Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per l'economia e l'impresa


# Esiste un teorema Pitagorico per le vittorie nel tennis?

Relatore Prof. Francesco Lisi
Dipartimento di Scienze Statistiche

Laureando Alessandro Pittalis
Matricola 2002695

Anno Accademico 2022/2023

# Indice

# Introduzione

La trattazione di questo elaborato verterà sulla specificazione e sull'adattamento di un modello in grado di stimare il numero di vittorie di una determinato giocatore per un intera stagione tennistica. Questo modello definito "Pitagorico" per la prima volta da James (1981) venne utilizzato inizialmente nel baseball e successivamente venne applicato in altri sport. Nel tennis e in generale negli sport individuali,il modello Pitagorico viene introdotto per la prima volta da Kovalchik (2016).

Il principale obbiettivo affrontato in questo elaborato è quello di specificare e stimare il modello Pitagorico, utilizzando dati più attuali ripetto a quelli utilizzati da Kovalchik (2016), in modo da essere in grado di evidenziare possibili differenze in relazione ai due periodi temporali. In un secondo momento si vorrà analizzare se il modello presenta differenze sostanziali in termini di precisione di risultati e di stime dei parametri in funzione della divisione delle superfici in 3 gruppi: terra, cemento, erba.

Nel primo capitolo di questo elaborato, verrà dedicata attenzione alla storia della rivoluzione sabermetrica[1] negli sport, con un focus particolare sull'introduzione del modello Pitagorico nel baseball e la sua successiva adozione in altri sport. Sarà quindi fornita una descrizione generale del modello Pitagorico e verranno delineati gli obbiettivi finali specifici per questo elaborato, concentrandosi sul contesto del tennis e sulla valutazione dell'applicazione del modello Pitagorico nei dati più recenti.

Nel secondo capitolo, si passerà alla descrizione del dataset di stima utilizzato nella ricerca. Saranno illustrati i dettagli su come il dataset è stato reperito e strutturato, inclusi i criteri di selezione e i metodi di raccolta dei dati. Verrà inoltre condotta un'analisi delle correlazioni all'interno del dataset, al fine di identificare eventuali relazioni tra le variabili che potrebbero influenzare i risultati. Infine, verrà eseguita un'analisi fattoriale confermativa sulla matrice delle covariate, per esplorare ulteriormente la struttura dei dati e identificare eventuali fattori latenti che potrebbero influire sul modello Pitagorico.

[1]La sabermetrica è l'analisi del baseball attraverso le statistiche. Il termine deriva dall'acronimo SABR, che sta per "Society for American Baseball Research"

Nel terzo capitolo, verranno specificati e spiegati in maniera precisa i modelli e gli indici di determinazione che verranno in seguito adattati ai diversi dataset.

Nel quarto capitolo, approfondiremo i risultati ottenuti dalla nostra analisi condotta nei capitoli precedenti. Questa sezione sarà dedicata a fornire spiegazioni dettagliate sui risultati emersi e a condurre un'analisi approfondita delle differenze osservate tra le diverse superfici di gioco, concentrandoci sulla loro influenza nella precisione del modello Pitagorico.

Infine, nelle conclusioni, verranno riassunte tutte le procedure eseguite e i risultati ottenuti, mettendo in evidenza l'evoluzione del modello Pitagorico nel contesto del tennis e sottolineando l'importanza delle diverse superfici di gioco nella precisione dei modelli predittivi. Saranno inoltre fornite considerazioni finali sull'applicabilità e le limitazioni del modello Pitagorico nel tennis e potenziali direzioni future di ricerca per ulteriori miglioramenti e sviluppi nel campo dell'analisi statistica degli sport individuali.

# Capitolo 1

# Introduzione

## 1.1 La rivoluzione sabermetrica negli sport

Bill James è universalmente riconosciuto come il pioniere della rivoluzione sabermetrica nel baseball. Fu lui a coniare il termine "sabermetrica", questo vocabolo deriva dall'acronimo SABR, che sta per "Society for American Baseball Research" (Società americana per la ricerca sul baseball). La sabermetrica può essere definita come l'approccio statistico avanzato che mira a ottenere una conoscenza oggettiva del baseball. Questa rivoluzione non ha influenzato solo il baseball, ma successivamente si è estesa anche ad altri sport di squadra come ad esempio l'hockey (Cochran & Blackstock (2009)) e il calcio (Hamilton (2011); Caro & Machtmes (2013)).

Nel tennis e negli sport individuali, tuttavia, la rivoluzione verso un'analisi più oggettiva ha impiegato maggior tempo per svilupparsi; Infatti solo nel corso degli anni 2000 si è verificata una rapida crescita nell'uso dei dati al fine di interpretare le dinamiche del gioco. Attualmente grazie alle moderne tecnologie e agli strumenti di rilevamento di informazioni, è possibile registrare e analizzare una vasta gamma di statistiche nel tennis. Sensori posizionati a bordo campo, telecamere e dispositivi indossabili offrono un controllo quasi completo sulle dinamiche di questo sport.

