOV + PF, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -618.88 -105.45    2.49  114.65  429.92
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 513.82999  629.50509   0.816  0.41576
## X3PA          0.69129    0.11843   5.837 3.58e-08 ***
## X2PA          0.36184    0.12259   2.952  0.00371 **
## FGA                NA         NA      NA       NA
## FTA           0.78122    0.09658   8.089 2.67e-13 ***
## ORB          -0.20277    0.22138  -0.916  0.36129
## DRB           0.61064    0.18204   3.354  0.00103 **
## AST           0.91510    0.13105   6.983 1.09e-10 ***
## STL           0.50506    0.29140   1.733  0.08528 .
## BLK           0.07975    0.27270   0.292  0.77039
## TOV          -0.62701    0.20417  -3.071  0.00257 **
## PF            0.30420    0.15811   1.924  0.05640 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184.4 on 139 degrees of freedom
## Multiple R-squared:  0.7944, Adjusted R-squared:  0.7796
## F-statistic: 53.69 on 10 and 139 DF,  p-value: < 2.2e-16
```

(a) X3PA, X2PA, FTA, DRB, AST, and TOV are all statistically significant at the 0.05 level.

(b) The coefficient on X3PA implies that an increase in X3PA of 1 corresponds to an increase in PTS of 0.69129. This implies a probability of success of a 3-point attempt of $0.69129/3 = 0.23043$.

(c) FGA does not have a coefficient in the model since it is perfectly collinear with X3PA, X2PA, and FTA since all field goal attempts must be one of those types.

```
# use regression to predict test set
pred = predict(lm_stats_2, newdata=test)
# compute SSE and SST
SSE = sum((test$PTS - pred)^2)
SST = sum((mean(train$PTS) - test$PTS)^2)
pred_r2 = 1 - SSE/SST
```

```
pred_r2
```

```
## [1] 0.9327728
```

(d) The test set R2 is 0.9327728.

See Part 3 in Python.