PREDICTION OF WINNING IN WOMEN'S PROFESSIONAL TENNIS

Siddhartha Pachhai

Jyoti Chaudhary

Karthik Keertipati

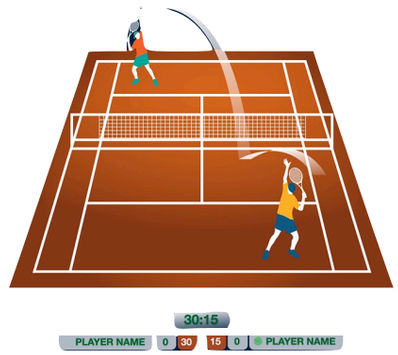
RUTGERS
Professional Science
Master's Program

Master Of Business And Science

Regression - Spring 2018

Introduction

Stats and Sports:



When it comes to statistically representing sports, tennis is remarkably well-suited.
Just consider, for a moment, the complexities in other sports that render them statistically unwieldy: football has over twenty people on the field at a given time; basketball has ten players on the court, all of whom affect the game simultaneously, even when playing off the ball.

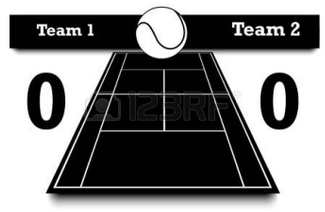
Tennis - mathematically and statistically straightforward game.

Analysis:

"Tennis Major Tournament Match Statistics Data Set"

women's game dataset.

Goal:



From our analysis, we wish to calculate the probability that one player A, wins a tennis match against another player B. It is not enough to know the rankings of A and B, as it is ambiguous to translate rankings into probabilities of winning.

Data Description

Source - UCI Machine Learning repository.

42 variables

452 observations

After Data Cleaning:

209 observations

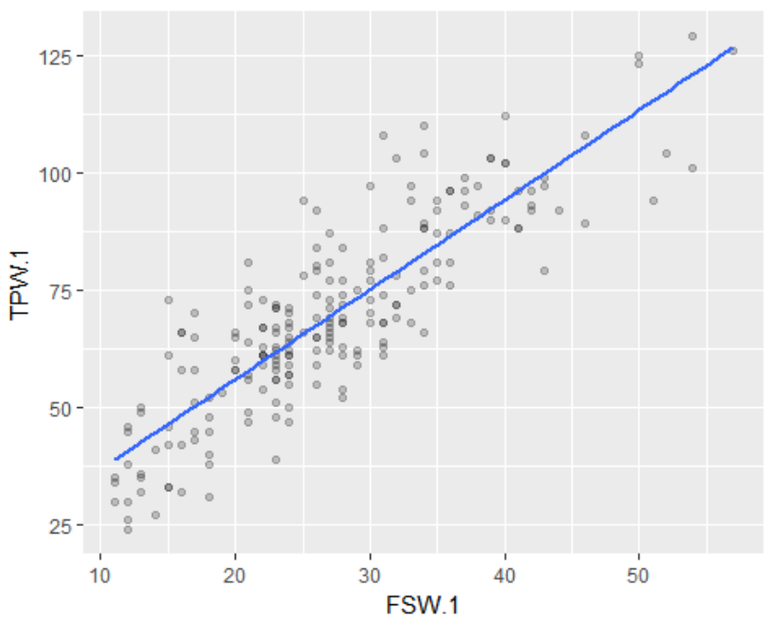
Same set of variables exists for player 1 and player 2 on the same record.

