Data Analysis 2 & Coding 1 - Term Project

Analyzing the relationship between rainfall and the number of goals scored in a football match

Haaris Afzal Cheema

12/22/2021

Introduction

This analysis will explore the relationship between rainfall and the number of goals scored in a football match. The analysis will be restricted to the English Premier League for the season 2018-2019. In England, rainfall is often observed during football matches. Due to this, playing conditions tend to get more difficult as the ball tends to skid more and it is expected that players, especially goalkeepers, will be prone to making more errors, leading to a greater number of goals in the match. Therefore, the primary purpose of this analysis will be to investigate whether a relationship exists between rainfall and the number of goals scored in a match.

Data Collection

Data was collected from three different sources for this analysis. Firstly, the data on the premier league (football) was obtained from here in a csv format. Next, data was collected manually to identify the city for each team (20 obs, 2 variables) and this was joined with the football data. Finally, weather data was collected from here using their query builder, for each of the city in the UK which was associated with each team.

The cleaned dataset for weather contained 3864 observations for three variables, city, date and and precipitation (in mm). The cleaned data for football contained 380 observations for nine variables, the important ones from which include the total goal count, total corners in the match, number of yellow cards, number of red cards, total shots, and total shots on target. These two datasets were joined using an inner join, and the result was 380 observations.The joined data was checked for missing values, and subsequently, none were found.

Before we delve into the analysis, it is important to mention some problems, challenges and data quality issues. The primary source of restriction for preparing a more accurate analysis was the lack of a free API on historical data. The data that is currently being used for weather, retrieves historical data based on the nearest weather station to our selected location. A possible issue could be that while the weather station would be experiencing rainfall on a certain day, it is possible that it did not rain on that day at the stadium. Secondly, the data measures the millimeters of rainfall in a day. However, what we would be more interested in, would be whether it would rain during the match, so at a particular time interval.So this analysis is being carried out with the assumption that the rainfall at the weather station reflects the rainfall experienced by players at the stadium and weather data for the day reflects the conditions during match time. Lastly, as mentioned before, there were was no data on premier league teams and their corresponding cities for the season of 2018-2019, so the data had to be manually collected, and so this was a different kind of difficulty which was faced in the data gathering process.

Data Descriptives

In the data discovery phase, the key variables (the dependent variable, the causal variable, and the conditioning variables) were analyzed. Fig 1 (Appendix) can be used along with Fig 2 (appendix) can be used to discover the variable distributions. The dependent variable, total goals scored (tgc) has a fairly symmetrical distribution. The rain variable distribution indicates that rain was observed on 74% of the observations. The distribution for number of corners (cnr) in a match has a symmetrical distribution. The number of yellow cards (yc) in a match, however, has a slightly skewed distribution with a right tail. Moreover, the mean for red card (rc) is 0.12, indicating that in 12% of the matches, one or more red cards was given. Lastly, the number of shots and the number of shots on target (shots & shots_t) in a match also have symmetrical distributions. At an overall level we can see that mostly the variable distributions are fairly symmetrical and there are barely any extreme values. Therefore, we will not exclude any value from the data and proceed to check for patterns of association.

Checking Patterns of Association

Now that we have an idea about our variable distributions, it is also important to identify the patterns of association between the dependent variable (tgs) and the causal variable (rain) as well as the other conditioning variables. To do this, the method used was a non-parametric lowess curve, which would aid in identifying how the variables are associated with the dependent variable and in what functional form they need to be included in the regressions. The lowess curve for rain vs total goals scored (Appendix : Fig 3a) shows that on average, there is barely any change in the number of goals scored on days where there is rain, as opposed to when it is not raining. We can, however, observe that the maximum number of goals scored is 6 on days when it does not rain, whereas on days where there was rain, the maximum number of goals was 8.The nature of data in this case, limits the analysis that can be done. For instance, we could have used precipitation amounts in absolute terms rather than converting it to a binary variable indicating whether it rained or not.However, since precipitation had a very skewed distribution to the right, with a lot of zero values as well, we could not use log-transformations and hence converted the variable to binary. Similarly, based on the pattern of association displayed by the lowess curve for yellow cards (Appendix: Fig 3a), we can observe a change in convexity at two points, hence for our regressions, we will model yellow cards as a linear spline, with knots at 4 and 6 cards per match. In the same manner, based on the patterns of associations exhibited, we will include total shots in a linear manner whereas shots on target will be modelled as a linear spline with a knot at 17 shots on target in a match (please refer to Appendix: Fig 3b).

