Analysis about Change in Football Player Value

DSCI510: Final Project
Alain Tamazian

This project is about determining what "To what extent does a footballer's performance in a single season affect their overall financial, value". Due to web-scraping challenges – like query limits, slow extraction time, and my own skill level – I narrowed down my analysis to focus on Spanish La Liga players from the 2018 and 2019 seasons. My motivation for choosing this topic is simple – I love football (i.e. soccer). As an avid football player and view, I was always interested about what performance contribution is objectively valued the highest. This analysis was, in part, designed to address that. Overall, this the question of how much a change in performance between seasons will affect a player's value is also has great practical relevance among football professionals since it directly affects their financial value. I collected my data from three sources: "SoFIFA, "API-Football", and "Transfermarkt". The SoFIFA website includes various data from the FIFA video game, which after rigorous internal research and analysis quantifies player skills and value based on past performances. I primarily used this source to get the 2019 and 2018 values (in millions of euros) for the La Liga players; the change between the two is my dependent variable. The API-Football source is a public API hosted by RapidAPI, which I used to gather the specific performance statistics for all players in the 2018 and 2019 seasons. Transfermarkt is a German website dedicated to recording all sorts of football information, such as match predictions, transfer rumors, and historical statistics. I used this source for the historical statistics.

Now, regarding my project specifically, there was a lot of data manipulation and pre-processing to be done before the actual analysis. I combined the 4 datasets from SoFIFA and Transfermarkt into a single table, creating a dataset with only the players that appeared in all 4 tables. Then based on this, I extracted the historical data from Transfermarket for the remaining players. Creating the final dataset, however, was not as straight forward since it involves attribute construction: transforming the performance statistics in a change value. The final, regression analysis will be mainly using the change in various performance statistics as the independent variables for the model; therefore, the accuracy of my results is directly dependent on what approach I used to compute the change.

There are the two main datasets at this stage: "careerStats" for historical data and "allCurrentStats" for the data on the 2018 and 2019 La Liga seasons. Between the two of them, there are dozens of different player performance attributes. I thought I may be able to get a lot of detailed analysis with them. However, most the attributed had to be converted into their "change" counterpart, and the method for doing so had to be uniform. In other words, "change in goals" can't be defined as "goals in 2019-historical average goals", while at the same time "change in saves" is "saves in 2019-saves-saves in 2018". Only about 12 attributes coincided among the datasets. Regardless, it is more important for our results to be statistically credible (using consistent metrics) than rather than detailed and have a robust array of independent variables. There where 3 different methodologies I was considering for how to calculate change. The performance stats from our target (2019-2020) season could be subtracted by one of the following: the players' performance stats from the 2018-2019 season, the average of the player's historical performances, or a weighted average of the historical performance stats – calculated with a greater weight placed on the most recent (2018) season results. I set the weight for the 2018 values as 3 times greater than the rest. I intuitively expected that a football player's most recent displays of skills would be evaluated as more relevant than his performances from further

Regression Table: Method 1

OLS Regression Results

Dep. Variable:	y	R-squared:	0.176
Model:	OLS	Adj. R-squared:	0.153
Method:	Least Squares	F-statistic:	7.378
Date:	Mon, 10 May 2021	Prob (F-statistic):	9.26e-10
Time:	09:37:49	Log-Likelihood:	-932.78
No. Observations:	320	AIC:	1886.
Df Residuals:	310	BIC:	1923.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.6095	1.740	6.672	0.000	8.186	15.033
age_20	-0.4017	0.063	-6.357	0.000	-0.526	-0.277
mins_chg	0.0010	0.000	2.637	0.009	0.000	0.002
conceded_chg	-0.0442	0.021	-2.102	0.036	-0.086	-0.003
goals_chg	0.0375	0.116	0.323	0.747	-0.191	0.266
red_chg	-0.3269	0.661	-0.495	0.621	-1.627	0.973
assists_chg	0.2182	0.146	1.496	0.136	-0.069	0.505
goalkeeper	-0.8793	0.981	-0.896	0.371	-2.810	1.051
midfielder	-0.8985	0.623	-1.443	0.150	-2.123	0.326
attacker	-1.2812	0.725	-1.766	0.078	-2.709	0.146

Omnibus:	98.654	Durbin-Watson:	2.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	682.944
Skew:	-1.075	Prob(JB):	5.02e-149
Kurtosis:	9.827	Cond. No.	5.85e+03

