

Predicting football results

Fadi Farhat, Loïc Wamster

ENAC

April 22, 2018

Table of contents

1 Introduction

- Context - Why this subject?
- Teams selected for the project
- Variables used

2 General model

- Linear regression
- Marginal effects
- F-test and Jarque bera
- Results and analysis

3 Model based on each team

- Regressions output
- Marginal effects
- Results and analysis

4 Conclusion

Introduction - Why this subject?

Can we make money with sport bets?

- The sport bets industry has reached 2 081 billion euros with an increase by 45% in only one year in 2016 (les echos).
- Harder than expected to build the perfect algorithm to make easy money...

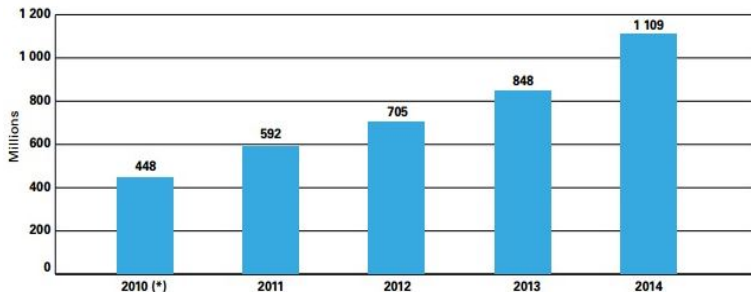


Table - Evolution of sport bets

Introduction - Teams selected for the project

First general model - based on all the L1 teams :

Time period : Every matches from 2012 to now.

Second model more specific - focus on 3 particular teams :

- Monaco (L1 top ranking)
- Bordeaux (Mid-table team)
- Toulouse (Bottom of the L1 ranking)

Why these teams ?

- Heterogeneity in their ranking
- To see if the results quality depend on the ranking
- To analyze for which teams is our model better
- To see if we have the same significant variables for each teams

Introduction - Variables used for this project

Variable	Description
FTR	Full Time Result
BWH	Bet&Win home win odds
BWD	Bet&Win draw odds
BWA	Bet&Win away win odds
WR	Win ratio of the 20 last matches played by the home team
RECENT	Result of the most recent match played by the home team
LAST1*LAST0	Result of the last match between home and away team last season
GHR	Goal home rate of the home team
GAR	Goal away rate of the home team
HY	Number of yellow card of the home team in the previous match
AY	Number of yellow card of the away team in the previous match
HR	Number of red card of the home team in the previous match
AWR	Number of red card of the away team in the previous match

Number of observations and variables

$n = 225$ and $k = 13$

Linear equation of the general model

$$\widehat{FTR} = \widehat{\beta}_0 + \widehat{\beta}_1.BWH + \widehat{\beta}_2.BWD + \widehat{\beta}_3.BWA + \widehat{\beta}_4.WR + \widehat{\beta}_5.RECENT + \widehat{\beta}_6.LAST1.LAST0 + \widehat{\beta}_7.GHR + \widehat{\beta}_8.GAR + \widehat{\beta}_9.HY + \widehat{\beta}_{10}.AY + \widehat{\beta}_{11}.HR + \widehat{\beta}_{12}.AWR$$

General model - Linear regression with $k=13$

Dependent Variable: FTR
Method: Least Squares
Date: 04/09/18 Time: 17:20
Sample: 1 255
Included observations: 255

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.324302	0.350328	0.925712	0.3555
BWH	-0.106414	0.041431	-2.568436	0.0108
BWD	0.155296	0.101103	1.536007	0.1258
BWA	-0.005638	0.035721	-0.157822	0.8747
WR	-0.023450	0.454150	-0.051635	0.9589
RECENT	-0.094160	0.088186	-1.067747	0.2867
LAST1*LAST0	0.147492	0.077540	1.902148	0.0583
GHR	-0.027126	0.089836	-0.301953	0.7629
GAR	0.058596	0.123534	0.474327	0.6357
HY	-0.017765	0.021560	-0.823984	0.4108
AY	0.020788	0.019942	1.042470	0.2982
HR	-0.122536	0.072299	-1.694860	0.0914
AWR	0.011907	0.058404	0.203878	0.8386
R-squared	0.189449	Mean dependent var		0.641176
Adjusted R-squared	0.149256	S.D. dependent var		0.393939
S.E. of regression	0.363352	Akaike info criterion		0.862747
Sum squared resid	31.95003	Schwarz criterion		1.043282
Log likelihood	-97.00029	Hannan-Quinn criter.		0.935366
F-statistic	4.713515	Durbin-Watson stat		2.020327
Prob(F-statistic)	0.000001			

