Probability and Statistics II Final Project

Jake Bodea

28 April 2021

INTRODUCTION

As an avid soccer fan my whole life, I've come to appreciate the joy of rooting for a team and living under the anticipation of how well they will perform in their league. European soccer leagues, which will be the basis for my data, are typically a league of about 20 teams from each country, where each team plays each other twice, once at home and once away. A win is worth 3 points, a draw 1, and a loss 0. Over the course of roughly 38 games, teams try to collect the highest score possible in an effort to stand alone atop the league table and win the trophy.

For this project, I hope to use data collected from the 2019-2020 season of soccer across the top 5 European leagues (Spain, England, Germany, France, and Italy) to predict the league position of a European team, given it's previous season's total **xG** and **xGA**, and also the **value** of the squad before the new season begins. These statistics will be explained later in the paper.

The data set used in this research was gathered by me via the use of various sources, namely Understat.com and TransferMarkt.us. Both of these sites are unbiased, third-party sources simply presenting statistics generated by higher-level computers. Observe below a sample of a couple rows of the data frame.

```
## # A tibble: 8 x 6
##   team_name   country league_pos    xG   xGA value
##   <chr>       <chr>        <dbl> <dbl> <dbl> <dbl>
## 1 Burnley     England         10  49.4  53.8 226.
## 2 Genoa       Italy           17  49.9  59.5 204.
## 3 Montpellier France           8  35.0  32.3 101.
## 4 Eibar       Spain           14  42.4  54.3  82.8
## 5 RB Leipzig  Germany          3  76.1  37.6 618.
## 6 Levante     Spain           12  49.4  64.2 117.
## 7 Leccce      Italy           18  53.5  92.0  77.4
## 8 Schalke 04  Germany         12  35.9  54.2 253.
```

Explanation of Variables

The first three variables **team_name**, **country**, and **leage_pos** are quite self-explanatory. Allow me now to explain the other three.

- First, **xG** is a statistic that measures the percentage of an average player scoring a goal from a given position. For example, a penalty kick is worth 0.76 xG, meaning that there is a 76% chance of scoring a penalty. If a team had no shots all game except two penalties, their xG for the match would be the addition of those two percentages, or 1.52 xG. If that team scored both penalties, they outperformed their xG by $2 - 1.52 = 0.48$. xG takes into account the location of the shooter, the body part shot with, the type of pass before the goal, and the type of attack. The use of **xG** in the data frame is the season total **xG** of each team mentioned, so values of, say 65.61 xG would refer to a team being expected to score roughly that many goals in the given league season.

- Next, **xGA** is very similar to **xG**, except that it covers more of the defensive perspective of a team. While a team with a great offense would have a high **xG**, a team with a great defense would have a low **xGA**. Thus, the benefit of the two stats has an inverse relationship.
- Finally, the transfer market **value** of a team is the sum of all the individual values of the players in the squad. This value is determined by a player's form, age, skill, position, etc. Practically, this means that the best teams should be worth the most money, for they are made up of the best players.

PROCEDURES

The goal of this project is to use multiple regression analysis to create an algorithm that would predict the **league_pos** of a team based on it's previous season's **xG** and **xGA**, as well as the current season's market **value**. In other words, considering independence between the variables, I hope to derive a formula in the structure of $\widehat{y = league\_pos} = \beta_0 + \beta_{xG} \cdot \text{xG} + \beta_{xGA} \cdot \text{xGA} + \beta_{value} \cdot \text{value}$, where β refers to the slope of a given variable. This should be done validated via inference of regression.

Additionally, a hypothesis test would be done on each variable's slope to determine whether or not it is statistically sound to consider it a factor in **league_pos**. This would appear as $H_0 : \beta = 0$ against a two-sided alternative hypothesis $H_1 : \beta \neq 0$ with a value of $\alpha = 0.05$.

RESULTS

Multiple Regression and Inference of Regression

In order to accept a multiple linear regression model for the data, we must first analyze the variables and ensure that there is no bias or issue with the relationships between each other and the potential prediction formula. Recall that Inference of Regression requires

1. **L**inearity of relationship between variables
2. **I**ndependence of the residuals
3. **N**ormality of the residuals
4. **E**quality of variance of the residuals

In order to test the variables, we first need to find a model (formula) by which we can test different aspects of the data.

```
pos_model <- lm(league_pos ~ xG + xGA + value, data = soccer_stats)
get_regression_table(pos_model)
```