Regression Table: Method 2

OLS Regression Results

Dep. Variable:	value_mil_chg	R-squared:	0.319
Model:	OLS	Adj. R-squared:	0.299
Method:	Least Squares	F-statistic:	16.12
Date:	Mon, 10 May 2021	Prob (F-statistic):	1.15e-21
Time:	09:37:49	Log-Likelihood:	-902.40
No. Observations:	320	AIC:	1825.
Df Residuals:	310	BIC:	1862.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	11.4910	1.562	7.358	0.000	8.418	14.564
age_20	-0.3685	0.057	-6.495	0.000	-0.480	-0.257
mins_chg	0.0014	0.000	3.686	0.000	0.001	0.002
conceded_chg	0.0051	0.052	0.097	0.923	-0.098	0.108
goals_chg	0.2697	0.116	2.324	0.021	0.041	0.498
red_chg	-0.4503	0.805	-0.559	0.576	-2.035	1.134
assists_chg	0.7425	0.164	4.519	0.000	0.419	1.066
goalkeeper	-0.2662	0.785	-0.339	0.735	-1.811	1.278
midfielder	-0.2339	0.571	-0.409	0.683	-1.358	0.890
attacker	0.5271	0.753	0.700	0.485	-0.955	2.009

Omnibus:	96.110	Durbin-Watson:	2.174
Prob(Omnibus):	0.000	Jarque-Bera (JB):	720.380
Skew:	-1.013	Prob(JB):	3.73e-157
Kurtosis:	10.066	Cond. No.	5.92e+03

Regression Table: Method 3

OLS Regression Results

Dep. Variable:	value_mil_chg	R-squared:	0.336
Model:	OLS	Adj. R-squared:	0.316
Method:	Least Squares	F-statistic:	17.40
Date:	Mon, 10 May 2021	Prob (F-statistic):	2.90e-23
Time:	09:37:49	Log-Likelihood:	-898.43
No. Observations:	320	AIC:	1817.
Df Residuals:	310	BIC:	1855.
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	9.9149	1.556	6.371	0.000	6.853	12.977
age_20	-0.3375	0.056	-5.985	0.000	-0.448	-0.227
mins_chg	0.0019	0.000	5.148	0.000	0.001	0.003
conceded_chg	-0.0326	0.053	-0.611	0.542	-0.138	0.072
goals_chg	0.2436	0.118	2.067	0.040	0.012	0.475
red_chg	-0.5573	0.777	-0.717	0.474	-2.086	0.972
assists_chg	0.7350	0.168	4.378	0.000	0.405	1.065
goalkeeper	-0.1212	0.766	-0.158	0.874	-1.629	1.386
midfielder	-0.3308	0.562	-0.589	0.556	-1.437	0.775
attacker	0.1787	0.716	0.250	0.803	-1.229	1.587

Omnibus:	106.636	Durbin-Watson:	2.150
Prob(Omnibus):	0.000	Jarque-Bera (JB):	836.107
Skew:	-1.141	Prob(JB):	2.76e-182
Kurtosis:	10.583	Cond. No.	5.45e+03

back in time (e.g. 8 years ago). The specific metric I used in my program of "3 times greater" is effectively arbitrary; its purpose is simply to indicate a trend regarding whether using a weighted average would give better results, worse, or simply insignificant.

The first approach can be tested using the existing "allCurrentStats" dataset. However, the second and third methods required creating new datasets through attribute construction. The resulting dataset for using the unweighted average is "allStatsHist_Unweighted", and table "allStatsHist" for the calculating change based on a weighted historical average. The independent variables chosen here are attributes I thought would be relevant for this test – such as change in goals, assists, and goals conceded. These attributes do not necessarily represent the final regression model's configuration. Also, other than how change is calculated, the overall attributes are all same for each of the 3 regression; this is done in order to maintain statistical comparability.

To determine which methodology is most effective, I ran three linear regressions. An effective and simple metric to determine which of the 3 methods delivers the best results is the R^2, which measures the much of the variance in our dependent variable ("change in value"), is accounted for by these change attributes. So, simply, this means that the regression for the method with the highest R^2 is most representative of how value is determined, and thus is our choice. Since these intermediate outputs are simply for comparing of the efficacy of the 3 techniques, specific (in depth) analysis of the variables and their coefficients is unnecessary at this stage. As you can see from the regression tables on the left, the R^2 of the 3 methods are 0.176, 0.319, and 0.336. Overall, unfortunately, none of them are as high as I would have hoped. However, regarding the task at hand, the results verified my assumption that using a weighted historical average would be most effect; 0.336 is clearly the biggest R^2. So, the final regression will be conducted using this methodology, and the "allStatsHist" dataset.

