Final Project - Winning Factors for Home Teams in Soccer Games

Liyao Zhang, Haochun Wang

2019/11/26

Abstract

In this data analysis, we are exploring what factors are affecting the odds of winning for the home teams in soccer games worldwide. In the data analysis, we are fitting multiple logistic regression models and compare the models using likelihood ratio test, AIC, BIC and McFadden r-squared. We also check the linearity assumption for each variable in our models.

We both fit models for the whole dataset and the five tiers of the teams that are divided by the criteria of team strength. For the whole dataset in general, the "best" model is the multiple logistic model using the team strength of both teams and the importance of the game for both teams. For different tiers of teams, some best models are using the team strength of both teams and the importance of the game for both teams and some best models are using only the team strength of both teams.

We find that generally, the (log) odds of winning for the home team is increasing as the home team is stronger and is decreasing as the opponent is stronger; the (log) odds of winning for the home team is increasing as the importance of the game is higher and is decreasing as the opponent has higher importance of the game. This coincides with our common knowledge about soccer.

Introduction

Soccer is one of the most popular sports in the world, especially in Europe and South America. After decades of development, billions of dollars are generated in this industry annually. From commercial advertisements and shares in broadcast rights to soccer equipment and even betting companies, millions of people make a living from soccer. But ultimately, the nature of the industry is soccer matches that take place everyday around the globe. Therefore, we are interested in the games themselves.

Soccer game is well-known for its uncertainty as there are too many factors for people to determine what the final result would be, which makes it even more fascinating. As a result, we decided to study what factors make home team win in a soccer game. After a brief search, we found an online dataset from FiveThirtyEight called "Match-by-match SPI ratings and forecasts back to 2016". This dataset is actually part of the project "Club Soccer Predictions" from this data analysis website and updated after each game is played. Two featured variables in the dataset are SPI ratings and match importance. They are both calculated under the unique algorithms provided by FiveThirtyEight. For SPI, it is an estimate of team strength that combines offensive and defensive ratings. And for importance, it represents how the outcome of a game would change the statistical outlook of each team.

Methods

The dataset we are using in this data analysis is "spi_matches" from package "fivethirtyeight". There are 32290 observations in total (till 10 p.m, Dec.5th, this dataset is updating daily) and each observation refers to a single soccer match. There are 16 effective variables in total (actually there exist 22, but some of the variables are not provided useful information by the provider of the dataset): *date* refers to the date that the match took place, league_id refers to the numerical identifier of the league within which the match was played, *league*

refers to the league name, *team*1 refers to one team that participated in the match, *team*2 refers to the other team that participated in the match, *spi* refers to the SPI score of team1 (i.e: Team strength of team1). *spi*2 refers to the SPI score of team2, *prob*1 refers to the probability that team1 would have won the match, *prob*2 refers to the probability that team2 would have won the match, *probtie* refers to the probability that the match would have resulted in a tie, proj_score1 refers to the predicted number of goals that team1 would have scored, proj_score2 refers to the predicted number of goals that team2 would have scored, *importance*1 refers to the importance of the match for team1, *importance*2 refers to the importance of the match for team2, *score*1 refers to the number of goals that team1 scored and *score*2 refers to the number of goals that team2 scored. By general knowledge in soccer matches, *team*1 refers to home team and *team*2 refers to away team.

There are a lot of missing data in the dataset, which are referring to the future matches that have not happened and the matches that the importances for both teams are not able to be evaluated. To avoid possible consequences of influenced by some mishandling of those missing data, we choose the relatively safe way: we clean them up and our dataset that will be used contains 23215 observations.

The next step we do is clean and add some of the variables. We are interested in the discussing the "home advantages" in soccer games, so we add a new variable of *result* which is assigned 1 if the home team wins the game and assigned 0 if they ties or loses the game. We set that to be our response variable. Then, we take out the variables of *date*, league_id, *league*, *team*1, *team*2 and other variables that the information are not provided by the provider of the dataset.

In this data analysis, we firstly fit three general models that work for the whole final dataset and then we split the home teams into five tiers of strength (by 0.2, 0.4, 0.6, 0.8) and fit new models for those different tiers of home teams and their matches. Since the response variable, *result*, is binary, we are using three (multiple) logistic regression models. No matter for the

general models and the models for each tier, the first model we use is using $spi2$ only as our explanatory variable, since the game outcome highly depends on how good the opponent team is; the second model we use is using $spi1$ and $spi2$ together, since how good the home team is also influence the outcome of games a lot; the third model we use is using $spi1$, $spi2$, $importance1$ and $importance2$ together, since the importance for the two teams in the match also are important for the final game result: for example, if a team does not think the game is important, the team is likely to play lazily on the court.

For the three models (for general and for each tier), since they are (multiple) logistic regression models, the assumptions are independence, randomness and linearity. For independence and randomness, we could not really have formal kind of testing. However, soccer matches and the outcomes of them, by common knowledge, could be regarded as random and independent. For linearity, we could plot the empirical logits and check if linearity is met or not.

When doing model comparisons, we would apply likelihood ratio test to see if adding variables will get us a better model or not. We will also use AIC, BIC and McFadden r-squared to compare the models: with lower AIC and BIC and higher McFadden r-squared values, the model is better.

Using the criteria above, our final model is selected by the information provided by the likelihood ratio test and optimally with lowest AIC and BIC and highest McFadden r-squared.

Results:

In our data analysis, all analysis was completed using R markdown and we have used packages 'knitr', 'fivethirtyeight', 'tidyverse', 'broom', 'readr' and 'dplyr'.

Firstly we check for the assumption of linearity: (note that independence and randomness could not be formally checked, but for soccer matches, they could be regarded as holding because of nature of soccer matches)

Table 1: Quantiles and means

spi1_cat	mean(spi1)
quantile 1	20.30037
quantile 2	34.58922
quantile 3	44.51549
quantile 4	56.73125
quantile 5	74.55088

Table 2: Empirical logits and other information

spi1_cat	result	n	prop	odds	log_odds
quantile 1	1	2148	0.3901199	0.6396665	-0.4468084
quantile 2	1	2207	0.4234459	0.7344426	-0.3086434
quantile 3	1	2259	0.4394942	0.7841027	-0.2432152
quantile 4	1	2364	0.4612683	0.8562115	-0.1552378
quantile 5	1	2937	0.5638318	1.2926937	0.2567282