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    9.82      1.72      5.71    0        6.41    13.2
## 2 xG          -0.203     0.031    -6.62    0       -0.264   -0.142
## 3 xGA          0.221     0.027     8.33    0        0.169    0.274
## 4 value       -0.001     0.002    -0.767   0.445   -0.004    0.002
```

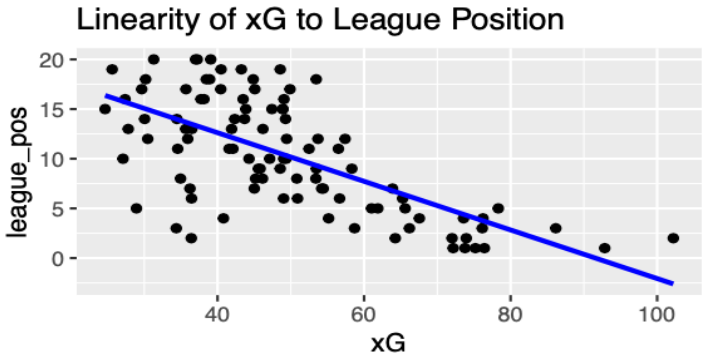
According to the table, our previous equation $\widehat{y = league\_pos} = \beta_0 + \beta_{xG} \cdot \text{xG} + \beta_{xGA} \cdot \text{xGA} + \beta_{value} \cdot \text{value}$ can now be simplified to

$$\widehat{league\_pos} = 9.818 - 0.203 \cdot \text{xG} + 0.221 \cdot \text{xGA} - 0.001 \cdot \text{value}$$

We may now proceed to ensure that each of the four conditions above are met.

Linearity of Relationship

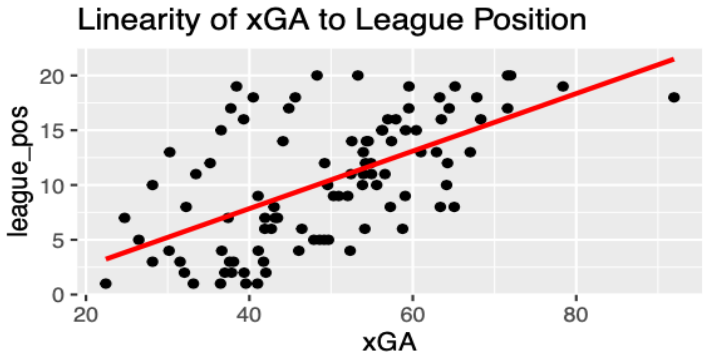
Let us determine that each of the variables do in fact share a linear relationship with **league_pos**.



xG and League Position

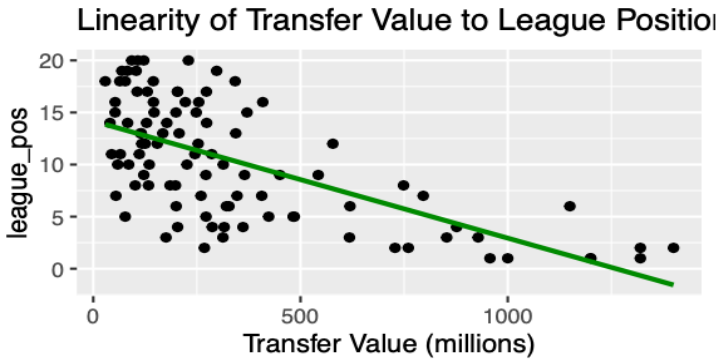
Observing the graph above, it would appear that `leage_pos` ~ `xG` have a negative linear relationship. This would mean that the higher the `xG`, the "lower" the league position.

Note the interesting contradiction that in terms of mathematics, being 1st in the league is equivalent to being the lowest, so negative correlations are beneficial and preferred.



xGA and League Position

Clearly, `xGA` and `league_pos` maintain a positive linear relationship. This means that as `xGA` increases (meaning that more goals are allowed, which means that a team's defense is performing more poorly), the team's league position rises. Thus, an increasing `xGA` is something teams try to avoid.



Tranfermarket Value and League Position

Finally, there appears to be a negative linear relationship between `value` and `league_pos`, meaning that according to this data, the greater the transfer `value` of a team, the smaller the league position would be.

Correlation between Variables Notice below the table detailing the correlations between the four variables.