L'analisi di queste variabili, come la percentuale di punti vinti al servizio o il numero di break point convertiti, consente di individuare i punti di forza e di debolezza di un giocatore e, soprattutto, di identificare i momenti cruciali dell'incontro, adattandosi al contesto dell'avversario e quindi permettendo una preparazione psicologica e fisica mirata a fronteggiare tali momenti.

L'analisi statistica e matematica applicata allo sport è in costante evoluzione, queste ultime rientrano nel contesto dello "sport analytics". Questa evoluzione apre la strada a

una comprensione ancora più profonda del gioco stesso, mostrando come i modelli statistici impiegati nello sport presentano un potenziale notevole nel plasmare le strategie di gioco e le metodologie di allenamento, in particolare quando riescono a rivelare le dinamiche di gioco più influenti nel determinare l'esito di una partita.

## 1.2 "Teorema" Pitagorico

Uno dei grandi teoremi che vennero introdotti da James (1981), assume la denominazione di "teorema Pitagorico" poiché la sua formula si fonda sulla somma dei quadrati dei punti accumulati da una squadra. Questo totale viene poi diviso per la somma dei quadrati dei punti realizzati dalla squadra stessa e dai suoi avversari.

Tale formulazione Pitagorica rappresenta uno dei contributi quantitativi di maggiore rilievo nel campo dell'analisi sportiva. In origine applicata al baseball, ha dimostrato notevole efficacia nell'offrire una stima accurata della percentuale di vittorie di una squadra, basata sui "runs[1]" conseguiti e subiti durante le partite.

In generale, è possibile fare riferimento all'aspettativa Pitagorica, poiché questo modello è capace di stimare il numero di partite vinte all'interno di una stagione, fornendo quindi una prospettiva sull'aspettativa di vittorie di una squadra in relazione solamente al numero di "runs" realizzati e subiti.

La formula generale del modello pitagorico nel baseball, proposta da James, è la seguente:

$$P(\text{Vittoria}) = \frac{X^{\alpha}}{X^{\alpha} + Y^{\alpha}}, \tag{1.1}$$

dove $X$ rappresenta i runs conseguiti dalla squadra in casa e $Y$ rappresenta i runs conseguiti dalla squadra in trasferta. Inizialmente il parametro ($\alpha$), chiamato "esponente pitagorico" veniva definito con un valore di 2, in conformità con quanto definito nel teorema Pitagorico. Tuttavia, uno studio successivo condotto da Davenport & Woolner (1999) dimostrò che attraverso la trasformazione della formula pitagorica in un modello statistico e quindi la trasformazione di $\alpha$ da una costante nota ad un parametro da stimare sui dati effettivi, portava all'ottenimento di risultati più accurati. Nello specifico, è emerso che l'uso di un esponente pari a 1.85 produceva risultati migliori e più coerenti con le evidenze empiriche. Questa ricalibrazione dell'esponente ha consentito di ottenere stime più precise e in linea con le osservazioni reali.

[1]Il Runs o Punto è quello conseguito da un giocatore in attacco che da battitore diventa corridore e tocca nell'ordine prima, seconda, terza e casa base

Basata sul principio matematico del teorema di Pitagora, questa formula consente di calcolare la probabilità di vittoria della squadra di casa, tenendo conto delle performance sia offensive che difensive della squadra stessa.

Attraverso la seguente dimostrazione matematica possiamo ottenere una trasformazione interessante della formula Pitagorica:

Partiamo dalla Formula 1.1:

$$P(\text{Vittoria}) = \frac{X^\alpha}{X^\alpha + Y^\alpha}$$

Calcoliamo il log-odds (logit) della probabilità di vittoria:

$$\text{logit}[P(\text{Vittoria})] = \ln\left(\frac{P(\text{Vittoria})}{1 - P(\text{Vittoria})}\right)$$

Sostituendo $P(\text{Vittoria})$ dalla Formula 1.1:

$$\text{logit}[P(\text{Vittoria})] = \ln\left(\frac{\frac{X^\alpha}{X^\alpha+Y^\alpha}}{1 - \frac{X^\alpha}{X^\alpha+Y^\alpha}}\right)$$

Semplificando l'espressione al denominatore:

$$\text{logit}[P(\text{Vittoria})] = \ln\left(\frac{X^\alpha}{Y^\alpha}\right)$$

Applichiamo la proprietà del logaritmo del rapporto:

$$\text{logit}[P(\text{Vittoria})] = \ln(X^\alpha) - \ln(Y^\alpha)$$

Utilizzando la proprietà del logaritmo della potenza:

$$\text{logit}[P(\text{Vittoria})] = \alpha \ln(X) - \alpha \ln(Y)$$

Ora possiamo considerare la differenza tra i logaritmi:

$$\text{logit}[P(\text{Vittoria})] = \ln(X^\alpha) - \ln(Y^\alpha) = \ln(X^\alpha / Y^\alpha)$$

