Chen-Tze Tsai

STAT 8561 2018

Linear Statistical Analysis Final Project

Predicting the Odds of Winning a Match in the First Round of Australia Open 2017

# Introduction

Tennis is one of the most commercialized sports. More than 100 Association of Tennis Professionals (ATP) tours are hosted by different cities all over the world each year, and a ticket to one single game may cost more than $7,000 (per the record in 2018, the highest price was $7,882). The performance of players has been passionately investigated and discussed by many sport TV programs, professionals, and fans. In support of the discussion, ATP released a series of data regarding players' performance in the end of each match. In this project, I used the performance data from the first round of match in Australia Open 2017. There were 128 players, half of whom won the match, and the performance was recorded in 10 variables. The aim of this project is to identify the performance factors that are significantly associated with the result of match, i.e., win or loss.

# Data Manipulation and Summary Statistics

Except age (in years), all the performance data are counts of a specific event. Since the count may be affected by the number of games played, I computed the rates of (1) first serve winning, (2) second serve winning, and (3) break point saving following the formulas below:

If a player faced no break point, the break point saving rate was set to 1 based on the assumption that if this player had faced break points he could have save them all. This is a reasonable assumption because facing no break point implies this player is not stressed by his opponent and has greater performance.

The means of these variables were compared between winners and losers using two-sample Student's *t*-test (shown in the table below).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Summary Statistics of the Players' Performance Grouped by the Result of Match** | | | | | | | |
|  |  | | Result of the match | | | |  |
|  | Total | | Win | | Loss | |  |
| Variable | (n = 128) | | (n = 64) | | (n = 64) | | *p*-value |
| Age in years | 27.9 | (5.0) | 27.8 | (5.0) | 28.0 | (5.1) | 0.819 |
| Ace | 11.1 | (9.6) | 8.5 | (6.6) | 13.6 | (11.4) | 0.002 |
| Total serves per game | 17.4 | (5.8) | 17.1 | (5.9) | 17.6 | (5.8) | 0.598 |
| Total serve points won | 109.4 | (35.8) | 112.1 | (35.4) | 106.7 | (36.2) | 0.395 |
| First serve success | 65.4 | (23.8) | 67.0 | (24.0) | 63.7 | (23.6) | 0.430 |
| First serve points won | 47.6 | (18.7) | 45.3 | (19.1) | 50.0 | (18.2) | 0.163 |
| First serve winning rate | 0.73 | (0.10) | 0.66 | (0.08) | 0.79 | (0.08) | < 0.001 |
| Second serve points won | 22.2 | (8.1) | 21.0 | (8.3) | 23.4 | (7.9) | 0.104 |
| Second serve points lost | 17.1 | (6.9) | 18.7 | (6.2) | 15.4 | (7.2) | 0.006 |
| Second serve winning rate | 0.57 | (0.11) | 0.52 | (0.09) | 0.62 | (0.10) | < 0.001 |
| Double fault | 4.7 | (4.8) | 5.3 | (2.7) | 4.2 | (2.8) | 0.027 |
| Break points faced | 8.9 | (5.2) | 11.3 | (4.4) | 6.5 | (5.0) | < 0.001 |
| Break points saved | 5.3 | (3.8) | 6.2 | (3.6) | 4.5 | (3.7) | 0.009 |
| Break point saving rate | 0.60 | (0.23) | 0.51 | (0.17) | 0.70 | (0.25) | < 0.001 |
| All the data are presented as mean (SD). Two-sample independent Student's *t*-test with unequal variance was used to compare the means between the winners and the losers. | | | | | | | |

# Maximum Likelihood Estimation

Logistic regression model has a form of

which is equivalent to

Since the binary outcome, , follows a binomial distribution with probability , the likelihood function of the model parameters is

and the MLEs of the model parameters, , are found by maximizing the likelihood function.

# Model Selection

I started with the full model, which includes the covariates showing significant difference in the means between the winners and the losers.