Comparing Explanatory Variables

Before we move towards modelling our regressions, it is important to analyze how our independent variables (causal variable as well as the conditioning variables) are correlated with each other. This is basically to check for issues pertaining to multicollinearity. To combat this issue, I constructed a correlation matrix (Appendix: Fig 4) between all of the independent variables.The causal variable had very low correlation with all other independent variables. Two independent variables, shots and shots on target, however, had a somewhat high positive correlation. While at an overall level, no two variables are too strongly correlated, I will still perform some additional tests to make sure that we can eliminate the effect of multicollinearity. This is because it is possible for example, that three variables together could explain a high percentage of the variation in a fourth variable. For this, I constructed multiple linear regression models, where each independent variable was regressed on all other independent variables and the respective variance inflation factor of each variable was calculated. All VIF values were below 2, indicating no concern regarding multicollinearity.

Model Choice

The most appropriate model for this particular analysis has been shown below:

$tgc = \beta_0 + \beta_1 \times rain_{bin} + \beta_2 \times cnr + \beta_3 \times yc_{(yc<4)} + \beta_4 \times yc_{(4>=yc<6)} + \beta_5 \times yc_{(yc>=8)} + \beta_6 \times rc_{(bin)} + \beta_7 \times shots + \beta_8 \times shots.t_{shots.t<17} + \beta_9 \times shots.t_{shots.t>=18}$

The main interpretation for this regression will be regarding the causal variable. Based on the regression output, conditioning the corners, yellow cards, red cards and shots on target in a match, when it rains, the number of goals increases by 0.24 units. In other words, on average, we could expect an additional goal in a match when it rains on roughly 24% of the occasions. As far as the other variables are concerned, we can see that conditioned on the other independent variables in the regression, each additional corner reduces the number of goals by 0.07 units. In more meaningful terms, this would mean that on average, if a game has total 14 corners, that will lead to an additional goal scored. Similarly we can see that when the number of yellow cards in a game is less than 4 or more than 6, each additional yellow card is negatively associated with the number of goals scored in a match. This relationship is more negative for the cases where there are greater than 6 yellow cards. This makes sense because generally in games with intense competition and fewer goals, there is greater frustration amongst the players and as a result more yellow cards are issued to them. Similarly, for the shots_t variable, the coefficient indicates that conditioning upon all independent variables in the regression, when there are less than 17 shots on target, each additional shot on target would lead to a 0.27 units increase in the number of goals scored. A more meaningful interpretation would be that within this category, for every four shots on target roughly, the number of goals is higher by 1 (important to remember that we will condition on all other independent variables). In a similar fashion, the other variables can also be interpreted. Based on the MLR model, we can see that the number of corners in a match, the number of yellow cards in a match (when between 4 and 6), and the number of shots on target in a match (when less than 17), are statistically signficant in terms of their 'effect' on the total number of goals in a match.

$$H_0 := \beta_1 = 0$$
$$H_A := \beta_1 \neq 0$$

Based on the regression output for reg4, we can see that the t-value for the beta coefficient is 1.57, and hence we can say that at a level of significance equal to 0.05, we fail to establish that the coefficient on rain is statistically different from zero as the t value lies within the t-critical values of -2 and 2 (for a 95% confidence interval). When we condition on so many of the variables which are likely to impact the number of goals scored in a match, we get closer to a causal interpretation (this is still not cetris paribus). What we can conclude however is that when conditioned on all other independent variables in the regression, the confidence interval for rain does not include zero, so we can say that rain and the number of goals scored in a match are still positively associated.

Robustness checks and external validity

A total of five models were constructed and tested for this analysis in order to have robust results. Firstly, a regression was run where all variables were included linearly. This was because for all variables the pattern of association with the dependent variable was somewhat linear. In this case the R-squared value was 0.26. In the next regression, column(2), the variable shots was dropped and the same regression otherwise was run. This was done to analyze the impact of any existing multicollinearity. In this case, there was a neglible change in R-squared or the adjusted R-squared. For the third and fourth regressions, the same models were tested as models 1 and 2, but the variable yellow card was included as a linear spline with knots at 4 and 6, whereas shots on target were also modelled as linear splines with a knot at 17. This was done based on the patterns of association observed. We can see that in the case roughly 2 percent greater variation in the dependent variable is explained by these two models. Lastly, due to the pattern of association exhibited between shots on target and the dependent variable, total goals scored per match, an identical regression as regression 4 was run, but shots on target was included as a squared term, rather than as a spline.In this case, the R-squared and the adjusted R-squared were both lower and hence this was not our preferred model choice.