Applichiamo la proprietà dell'esponente nei logaritmi:

$$\text{logit}[P(\text{Vittoria})] = \ln\left(\left(\frac{X}{Y}\right)^\alpha\right)$$

Semplifichiamo ulteriormente:

$$\text{logit}[P(\text{Vittoria})] = \alpha \ln\left(\frac{X}{Y}\right) \tag{1.2}$$

Questa formulazione (Rosenfeld et al. (2010)), basata sul legame logit, mette in luce il fatto che la probabilità di vittoria per tutta la stagione dipende da un indice relativo di forza della squadra indipendente dal avverario incontrato, infatti: La X è la frequenza cumulata di una determinata statistica calcolata per la squadra in considerazione in tutte le partite giocate in un fissato periodo di tempo, mentre la Y è sempre la frequenza cumulata della stessa statistica che invece la squadra ha subito nelle diverse partite dello stesso periodo temporale di riferimento.

È stata riconosciuta l'importanza di questo modello nel baseball dagli analisti del settore in quanto fornisce uno strumento fondamentale per valutare le prestazioni delle squadre, infatti attraverso il confronto tra le vittorie attese calcolate tramite il modello e quelle effettive è possibile valutare l'efficacia delle squadre nel tradurre i risultati ottenuti in campo in vittorie effettive. Questo tipo di analisi ha aperto nuove prospettive per una maggiore comprensione del gioco e ha consentito agli allenatori e ai dirigenti sportivi di prendere decisioni più consapevoli basate su dati quantitativi.

## 1.3 Obiettivi dell'analisi

Nel contesto del tennis, estendere il modello pitagorico è stata una sfida complessa, poiché bisognava adattare un modello ideato per gli sport di squadra ad uno sport individuale.

L'obiettivo di questo elaborato sarà quello di replicare quanto fatto da Kovalchik (2016), utilizzando lo stesso periodo di partite ma cercando di fare più luce sulle metodologie utilizzate. In un secondo momento, si è applicato il modello a un nuovo set di dati costituito dalle partite più recenti al fine di analizzare se e come il gioco del tennis sia cambiato nel tempo. Un'ulteriore approfondimento che si desidera effettuare riguarda l'adattamento del modello su diverse superfici. Infatti, è noto agli appassionati di questo sport che il tipo di superficie (terra, erba, cemento) determina in modo significativo il gioco. Attraverso questo studio, si vuole contribuire alla comprensione delle prestazioni degli atleti nel tennis e valutare se vi siano differenze significative tra le diverse superfici, aprendo così nuove prospettive nell'analisi quantitativa delle prestazioni nel tennis.

# Capitolo 2

# Il Dataset

## 2.1 I dati

Come accennato nel capitolo precedente, lo studio si suddividerà in due parti: la prima verterà sull'analisi di un dataset contenente le partite dal 2004 al 2015, al fine di ottenere risultati comparabili con quelli riportati da Kovalchik (2016). Nella seconda parte, invece, i modelli saranno adattati utilizzando dati più recenti, infatti il set di dati considererà tutti gli incontri giocati dal 2010 al 2021.

Per condurre questa ricerca, sono stati utilizzati due dataset provenienti dalle librerie gestite e organizzate da Jeff Sackmann[1]. Le librerie di Sackmann sono una fonte aperta che non ha una collaborazione diretta con l'ATP(Association of Tennis Professionals). Esse si basano semplicemente sulle informazioni disponibili sul sito web ATP Tour, senza coinvolgimento diretto da parte dell'associazione nella loro gestione. Queste librerie rappresentano una risorsa preziosa per gli studiosi e gli appassionati di tennis interessati ad analizzare le prestazioni dei giocatori e studiare le tendenze nel corso degli anni.

I dati presenti nelle librerie di Sackmann sono stati sottoposti a un'operazione preliminare di pulizia. In particolare, sono state eliminate le partite in cui mancavano informazioni complete su una o più variabili utilizzate nell'analisi, è stata quindi applicata la procedura del listwise delection, successivamente sono state escluse le partite in cui il giocatore vincitore ha ottenuto più del 60% dei punti totali. Queste procedure comportano una riduzione del numero di partite disponibili per l'analisi rispetto a quanto riportato da Kovalchik. Nonostante questa sottile variazione, le conclusioni raggiunte avranno lo stesso significato.

[1] https://github.com/JeffSackmann/tennis_atp

I due dataset utilizzati inizialmente per le analisi, indipendentemente dal periodo considerato, presentano una struttura comune. Essi includono tutte le partite dei giocatori del circuito ATP, comprese le partite giocate in tornei esterni, come ad esempio: Olimpiadi, Davis Cup e Finals.

Nella selezione delle partite finali, sono state considerate solamente quelle appartenenti ai tornei del circuito professionistico maschile con una classificazione di almeno 250 punti (assegnati al vincitore del torneo) e in cui almeno un giocatore dei due appartenesse ai primi 100 in classifica (secondo l'ATP ranking). Al contrario, sono state escluse tutte le partite degli altri eventi menzionati in precedenza.