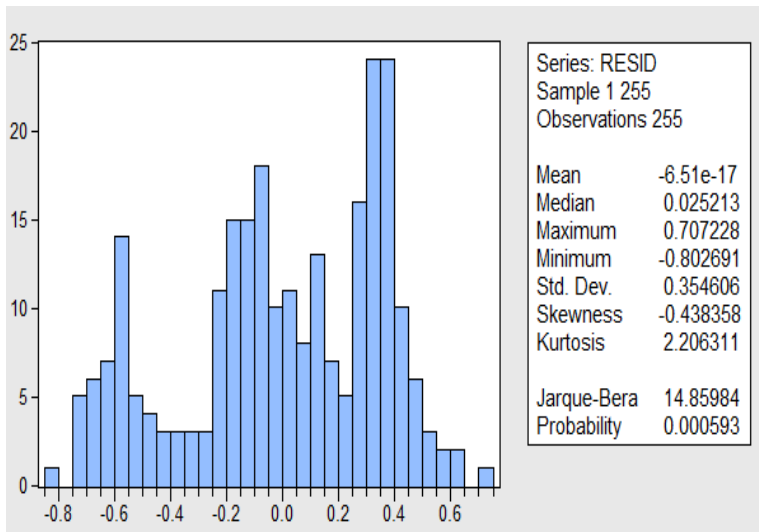
Table - General regression

General model - Marginal effects

Table: Marginal effects table

Variable	Marginal effect	Significant
BWH	-	yes
BWD	+	almost
BWA	-	no
WR	-	no
RECENT	-	no
LAST1*LAST0	+	yes
GHR	-	no
GAR	+	no
HY	-	no
AY	+	no
HR	-	yes
AWR	+	no

General model - Jarque bera test



Graph - Jarque bera test

Table: Results summary

	Test value
R^2	18.9%
Adjusted R^2	14.9%
$\left(\frac{n-k}{n}\right) JB_e$	14
$\chi^2_{0,95}(2)$	5.99
Reject normality ?	yes

General model - Linear regression with $k=9$

Dependent Variable: FTR
Method: Least Squares
Date: 04/09/18 Time: 18:08
Sample: 1 255
Included observations: 255

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.313344	0.192048	1.631597	0.1040
BWH	-0.097178	0.018197	-5.340414	0.0000
BWD	0.129071	0.030998	4.163807	0.0000
RECENT	-0.096114	0.084565	-1.136571	0.2568
LAST1*LAST0	0.140187	0.073016	1.919949	0.0560
GAR	0.064610	0.097225	0.664540	0.5070
HY	-0.016519	0.021057	-0.784503	0.4335
AY	0.020177	0.019630	1.027823	0.3050
HR	-0.123706	0.071120	-1.739390	0.0832
R-squared	0.188283	Mean dependent var		0.641176
Adjusted R-squared	0.161885	S.D. dependent var		0.393939
S.E. of regression	0.360645	Akaike info criterion		0.832812
Sum squared resid	31.99598	Schwarz criterion		0.957798
Log likelihood	-97.18354	Hannan-Quinn criter.		0.883087
F-statistic	7.132652	Durbin-Watson stat		2.008680
Prob(F-statistic)	0.000000			

Table - General regression $k=9$

F-test results

The resids are not normal so we use the formula

$$qF = q \left(\frac{\hat{u}'_R \hat{u}_R}{\hat{u}' \hat{u}} - 1 \right) (n - k) \sim \chi^2(q)$$

Table: Rigorous test to simplify first equation into the second one

	test
q	4
qF	0.31
$\chi^2_{0,95}(q)$	9.49
Reject the null	no

Result of the general model

- We have an R^2 equal to 18.8%
- Most significant variables: BWH, BWD, LAST1*LAST0, HR.
- The dream of making a perfect algorithm to make easy money seems to be only a dream.