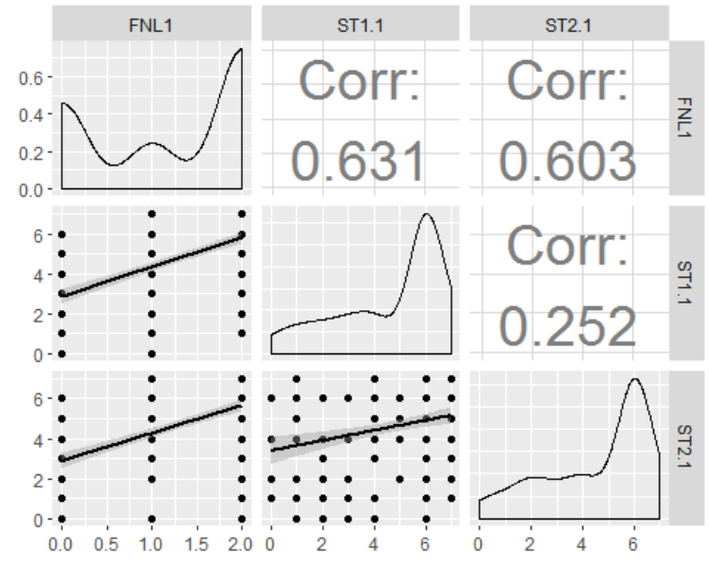
Player1/2	Player in the game
Round	Round of the tournament at which game is played (Numeric-Integer)
Result	A value of 1 or 0. "1" indicates that player1 won the game. "0" indicates that player2
FNL	Final Number of Set Won (Numeric-Integer)
FSP	First Serve Percentage (Real Number)
FSW	First Serve Won (Real Number)
SSP	Second Serve Percentage (Real Number)
SSW	Second Serve Won (Real Number)
ACE	Aces won (Numeric-Integer)
DBF	Double Faults committed (Numeric-Integer)
WNR	Winners earned (Numeric)
UFE	Unforced Errors committed (Numeric)
BPC	Break Points Created (Numeric)
BPW	Break Points Won (Numeric)
NPA	Net Points Attempted (Numeric)
NPW	Net Points Won (Numeric)
TPW	Total Points Won (Numeric)
ST1	Set 1 result (Numeric-Integer)
ST2	Set 2 Result (Numeric-Integer)
ST3	Set 3 Result (Numeric-Integer)
ST4	Set 4 Result (Numeric-Integer)
ST5	Set 5 Result (Numeric-Integer)

Correlation Analysis

Sensible step to understand how your different variable interact together.

Correlation - trends shared between two variables:





Considering the correlation that is high:

- strong correlation between FSW and TPW ==> remove TPW from the model
- strong correlation between FNL & ST1 and FNL & ST2 ==> remove FNL from the model fit

Based on similar analysis and the fact that these regressors created errors in our GLM model we decided to drop:

FNL (Final number of set won by player), TPW (Total points won by player), SSP (Second Serve percentage that was playable), NPA (Net Points Attempted by player), BPC (Break Points Created by player), NPW (Net Points Won by player), BPW (Break Points Won by player).

Model Selection

Full model: = *Result ~ ST1.1 + ST2.1 + FSP.1 + FSW.1 + SSW.1 + ST1.2 + ST2.2 + FSP.2 + FSW.2 + SSW.2 + ACE.1 + DBF.1 + WNR.1 + UFE.1 + ACE.2 + DBF.2 + WNR.2 + UFE.2*

Stats on the full model:

P-value for Null - Partial :0 Deviance 42.84735 Deviance / df : 0.2255124 P-Value 1
Pearson: 509.9543
Hosmer and Lemeshow goodness of fit (GOF) test
X-squared = 24.426, df = 8, p-value = 0.001943

Backward elimination on the full model(step):

Full model = *Result ~ ST1.1 + ST2.1 + FSP.1 + FSW.1 + SSW.1 + ST1.2 + ST2.2 + FSP.2 + FSW.2 + SSW.2 + ACE.1 + DBF.1 + WNR.1 + UFE.1 + ACE.2 + DBF.2 + WNR.2 + UFE.2*

Output = *FSW.1 + SSW.1 + FSW.2 + SSW.2 + WNR.1 + UFE.1 + ACE.2 + WNR.2 + UFE.2*

After forward, backward and stepwise selection:

- We found that forward selection is not efficient as the data has many variables.
- Backward and Stepwise selection has given a total of 9 same variables.
- Built our final model using variables from backward selection.

Backward selection:

- Start with all variables in the model
- Remove the variable with the largest p-value(Pr(>Chi)) ==> the variable that is the least statistically significant.
- Continued…….. until a stopping rule is reached.

*But in this process, there was one variable ACE.2, which has a p value greater than 0.05 but wasn't removed during the process and was included in the model. It could be because the algorithm worked on reducing AIC of the model rather than looking at p-values.

Model Selection

As we saw that one of the regressors in our model seems to have a high p-value we **test its significance to the model**

The statistics for the model: *FSW.1 + SSW.1 + FSW.2 + SSW.2 + WNR.1 + UFE.1 + ACE.2 + WNR.2 + UFE.2*

P-value for null-partial:0 Deviance / df :0.2497509 P-value : 1
Pearson:388.6017
Hosmer and Lemeshow goodness of fit (GOF) test
X-squared = 12.899, df = 8, p-value = 0.1154

$$H_0 : \beta_{ACE_2} = 0 \text{ vs } H_A : \beta_{ACE_2} \neq 0$$

Deviace of model *with* ACE2 **49.70043** Deviance of model *without* ACE2 **52.2205**