Nel periodo compreso tra il 2004 e il 2014, sono state selezionate un totale di 28.669 partite giocate da 318 giocatori. Per la validazione esterna, è stato utilizzato un dataset relativo alla stagione 2015, che comprende 2.716 partite e 120 giocatori. Questo dataset è stato considerato un insieme di dati indipendente per la verifica e la validazione dei modelli.

Per il periodo dal 2010 al 2020, è stato creato un dataset che include 26.397 partite giocate da 306 giocatori. Allo stesso modo, per la validazione esterna, è stata considerata la stagione 2021, con un dataset che contiene 2.732 partite e 109 giocatori. Ciò consente di testare l'efficacia dei modelli su un periodo più recente.

Per ogni partita, sono state registrate e analizzate 49 variabili, che includono non solo il risultato, ma anche statistiche dettagliate sul servizio e sulla risposta, oltre a informazioni fisiche sui singoli giocatori. Tutte queste variabili sono state attentamente controllate e documentate nel sito web che si è occupato della pubblicazione dei dati in forma grezza, ovvero l'ATP TOUR.

In linea con Kovalchik (2016), il processo di selezione delle variabili prevede inizialmente l'individuazione di un totale di 27 variabili diverse. Successivamente, da questa vasta lista iniziale, ne vengono scelte solo 10. Queste 10 variabili selezionate saranno poi sottoposte a un processo di calcolo bidirezionale in relazione al giocatore preso in considerazione.

Per essere più precisi, per ciascun giocatore oggetto dell'analisi e per ciascuna delle 10 statistiche scelte, verranno calcolati due valori distinti. Il primo valore rappresenterà la quantità prodotta, cioè ciò che il giocatore ha realizzato in termini di prestazioni specifiche in tutte le partite della stagione. Il secondo valore indicherà la quantità subita, ovvero ciò che il giocatore ha subito a causa delle prestazioni degli avversari in tutte le partite del anno indipendentemente dai giocatori affrontati. Questi valori verranno calcolati all'interno di uno specifico intervallo temporale, fornendo quindi una misura dell'efficacia sia delle azioni eseguite dal giocatore che di quelle affrontate dagli

avversari durante questo periodo.

Successivamente, sarà calcolato il logaritmo naturale del rapporto tra il valore prodotto dal giocatore e il valore prodotto dagli avversari, per ciascuna delle 10 statistiche. Questa operazione crea quelli che definamo come "indici di forza" che sono relativi quindi ad ogni singolo giocatore e hanno valori diversi per ogni singola statistica.

Ricordiamo che la scelta delle variabili è un assunzione ricavata dallo studio di Kovalchik (2016), secondo il quale le 10 statistiche scelte siano sufficentemente incorrelate tra di loro. Avendo dei dataset iniziali differenti si è impossibilitati dalla creazione delle 27 variabili quindi assumiamo che anche per i nostri dati le variabili sufficentemente incorrelate siano le stesse presentate nello studio di riferimento.

Si può osservare che nel momento in cui andremo a creare gli indicatori di forza, sette saranno relativi alla risposta, mentre gli altri tre saranno relativi al servizio.

Il dataset avrà le seguenti colonne, sottolineiamo come ogni statistica sia riferita all'intera stagione:

| **Nome della variabile** | **Tipo di variabile** |
|---|---|
| Nome del giocatore | Variabile qualitativa non ordinale |
| N. di partite giocate | Variabile quantitativa discreta |
| N. di partite vinte | Variabile quantitativa discreta |
| Ranking più alto del giocatore | Variabile quantitativa discreta |
| N. di Aces fatti | Variabile quantitativa discreta |
| N. di Aces subiti | Variabile quantitativa discreta |
| N. di punti vinti in risposta alla prima di servizio | Variabile quantitativa discreta |
| N. di punti subiti in risposta alla prima di servizio | Variabile quantitativa discreta |
| N. di punti vinti in risposta alla seconda di servizio | Variabile quantitativa discreta |
| N. di punti subiti in risposta alla seconda di servizio | Variabile quantitativa discreta |
| N. di break point vinti | Variabile quantitativa discreta |
| N. di break point subiti | Variabile quantitativa discreta |
| N. di opportunità di break point avute | Variabile quantitativa discreta |
| N. di opportunità di break point subite | Variabile quantitativa discreta |
| N. di punti vinti nel tiebreak | Variabile quantitativa discreta |
| N. di punti subiti nel tiebreak | Variabile quantitativa discreta |
| N. di punti vinti al servizio | Variabile quantitativa discreta |
| N. di punti subiti al servizio | Variabile quantitativa discreta |
| N. di doppi falli fatti | Variabile quantitativa discreta |
| N. di doppi falli fatti dagli avversari | Variabile quantitativa discreta |
| Percentuale di primi punti vinti con la prima di servizio | Variabile quantitativa continua |
| Percentuale di primi punti subiti con la prima di servizio | Variabile quantitativa continua |
| Percentuale di primi punti vinti con la seconda di servizio | Variabile quantitativa continua |
| Percentuale di primi punti subiti con la seconda di servizio | Variabile quantitativa continua |

Come spiegato in precedenza ricordiamo che da questo dataset si passerà poi alla costruzione della struttura finale formata dagli indici di forza.