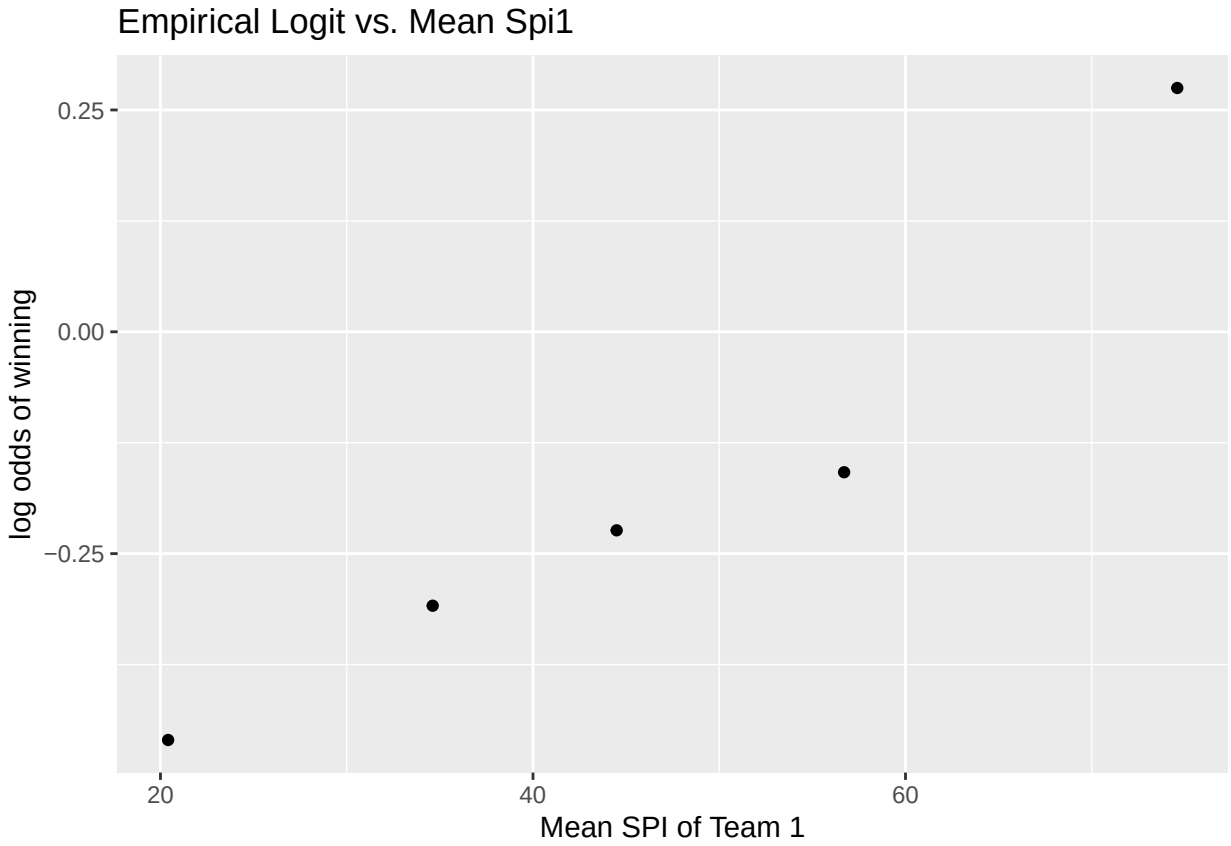


Table 3: Quantiles and means

spi2_cat	mean(spi2)
quantile 1	20.28935
quantile 2	34.51455
quantile 3	44.40641
quantile 4	56.62678
quantile 5	74.52770

Table 4: Empirical logits and other information

spi2_cat	result	n	prop	odds	log_odds
quantile 1	1	2756	0.5018208	1.0073099	0.0072834

spi2_cat	result	n	prop	odds	log_odds
quantile 2	1	2548	0.4880291	0.9532361	-0.0478927
quantile 3	1	2405	0.4670810	0.8764577	-0.1318668
quantile 4	1	2335	0.4541918	0.8321454	-0.1837481
quantile 5	1	1871	0.3605704	0.5638939	-0.5728891

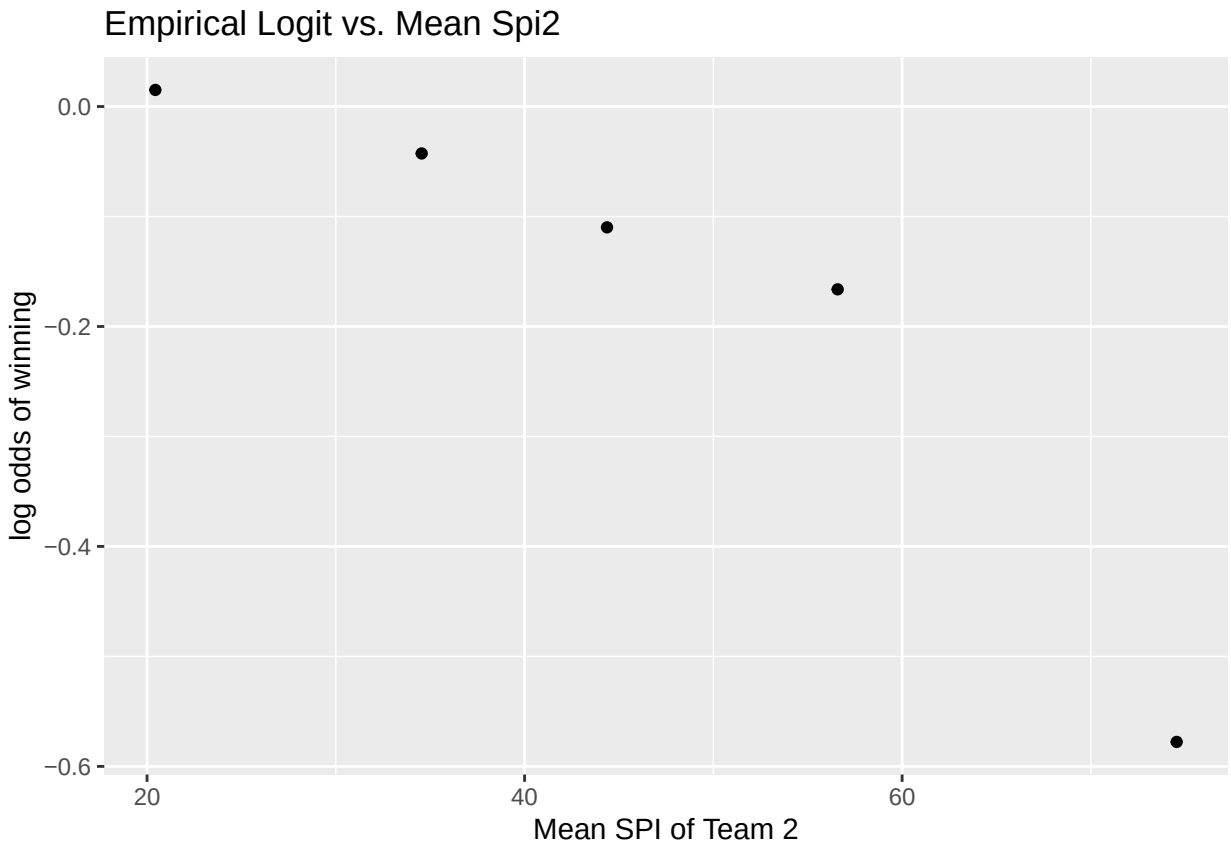


Table 5: Quantiles and means

importance1_cat	mean(importance1)
quantile 1	3.151754
quantile 2	18.735077
quantile 3	34.974006
quantile 4	67.732012

Table 6: Empirical logits and other information

importance1_cat	result	n	prop	odds	log_odds
quantile 1	1	2846	0.4360349	0.7731595	-0.2572700
quantile 2	1	2924	0.4470948	0.8086283	-0.2124159
quantile 3	1	2681	0.4130971	0.7038593	-0.3511768
quantile 4	1	3464	0.5220799	1.0923999	0.0883770

