W203 Lab 2

2022-04-01

Section 4. Results

Stargazer regression table:

```
##
## =====================================================
##                              Dependent variable:
##                           ---------------------------
##                                   log(Value)
##                              (1)       (2)      (3)
## -----------------------------------------------------
## Age                       0.306***  0.447***  0.742***
##                            (0.043)   (0.039)   (0.043)
## I(Age2)                   -0.008*** -0.010*** -0.015***
##                            (0.001)   (0.001)   (0.001)
## Height                    0.030***  0.016***   0.012
##                            (0.007)   (0.006)   (0.007)
## Weight                     0.002     0.0003    -0.006
##                            (0.005)   (0.004)   (0.005)
## Special                   0.007***  0.006***
##                           (0.0002)  (0.0002)
## Contract.Years                      -0.193*** -0.263***
##                                      (0.019)   (0.021)
## Agility                                        0.003
##                                                (0.003)
## Strength                                      0.013***
##                                                (0.003)
## Jumping                                       0.014***
##                                                (0.003)
## Acceleration                                  0.009***
##                                                (0.003)
## Stamina                                        0.008*
##                                                (0.004)
## Weak.Foot                                     0.121***
##                                                (0.036)
## International.Reputation            0.847***  1.142***
##                                      (0.053)   (0.059)
## Constant                  -2.395**   -0.867    -0.889
##                            (1.149)   (1.023)   (1.279)
## -----------------------------------------------------
## Observations                1,358     1,358     1,358
## R2                          0.557     0.655     0.543
## Adjusted R2                 0.555     0.653     0.539
## =====================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Statistical Significance:

```
## [1] "Model(1) VIF"
```

```
##        Age   I(Age^2)     Height     Weight    Special
## 102.765687  96.436092   1.487106   1.595312   1.749680
```

```
## [1] "Model(2) VIF"
```

```
##                      Age                 I(Age^2)                   Height
##               107.198673               100.997117                 1.517401
##                   Weight                  Special           Contract.Years
##                 1.601764                 2.084273                 1.147207
## International.Reputation
##                 1.285620
```

```
## [1] "Model(3) VIF"
```

```
##                      Age                 I(Age^2)                   Height
##                97.672622                95.391066                 1.693843
##                   Weight International.Reputation           Contract.Years
##                 1.796055                 1.201249                 1.116283
##                  Agility                 Strength                  Jumping
##                 1.760678                 1.561256                 1.736157
##             Acceleration                  Stamina                Weak.Foot
##                 2.044777                 1.513728                 1.021232
```

Our Model(1) includes only key variables based on our research question and preliminary EDA: `Age`, `Height`, `Weight`, and `Special`. Our EDA revealed that `Age` had a polynomial relationship with `log(Value)`, and so our linear model includes both `Age` and `Age^2`. Model(1) has a high VIF for `Age` (102.76) and `Age^2` (96.43). Although high VIFs are typically a concern, it makes sense that `Age` and `Age^2` have collinearity and the model does not aim to differentiate between `Age` and `Age^2`.

```
## model1 msr model2 msr model3 msr
##  0.8229954  0.6410832  0.8491938
```

All of the variables except `Weight` have statistical significance. Our initial Model(1) has an `R^2` of 0.557 and MSR of 0.823. Model(2) contains the key variables in addition to `Contract.Years` and `International.Reputation', which may also influence market value. Model(2) has anR^2of 0.655, MSR of 0.641 and stable VIFs. The stargazer regression model shows that`Contract.YearsandInternationa are both significant variables in addition to the significant variables in Model(1). When comparing Model(1) and Model(2) through the F-test, Model(2) has a significant p-value less than 2.2e-16 and thus improved the model's fit.

```r
anova(model1, model2, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: log(Value) ~ Age + I(Age^2) + Height + Weight + Special
## Model 2: log(Value) ~ Age + I(Age^2) + Height + Weight + Special + Contract.Years +
##     International.Reputation
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   1352 1117.63
## 2   1350  870.59  2    247.04 191.54 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After creating the linear model for our key variables, Model(3) investigates the impact of the omitted variable bias involved with using `Special` in place of independent performance variables. From the collinearity matrix in our exploratory data analysis, many performance metrics and scoring are collinear. To maintain independence as best as possible, `Agility`, `Strength`, `Jumping`, `Acceleration`, `Stamina`,and `Weak.Foot` were selected due to their weak correlation with each other. The stargazer regression table shows that all the variables except `Agility` and `Weight` are significant. Model(3)'s VIF's confirmed that these variables do not cause multicollinearity problems. Compared to Model(2), Model(3) had a higher MSR of 0.849 and lower `R^2` of 0.492. An F-test comparing Model(2) and Model(3) in the code below did not produce a significant p-value and so Model(3) did not improve Model(2). Although Model(3) may have reduced omitted variable bias, performance measurements are inherently related to each other because it is a measurement of a player's physical ability and can create causality problems. Additionally, our variable selection may not match FIFA's `Special` scoring process and contribute to Model(3)'s inaccuracy.

```
anova(model2, model3, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: log(Value) ~ Age + I(Age^2) + Height + Weight + Special + Contract.Years +
##     International.Reputation
## Model 2: log(Value) ~ Age + I(Age^2) + Height + Weight + International.Reputation +
##     Contract.Years + Agility + Strength + Jumping + Acceleration +
##     Stamina + Weak.Foot
##   Res.Df     RSS Df Sum of Sq F Pr(>F)
## 1   1350  870.59
## 2   1345 1153.21  5   -282.61
```

After evaluating the statistical significance of each model, Model(2) appears to be the best linear regression model among the three because of its higher `R^2`, lower MSR, and variable selection. When applying the t-test, all coefficients are significant except `Weight` and we can reject the null based on the null hypothesis that the p-value must be less than 0.1.

```
coeftest(model2, vcovHC)
```

```
##
## t test of coefficients:
##
##                             Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)              -0.86699314  1.13153126  -0.7662  0.443684
## Age                       0.44657651  0.04398013  10.1541 < 2.2e-16 ***
## I(Age^2)                 -0.01036027  0.00078824 -13.1436 < 2.2e-16 ***
## Height                    0.01613469  0.00623590   2.5874  0.009774 **
## Weight                    0.00033100  0.00444658   0.0744  0.940672
## Special                   0.00567491  0.00024287  23.3661 < 2.2e-16 ***
## Contract.Years           -0.19314955  0.02077716  -9.2962 < 2.2e-16 ***
## International.Reputation  0.84687202  0.05826756  14.5342 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Practical Significance:

$Log(Value) = 0.447 * Age + -0.10 * Age^2 + 0.16 * Height + 0.003 * Weight + 0.006 * Special + 0.193 * Contract.Years + 0.847 * International.Reputation$

The Model(3) linear regression can be interpreted as how Log(Value) will change with increases in each variable. For example, a one-point increase in international reputation while keeping all else constant will lead to 0.847 increase in Log(Value) (or \$2.33). The coefficients reveal how much each weight each variable carries in the determination of a player's market value. From this linear regression, international reputation has the largest impact, followed by Age, Contract.years, Height, Special, and Weight. Based on the selected linear model, it is surprising that `Special`, which measures a player's skill, does not play a large role in market value compared to other factors. International reputation and player skill can have a large influence on a club's revenue and performance while weight did not influence market value as much as we hypothesized. As a team manager or scout, this regression can support the determination of whether it is worth recruiting a high-market value player based on the team's priorities. A highly skilled player with low international reputation may have a lower market value compared to a player with high international reputation but is not as skilled. A team looking for a highly skilled player may find that recruiting the player with the lower market value is more beneficial.