Why is it so bad?

- We choose three teams with different ranking level.
- Predicting the result of matches seem to depend on the level of the considered team.

Number of observations and variables

- Try the same model to each team in order to compare them with the general model.
- The number of observation n depend on the team.
- $k = 13$ for each team.

Model for each team

$$\widehat{FTR} = \widehat{\beta}_0 + \widehat{\beta}_1.BWH + \widehat{\beta}_2.BWD + \widehat{\beta}_3.BWA + \widehat{\beta}_4.WR + \widehat{\beta}_5.RECENT + \widehat{\beta}_6.LAST1.LAST0 + \widehat{\beta}_7.GHR + \widehat{\beta}_8.GAR + \widehat{\beta}_9.HY + \widehat{\beta}_{10}.AY + \widehat{\beta}_{11}.HR + \widehat{\beta}_{12}.AWR$$

Model for each team - With all the variables

Monaco

Included observations: 73

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.273147	0.766466	1.661061	0.1019
BWH	-0.167569	0.081936	-2.045131	0.0452
BWD	0.110250	0.130964	0.841830	0.4032
BWA	-0.022230	0.045128	-0.492590	0.6241
WR	-2.092308	1.256750	-1.664856	0.1012
RECENT	-0.034563	0.141010	-0.245109	0.8072
LAST1*LAST0	0.126835	0.108760	1.166199	0.2481
GHR	0.261783	0.180493	1.450382	0.1522
GAR	0.627552	0.258742	2.425396	0.0183
HY	-0.056237	0.032082	-1.752899	0.0847
AY	0.044782	0.030977	1.445004	0.1537
HR	0.157283	0.125925	1.249019	0.2165
AWR	0.090761	0.111720	0.770992	0.4437

Bordeaux

Included observations: 91

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.485958	0.880878	0.551675	0.5827
BWH	-0.101731	0.117648	-0.864702	0.3899
BWD	0.065092	0.233020	0.279339	0.7807
BWA	0.022803	0.081755	0.278923	0.7810
WR	-0.723396	1.060876	-0.681886	0.4973
RECENT	-0.272668	0.165770	-1.644858	0.1040
LAST1*LAST0	0.271841	0.152231	1.785712	0.0780
GHR	0.099028	0.268440	0.368902	0.7132
GAR	0.540777	0.369638	1.462992	0.1475
HY	-0.054434	0.036112	-1.507368	0.1358
AY	-0.029692	0.038739	-0.766460	0.4457
HR	-0.343389	0.125100	-2.744907	0.0075
AWR	0.010172	0.116290	0.087475	0.9305

Toulouse

Included observations: 91

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.995041	0.740945	1.342936	0.1832
BWH	-0.116100	0.072040	-1.611608	0.1111
BWD	0.191762	0.189133	1.013903	0.3138
BWA	-0.065555	0.081126	-0.808068	0.4215
WR	-2.488859	1.153841	-2.157022	0.0341
RECENT	0.000185	0.158528	0.001164	0.9991
LAST1*LAST0	0.069423	0.157323	0.441274	0.6602
GHR	0.335691	0.221216	1.517484	0.1332
GAR	0.001781	0.199542	0.008928	0.9929
HY	0.090203	0.043743	2.062128	0.0425
AY	0.069280	0.036548	1.895601	0.0617
HR	-0.016303	0.121592	-0.134080	0.8937
AWR	-0.091722	0.087748	-1.045291	0.2991

Table - Regressions output for each teams

General model - Marginal effects

Table: Marginal effects table

Variable	Monaco	Bordeaux	Toulouse
BWH	-	-	-
BWD	+	+	+
BWA	-	+	-
WR	-	-	-
RECENT	-	-	+
LAST1*LAST0	+	+	+
GHR	+	+	+
GAR	+	+	+
HY	-	-	+
AY	-	-	+
HR	+	-	-
AWR	+	+	-