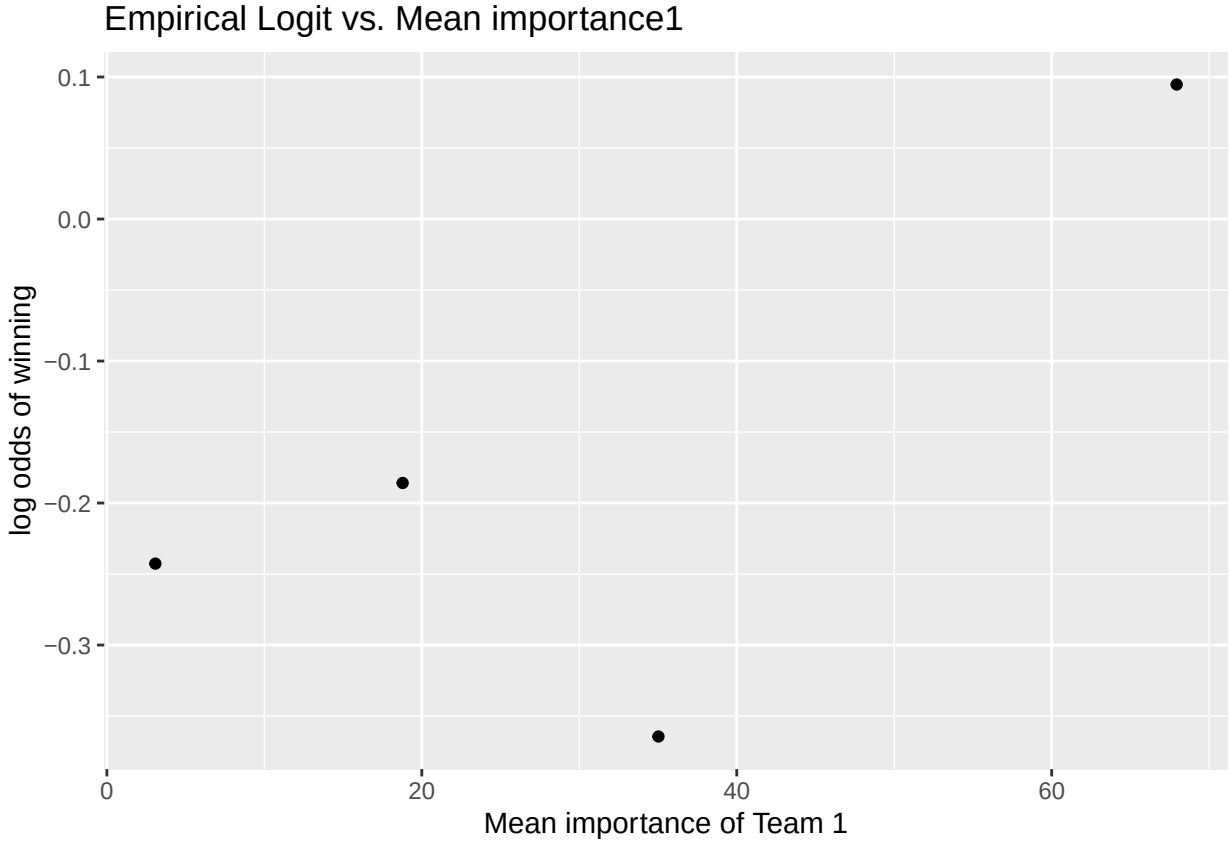


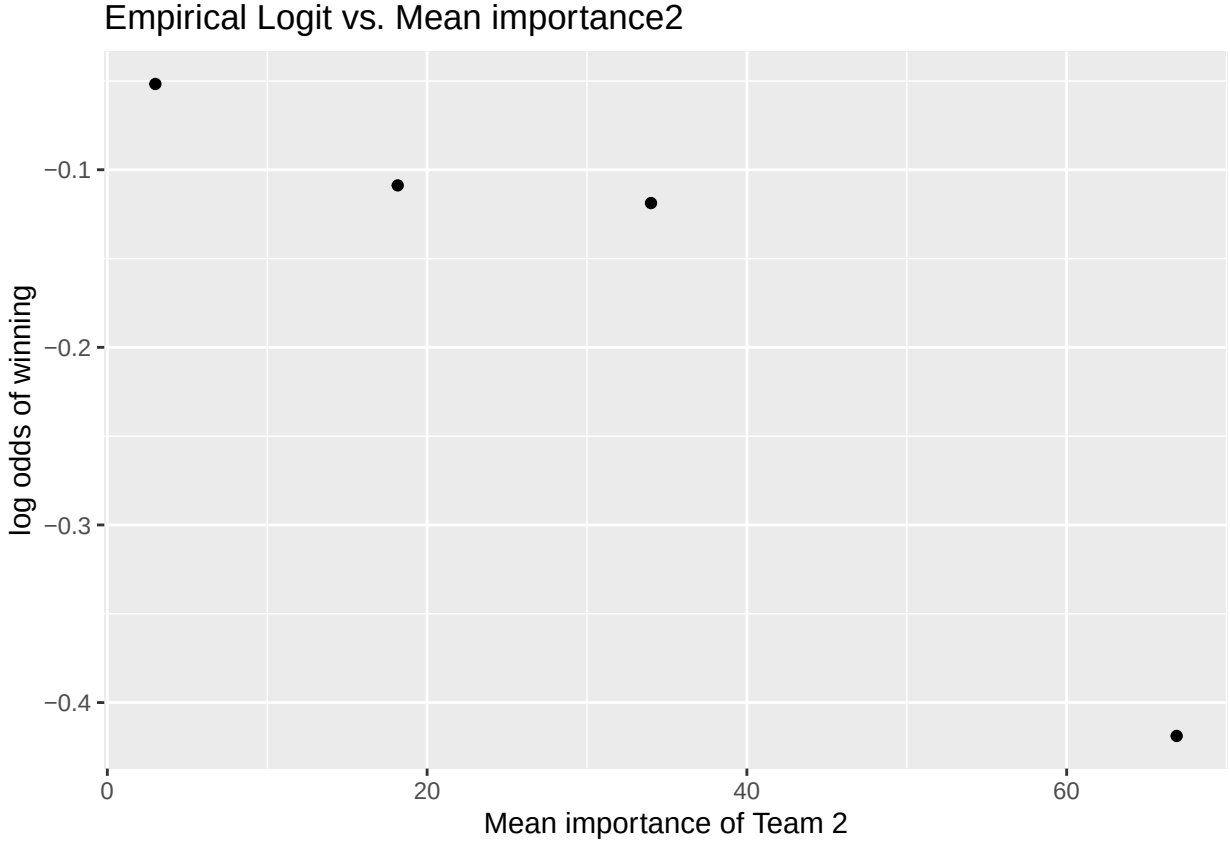
Table 7: Quantiles and means

importance2_cat	mean(importance2)
quantile 1	3.110691
quantile 2	18.126435

importance2_cat	mean(importance2)
quantile 3	33.951208
quantile 4	66.712931

Table 8: Empirical logits and other information

importance2_cat	result	n	prop	odds	log_odds
quantile 1	1	3181	0.4830676	0.9344888	-0.0677556
quantile 2	1	3039	0.4700696	0.8870403	-0.1198649
quantile 3	1	3064	0.4714571	0.8919942	-0.1142957
quantile 4	1	2631	0.3960560	0.6557827	-0.4219259



Empirical logits plots are made for each explanatory variable in our model. For both $Spi1$

and $Spi2$, the points are linear for the first four quantiles. But the slope suddenly becomes steeper for the best teams in the fifth quantile. We can conclude that SPI variables slightly violated the linearity assumption. However, $importance1$ and $importance2$ don't meet the assumption since the log odds of winning unexpectedly drops for some quantiles.

Then we start to fit our models. Firstly, we fit a model for our entire dataset, with response variable of result (binary, 1 refers to win for the home team and 0 refers to not win for the home team). Here, the three models we would fit are:

Model1: $log(\hat{Odds for HomeTeamWin}) = \hat{\beta}_0 + \hat{\beta}_1 * spi2$

Model2: $log(\hat{Odds for HomeTeamWin}) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2$

Model3: $log(\hat{Odds for HomeTeamWin}) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2 + \hat{\beta}_3 * importance1 + \hat{\beta}_4 * importance2$

Table 9: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.3060994	0.0319631	9.57666	0	0.2434860	0.3687824
spi2	-0.0106911	0.0006491	-16.47070	0	-0.0119648	-0.0094203

Table 10: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-17910.66	35825.32	35841.67	35821.32	26190

```
## [1] 5.478076e-57
```

Table 11: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2613003	0.0351626	-7.431208	0	-0.3302519	-0.1924130
spi1	0.0504740	0.0010997	45.896262	0	0.0483267	0.0526377
spi2	-0.0492123	0.0011083	-44.402699	0	-0.0513930	-0.0470484

Table 12: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-16683.64	33373.28	33397.8	33367.28	26189

```
## [1] 0
```

Table 13: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2576867	0.0370254	-6.959719	0.0000000	-0.3302834	-0.1851419
spi1	0.0491036	0.0011423	42.988263	0.0000000	0.0468727	0.0513504
spi2	-0.0477994	0.0011485	-41.617500	0.0000000	-0.0500587	-0.0455563
importance1	0.0021015	0.0005652	3.718455	0.0002004	0.0009942	0.0032097
importance2	-0.0023301	0.0005781	-4.030915	0.0000556	-0.0034639	-0.0011979

Table 14: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-16672.64	33355.28	33396.15	33345.28	26187

```
## [1] 2.938577e-05
```

```
## [1] 0.007906289
```

```
## [1] 0.0771816
```

```
## [1] 0.0778339
```

Model1 uses only $spi2$ as explanatory variable. Here, we use likelihood ratio test to compare this model with the intercept-only model: firstly set up null hypothesis: $\hat{\beta}_1 = 0$ and alternative hypothesis $\hat{\beta}_1 \neq 0$. Then, we put drop in deviance and change in degrees of freedom, we get p-value equals to 5.478076e-57, which is smaller than 0.05. Which means, null hypothesis is rejected and $\hat{\beta}_1 \neq 0$.