```
soccer_stats %>%
  select(league_pos, xG, xGA, value) %>%
  cor()
```

```
##            league_pos         xG        xGA      value
## league_pos  1.0000000 -0.6670457  0.5942260 -0.6442437
## xG         -0.6670457  1.0000000 -0.1276957  0.7361401
## xGA         0.5942260 -0.1276957  1.0000000 -0.3381410
## value      -0.6442437  0.7361401 -0.3381410  1.0000000
```

As expected, `xG` and `value` maintained a moderately negative correlation with `leage_pos`, evidenced by values of about -0.6667 and -0.644 respectively. In addition, `xGA` follows a moderately positive correlation with `leage_pos` as seen by the value of 0.594 . Thus, each of the variables seem to have a moderately linear relationship with the outcome variable `leage_pos`.

Independence of the Residuals

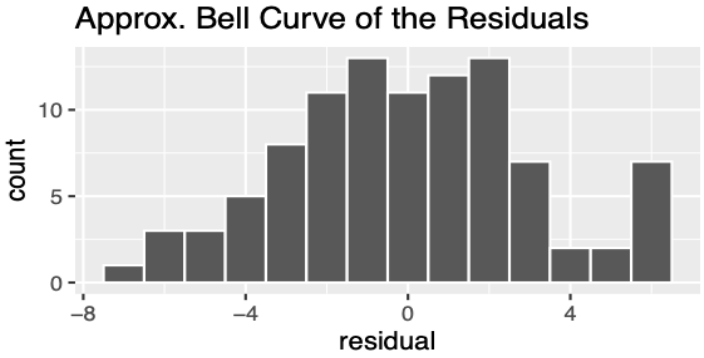
Given that the data represents each team only once, there is no repetition in any of the teams presented. This can be guaranteed because I was the one who collected the data.

Normality of Residuals

```
pos_model_reg_pts <- get_regression_points(pos_model)
pos_model_reg_pts %>% sample_n(size = 4)
```

```
## # A tibble: 4 x 7
##      ID league_pos    xG   xGA value league_pos_hat residual
##   <int>      <dbl> <dbl> <dbl> <dbl>          <dbl>    <dbl>
## 1    22          2 102.   37   1400           -4.39    6.39
## 2    34         14  34.4  57.4  274.           15.2    -1.20
## 3    45          5  61.9  49.7  485.            7.68   -2.68
## 4    63          5  78.3  49.2  482.            4.23    0.772
```

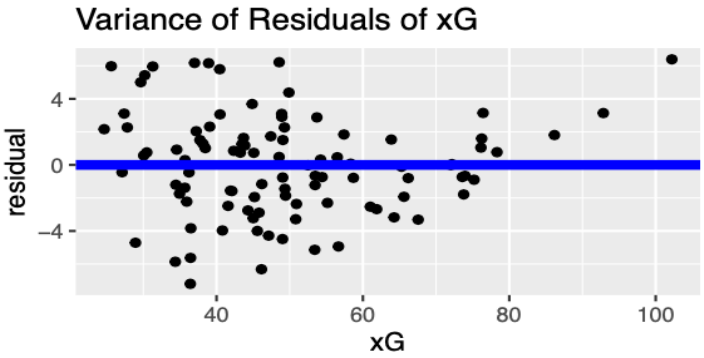
The regression points table modifies the original data frame to only select the variables we are working with, then uses the formula in the model given (`lm(leage_pos ~ xG + xGA + value)`) to predict `league_pos_hat` and calculate the residual, which is the value of `leage_pos` - `league_pos_hat`. We can now use these data points to graph a histogram of the residuals.



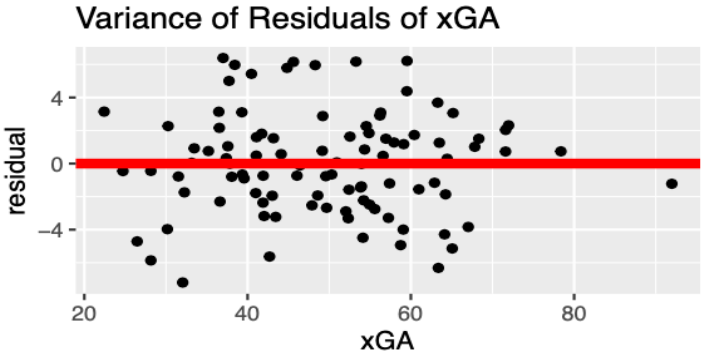
While not a perfect normal distribution, it would seem that the values do follow close to a bell curve, with the mean being somewhere around 0. I would say that the residuals do indeed follow a normal distribution.

Equality of Variance

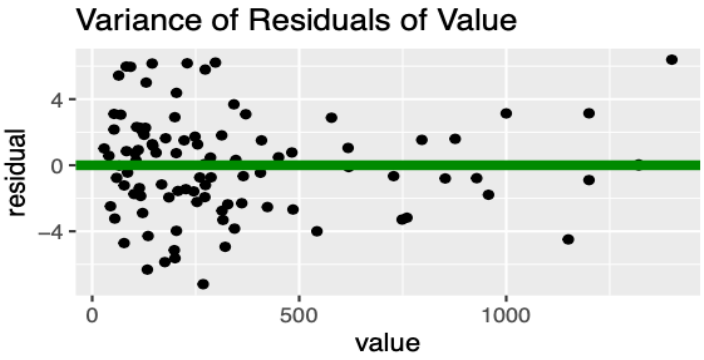
Since there is no "all encompassing" value by which the residual may be tested, we will analyze the equality of variance in each of the explanatory variables.



The graph above does not seem to change the variance of the `residual` based on the value of `xG`.



Similarly, `xGA` does not seem to follow any trend of a changing variance due to a change in `xGA`.



Finally, the graph once again demonstrates that there is no real change in the variance of the residuals as the `value` changes.

Thus, the equality of variance is maintained in each of the explanatory variables.

Inference of Regression Conclusion:

Thus, the findings are as follows:

1. Linearity of Variables: Yes
2. Independence of Residuals: Yes
3. Normality of Residuals: Yes
4. Equality of Variances: Yes

Thus, we found that the preliminary conditions of our analysis have been met, so we can put more faith in our model and trust in the p-values and confidence intervals explored in the next sections.

Hypothesis Tests of Regression

Now that we have determined the solidity of our model, we may begin analysis of the multiple linear regression model we had discovered above:

$$\hat{y} = \widehat{league\_pos} = 9.818 - 0.203 \cdot \text{xG} + 0.221 \cdot \text{xGA} - 0.001 \cdot \text{value}$$

.

Recall that we are trying to verify that our slopes are each valid estimates for the `leage_pos`. That is, $H_0 : \beta = 0$ against a two-sided alternative hypothesis $H_1 : \beta \neq 0$ with a value of $\alpha = 0.05$.

Observe the regression table below:

```
## # A tibble: 4 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    9.82      1.72      5.71    0         6.41    13.2
## 2 xG          -0.203     0.031    -6.62    0        -0.264   -0.142
## 3 xGA          0.221     0.027     8.33    0         0.169    0.274
## 4 value       -0.001     0.002    -0.767   0.445    -0.004    0.002
```

Looking at the p-values for each of the columns above, we reject H_0 (because the p-values are $< 0.05 = \alpha$) for all but β_{value}. In other words, we can say with 95% confidence that each of these data values listed below have a linear relationship with `leage_pos` and that:

- $\beta_0 \Rightarrow [6.406, 13.231]$
- $\beta_{xG} \Rightarrow [-0.264, -0.142]$
- $\beta_{xGA} \Rightarrow [0.169, 0.274]$

Since the p-value for $\beta_{value} = 0.445 > 0.05 = \alpha$, we fail to reject H_0, so we cannot say that it is a factor in the multiple linear regression formula. Therefore, our hypothesis testing found that a statistically sound prediction for `league_pos` is:

$$\widehat{league\_pos} = 9.818 - 0.203 \cdot \text{xG} + 0.221 \cdot \text{xGA}$$

CONCLUSION

Using multiple regression, inference for regression, and hypothesis testing for regression, we were able to determine the following formula:

$$\widehat{league\_pos} = 9.818 - 0.203 \cdot \text{xG} + 0.221 \cdot \text{xGA}$$

Inference for regression allowed us to verify that the results of our hypothesis tests and confidence intervals do indeed have a valid meaning. Hypothesis testing of $H_0 : \beta = 0$ against the two-sided alternate hypothesis revealed that we can say with 95% confidence that $\beta \neq 0$ for all but β_{value}, which means that the slopes between each of the remaining explanatory variables and the outcome variable are indeed statistically different from 0.

Finally, confidence intervals demonstrated that with 95% confidence, the true value of β is:

- $\beta_0 \Rightarrow [6.406, 13.231]$
- $\beta_{xG} \Rightarrow [-0.264, -0.142]$
- $\beta_{xGA} \Rightarrow [0.169, 0.274]$

Final Remarks

It is important to note that there are MANY other factors that could go into such a predictive algorithm, such as injuries, contract tensions, a global pandemic putting a stop to the entire world, etc. However, these findings seem to be quite accurate given the data with all other things held constant. Despite this, it is important to recall that correlation does not imply causation, so even though these explanatory variables logically make sense to have an affect of a team's league position, there is always the possibility that there is no causation between the variables.

I'll close with this: regardless of how complex our algorithms and supercomputers are, sport is one of those things that consistently goes beyond the statistics. If this were not the case, an advanced super computer would be able to predict the brackets of March Madness, for example, yet nobody in history has come up with a correct bracket. There seems to be a beauty in the space that won't allow itself to be governed by the laws of the statistics.