It is important to gauge whether the general pattern that is represented by our data, would hold true for the general pattern in the situation that we truly care about, which is that at a more global scale, if rainfall affects the number of goals in a game.Therefore,in regards to external validity of our results, we can examine it based on three criterion. Firstly, for this particular modelling, where rain is included as a binary variable, we expect results to be stable in time. Rain patterns in the UK are fairly similar each year. Even days, we very extreme rainfall, the number of goals in that match were not exceptionally high, so our conclusions are not susceptible to extreme values. It must be noted however that only a handful of these extreme values took place so we cannot be certain that this will be the case. Nonetheless, we can expect stability in time. We can also expect stability in subgroups. For instance, if we run the same models on a women's football league, we would expect similar results because there is no particular impact that rainfall would have on a woman's performance on a rainy day as opposed to a man's. It would be interesting though, to compare leagues with higher percentage of older players, like MLS league and the Indian and Chinese football leagues which focus on recruiting football legends to promote their leagues. We could speculate that older players might be prone to more errors on a rainy day and that would lead to more goals. Finally, to gauge whether the results would be stable in space is a tougher task. However, we may exhibit a different pattern in lets say the German league or the French league where even though there may be fewer rainy days, but with greater amounts of rainfall which would lead to perhaps different results. Therefore, the external validity for our results in terms of space, would likely be low.

Causal Interpretation

As mentioned above, by conditioning on as many variables which affect our dependent variables as possible, the aim was to bring our analysis close to a causal one. While there was an initial expectation that rain would cause an increase in the number of goals in the game, based on the results, we can conclude that this is not the case. While the two variables may have a positive correlation, we have failed to identify a somewhat causal relationship between the two variables. However, we can see that through our robustness checking and including the correct functional forms in the regression, the coefficient on the causal variable increases from 0.1934 to 0.2376(in case of the chosen model). It is possible that if we are able to add more conditioning variables, we may be able to establish a somewhat causal link between rain and the total number of goals in a match.

Conclusion

Due to the nature of research question, there is no policy recommendation per se. However, some key findings have been gathered from this exercise. There are some concerns over the data quality as there can be discrepancies between the weather station and the stadium. Also, the data gives the total rainfall expected for the day and not the rainfall experienced during matchtime. Therefore, there is a measurement error associated with our causal variable. It is possible that when exactly measured for the location as well as the time, we may observe a stronger association between rainfall and the number of goals scored in a game. By overcoming the measurement error, we can also move closer towards discussing causality between the two variables.

Appendix

Fig 1: Data Summary Table

	Min	Max	P25	Median	P75	Mean	SD	P95	N
tgc	0.00	8.00	2.00	3.00	4.00	2.82	1.60	6.00	380
rain	0.00	1.00	0.00	1.00	1.00	0.74	0.44	1.00	380
cnr	2.00	21.00	8.00	10.00	13.00	10.28	3.18	15.00	380
yc	0.00	9.00	2.00	3.00	4.00	3.26	1.84	6.05	380
rc	0.00	1.00	0.00	0.00	0.00	0.12	0.32	1.00	380
shots	7.00	32.00	18.00	21.00	23.00	20.53	4.27	28.00	380
shots_t	2.00	21.00	8.00	10.00	13.00	10.65	3.26	16.00	380