Model2 use $spi1$ and $spi2$ at the same time. Here, we apply likelihood ratio test to compare those models at first: our null hypothesis is $\hat{\beta}_2 = 0$ and the alternative hypothesis is $\hat{\beta}_2 \neq 0$. By putting in the drop in deviance and change in degrees of freedom, we get p-value equals to approximately 0, which is smaller than 0.05. So, it indicates that the null hypothesis is rejected, so $\hat{\beta}_2 \neq 0$. Then we look at the AIC and BIC: we note that no matter for AIC and BIC, model2 has smaller values. So, we may conclude that model2 is better than model1.

Then we fit our model3 and compare it with model2: still we apply the likelihood ratio test at first. Firstly we set the null hypothesis to be $\hat{\beta}_3 = \hat{\beta}_4 = 0$ and the alternative hypothesis to be $\hat{\beta}_3 \neq 0$ or $\hat{\beta}_4 \neq 0$. By the likelihood ratio test, the p-value is 2.938577e-05, which is smaller than 0.05. That means, null hypothesis is rejected. Then we look at AIC and BIC, we find model3 has smaller AIC and BIC values. Finally, we compare the Mcfadden r-squared: we see that the model3 gets the highest McFadden r-squared value, 0.07787081. So, here, we pick model3 as our "best" model. It has AIC of 29528.51, BIC of 29568.77 and McFadden r-squared of 0.0778339.

By looking at the table, our model3 is: $log(Odds for \hat{Home} Team Win)$ = -0.2542666 + 0.0495105 * $spi1$ + -0.0479789 * $spi2$ + 0.0020808 * $importance1$ + -0.0025197 * $importance2$.

The intercept in the model is $\hat{\beta}_0 = -0.2542666$, with 95% CI of (-0.3324844, -0.1761076), which means as all other variables equal to 0, the expected log-odds for the home team to win the game is -0.2542666. The coefficient of $spi1$ in the model is 0.0495105, with 95% CI of (0.0471486, 0.0518905), which means holding other variables constant, one-unit increase in $spi1$ increases the expected log-odds of home team to win the game by 0.0495105. The coefficient of $spi2$ in the model is -0.0479789, with 95% CI of (-0.0503698, -0.0456062), which means holding other variables constant, one-unit increase in $spi2$ decreases the expected log-odds of home team to win the game by 0.0479789. The coefficient of $importance1$ is 0.0020808 with 95% CI of (0.0008989, 0.0032635), which means holding other variables constant, one-unit increase in $importance1$ increases the expected log-odds of home team to win the game by 0.0020808. The coefficient of $importance2$ is -0.0025197 with 95% CI of (-0.0037298, -0.0013115), which means holding other variables constant, one-unit increase in $importance2$ decreases the expected log-odds of home team to win the game by 0.0025197.

By the four plots of empirical logits, we see that the linearity assumptions are not quite satisfied since the points are not approximately lying on a line. However, among all the models, model3 is the "best" according to previous analysis.

Nextly, we divide the home teams into five tiers according to the team strength and fit new models for those different tiers of home teams and their matches.

Tier1: the weakest teams, $spi1 < 29.13$

```
## [1] 1.163229e-36
```

```
## [1] 5.073025e-15
```

```
## [1] 0.04836398
```

Table 15: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2198635	0.1048973	-2.0959879	0.0360833	-0.4258856	-0.0146128
spi1	0.0447427	0.0051076	8.7599708	0.0000000	0.0347663	0.0547906
spi2	-0.0429159	0.0029239	-14.6776843	0.0000000	-0.0486917	-0.0372287
importance1	0.0010954	0.0014268	0.7677564	0.4426319	-0.0017031	0.0038913
importance2	-0.0030213	0.0013412	-2.2527236	0.0242766	-0.0056587	-0.0004001

Table 16: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7364.844	5505	-3551.211	7112.421	7145.489	7102.421	5501

```
## [1] 0.03668244
```

Still, we fit three models here with model1 one using $spi2$ only, model2 using $spi1$ and $spi2$ and model3 using $spi1$, $spi2$, $importance1$ and $importance2$:

Model1: $log(\hat{Odds for Tier1 Home Team Win}) = \hat{\beta}_0 + \hat{\beta}_1 * spi2$

Model2: $log(\hat{Odds for Tier1 Home Team Win}) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2$

Model3: $log(\hat{Odds for Tier1 Home Team Win}) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2 + \hat{\beta}_3 * importance1 + \hat{\beta}_4 * importance2$

Model1 uses only $spi2$ as explanatory variable. Here, we use likelihood ratio test to compare this model with the intercept-only model: firstly set up null hypothesis: $\hat{\beta}_1 = 0$ and alternative hypothesis $\hat{\beta}_1 \neq 0$. Then, we put drop in deviance and change in degrees of freedom, we get p-value equals to 1.163229e-36, which is smaller than 0.05. Which means, null hypothesis is

rejected and $\hat{\beta}_1 \neq 0$.

Model2 use $spi1$ and $spi2$ at the same time. Here, we apply likelihood ratio test to compare those models at first: our null hypothesis is $\hat{\beta}_2 = 0$ and the alternative hypothesis is $\hat{\beta}_2 \neq 0$. By putting in the drop in deviance and change in degrees of freedom, we get p-value equals to 5.073025e-15, which is smaller than 0.05. So, it indicates that the null hypothesis is rejected, so $\hat{\beta}_2 \neq 0$.

Then we fit our model3 and compare it with model2: still we apply the likelihood ratio test at first. Firstly we set the null hypothesis to be $\hat{\beta}_3 = \hat{\beta}_4 = 0$ and the alternative hypothesis to be $\hat{\beta}_3 \neq 0$ or $\hat{\beta}_4 \neq 0$. By the likelihood ratio test, the p-value is 0.04836398, which is smaller than 0.05. That means, null hypothesis is rejected.

Comparing the three models, we notice that the model3 has lowest AIC but model2 has lowest BIC. Here, we still pick model3 to be the "best" model because of the result of the likelihood ratio test and its highest McFadden r-squared value of 0.03668244. In summary, model3 has AIC of 5977.392, BIC of 6009.605 and McFadden r-squared value of 0.03668244.

By looking at the table, our model3 is: $log(Odds for Tier\hat{1} Home Team Win)$ = -0.1461877 + 0.0424858 * $spi1$ + -0.0430621 * $spi2$ + 0.0006065 * $importance1$ + -0.0036273 * $importance2$. The intercept in the model is $\hat{\beta}_0 = -0.1461877$, with 95% CI of (-0.3764947, 0.0832603), which means as all other variables equal to 0, the expected log-odds for the home team to win the game is -0.1461877. The coefficient of $spi1$ in the model is 0.0424858, with 95% CI of (0.0316373, 0.0534193), which means holding other variables constant, one-unit increase in $spi1$ increases the expected log-odds of home team to win the game by 0.0424858. The coefficient of $spi2$ in the model is -0.0430621, with 95% CI of (-0.0492953, -0.0369315), which means holding other variables constant, one-unit increase in $spi2$ decreases the expected log-odds of home team to win the game by 0.0430621. The coefficient of $importance1$ is 0.0006065 with 95% CI of (-0.0024770, 0.0036848), which means holding other variables constant, one-unit increase in $importance1$ increases the expected log-odds of home team to

win the game by 0.0006065. The coefficient of *importance*2 is -0.0036273 with 95% CI of (-0.0065552, -0.0007216), which means holding other variables constant, one-unit increase in *importance*2 decreases the expected log-odds of home team to win the game by 0.0036273.

By the four plots of empirical logits, we see that the linearity assumptions are not quite satisfied since the points are not approximately lying on a line. However, among all the models in this tier, model3 is the "best" according to previous analysis.