Teams model - Normality test

	Monaco	Bordeaux	Toulouse
n	73	91	91
k	13	13	13
R ²	38.9%	25.9%	23.6%
Adjusted R ²	26.7%	15.5%	11.8 %
$\left(\frac{n-k}{n}\right) JB_e$	8.11	1.75	3.16
$\chi^2_{0,95}(2)$	5.99	5.99	5.99
Reject normality ?	yes	no	no

Table: Results summary

Model for each team - Without insignificant variables

Monaco

Included observations: 73

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.346032	0.745611	1.805273	0.0758
BWH	-0.135677	0.054433	-2.492531	0.0153
BWD	0.050008	0.050496	0.990339	0.3258
WR	-2.146013	1.210492	-1.772844	0.0811
LAST1*LAST0	0.115351	0.085006	1.356977	0.1796
GHR	0.281532	0.174623	1.612230	0.1119
GAR	0.589392	0.247926	2.377286	0.0205
HY	-0.054826	0.031473	-1.741998	0.0864
AY	0.040910	0.028335	1.443796	0.1538
HR	0.167109	0.123106	1.357441	0.1795

Bordeaux

Included observations: 91

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.005292	0.579613	1.734419	0.0866
BWH	-0.139067	0.050618	-2.747364	0.0074
WR	-0.682163	0.671969	-1.015170	0.3130
RECENT	-0.240256	0.157797	-1.522560	0.1317
LAST1*LAST0	0.259841	0.147428	1.762499	0.0817
GAR	0.528327	0.341632	1.546480	0.1258
HY	-0.060093	0.033529	-1.792252	0.0768
AY	-0.023473	0.036320	-0.646267	0.5199
HR	-0.350847	0.121241	-2.893805	0.0049

Toulouse

Included observations: 91

	Coefficient	Std. Error	t-Statistic	Prob.
C	0.970631	0.547727	1.772106	0.0801
BWH	-0.123220	0.067324	-1.830247	0.0708
BWD	0.213417	0.176243	1.210927	0.2294
BWA	-0.078868	0.073326	-1.075580	0.2853
WR	-2.496095	1.018496	-2.450765	0.0164
GHR	0.362136	0.204841	1.767891	0.0808
HY	0.091159	0.041711	2.185514	0.0317
AY	0.073296	0.033960	2.158274	0.0338
AWR	-0.082738	0.083432	-0.991682	0.3243

Table - Regressions output for each teams

Model for each team - F-test

	Monaco	Bordeaux	Toulouse
q	3	4	4
F (or qF)	0.86	0.29	0.07
$\chi^2_{0,95}(q)$	7.82	9.49	9.49
Reject the null	no	no	no

Table: Rigorous test to simplify first equation into the second one

Conclusion of the results - 1/2

Table: Results summary

Variables	General model		Monaco		Bordeaux		Toulouse	
	M.E	Sign	M.E	Sign	M.E	Sign	M.E	Sign
BWH	-	yes	-	yes	-	yes	-	yes
BWD	+	yes	+	no	+	no	+	no
BWA	-	no	-	no	+	no	-	no
WR	-	no	-	yes	-	no	-	yes
RECENT	-	no	-	no	-	no	+	no
LAST0*LAST1	+	yes	+	no	+	yes	+	no
GHR	+	no	+	yes	+	yes	+	yes
GAR	+	no	+	yes	+	yes	+	no
HY	-	no	-	yes	-	yes	+	yes
AY	+	no	-	no	-	no	+	yes
HR	+	yes	+	no	-	no	-	no
AWR	+	no	+	no	+	no	-	no

Conclusion of the results - 2/2

General model

- The first model (general model) can't be used for prediction because of its low R^2 .
- Marginal effects change from one team to another.
- Each team's model has different significant variables.

Team's model

- Linear regression on a precise team gives us a bigger R^2 than for the general model.
- Our model forecasts more precisely Monaco's results : more efficient on a top ranking team.

Thank you for your attention !