##   
## Call:  
## glm(formula = result ~ ., family = "binomial", data = firstwork)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.00565 -0.36780 -0.00382 0.40536 2.51887   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.560e+01 7.875e+00 -3.251 0.001149 \*\*   
## ace -4.280e-02 4.463e-02 -0.959 0.337513   
## df -3.588e-01 1.297e-01 -2.766 0.005683 \*\*   
## bp\_saved 7.760e-02 3.694e-01 0.210 0.833614   
## bp\_faced -8.089e-02 3.060e-01 -0.264 0.791498   
## second\_lose -8.086e-04 7.693e-02 -0.011 0.991614   
## save\_rate 4.177e+00 2.619e+00 1.595 0.110733   
## first\_rate 2.938e+01 8.643e+00 3.400 0.000675 \*\*\*  
## second\_rate 7.835e+00 4.968e+00 1.577 0.114781   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 177.446 on 127 degrees of freedom  
## Residual deviance: 75.943 on 119 degrees of freedom  
## AIC: 93.943  
##   
## Number of Fisher Scoring iterations: 7

Using Akaike Information Criterion (AIC) as the selection criterion, I performed backward and bi-directional stepwise selection. The two procedure gave the same final model, which includes (1) double fault, (2) break point saving rate, (3) first serve winning rate, and (4) second serve winning rate.

##   
## Call:  
## glm(formula = result ~ df + save\_rate + first\_rate + second\_rate,   
## family = binomial, data = firstwork)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.00705 -0.39563 -0.00291 0.41676 2.53508   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -25.1446 4.7689 -5.273 1.35e-07 \*\*\*  
## df -0.3988 0.1161 -3.435 0.000592 \*\*\*  
## save\_rate 4.0404 1.6715 2.417 0.015638 \*   
## first\_rate 27.7174 5.6122 4.939 7.86e-07 \*\*\*  
## second\_rate 8.1749 3.6613 2.233 0.025564 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 177.45 on 127 degrees of freedom  
## Residual deviance: 77.18 on 123 degrees of freedom  
## AIC: 87.18  
##   
## Number of Fisher Scoring iterations: 6

# Interpreting the Model

Since all the covariates in this dataset are continuous variable, the estimated regression coefficients represent the log-transformed odds ratio (OR) of winning a match as the corresponding covariates increase in one unit. However, break point saving rate (save\_rate), first serve winning rate (first\_rate), and second serve winning rate (second\_rate) range from 0 to 1, so I chose to calculate the ORs as these covariates increase in 0.1 instead.

## df (+1) save\_rate (+0.1) first\_rate (+0.1)   
## 0.6711119 1.4978711 15.9864862   
## second\_rate (+0.1)   
## 2.2648084

One unit of increase in double fault decreases the OR of winning a match about two-third, while 10% of increase in break point saving rate, first serve winning rate, and second serve winning rate increases the OR around 1.5, 16, and 2.2 folds.

# Goodness of Fit

I used likelihood ratio test to investigate the contribution of each covariate.

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: result  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 127 177.45   
## df 1 5.000 126 172.45 0.02535 \*   
## save\_rate 1 23.473 125 148.97 1.267e-06 \*\*\*  
## first\_rate 1 65.943 124 83.03 4.643e-16 \*\*\*  
## second\_rate 1 5.850 123 77.18 0.01558 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

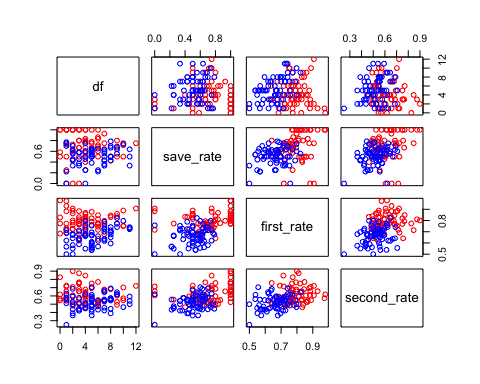
The entry of each covariate significantly improves the model fitting according to the deviance, which approximately follows a distribution (by Wilk’s theorem). I then used Hosmer-Lemeshow Test to test the goodness of fit.

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: final$y, fitted(final)  
## X-squared = 5.6052, df = 8, p-value = 0.6914

The -value indicates there is no significant disagreement between the observed and expected outcomes.

# Collinearity

I first explored the correlation between covariates using scatter plots and Pearson correlation coefficient.



## df save\_rate first\_rate second\_rate  
## df 1.00 -0.04 0.05 -0.05  
## save\_rate -0.04 1.00 0.41 0.43  
## first\_rate 0.05 0.41 1.00 0.41  
## second\_rate -0.05 0.43 0.41 1.00  
##   
## n= 128   
##   
##   
## P  
## df save\_rate first\_rate second\_rate  
## df 0.6553 0.5588 0.5865   
## save\_rate 0.6553 0.0000 0.0000   
## first\_rate 0.5588 0.0000 0.0000   
## second\_rate 0.5865 0.0000 0.0000

There are significant correlations among break point saving rate, first serve winning rate, and second serve winning rate. The correlation coefficients are around 0.4. To investigate the impact of these collinearities, I calculated the variance inflation factor (VIF) for each covariate.