Tier2: weak teams, *spi*1 between 29.13 and 39.6

```
## [1] 1.089375e-54
```

```
## [1] 1.932979e-07
```

```
## [1] 0.2047429
```

Table 17: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.4181164	0.3306271	-1.264616	0.2060089	-1.0665301	0.2296661
spi1	0.0493436	0.0098368	5.016243	0.0000005	0.0300974	0.0686618
spi2	-0.0430006	0.0025815	-16.656966	0.0000000	-0.0481022	-0.0379812

Table 18: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7102.704	5211	-3394.334	6794.668	6814.344	6788.668	5209

```
## [1] 0.04261846
```

Still, we fit three models here with model1 one using *spi*2 only, model2 using *spi*1 and *spi*2

and model3 using $spi1$, $spi2$, $importance1$ and $importance2$:

Model1: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi2$

Model2: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2$

Model3: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2 + \hat{\beta}_3 * importance1 + \hat{\beta}_4 * importance2$

We apply the similar process as what we did in the tier1: by likelihood ratio tests between model1 and intercept only model, we get p-value equals to 1.089375e-54 < 0.05, which means $\hat{\beta}_1 \neq 0$; by likelihood ratio tests between model2 and model1, we get p-value equals to 1.932979e-07 < 0.05, which means $\hat{\beta}_2 \neq 0$; by likelihood ratio tests between model3 and model2, we get p-value equals to 0.2047429, which is larger than 0.05, so that indicates $\hat{\beta}_3 = 0$ or $\hat{\beta}_4 = 0$. By AIC and BIC, we notice that model2 has lowest BIC but model3 has lowest AIC. However, by the likelihood ratio test result, we believe model2 is the "best" model here. Here, model2 has AIC of 6063.601, BIC of 6082.93 and McFadden r-squared of 0.04261846.

By looking at the table, our model2 is: $log(Odds for Tier\hat{2} Home Team Win)$ = -0.6018630 + 0.0540566 * $spi1$ + -0.0423386 * $spi2$. The intercept has 95% CI of (-1.2900333, 0.0853970); the coefficient of $spi1$ has 95% CI of (0.0336511, 0.0745473); the coefficient of $spi2$ has 95% CI of (-0.0477681, -0.0370014). Those coefficients could be interpreted similarly like the previous models.

Tier3: moderate teams: $spi1$ between 39.6 and 50.084

```
## [1] 2.922072e-70
```

```
## [1] 8.050011e-08
```

```
## [1] 0.06610572
```

Table 19: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.6843993	0.4352045	-1.572592	0.1158132	-1.5376593	0.1685266
spi1	0.0538281	0.0099320	5.419651	0.0000001	0.0343911	0.0733290
spi2	-0.0438223	0.0024769	-17.692466	0.0000000	-0.0487175	-0.0390065

Table 20: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7057.831	5144	-3349.01	6704.021	6723.658	6698.021	5142

```
## [1] 0.05371319
```

Model1: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi2$

Model2: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2$

Model3: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2 + \hat{\beta}_3 * importance1 + \hat{\beta}_4 * importance2$

We apply the similar process as what we did in the tier1 and tier2: by likelihood ratio tests between model1 and intercept only model, we get p-value equals to 2.922072e-70 < 0.05, which means $\hat{\beta}_1 \neq 0$; by likelihood ratio tests between model2 and model1, we get p-value equals to 8.050011e-08 < 0.05, which means $\hat{\beta}_2 \neq 0$; by likelihood ratio tests between model3 and model2, we get p-value equals to 0.06610572, which is larger than 0.05, so that indicates $\hat{\beta}_3 = 0$ or $\hat{\beta}_4 = 0$. By AIC and BIC, we notice that model2 has lowest BIC but model3 has lowest AIC. However, by the likelihood ratio test result, we believe model2 is the "best" model here. Model2 has AIC of 6043.884, BIC of 6076.102 and McFadden r-squared of 0.05371319.

By looking at the table, our model2 is: $log(Odds for Tier\hat{2} Home Team Win)$ = -0.7129158 + 0.0559419 * $spi1$ + -0.0447885 * $spi2$. The intercept has 95% CI of (-1.6111243, 0.1848869); the coefficient of $spi1$ has 95% CI of (0.0354644, 0.0764931); the coefficient of $spi2$ has 95% CI of (-0.0499340, -0.0397326). Those coefficients could be interpreted similarly like the previous models.

Tier4: Good teams, $spi1$ between 50.084 and 63.74

```
## [1] 1.058831e-107
```

```
## [1] 3.41594e-12
```

```
## [1] 0.3171121
```

Table 21: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.7425512	0.4349059	-1.707384	0.0877508	-1.5954734	0.1095556
spi1	0.0602889	0.0079902	7.545318	0.0000000	0.0446705	0.0759961
spi2	-0.0510837	0.0022796	-22.409285	0.0000000	-0.0555906	-0.0466533

Table 22: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7066.545	5119	-3235.453	6476.906	6496.529	6470.906	5117

```
## [1] 0.08344526
```

Model1: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi2$

Model2: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2$

Model3: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2 + \hat{\beta}_3 * importance1 + \hat{\beta}_4 * importance2$

We apply the similar process as what we did in the previous tiers: by likelihood ratio tests between model1 and intercept only model, we get p-value equals to 1.058831e-107 < 0.05, which means $\hat{\beta}_1 \neq 0$; by likelihood ratio tests between model2 and model1, we get p-value equals to 3.41594e-12 < 0.05, which means $\hat{\beta}_2 \neq 0$; by likelihood ratio tests between model3 and model2, we get p-value equals to 0.3171121, which is much larger than 0.05, so that indicates $\hat{\beta}_3 = 0$ or $\hat{\beta}_4 = 0$. By AIC and BIC, we notice that model2 has lowest BIC but model3 has lowest AIC. However, by the likelihood ratio test result, we believe model2 is the "best" model here. Model2 has AIC of 5877.225, BIC of 5896.554 and McFadden r-squared of 0.08344526.

By looking at the table, our model2 is: $log(Odds for Tier\hat{2} Home Team Win)$ = -0.6412210 + 0.0581916 * $spi1$ + -0.0507153 * $spi2$. The intercept has 95% CI of (-1.5436982, 0.2604930); the coefficient of $spi1$ has 95% CI of (0.0417118, 0.0747666); the coefficient of $spi2$ has 95% CI of (-0.0554389, -0.0460759). Those coefficients could be interpreted similarly like the previous models.

Tier5: Outstanding teams, $spi1 >$ 63.74 or equal to 63.74

```
## [1] 7.673749e-115
```

```
## [1] 1.39756e-63
```

```
## [1] 0.4413134
```

Table 23: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.1472496	0.3081183	-3.723406	0.0001966	-1.7523372	-0.5442873

term	estimate	std.error	statistic	p.value	conf.low	conf.high
spi1	0.0703125	0.0041093	17.110571	0.0000000	0.0623131	0.0784246
spi2	-0.0579393	0.0024161	-23.980109	0.0000000	-0.0627195	-0.0532466

Table 24: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7136.079	5208	-3125.129	6256.259	6275.933	6250.259	5206

```
## [1] 0.1263026
```

Model1: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi2$

Model2: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2$

Model3: $log(Odds for Tier\hat{2} Home Team Win) = \hat{\beta}_0 + \hat{\beta}_1 * spi1 + \hat{\beta}_2 * spi2 + \hat{\beta}_3 * importance1 + \hat{\beta}_4 * importance2$

We apply the similar process as what we did in the previous tiers: by likelihood ratio tests between model1 and intercept only model, we get p-value equals to 7.673749e-115 < 0.05, which means $\hat{\beta}_1 \neq 0$; by likelihood ratio tests between model2 and model1, we get p-value equals to 1.39756e-63 < 0.05, which means $\hat{\beta}_2 \neq 0$; by likelihood ratio tests between model3 and model2, we get p-value equals to 0.4413134, which is much larger than 0.05, so that indicates $\hat{\beta}_3 = 0$ or $\hat{\beta}_4 = 0$. By AIC and BIC, we notice that model2 has lowest AIC and BIC values at the same time. Here, Model2 has AIC of 5554.963, BIC of 5574.293 and McFadden r-squared of 0.1263026.

By looking at the table, our model2 is: $log(Odds for Tier\hat{2} Home Team Win)$ = -1.0271808 + 0.0697231 * $spi1$ + -0.0588753 * $spi2$. The intercept has 95% CI of (-1.6722365, -0.3843687); the coefficient of $spi1$ has 95% CI of (0.0612481, 0.0783251); the coefficient of $spi2$ has 95%

CI of (-0.0639828, -0.0538670). Those coefficients could be interpreted similarly like the previous models.

Conclusion and Discussion

From the models, we learned that for the general dataset, using SPI's and Importances together makes the "best" model. However, it is not necessarily true for some tiers of the teams. When the home teams are stronger, the importances of the games usually do not contribute as much for the weaker teams. Also, by the McFadden r-squared, we see that when the home teams are stronger, the models we fit are more "accurate" and useful.