(Esempio di indicatori di forza: $\ln\left(\frac{\text{break point fatti}}{\text{break point subiti}}\right)$)

### 2.1.1 Differenze rispetto al dataset di Kovalchik

Come accennato in precedenza il dataset iniziale utilizzato da Kovalchik (2016) è differente da quello che viene utilizzato per le analisi in questo elaborato, questa differenza è dovuta dalle operazioni preliminari di pulizia condotte da Sackman e da un ulteriore differenza: Kovalchik (2016) aveva selezionato solamente le partite disputate dai primi 100 giocatori nel ranking all'inizio di ogni stagione, esludendo da questa classifica anche chi avesse giocato un numero non sufficiente di punti nella stagione in considerazione, nel presente studio invece non è stata effettuata questa distinzione: sono stati incluse le partite riferite ai giocatori che sono entrati almeno una volta nella top 100 durante il periodo di riferimento.

Queste differenze potrebbero influire leggermente sulle stime, ma non avranno un impatto significativo sui risultati attesi.

## 2.2 Analisi esplorativa

È stata condotta un'analisi esplorativa al fine di dimostrare la sufficiente incorrelazione degli indici di forza. Dal articolo di riferimento, Kovalchik (2016), sappiamo che sono state selezionate le variabili in modo che il loro VIF (Variance Inflation Factor) fosse inferiore a 100 e che gli indici di forza potessero essere raggruppati in 2 gruppi: variabili inerenti alla risposta e variabili inerenti al servizio.

Pertanto, in questo elaborato, verrà analizzata la correlazione tra le variabili e attraverso un analisi dei fattori si vedrà se queste possano essere riassunte in 2 fattori generali come:"risposta" e "servizio", in seguito si analizzerà la presenza di multicollinearità e mostreremo come le 10 variabili scelte abbiano un $VIF<100$ in linea con quanto espresso nell'articolo di riferimento. È necessario sottolineare che la multicollinearità si verifica quando due o più variabili indipendenti in un modello di regressione lineare sono altamente correlate tra loro. Questa correlazione può causare problemi nell'interpretazione dei coefficienti di regressione e può influenzare la stabilità e l'accuratezza del modello.

Nella tabella .1 in Appendice sono spiegati i significati dei nomi delle variabili.

Come ci aspettavamo, i risultati nel Grafico 2.1 mostrano che ci sono dei gruppi di variabili che mostrano una forte correlazione tra di loro, ma non sembra esserci una netta divisione tra le variabili riguardanti i 2 ipotetici fattori: risposta e servizio.

Per quanto riguarda la multicollinearità così come accade nel lavoro di Kovalchik anche in questo elaborato viene dimostrato che i valori del VIF delle variabili selezionate sono tutti minori di 100 (vedi Tabella 2.1).