Final Regression Output

OLS Regression Results

Dep. Variable:	value_mil_chg	R-squared:	0.388
Model:	OLS	Adj. R-squared:	0.366
Method:	Least Squares	F-statistic:	17.72
Date:	Mon, 10 May 2021	Prob (F-statistic):	3.39e-27
Time:	14:31:49	Log-Likelihood:	-885.37
No. Observations:	320	AIC:	1795.
Df Residuals:	308	BIC:	1840.
Df Model:	11		
Covariance Type:	nonrobust		

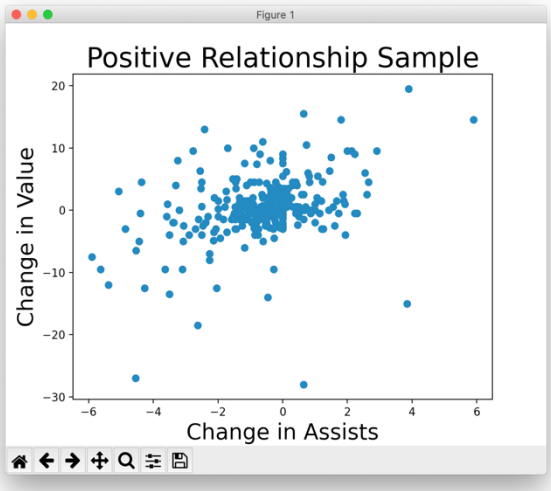
	coef	std err	t	P>|t|	[0.025	0.975]
const	20.6105	3.037	6.785	0.000	14.634	26.587
age_20	-0.1740	0.068	-2.556	0.011	-0.308	-0.040
overall_19	-0.2034	0.050	-4.092	0.000	-0.301	-0.106
mins_chg	0.0046	0.001	6.056	0.000	0.003	0.006
apps_chg	-0.1826	0.069	-2.630	0.009	-0.319	-0.046
goals_chg	0.2239	0.114	1.966	0.050	-0.000	0.448
conceded_chg	-0.0606	0.053	-1.149	0.251	-0.164	0.043
yellow_chg	-0.3176	0.118	-2.689	0.008	-0.550	-0.085
assists_chg	0.6448	0.166	3.878	0.000	0.318	0.972
goalkeeper	-0.5228	0.878	-0.596	0.552	-2.250	1.204
defender	-0.5824	0.694	-0.839	0.402	-1.949	0.784
midfielder	-0.5296	0.671	-0.789	0.431	-1.851	0.791

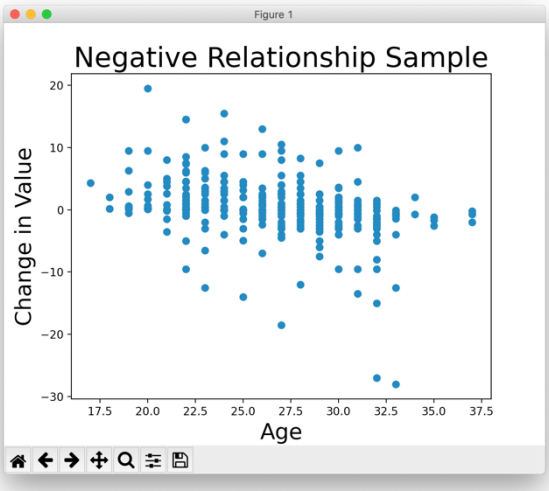
Omnibus:	62.869	Durbin-Watson:	2.093
Prob(Omnibus):	0.000	Jarque-Bera (JB):	522.363
Skew:	-0.501	Prob(JB):	3.72e-114
Kurtosis:	9.178	Cond. No.	1.09e+04

On the left-hand side, we can see the output of final regression model. After some guess-and-check attempts, I found this set of attributes have the most informative results and to yield the highest R^2. At 0.388, it is still considered "low effect size", but it is by no means negligible. It means that almost 40% of the variable in the "change in value" of La Liga players (in 2019) can be explained by this model – with this sets of attributes. This result is more than twice as good as the regression for method 1 that was tested. Using the typical significance level of 5%, any variable with a p-value ≤ 0.05 is significant. So, we see that the intercept as well as 7 out of our 11 independent variables are statistically significant. Significance tells us that there is a correlation between the shifts of "change in value" of a player and its significant factors.