Also, generally, we find the (log) odds of winning for the home team is increasing as the home team is stronger and is decreasing as the opponent is stronger; the (log) odds of winning for the home team is increasing as the importance of the game is higher and is decreasing as the opponent has higher importance of the game. This coincides with our common knowledge about soccer.

Citations

Boice, Jay. "How Our Club Soccer Predictions Work." FiveThirtyEight, 10 Aug. 2018, https://fivethirtyeight.com/methodology/how-our-club-soccer-predictions-work/.

Yihui Xie (2019). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.24.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible

Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595

Kim, A. Y. and Ismay, C. and Chunn, J. (2018) The fivethirtyeight R Package: 'Tame Data' Principles for Introductory Statistics and Data Science Courses, Technology Innovations in Statistics Education. 11(1)

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. http://tidyverse.tidyverse.org, https://github.com/tidyverse/tidyverse.

David Robinson and Alex Hayes (2019). broom: Convert Statistical Analysis Objects into Tidy Tibbles. R package version 0.5.2. http://github.com/tidyverse/broom

Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read Rectangular Text Data. R package version 1.3.1. https://CRAN.R-project.org/package=readr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. http://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Appendix

Table 25: Intercept only model information

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1808517	0.0124084	-14.57489	0

Table 26: Intercept only model Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-18048.26	36098.53	36106.7	36096.53	26191

Table 27: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.3060994	0.0319631	9.57666	0	0.2434860	0.3687824
spi2	-0.0106911	0.0006491	-16.47070	0	-0.0119648	-0.0094203

Table 28: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-17910.66	35825.32	35841.67	35821.32	26190

Table 29: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2613003	0.0351626	-7.431208	0	-0.3302519	-0.1924130
spi1	0.0504740	0.0010997	45.896262	0	0.0483267	0.0526377
spi2	-0.0492123	0.0011083	-44.402699	0	-0.0513930	-0.0470484

Table 30: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-16683.64	33373.28	33397.8	33367.28	26189

Table 31: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2576867	0.0370254	-6.959719	0.0000000	-0.3302834	-0.1851419
spi1	0.0491036	0.0011423	42.988263	0.0000000	0.0468727	0.0513504
spi2	-0.0477994	0.0011485	-41.617500	0.0000000	-0.0500587	-0.0455563
importance1	0.0021015	0.0005652	3.718455	0.0002004	0.0009942	0.0032097
importance2	-0.0023301	0.0005781	-4.030915	0.0000556	-0.0034639	-0.0011979

Table 32: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
36096.53	26191	-16672.64	33355.28	33396.15	33345.28	26187

Table 33: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.3461520	0.0667111	5.188819	2e-07	0.2157093	0.477246
spi2	-0.0313675	0.0024526	-12.789539	0e+00	-0.0362053	-0.026590

Table 34: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7364.844	5505	-3594.536	7193.073	7206.3	7189.073	5504

Table 35: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2650628	0.0963508	-2.751018	0.005941	-0.4543598	-0.0766004
spi1	0.0453012	0.0050887	8.902401	0.000000	0.0353628	0.0553127
spi2	-0.0436901	0.0028983	-15.074122	0.000000	-0.0494175	-0.0380546

Table 36: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7364.844	5505	-3553.789	7113.578	7133.419	7107.578	5503

Table 37: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.2198635	0.1048973	-2.0959879	0.0360833	-0.4258856	-0.0146128
spi1	0.0447427	0.0051076	8.7599708	0.0000000	0.0347663	0.0547906
spi2	-0.0429159	0.0029239	-14.6776843	0.0000000	-0.0486917	-0.0372287
importance1	0.0010954	0.0014268	0.7677564	0.4426319	-0.0017031	0.0038913
importance2	-0.0030213	0.0013412	-2.2527236	0.0242766	-0.0056587	-0.0004001

Table 38: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7364.844	5505	-3551.211	7112.421	7145.489	7102.421	5501

Table 39: Intercept only model information

term	estimate	std.error	statistic	p.value
(Intercept)	-0.4468084	0.0276287	-16.17189	0

Table 40: Intercept only model information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7364.844	5505	-3682.422	7366.844	7373.458	7364.844	5505

Table 41: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.1752563	0.0951167	12.35594	0	0.9897975	1.3627063
spi2	-0.0399266	0.0024820	-16.08636	0	-0.0448294	-0.0350985

Table 42: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7102.704	5211	-3407.016	6818.031	6831.149	6814.031	5210

Table 43: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.4181164	0.3306271	-1.264616	0.2060089	-1.0665301	0.2296661
spi1	0.0493436	0.0098368	5.016243	0.0000005	0.0300974	0.0686618

term	estimate	std.error	statistic	p.value	conf.low	conf.high
spi2	-0.0430006	0.0025815	-16.656966	0.0000000	-0.0481022	-0.0379812

Table 44: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7102.704	5211	-3394.334	6794.668	6814.344	6788.668	5209

Table 45: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.3610630	0.3337093	-1.081969	0.2792664	-1.0154766	0.2928068
spi1	0.0477650	0.0098623	4.843199	0.0000013	0.0284678	0.0671322
spi2	-0.0421771	0.0026134	-16.138565	0.0000000	-0.0473390	-0.0370929
importance1	0.0017903	0.0014027	1.276281	0.2018562	-0.0009587	0.0045412
importance2	-0.0028682	0.0013164	-2.178862	0.0293420	-0.0054567	-0.0002954

Table 46: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7102.704	5211	-3391.781	6793.562	6826.356	6783.562	5207

Table 47: Intercept only model information

term	estimate	std.error	statistic	p.value
(Intercept)	-0.3086434	0.0280336	-11.00977	0

term	estimate	std.error	statistic	p.value

Table 48: Intercept only model information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7102.704	5211	-3551.352	7104.704	7111.263	7102.704	5211

Table 49: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.6059242	0.1102236	14.56970	0	1.3911712	1.8233116
spi2	-0.0414238	0.0024159	-17.14643	0	-0.0461967	-0.0367249

Table 50: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7057.831	5144	-3363.801	6731.603	6744.694	6727.603	5143

Table 51: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.6843993	0.4352045	-1.572592	0.1158132	-1.5376593	0.1685266
spi1	0.0538281	0.0099320	5.419651	0.0000001	0.0343911	0.0733290
spi2	-0.0438223	0.0024769	-17.692466	0.0000000	-0.0487175	-0.0390065

Table 52: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7057.831	5144	-3349.01	6704.021	6723.658	6698.021	5142

Table 53: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.7117575	0.4365037	-1.630588	0.1029774	-1.5675707	0.1437099
spi1	0.0537540	0.0099345	5.410842	0.0000001	0.0343121	0.0732598
spi2	-0.0426196	0.0025429	-16.760537	0.0000000	-0.0476416	-0.0376720
importance1	0.0018397	0.0012710	1.447473	0.1477644	-0.0006501	0.0043332
importance2	-0.0025846	0.0013021	-1.984969	0.0471480	-0.0051436	-0.0000382

Table 54: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7057.831	5144	-3346.785	6703.57	6736.299	6693.57	5140

Table 55: Intercept only model information

term	estimate	std.error	statistic	p.value
(Intercept)	-0.2417931	0.0280866	-8.60885	0

Table 56: Intercept only model information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7057.831	5144	-3528.916	7059.831	7066.377	7057.831	5144

Table 57: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	2.4321433	0.1228211	19.80232	0	2.1930989	2.674626
spi2	-0.0466493	0.0021667	-21.52995	0	-0.0509307	-0.042436

Table 58: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7066.545	5119	-3264.432	6532.863	6545.945	6528.863	5118

Table 59: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.7425512	0.4349059	-1.707384	0.0877508	-1.5954734	0.1095556
spi1	0.0602889	0.0079902	7.545318	0.0000000	0.0446705	0.0759961
spi2	-0.0510837	0.0022796	-22.409285	0.0000000	-0.0555906	-0.0466533

Table 60: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7066.545	5119	-3235.453	6476.906	6496.529	6470.906	5117