Fig 2: Variable Distributions

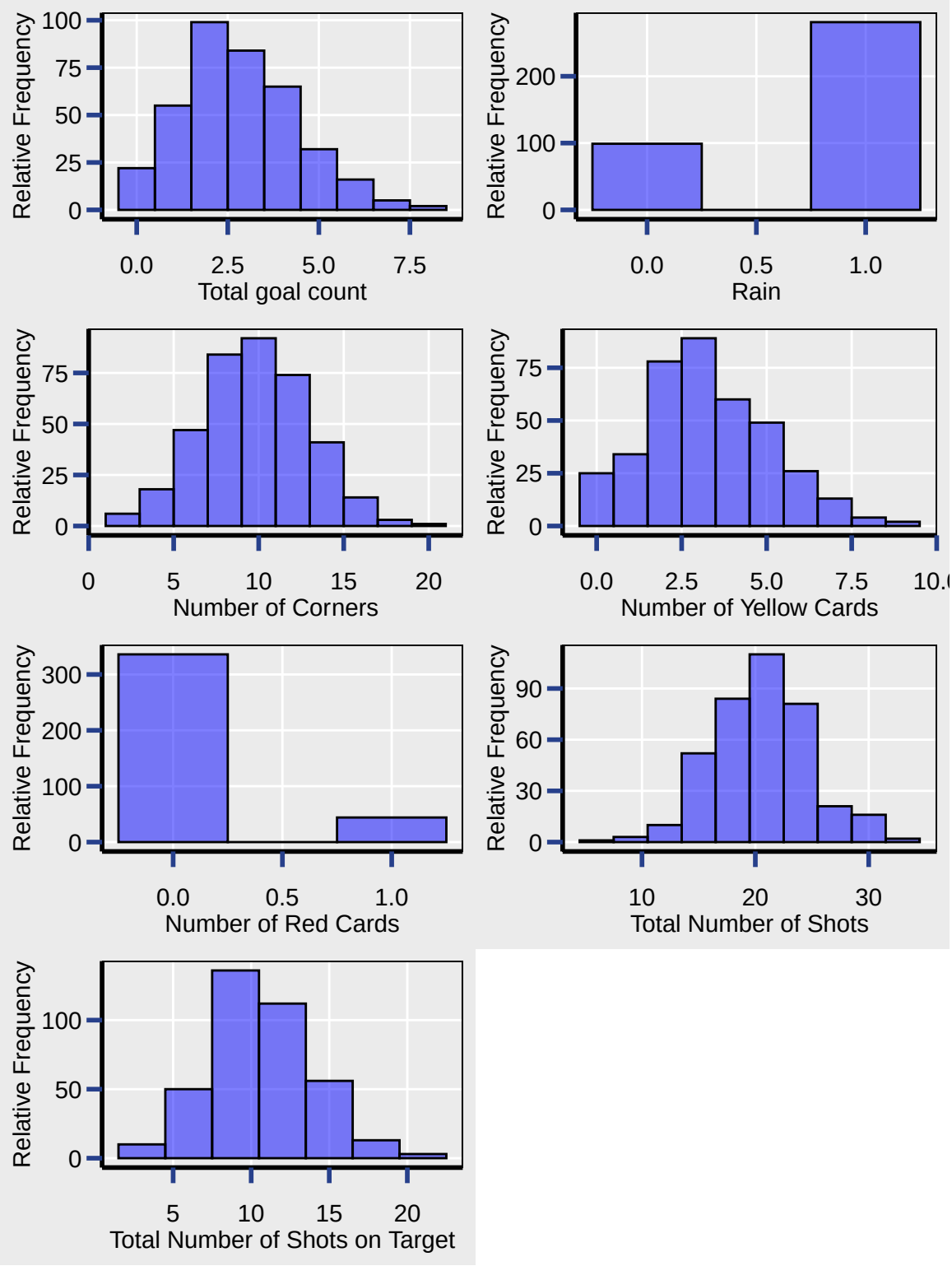

Fig 3a: Patterns of association

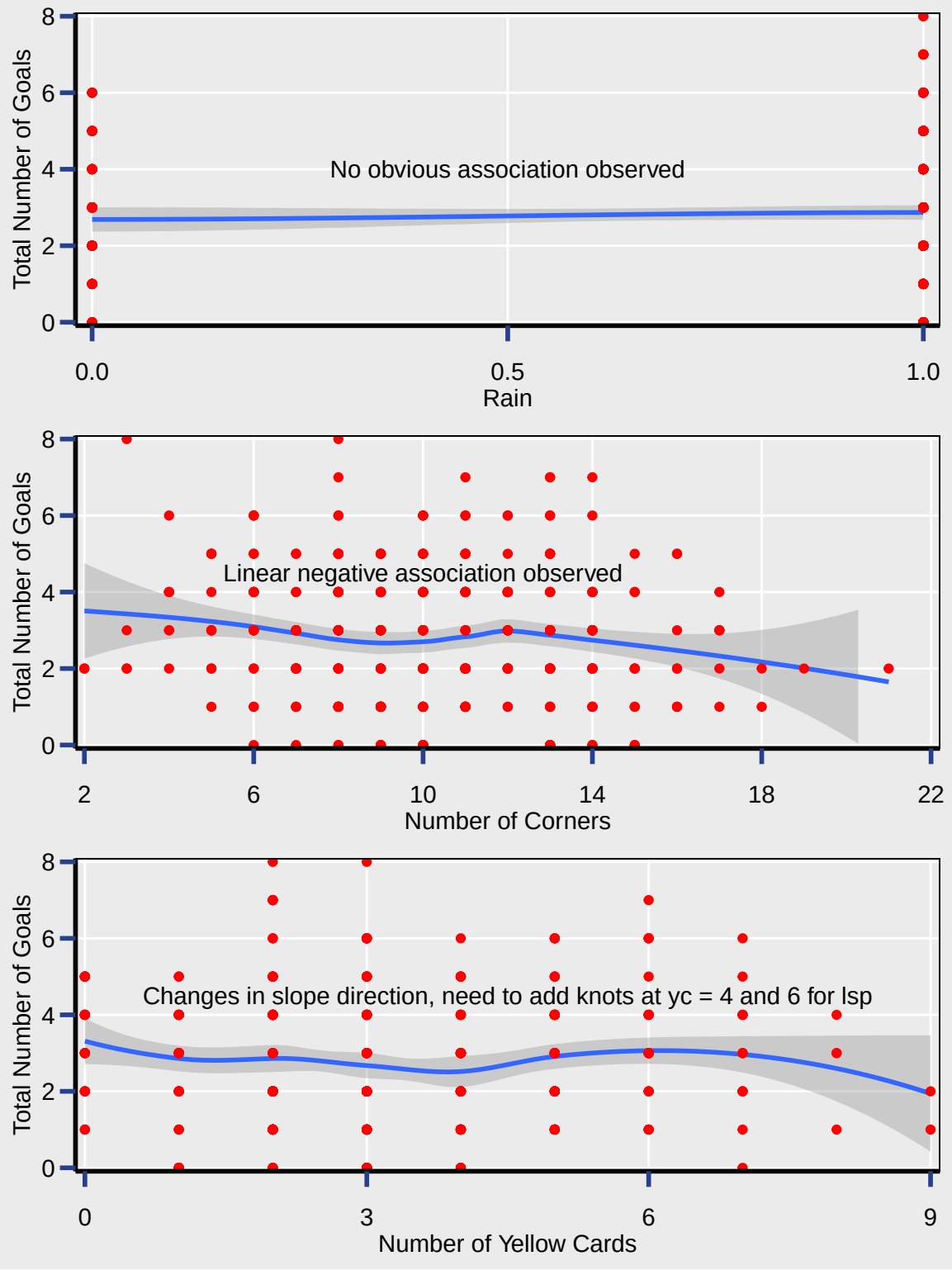

Fig 3b: Patterns of association

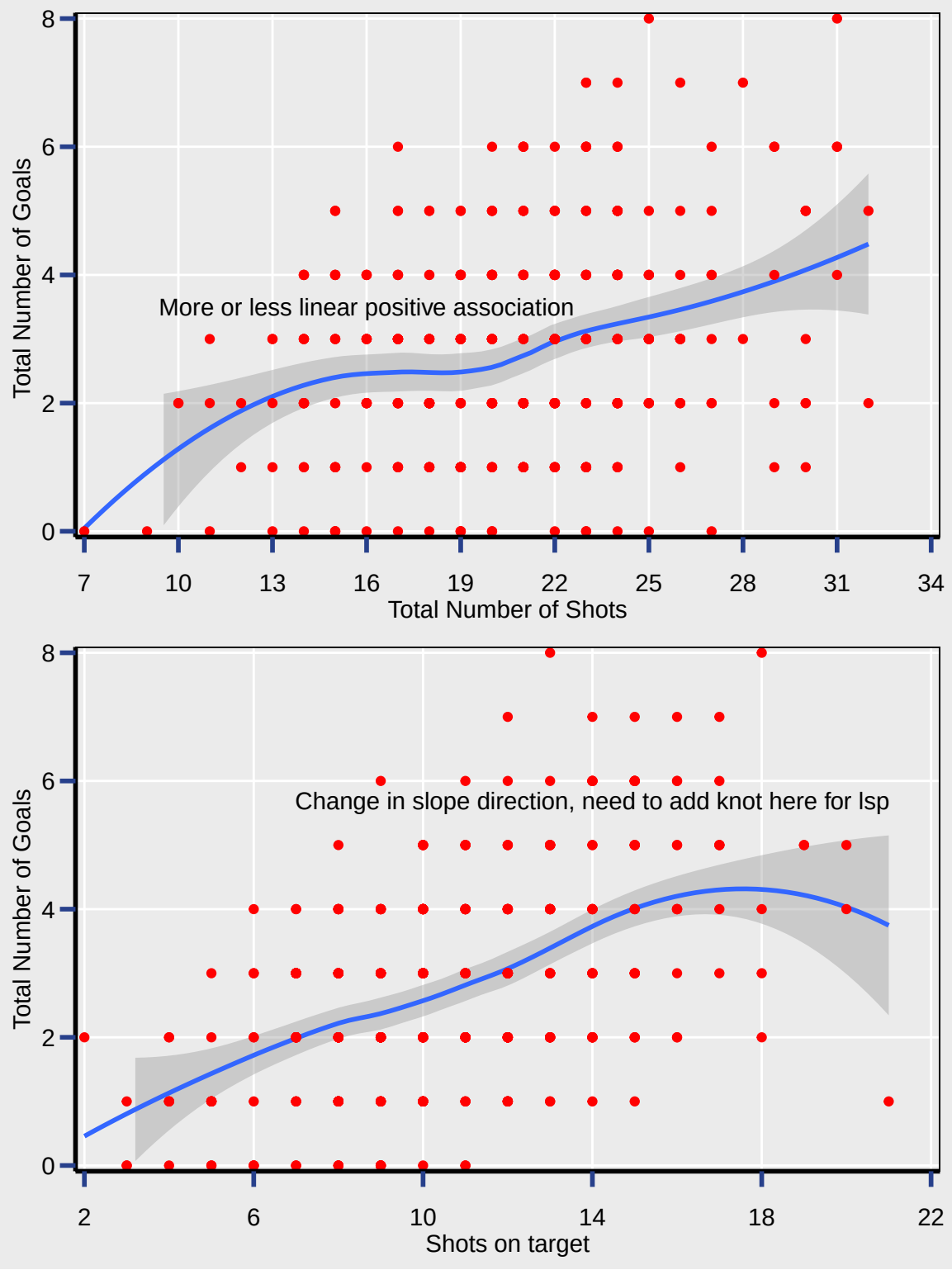

Fig 4: Correlation Matrix

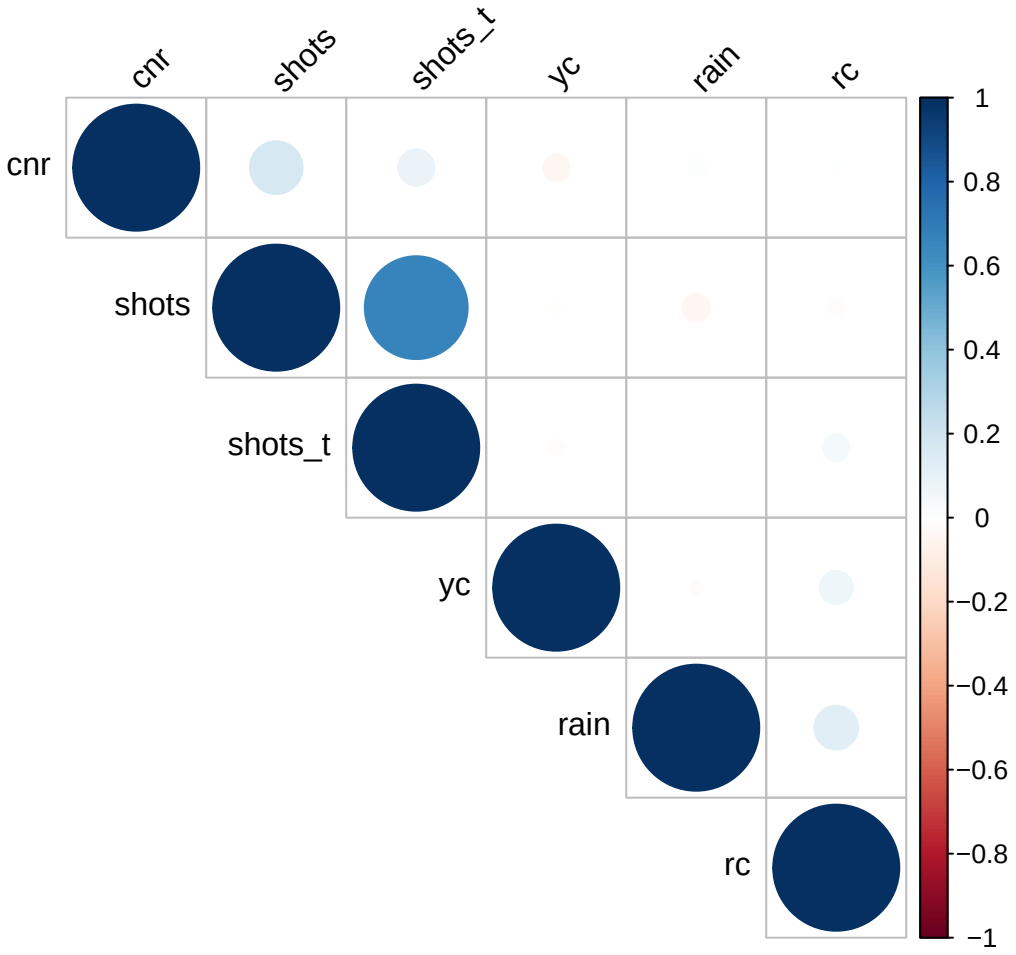

Fig 5: Comparing Regression Outputs

	reg1	reg2	reg3	reg4	reg5
Dependent Var.:	tgc	tgc	tgc	tgc	tgc
(Intercept)	0.8640* (0.4270)	0.7557* (0.3601)	0.9388* (0.4288)	0.7950* (0.3645)	2.379*** (0.4386)
rain	0.1957 (0.1547)	0.2000 (0.1532)	0.2322 (0.1527)	0.2376 (0.1513)	0.1934 (0.1521)