The intercept is highly significant and represents the expected change in value when all the other variables are 0. So, since "attacker" was set as our reference dummy, it's effect on value is determined within the constant. When all else is 0, the average player with the attacker position increased their value by 20 million euros. Since the trend of the other positional dummies depends on the reference dummy, it makes sense that the effect of the other positions like goalkeeper and defender has a negative effect on value. This makes sense since the attacker position is known to attract the most attention and money. However, despite the negative trends of the other dummies, their correlation with value is not statistically significant and therefore not specifically credible. The player's age at the end of the 2019 season is significant for their designated value. It has negative relationship that is approximately interpreted as, for a 1-year increase in age a player's value (in 2019 La Liga) decreases by 174,000 euros – when all else the same. This makes sense, since young players have a longer career and therefore greater potential. FIFA's metric of overall skill (before the beginning of the 2019 season) is statistically significant with a negative effect on change in value. Honestly, this was very surprising to me, since I though people with a high "Overall" rating would increase in value – even just based on reputation. Similarly surprising is the negative effect an increase in games played (ie apps) has on value, all else ghe same. Since all else is the same, maybe this is interpreted as: players who need more days to score the same number of goals (for example) are less valuable. Also, there might be a slight bias in the dataset due to potential multicollinearity – such as between "age_20" with "overall_19" and "mins_chg" with "apps_chg". This could, potential also explain both surprising trends. As expected, an increase in the number of minutes played has a positive impact of the change in value. It makes sense that the coefficient is so small relative to the coefficients

of the other attributes, since it signifies the effect on value per single minute. The change in minutes played attribute has a much high magnitude than all others, with its values being in the hundreds and thousands digit – as opposed to ones and tens. Change in goals scored has an expectedly positive (and relatively large) influence on a player's value determination; it is just barely significant at p-value = 0.05. While it makes sense that an increase in goals conceded to have a negative effect on change in value, I expected it be more significant. I believe this is because, there are only 39 goalkeepers in this dataset of 320 players, which biases the results against goalies since the other positions naturally have 0 goals conceded. However, analyzing the positions separately would require a much larger web-scraping effort with both more rows and matching attributes. Moving on, receiving more yellow cards per season has a significant, negative consequence for a player's value after the season. Based on the regression table, a single extra yellow card compared to the historical weighted average – all else even – results in an almost €320,000 deduction in a player's value; this is the largest negative relationship within our model – which is statistically significant. Finally, this last factor is also the most relevant according to the regression table. An improvement in assists – from the historical weighted average – has a tremendous positive and significant effect on the change in value. All else equal, a single additional assist is estimated to raise a players' value by €644,800. This is almost 3 times the payoff from scoring one more goal, which is a pleasant surprise. Below, you can also see a scatterplot visualization between change in assists and change in values – as well as the dependent variable's relationship with age. These scatterplots are tangential to my actual analysis. They are simply an intuitive visual aid to help interpret the model – displaying the relationship between two significant attributes with change in value.





Although some results weren't as accurate and significant as I would have hoped, the analysis was still fun and informative. This analysis helped answer, "To what extent a La Liga player's performance in a single season effects their value in football?" Though not fully complete, my regression analysis shows the direct benefit or consequence – to a footballer's career – that a specific play in a season can have. If this analysis was effectively scaled up – with some statistical support – I believe its results would be very beneficial to football player by helping them increase their financial potential. In hindsight, my regression model had a few pitfalls that may or may not have affected its accuracy. Firstly, it would have been preferrable to

have many more observations and matching attributes; then it may have even been possible to run a different analysis on players of each position. Unfortunately, even this amount requires a long program run time (of ~25 min) and already risks getting blocked by the source host for excessive querying. Also, other than deciding on the appropriate variables for the analysis, there are many other statistical parameters that could be considered during a multiple regression. For example, other than the typical linear (OLS) model, there are other types of regressions (e.g. polynomial, log-linear, Lasso, and ridge) that may have been more effective. However, most of these – other than just a lot of time and effort – would require receiving assistance from people with a lot more knowledge in both football and statistics.