FIGURA 2.1: Scatterplot matrix per gli indici di forza 2004-2014

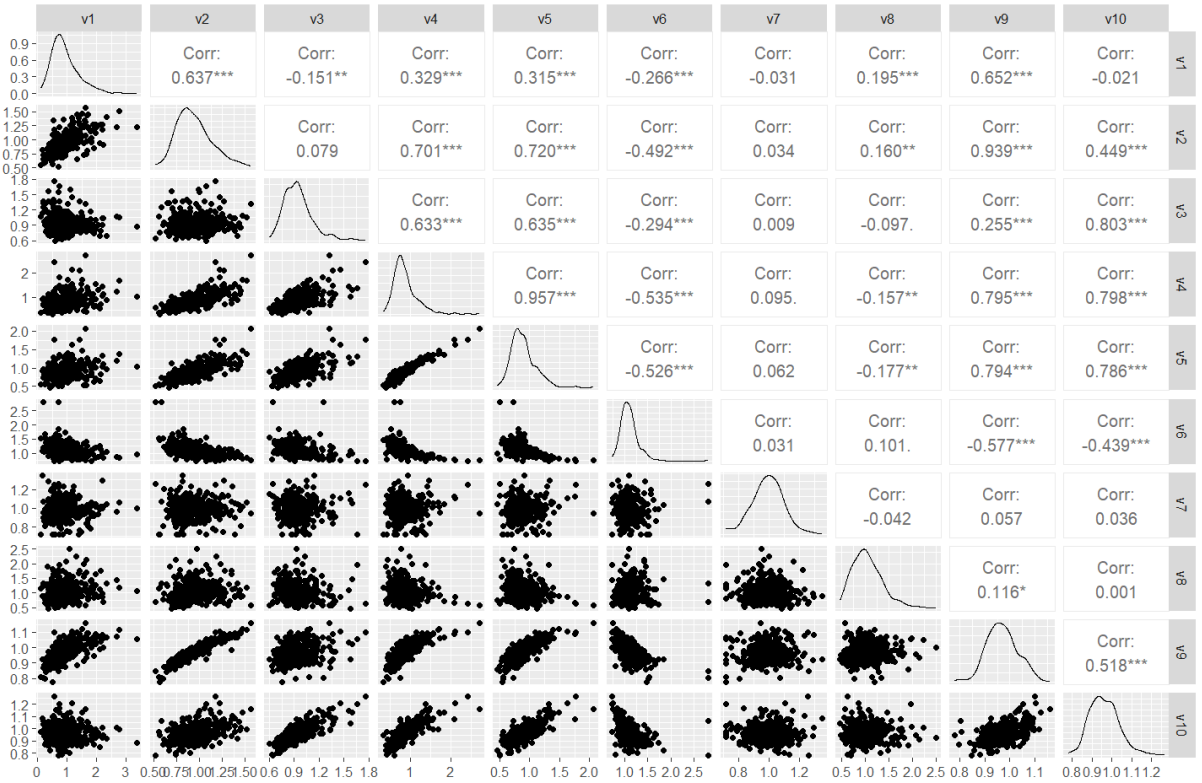


TABELLA 2.1: Valori VIF 2004-2014

| Rapporti di forza delle seguenti statistiche | Valore VIF |
|---|---|
| Aces | 2.647 |
| Punti vinti in risposta alla prima di servizio | 19.428 |
| Punti vinti in risposta alla seconda di servizio | 7.717 |
| Break point vinti | 15.804 |
| Opportunità di Break point | 18.522 |
| Punti vinti nel tiebreak | 1.721 |
| Punti vinti con la prima di servizio | 1.056 |
| Percentuale di doppi falli | 1.844 |
| Percentuale di punti vinti con la prima di servizio | 19.334 |
| Percentuale di punti vinti con la seconda di servizio | 7.612 |

### 2.2.1 Analisi Fattoriale

L'analisi fattoriale come anticipato è un metodo che permette di poter riassumere p variabili in m fattori non specificati dove il numero di fattori è nettamente inferiore al numero di variabili. Questo metodo si basa nel trovare un modello che permetta di riassumere in maniera statisticamente significativa la matrice di covarianza delle p variabili, in modo che il modello ottenuto in seguito abbia dimensione m. Nel presente studio è stata fatta un analisi fattoriale confermativa, ovvero l'obiettivo di questa analisi

era quello di confermare o rifiutare quanto era definito a priori, ovvero che le statistiche potessero essere divise in 2 gruppi.