cnr	-0.0739*** (0.0220)	-0.0757*** (0.0213)	-0.0720** (0.0223)	-0.0743*** (0.0216)	-0.0730** (0.0223)
yc	0.0077 (0.0362)	0.0075 (0.0363)			
rc	-0.0227 (0.2143)	-0.0170 (0.2146)	-0.0426 (0.2120)	-0.0351 (0.2122)	0.0134 (0.2128)
shots	-0.0109 (0.0218)		-0.0141 (0.0219)		-0.0035 (0.0224)
shots_t	0.2604*** (0.0289)	0.2511*** (0.0238)			
lspline(yc,c(4,6))1			-0.1031 (0.0625)	-0.1019 (0.0626)	-0.1115. (0.0638)
lspline(yc,c(4,6))2			0.3146* (0.1329)	0.3104* (0.1333)	0.3861** (0.1336)
lspline(yc,c(4,6))3			-0.3086 (0.2086)	-0.3063 (0.2083)	-0.4045* (0.2008)
lspline(shots_t,17)1			0.2788*** (0.0285)	0.2666*** (0.0232)	
lspline(shots_t,17)2			-0.2410 (0.2793)	-0.2520 (0.2793)	
shots_t_sq					0.0107*** (0.0016)
S.E. type	Heteroskedast.-rob.	Heteroskedast.-rob.	Heteroskedas.-rob.	Heteroskedast.-rob.	Heteroskedas.-rob.
Observations	380	380	380	380	380
R2	0.27451	0.27406	0.29821	0.29745	0.27309
Adj. R2	0.26284	0.26435	0.28114	0.28230	0.25741

Fig 6: Summary Output for chosen model

```
## OLS estimation, Dep. Var.: tgc
## Observations: 380
## Standard-errors: Heteroskedasticity-robust
##                          Estimate Std. Error   t value   Pr(>|t|)
## (Intercept)              0.795009   0.364469  2.181279 0.02979007 *
## rain                     0.237554   0.151334  1.569732 0.11732993
## cnr                     -0.074326   0.021638 -3.435016 0.00065952 ***
## lspline(yc, c(4, 6))1 -0.101867   0.062621 -1.626728 0.10464396
## lspline(yc, c(4, 6))2  0.310443   0.133301  2.328886 0.02040154 *
## lspline(yc, c(4, 6))3 -0.306339   0.208259 -1.470950 0.14215221
## rc                      -0.035091   0.212192 -0.165375 0.86873925
## lspline(shots_t, 17)1  0.266643   0.023169 11.508410  < 2.2e-16 ***
## lspline(shots_t, 17)2 -0.251951   0.279266 -0.902193 0.36753974
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.34043   Adj. R2: 0.2823
```