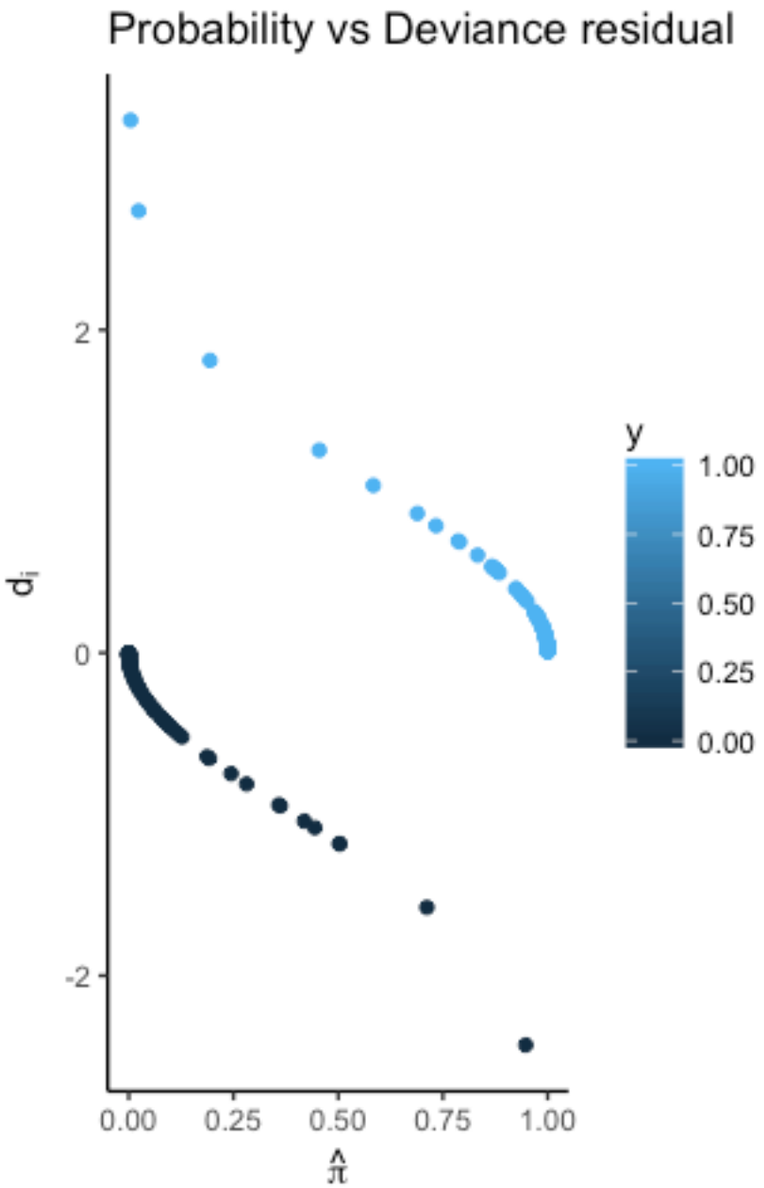
$$\text{Test statistic} = 2.52 < \chi^2_{,1,0.05}$$