Table 61: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.7831661	0.4362607	-1.795179	0.0726252	-1.6387825	0.0715598
spi1	0.0598549	0.0080007	7.481241	0.0000000	0.0442158	0.0755822
spi2	-0.0499307	0.0023974	-20.827161	0.0000000	-0.0546685	-0.0452694
importance1	0.0017583	0.0012764	1.377492	0.1683603	-0.0007400	0.0042646
importance2	-0.0016938	0.0013132	-1.289802	0.1971194	-0.0042708	0.0008781

Table 62: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7066.545	5119	-3234.178	6478.357	6511.061	6468.357	5115

Table 63: Intercept only model information

term	estimate	std.error	statistic	p.value
(Intercept)	-0.1565691	0.0280365	-5.584475	0

Table 64: Intercept only model information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7066.545	5119	-3533.273	7068.545	7075.086	7066.545	5119

Table 65: Model1 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	3.5539712	0.1554781	22.85834	0	3.2519872	3.8615453
spi2	-0.0499219	0.0022859	-21.83928	0	-0.0544411	-0.0454792

Table 66: Model1 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7136.079	5208	-3287.074	6578.147	6591.263	6574.147	5207

Table 67: Model2 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.1472496	0.3081183	-3.723406	0.0001966	-1.7523372	-0.5442873
spi1	0.0703125	0.0041093	17.110571	0.0000000	0.0623131	0.0784246
spi2	-0.0579393	0.0024161	-23.980109	0.0000000	-0.0627195	-0.0532466

Table 68: Model2 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7136.079	5208	-3125.129	6256.259	6275.933	6250.259	5206

Table 69: Model3 Information

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.1143652	0.3104737	-3.5892422	0.0003316	-1.7241374	-0.5068347
spi1	0.0687520	0.0043597	15.7699033	0.0000000	0.0602685	0.0773623
spi2	-0.0572345	0.0026840	-21.3241055	0.0000000	-0.0625461	-0.0520225
importance1	0.0012123	0.0011583	1.0466880	0.2952435	-0.0010598	0.0034815
importance2	-0.0004999	0.0012665	-0.3946965	0.6930669	-0.0029807	0.0019850

Table 70: Model3 Information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7136.079	5208	-3124.581	6259.162	6291.953	6249.162	5204

Table 71: Intercept only model information

term	estimate	std.error	statistic	p.value
(Intercept)	0.2567281	0.0279393	9.188789	0

