LinearModel

Pratik Kulkarni

22/10/2021

```r
data1 <- read.csv("fifa_AllFairnessScore.csv")

set.seed(1)

attach(data1)

#data1$Country = as.factor(data1$Country)

# Defining training set by taking data until 2014.
train1 <- data1[1:176, ]

# Defining testing set by taking remaining 2018 data
test1 <- data1[177:208, ]
```

```r
# Create a simple lin reg model
model1 <- lm(FinalPos ~ TeamRank, data = train1)

# View summary of model.
summary(model1)
```

```
##
## Call:
## lm(formula = FinalPos ~ TeamRank, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3390  -5.0205  -0.7666   6.0984  14.1479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.79699    0.81400   12.04   <2e-16 ***
## TeamRank     0.27757    0.02773   10.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.932 on 174 degrees of freedom
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3618
## F-statistic: 100.2 on 1 and 174 DF,  p-value: < 2.2e-16
```

```r
# Predict for testing dataset.
fifa_predicted1 <- predict(model1, newdata = test1)

# Fill vector with actual values from testing dataset.
```

```r
fifa_test <- test1[, "FinalPos"]


# Calculate MSE by taking the mean of squared error difference.
MSE <- mean((fifa_predicted1 - fifa_test) ^ 2)
print(MSE)
```

```
## [1] 50.1357
```

Successive predictive models with FinalPos as response variable were created, to assess and compare the predictive accuracy at alpha (alpha symbol) = 0.05.

TeamRank was entered as the predictor in Model 1 and it was significant, p-value < 2e-16, explaining 36.6% of variation in FinalPos outcome.

```r
# Create a simple lin reg model
model2 <- lm(FinalPos ~ Fairness, data = train1)

# View summary of model.
summary(model2)
```

```
##
## Call:
## lm(formula = FinalPos ~ Fairness, data = train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.165  -7.604  -1.598   7.483  10.418
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.392e+01  1.174e+00  11.855   <2e-16 ***
## Fairness    1.160e-03  5.344e-04   2.171   0.0313 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.587 on 174 degrees of freedom
## Multiple R-squared:  0.02637,    Adjusted R-squared:  0.02078
## F-statistic: 4.713 on 1 and 174 DF,  p-value: 0.03129
```

```r
# Predict for testing dataset.
fifa_predicted2 <- predict(model2, newdata = test1)


# Calculate MSE by taking the mean of squared error difference.
MSE <- mean((fifa_predicted2 - fifa_test) ^ 2)
print(MSE)
```

```
## [1] 72.59164
```

Model 2 had only Fairness as the predictor, and the model was significant, p-value = 0.0313, explaining 2.6% of variation in FinalPos outcome.

```r
# Create a simple lin reg model
model3 <- lm(FinalPos ~ TeamRank+Fairness, data = train1)

# View summary of model.
```

```r
summary(model3)
```

```
##
## Call:
## lm(formula = FinalPos ~ TeamRank + Fairness, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.4582  -5.0004  -0.7949   6.0332  14.1738
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.867e+00  1.040e+00   9.490   <2e-16 ***
## TeamRank     2.784e-01  2.896e-02   9.616   <2e-16 ***
## Fairness    -4.893e-05  4.505e-04  -0.109    0.914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.952 on 173 degrees of freedom
## Multiple R-squared:  0.3655, Adjusted R-squared:  0.3582
## F-statistic: 49.83 on 2 and 173 DF,  p-value: < 2.2e-16
```

```r
# Predict for testing dataset.
fifa_predicted3 <- predict(model3, newdata = test1)


# Calculate MSE by taking the mean of squared error difference.
MSE <- mean((fifa_predicted3 - fifa_test) ^ 2)
print(MSE)
```

```
## [1] 50.06966
```

Model 3 had TeamRank and Fairness as the predictors. The overall model was significant, p-value < 2.2e-16, explaining 36.6% of variation in FinalPos outcome. TeamRank was a significant predictor while Fairness was not significant.

```r
# Create a simple lin reg model
model4 <- lm(FinalPos ~ TeamRank*Fairness, data = train1)

# View summary of model.
summary(model4)
```

```
##
## Call:
## lm(formula = FinalPos ~ TeamRank * Fairness, data = train1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2249  -4.9781  -0.7494   5.5535  14.5968
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        5.833e+00  1.545e+00   3.774 0.000220 ***
## TeamRank           4.842e-01  6.600e-02   7.337 8.27e-12 ***
## Fairness           1.810e-03  6.942e-04   2.607 0.009950 **
## TeamRank:Fairness -8.407e-05  2.440e-05  -3.445 0.000717 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.743 on 172 degrees of freedom
## Multiple R-squared:  0.4065, Adjusted R-squared:  0.3961
## F-statistic: 39.27 on 3 and 172 DF,  p-value: < 2.2e-16
```

```
# Predict for testing dataset.
fifa_predicted4 <- predict(model4, newdata = test1)


# Calculate MSE by taking the mean of squared error difference.
MSE <- mean((fifa_predicted4 - fifa_test) ^ 2)
print(MSE)
```

```
## [1] 56.71364
```

Model 4 had TeamRank, Fairness and their interaction as the predictors. The overall model was significant, p-value < 2.2e-16, explaining 40.7% of variation in FinalPos outcome. The interaction term was significant in this model.

```
# Create a simple lin reg model
model5 <- lm(FinalPos ~ TeamRank + Fairness + NumGames, data = train1)

# View summary of model.
summary(model5)
```

```
##
## Call:
## lm(formula = FinalPos ~ TeamRank + Fairness + NumGames, data = train1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.300 -3.392  1.570  2.598  4.933
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.698e+01  1.142e+00  32.381  < 2e-16 ***
## TeamRank     5.723e-02  1.558e-02   3.673  0.00032 ***
## Fairness     3.042e-05  2.033e-04   0.150  0.88122
## NumGames    -5.476e+00  2.103e-01 -26.039  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.136 on 172 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8694
## F-statistic: 389.2 on 3 and 172 DF,  p-value: < 2.2e-16
```

```
# Predict for testing dataset.
fifa_predicted5 <- predict(model5, newdata = test1)


# Calculate MSE by taking the mean of squared error difference.
MSE <- mean((fifa_predicted5 - fifa_test) ^ 2)
print(MSE)
```

```
## [1] 10.08672
```

Model 5 had TeamRank, Fairness and NumGames as the predictors. The overall model was significant, p-value < 2.2e-16, explaining 87.2% of variation in FinalPos outcome. TeamRank and NumGames were significant predictors, while Fairness was not.

```
library("ggplot2")

data2 <- data1

data2$Year <- as.factor(data2$Year)

ggplot(data2) +
  aes(x = Fairness, y = TeamRank, color = Year) +
  geom_point(color = "grey") +
  geom_smooth(method = "lm") +
  labs(x = "Fairness Index",
       y = "Team Rank",
       title = "Relationship between Team Rank and Fairness by Year")
```

```
## `geom_smooth()` using formula 'y ~ x'
```