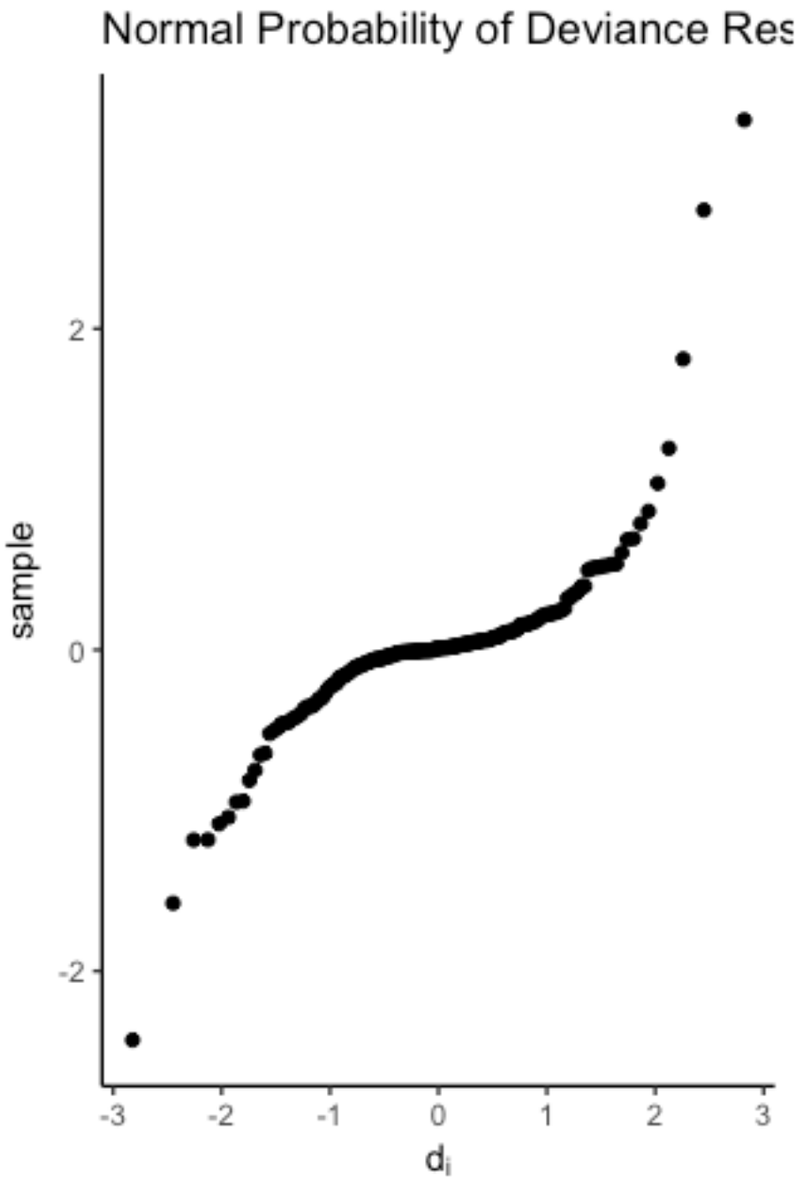
As the difference of deviance between the two models is less than the chi-square value, which also means adding ACE2 to existing model will result in a insignificant statistic, thus we don't have enough evidence to reject the null and say that ACE2 is a significant variable.

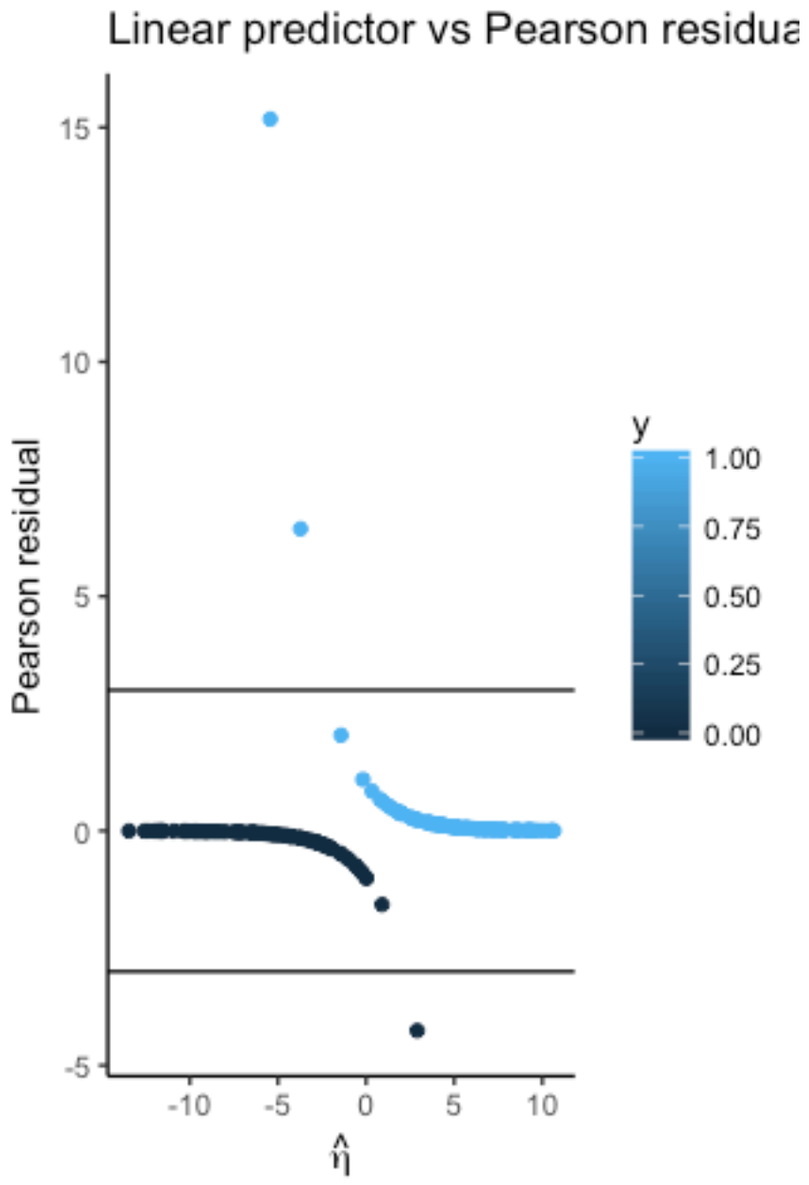
Confidence Interval	ACE.2	-0.78979783	0.07053231
CI for odds Ratio	ACE.2	4.539366E-01	1.0730792