Table 72: Intercept only model information

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual
7136.079	5208	-3568.039	7138.079	7144.637	7136.079	5208

```
library(knitr)
opts_chunk$set(echo = FALSE)
library(fivethirtyeight)
library(tidyverse)
library(broom)
library(readr)
library(dplyr)
spi_matches <-
  "https://projects.fivethirtyeight.com/soccer-api/club/spi_matches.csv" %>%
  read_csv()
SPI <- spi_matches %>%
  select(date, league, team1, team2, spi1, spi2, prob1, prob2, probtie, proj_score1, pro
  mutate(dif = spi1 - spi2, difImportance = importance1 - importance2)
SPI <- SPI %>%
  filter(score1!="NA", importance1!="NA") %>%
  mutate(result = case_when(
    score1 > score2 ~ 1,
    score1 <= score2 ~ 0
  ))
quantile(SPI$spi1,0.2)
quantile(SPI$spi1,0.4)
```

```r
quantile(SPI$spi1,0.6)
quantile(SPI$spi1,0.8)
SPI <- SPI %>%
  mutate(
  spi1_cat = case_when(
  spi1 < 29.13 ~ "quantile 1",
  spi1 >= 29.13 & spi1 < 39.6 ~ "quantile 2",
  spi1 >= 39.6 &  spi1 < 50.08 ~ "quantile 3",
  spi1 >= 50.08 & spi1 < 63.74 ~ "quantile 4",
  spi1 >= 63.74 ~ "quantile 5"
))
library(knitr)
SPI %>%
  group_by(spi1_cat) %>%
  summarise(mean(spi1)) %>%
  kable(caption = "Quantiles and means")
library(knitr)
SPI %>%
  group_by(spi1_cat, result) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n),
         odds = prop/(1-prop),
         log_odds = log(odds)) %>%
  filter(result == 1) %>%
  kable(caption = "Empirical logits and other information")
data1 <- data.frame(
  mean_spi1 = c(20.40914, 34.61565, 44.49413, 56.70562, 74.59321),
```

```
  emp_logit = c(-0.4600011, -0.3086957, -0.2237253, -0.1582810, 0.2747560)
)
ggplot(data1, aes(mean_spi1, emp_logit)) +
  geom_point() +
  labs(title = "Empirical Logit vs. Mean Spi1", x = "Mean SPI of Team 1", y = "log odds
quantile(SPI$spi2,0.2)
quantile(SPI$spi2,0.4)
quantile(SPI$spi2,0.6)
quantile(SPI$spi2,0.8)
SPI <- SPI %>%
  mutate(
  spi2_cat = case_when(
  spi2 < 29.15 ~ "quantile 1",
  spi2 >= 29.15 & spi2 < 39.5 ~ "quantile 2",
  spi2 >= 39.5 &  spi2 < 50.05 ~ "quantile 3",
  spi2 >= 50.05 & spi2 < 63.72 ~ "quantile 4",
  spi2 >= 63.72 ~ "quantile 5"
))
library(knitr)
SPI %>%
  group_by(spi2_cat) %>%
  summarise(mean(spi2)) %>%
  kable(caption = "Quantiles and means")
library(knitr)
SPI %>%
  group_by(spi2_cat, result) %>%
  summarise(n = n()) %>%
```

```
  mutate(prop = n / sum(n),
         odds = prop/(1-prop),
         log_odds = log(odds)) %>%
  filter(result == 1) %>%
  kable(caption = "Empirical logits and other information")
data2 <- data.frame(
  mean_spi2 = c(20.43950, 34.54871, 44.36776, 56.58594, 74.54673),
  emp_logit = c(0.0150897, -0.0426697, -0.1098586, -0.1662227, -0.5777565)
)
ggplot(data2, aes(mean_spi2, emp_logit)) +
  geom_point() +
  labs(title = "Empirical Logit vs. Mean Spi2", x = "Mean SPI of Team 2", y = "log odds
quantile(SPI$importance1,0.25)
quantile(SPI$importance1,0.5)
quantile(SPI$importance1,0.75)
SPI <- SPI %>%
  mutate(
  importance1_cat = case_when(
  importance1 < 10.9 ~ "quantile 1",
  importance1 >= 10.9 & importance1 < 26.1 ~ "quantile 2",
  importance1 >= 26.1 &  importance1 < 45.3 ~ "quantile 3",
  importance1 >= 45.3 ~ "quantile 4"
))
library(knitr)
SPI %>%
  group_by(importance1_cat) %>%
  summarise(mean(importance1)) %>%
```

```
  kable(caption = "Quantiles and means")
library(knitr)
SPI %>%
  group_by(importance1_cat, result) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n),
         odds = prop/(1-prop),
         log_odds = log(odds)) %>%
  filter(result == 1) %>%
  kable(caption = "Empirical logits and other information")
data3 <- data.frame(
  mean_importance1 = c(3.078908, 18.782330, 35.026542, 67.936947),
  emp_logit = c(-0.2426765, -0.1858654, -0.3643986, 0.0947188)
)
ggplot(data3, aes(mean_importance1, emp_logit)) +
  geom_point() +
  labs(title = "Empirical Logit vs. Mean importance1", x = "Mean importance of Team 1",
quantile(SPI$importance2,0.25)
quantile(SPI$importance2,0.5)
quantile(SPI$importance2,0.75)
SPI <- SPI %>%
  mutate(
  importance2_cat = case_when(
  importance2 < 10.475 ~ "quantile 1",
  importance2 >= 10.475 & importance2 < 25.2 ~ "quantile 2",
  importance2 >= 25.2 &  importance2 < 44.4 ~ "quantile 3",
  importance2 >= 44.4 ~ "quantile 4"
```

```
))
library(knitr)
SPI %>%
  group_by(importance2_cat) %>%
  summarise(mean(importance2)) %>%
  kable(caption = "Quantiles and means")
library(knitr)
SPI %>%
  group_by(importance2_cat, result) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n),
         odds = prop/(1-prop),
         log_odds = log(odds)) %>%
  filter(result == 1) %>%
  kable(caption = "Empirical logits and other information")
data4 <- data.frame(
  mean_importance2 = c(3.004135, 18.163344, 34.002503, 66.880323),
  emp_logit = c(-0.0517000, -0.1088066, -0.1187544, -0.4188015)
)
ggplot(data4, aes(mean_importance2, emp_logit)) +
  geom_point() +
  labs(title = "Empirical Logit vs. Mean importance2", x = "Mean importance of Team 2",
model1 <- glm(result ~ spi2, data = SPI, family = "binomial")
model1 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
model1 %>% glance() %>% kable(caption = "Model1 Information")


pchisq(32009.97 - 31756.87, df = 1, lower.tail = FALSE)
```

```r
model2 <- glm(result ~ spi1 + spi2, data = SPI, family = "binomial")
model2 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
model2 %>% glance() %>% kable(caption = "Model2 Information")

pchisq(31756.87 - 29539.38, df = 1, lower.tail = FALSE)

model12 <- glm(result ~ spi1 + spi2 + importance1 + importance2, data = SPI, family = "b
model12 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
model12 %>% glance() %>% kable(caption = "Model3 Information")

pchisq(29539.38 - 29518.51, df = 2, lower.tail = FALSE)

modelcheck <- glm(result ~ 1, data = SPI, family = "binomial")

1 - (-15878.44 / -16004.98)
1 - (-14769.69 / -16004.98)
1 - (-14759.25 / -16004.98)
df1 <- SPI %>% filter(spi1 < 29.13)
df2 <- SPI %>% filter(spi1 >= 29.13 & spi1 < 39.6)
df3 <- SPI %>% filter(spi1 >= 39.6 &  spi1 < 50.084)
df4 <- SPI %>% filter(spi1 >= 50.084 & spi1 < 63.74)
df5 <- SPI %>% filter(spi1 >= 63.74)
modeltry11 <- glm(result ~ spi2, data = df1, family = "binomial")
pchisq(6194.627 - 6034.682, df = 1, lower.tail = FALSE)

modeltry12 <- glm(result ~ spi1 + spi2, data = df1, family = "binomial")
```

```r
pchisq(6034.682 - 5973.45, df = 1, lower.tail = FALSE)

modeltry13 <- glm(result ~ spi1 + spi2 + importance1 + importance2, data = df1, family =
pchisq(5973.45 - 5967.392, df = 2, lower.tail = FALSE)

modeltry1check <- glm(result ~ 1, data = df1, family = "binomial")

modeltry13 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
modeltry13 %>% glance() %>% kable(caption = "Model3 Information")

1 - (-2983.696/-3097.313)

modeltry21 <- glm(result ~ spi2, data = df2, family = "binomial")
pchisq(6327.257 - 6084.7, df = 1, lower.tail = FALSE)

modeltry22 <- glm(result ~ spi1 + spi2, data = df2, family = "binomial")
pchisq(6084.7 - 6057.601, df = 1, lower.tail = FALSE)

modeltry23 <- glm(result ~ spi1 + spi2 + importance1 + importance2, data = df2, family =
pchisq(6057.601 - 6054.429, df = 2, lower.tail = FALSE)

modeltry2check <- glm(result ~ 1, data = df2, family = "binomial")

modeltry22 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry22 %>% glance() %>% kable(caption = "Model2 Information")
1 - (-3028.8 / -3163.629)
modeltry31 <- glm(result ~ spi2, data = df3, family = "binomial")
```

```r
pchisq(6382.121 - 6068.111, df = 1, lower.tail = FALSE)

modeltry32 <- glm(result ~ spi1 + spi2, data = df3, family = "binomial")
pchisq(6068.111 - 6039.317, df = 1, lower.tail = FALSE)

modeltry33 <- glm(result ~ spi1 + spi2 + importance1 + importance2, data = df3, family =
pchisq(6039.317 - 6033.884, df = 2, lower.tail = FALSE)

modeltry3check <- glm(result ~ 1, data = df3, family = "binomial")

modeltry32 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry32 %>% glance() %>% kable(caption = "Model2 Information")

1 - (-3019.658 / -3191.06)
modeltry41 <- glm(result ~ spi2, data = df4, family = "binomial")

pchisq(6405.656 - 5919.659, df = 1, lower.tail = FALSE)

modeltry42 <- glm(result ~ spi1 + spi2, data = df4, family = "binomial")

pchisq(5919.659 - 5871.225, df = 1, lower.tail = FALSE)

modeltry43 <- glm(result ~ spi1 + spi2 + importance1 + importance2, data = df4, family =

pchisq(5871.225 - 5868.928, df = 2, lower.tail = FALSE)

modeltry4check <- glm(result ~ 1, data = df4, family = "binomial")
```

```r
modeltry42 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry42 %>% glance() %>% kable(caption = "Model2 Information")
1 - (-2935.613 / -3202.878)
modeltry51 <- glm(result ~ spi2, data = df5, family = "binomial")

pchisq(6351.126 - 5832.314, df = 1, lower.tail = FALSE)

modeltry52 <- glm(result ~ spi1 + spi2, data = df5, family = "binomial")

pchisq(5832.314 - 5548.963, df = 1, lower.tail = FALSE)

modeltry53 <- glm(result ~ spi1 + spi2 + importance1 + importance2, data = df5, family =

pchisq(5548.963 - 5547.327, df = 2, lower.tail = FALSE)

modeltry5check <- glm(result ~ 1, data = df5, family = "binomial")

modeltry52 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry52 %>% glance() %>% kable(caption = "Model2 Information")

1 - (-2774.481 / -3175.563)

modelcheck %>% tidy() %>% kable(caption = "Intercept only model information")
modelcheck %>% glance() %>% kable(caption = "Intercept only model Information")
model1 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
model1 %>% glance() %>% kable(caption = "Model1 Information")
```

```
model2 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
model2 %>% glance() %>% kable(caption = "Model2 Information")
model12 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
model12 %>% glance() %>% kable(caption = "Model3 Information")

modeltry11 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
modeltry11 %>% glance() %>% kable(caption = "Model1 Information")
modeltry12 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry12 %>% glance() %>% kable(caption = "Model2 Information")
modeltry13 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
modeltry13 %>% glance() %>% kable(caption = "Model3 Information")
modeltry1check %>% tidy() %>% kable(caption ="Intercept only model information")
modeltry1check %>% glance() %>% kable(caption ="Intercept only model information")

modeltry21 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
modeltry21 %>% glance() %>% kable(caption = "Model1 Information")
modeltry22 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry22 %>% glance() %>% kable(caption = "Model2 Information")
modeltry23 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
modeltry23 %>% glance() %>% kable(caption = "Model3 Information")
modeltry2check %>% tidy() %>% kable(caption ="Intercept only model information")
modeltry2check %>% glance() %>% kable(caption ="Intercept only model information")

modeltry31 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
modeltry31 %>% glance() %>% kable(caption = "Model1 Information")
modeltry32 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry32 %>% glance() %>% kable(caption = "Model2 Information")
```

```
modeltry33 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
modeltry33 %>% glance() %>% kable(caption = "Model3 Information")
modeltry3check %>% tidy() %>% kable(caption ="Intercept only model information")
modeltry3check %>% glance() %>% kable(caption ="Intercept only model information")

modeltry41 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
modeltry41 %>% glance() %>% kable(caption = "Model1 Information")
modeltry42 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry42 %>% glance() %>% kable(caption = "Model2 Information")
modeltry43 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
modeltry43 %>% glance() %>% kable(caption = "Model3 Information")
modeltry4check %>% tidy() %>% kable(caption ="Intercept only model information")
modeltry4check %>% glance() %>% kable(caption ="Intercept only model information")

modeltry51 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model1 Information")
modeltry51 %>% glance() %>% kable(caption = "Model1 Information")
modeltry52 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model2 Information")
modeltry52 %>% glance() %>% kable(caption = "Model2 Information")
modeltry53 %>% tidy(conf.int = TRUE) %>% kable(caption = "Model3 Information")
modeltry53 %>% glance() %>% kable(caption = "Model3 Information")
modeltry5check %>% tidy() %>% kable(caption ="Intercept only model information")
modeltry5check %>% glance() %>% kable(caption ="Intercept only model information")
```