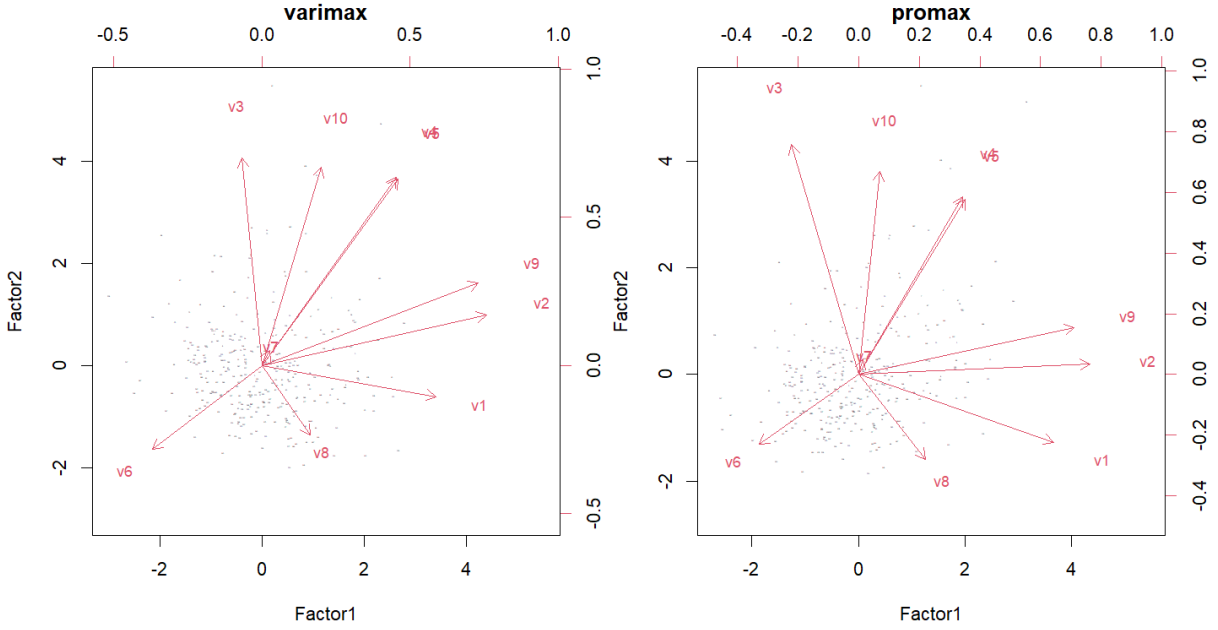


FIGURA 2.2: Grafici analisi dei fattori con rotazioni ortogonali

I nomi delle variabili sono spiegati nella tabella .1. L'analisi fattoriale, tramite un test di verosomiglianza, dimostra chiaramente che l'ipotesi secondo cui 2 fattori siano sufficienti per riassumere il comportamento delle 10 variabili è significativamente lontana dalla realtà, con un p-value prossimo allo 0.

Inoltre, se osserviamo i fattori che sono emersi (Figura 2.2), possiamo notare che le statistiche riguardanti il rapporto di punti vinti con la prima di servizio e gli ace realizzati si distinguono nettamente dagli altri indicatori. Questo suggerisce che il colpo della prima di servizio riveste un'importanza e una differenza significativa, tanto da essere considerato un fattore distaccato dal resto delle variabili come ad esempio i break point convertiti o i punti del tiebreack vinti .

# Capitolo 3

# Modelli Statistici e indici di valutazione

Nel corso di questo capitolo, l'impegno sarà volto nell'approfondimento e nella specificazione dei diversi modelli e degli indicatori di valutazione che saranno successivamente adattati ai nostri specifici dati. Attraverso questo processo, si portrà essere in grado di ottenere una comprensione più approfondita degli obbiettivi dell'analisi e trarre conclusioni significative che si rivelano essenziali per lo studio.

## 3.1 Modello Pitagorico

Si parla di modello Pitagorico quando l'esponente pitagorico ($\alpha$) nella formula 1.2, corrisponde al parametro del modello che deve essere stimato sui dati. Come precedentemente accennato l'utilità primaria del modello è di stimare la probabilità di vittoria di una squadra e, di conseguenza, il numero previsto di vittorie in una stagione: questa stima viene ottenuta moltiplicando la probabilità di vittoria stimata dal modello per il numero di partite giocate in una stagione. Tale probabilità di vittoria, Braunstein (2010), viene successivamente utilizzata come un indicatore di performance per ogni singola squadra, infatti se la squadra considerata sta ottenendo dei valori elevati per il rapporto $\frac{X}{Y}$ ma la sua percentuale di vittorie stimate rimane inferiore a quella osservata, potrebbe significare che la squadra stia faticando in alcuni aspetti del gioco, il che potrebbe indurre gli allenatori a portare delle modifiche nelle strategie (Vollmayr-Lee (2002); Cha et al. (2007)).

Un importante punto di vista emerso è espresso dalla formula 1.2. Infatti, l'applicazione di questa trasformazione al modello evidenzia come la risposta dipenda da una

singola statistica, analizzata bidirezionalmente rispetto alla squadra o al giocatore considerato (Esempio: runs conseguiti e runs subiti $\frac{X}{Y}$). La trasformazione logaritmica del rapporto precedente sarà la variabile concomitante del modello ed esso essendo indipendentemente dal avversario incontrato può essere definito come un indice di forza della squadra rispetto al terreno di gioco (Rosenfeld et al. (2010)). Questa nominazione vuole intendere che questo valore è indipendente dall'avversario, ma varia solo in funzione del singolo giocatore che stiamo prendendo in considerazione e dalla statistica selezionata inizialmente, ogni giocatore avrà quindi i propri indici di forza relativi ad un fissato periodo temporale, generato da tutte le partite giocate in quel periodo.

Il modello utilizzato è un modello lineare, in cui la variabile risposta è la trasformazione logit della probabilità di vittoria, definita come: $\log\left(\frac{x}{1-x}\right)$. Quest'ultima trasformazione permette di ottenere valori nell'intervallo completo dei numeri reali, i quali implicano poi l'utilizzo di un modello di regressione normale e non l'utilizzo di un modello binomiale. La variabile predittiva scelta per il modello sarà il logaritmo naturale degli indici di forza individuali.

Ciò che caratterizza il modello Pitagorico è la possibilità di fornire previsioni ottime in confronto anche ad un modello che utilizza diverse variabili come concomitanti. La scelta di un modello lineare con la trasformazione della risposta consente di ottenere valori di parametri più vicini al coefficiente ideale del modello Pitagorico, ovvero 2, motivo per cui viene preferito rispetto al modello binomiale logistico che non permetterebbe di ottenere un parametro vicino a quello ideale Pitagorico.

In future ricerche, sarebbe interessante valutare se l'applicazione di un modello binomiale logistico porti a risultati migliori rispetto al modello normale utilizzato in precedenza da Kovalchik, incluso anche in questo studio.