## df save\_rate first\_rate second\_rate   
## 1.414135 1.082992 1.363839 1.010885

None of the covariate has VIF larger than 5, indicating there is no severe collinearity.

# Conclusion

I analyzed the players’ performance data from the first round of match in Australia Open 2017. Using logistic regression, I identified the factors that are associated with the result of the match. Qualitatively, the prediction of these factors is intuitive, since the factors for ‘good’ performance (break point saving rate, first serve winning rate, and second serve winning rate) are positively correlated with the probability of winning a match, while double fault, the factor for ‘bad’ performance, is negatively correlated. Also, I see that three out of the four covariates included in the final model are indicators of the player’s performance on serving (double fault, first serve winning rate, and second serve winning rate). This agrees with the consensus that, unlike badminton or table tennis, the quality of serving in tennis is relatively critical for scoring.

# Appendix: R code

require(Hmisc)  
require(MASS)  
require(car)  
require(ResourceSelection)  
  
setwd('/Users/tiger/Dropbox/JT/Final Project/')  
auo <- read.table('./data/AUO 2017.txt', header = F, sep = ',')  
auo.1 <- auo[which(auo[, 7] <= 163), ] # First rounds  
win <- auo.1[, c(8:17, 32:40)]; colnames(win) <- paste0('V', 1:19)  
los <- auo.1[, c(18:27, 41:49)]; colnames(los) <- paste0('V', 1:19)  
auo.2 <- rbind(win, los)  
colnames(auo.2) <- c('id', 'seed', 'entry', 'name', 'hand', 'ht',  
 'ioc', 'age', 'rank', 'rank\_points', 'ace',  
 'df', 'svpt', 'first\_in', 'first\_won', 'second\_won',  
 'svgms', 'bp\_saved', 'bp\_faced')  
attach(auo.2)  
second\_lose <- svpt-first\_in-df-second\_won  
save\_rate <- ifelse(bp\_faced != 0, bp\_saved/bp\_faced, 1)  
first\_rate <- first\_won/first\_in  
second\_rate <- second\_won/(second\_lose+second\_won)  
result <- rep(c(1, 0), each = 64)  
detach(auo.2)  
  
auo.3 <- auo.2[, c('age', 'ace', 'df', 'svpt', 'first\_in', 'first\_won',  
 'second\_won', 'svgms', 'bp\_saved', 'bp\_faced')]  
firstrd <- cbind(auo.3, second\_lose, save\_rate, first\_rate, second\_rate, result)  
var.name <- names(firstrd)  
  
# Summary statistics (Table 1)  
out <- c()  
for (i in 1:14){  
 xx <- firstrd[, var.name[i]]  
 mt <- mean(xx)  
 st <- sd(xx)  
 mm <- tapply(xx, result, mean)  
 ss <- tapply(xx, result, sd)  
 test <- t.test(xx ~ result)  
 out <- rbind(out, c(mt, st, mm[1], ss[1], mm[2], ss[2], round(test$p.value, 3)))  
   
}  
rownames(out) <- var.name[1:14]  
colnames(out) <- c('Mean (total)', 'SD (total)', 'Mean (win)', 'SD (win)',  
 'Mean (lose)', 'SD (lose)', 'p.value')  
out  
  
# Correlations among independent variables  
firstsig <- firstrd[, c('ace', 'df', 'bp\_saved', 'bp\_faced', 'second\_lose',  
 'save\_rate', 'first\_rate', 'second\_rate')]  
pairs(firstsig, cex = 0.7, col = ifelse(result == 0, 4, 2))  
rcorr(as.matrix(firstsig))  
  
# Logistic regression  
firstwork <- cbind(firstsig, result)  
out.1 <- glm(result ~ ., data = firstwork, family = 'binomial')  
summary(out.1)  
  
# Backward selection  
stepAIC(out.1)  
# Bi-directional selection  
stepAIC(out.1, scope = list(upper = ~., lower = ~1))  
  
final <- glm(result ~ df+save\_rate+first\_rate+second\_rate,  
 data = firstwork, family = binomial)  
summary(final)  
  
# Interpretation of model (odds ratio)  
exp(final$coefficient)  
confint(final)  
  
# Goodness of fit  
anova(final, test = 'LRT')  
hoslem.test(final$y, fitted(final))  
  
# Collinearity between first\_rate and second\_rate  
firstreg <- firstrd[, c('df', 'save\_rate', 'first\_rate', 'second\_rate')]  
pairs(firstreg, col = ifelse(result == 0, 4, 2))  
  
# Variance inflation factor  
vif(final)