Model Adequacy

"The deviance and Pearson residuals are the most appropriate for conducting model adequacy checks. Plots of these residuals versus the estimated probability and a normal probability plot of the deviance residuals are useful in checking the fit of the model at individual data points and in checking for possible outliers." (Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J, 2010,p 156)







Model Adequacy

i) Probability against deviance residual: Probability is $\frac{1}{1+e^{-(x'\beta)}}$ and d_i is the deviance residual. The blue points are all the points where player 1 won the game and the black points are all the games where player 2 won. The plot seems alright, with some certain points that could be outliers.

ii) Normal Probability Plot: Looking at the graph above, it does not seem to look to great. There seems to be problems with a few outliers and possible problems with the distribution of the residuals.

iii) Linear predictor vs Pearson residual: Linear predictor $(\ln \frac{\pi_i}{1-\pi_i})$ versus the Pearson residual. When comparing residuals we have a cut off region between -2/2 or -3/3, for above graph I chose regions of -3/3. With the plot we find the existing outlier's which can be dropped.

*Dropping the outliers results in a slight improvement in the model(not shown here)

Final Model

$$\hat{y} = \hat{\pi} = \frac{1}{1+e^{-(-4.211+0.346UFE_2-0.267SSW_2+0.3169WNR_1-0.27FSW_2+0.273FSW_1-0.263UFE_1+0.227WNR_2+0.264SSW_1)}}$$

Final Model: Result = UFE.2 + SSW.2 + WNR.1 + FSW.2 + FSW.1 + UFE.1 + WNR.2 + SSW.1
Null Deviance: 285.4019
Residual Deviance: 51.89584
P-value for null - partial: 0 Deviance / df : 0.2634307 P-value : 1
Pearson Statt : 300.2774
Hosmer and Lemeshow goodness of fit (GOF) test
X-squared = 9.6132, df = 8, p-value = 0.3

Analysis of Deviance Table (Type II tests)

Response: Result

ANOVA	Df	Chisq	Pr(>Chisq)	Sig
UFE.2	1	25.8560	3.679E-07	***
SSW.2	1	6.7450	0.0094009	**
WNR.1	1	18.6898	1.538E-05	***
FSW.2	1	10.4413	0.0012323	**
FSW.1	1	11.6257	0.0006505	***
UFE.1	1	14.8547	0.0001161	***
WNR.2	1	11.2071	0.0008148	***
SSW.1	1	5.5438	0.0185466	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

All our variables and our model gave satisfactory results.

Conclusion

- The final model is the best working model because it consists of only significant variables.
- The full model has better deviance, but it was proven that adding extra variables were not significant.
- One more fact that we would like to comment about is that amongst the discussed models the last one had the best HL statistic, hence we would say that it fit the best as well.

FAQ

1. What libraries did we use?

Ans: LogisticDx, MASS, ggplot2 , latex2exp , MKmisc, ResourceSelection , car, gridExtra and corrplot.

2. What alpha level were the tests?

Ans: They were all done with alpha 0.05

3. What type of tests were done to find significance in the ANOVA table?

Ans: It was the Wald test.

4. How did you get residual information?

Ans: We used logistic DX package, which will find the residuals needed.

5. Why didn’t we show further information about dropping observations?

Ans: The outliers were dropped but after that when we tried to make graphs out of it, maybe due to problems with the package; it was regenerating the dropped observations. So we decided to not go in deeper, but statically the deviance dropped by a small fraction after dropping outliers.

References

- Myers, R. H., Montgomery, D. C., Vinning, G. G., & Robinson, T. J. (2010). Generalized linear models: With applications in engineering and the sciences. Hoboken (New Jersey): John Wiley & Son.

- 7.2.1 - Model Diagnostics. (n.d.). Retrieved from https://onlinecourses.science.psu.edu/stat504/node/161

- Multiple Logistic Regression. (n.d.). Retrieved from https://rcompanion.org/rcompanion/e_07.html

- Residual plots for Binary Logistic Regression. (n.d.). Retrieved from http://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/regression/how-to/binary-logistic-regression/interpret-the-results/all-statistics-and-graphs/residual-plots/

Questions?

Thank You.