## 3.2 Modello di regressione Lasso

Il metodo di regressione Lasso (Least Absolute Shrinkage and Selection Operator), introdotto da Tibshirani (1996), è un metodo statistico che permette la regolarizzazione e la selezione delle variabili.

La peculiarità di questo modello è l'introduzione di un termine di penalità basato sulla norma L1 dei coefficienti del modello, la norma L1 è una misura di grandezza definita come la somma dei valori assoluti degli elementi di un vettore, questa particolarità impone che i metodi di regolarizzazione non cerchino di modificare la complessità del modello ponendo un sottogruppo dei coefficienti di regressione $\beta_j$ uguale a 0, ma di contrarre (shrink) i coefficienti verso zero.

È importante evidenziare che l'obiettivo del modello Lasso è quello di minimizzare una funzione di costo che comprende due componenti: l'errore di regressione e il termine di penalizzazione L1. L'equilibrio tra questi due termini viene regolato da un parametro $\lambda$. Aumentando $\lambda$ si aumenta la penalizzazione, quindi più termini saranno vincolati a 0 o a valori simili, viceversa scegliendo un valore di $\lambda$ vicino allo zero i termini non saranno contratti. Per la stima dei coefficienti nel modello Lasso si utilizza l'algoritmo di ottimizzazione LAR (Least Angle Regression), che consente di creare un insieme di soluzioni che gradualmente portano al $\lambda$ ottimale per il modello, considerando sia l'errore di regressione che il termine di penalizzazione L1.

La funzione obiettivo che il modello Lasso vuole minimizzare è la seguente:

$$\min \left( \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right) \tag{3.1}$$

Nel contesto specifico della seguente ricerca, si utilizzerà il modello Lasso durante la fase di validazione esterna del modello Pitagorico. Lo scopo è dimostrare che l'uso del modello Lasso non porta a miglioramenti significativi rispetto a un modello con una singola variabile. Inizialmente, si avvierà la procedura Lasso utilizzando tutti e 10 gli indici di forza disponibili. Questo ci consentirà di ottenere un modello con un numero ridotto di variabili, m, che minimizza la radice del errore quadratico medio di previsione calcolato tramite la procedura LOOCV (Leave-One-Out Cross-Validation).

## 3.3 Indici per i confronti

Per confrontare i diversi modelli statistici che verranno adattati ai dati in seguito, bisogna definire gli indici che permetteranno di decretare quale sia il modello che mostra un miglior adattamento ai dati e nel caso della validazione esterna, quale sia il modello che avrà le previsioni migliori

### 3.3.1 $R^2$

Il coefficiente di determinazione, spesso indicato come $R^2$, è una misura che indica la proporzione di varianza della variabile dipendente che può essere spiegata dalle variabili indipendenti nel modello di regressione. La formula per calcolare $R^2$ è la seguente:

$$R^2 = \frac{SSR}{SST} \tag{3.2}$$

dove SSR (Sum of Squares Regression) rappresenta la somma dei quadrati delle differenze tra i valori stimati dal modello di regressione e la media della variabile dipendente, e SST (Sum of Squares Total) rappresenta la somma dei quadrati delle differenze tra i valori osservati della variabile dipendente e la sua media.

Formula per SSR:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3.3}$$

$\hat{y}_i$ sono i valori stimati, $\bar{y}$ è la media della risposta

Formula per SST:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.4}$$

$y_i$ sono i diversi valori della variabile risposta

Questo indicatore verrà utilizzato per confrontare i diversi modelli pitagorici all'interno dei set di adattamento e sancire quale tra questi porti a risultati migliori in termini di porzione di variabilità della variabile dipendente spiegata.

### 3.3.2 LOOCV RMSE

La radice quadratica dell'errore quadratico medio (RMSE) è un indicatore spesso utilizzato per misurare la distanza tra i valori osservati e i valori predetti dal modello. Questo viene calcolato come la radice quadrata della media degli errori quadratici tra i valori osservati e i valori effettivi del dataset:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{3.5}$$

La convalida incrociata Leave-One-Out(LOOCV) è una tecnica particolarmente utile, essa suddivide il dataset in modo che in ogni iterazione, uno dei campioni sia utilizzato per la stima e gli altri campioni costituiscano i set di addestramento. Questa procedura viene ripetuta fino a quando ogni campione del dataset è stato utilizzato come set di test esattamente una volta. Alla fine, otteniamo una stima delle prestazioni del modello basata sull'aggregazione dei risultati ottenuti dalle iterazioni. Nel nostro caso i campioni di addestramento saranno sempre costituiti da un unica unità statistica, mentre i set di addestramento sono costituiti da tutte le altre unità, quindi questa procedurà avrà un numero di iterazioni pari al numero di righe del dataset.

L'indicatore RMSE LOOCV verrà utilizzato per confrontare i modelli nella fase della validazione esterna.

# Capitolo 4

# Adattamento dei modelli e presentazione risultati

In questo capitolo, vengono presentati e commentati i risultati delle procedure descritte nei capitoli precedenti.

## 4.1 Adattamento dei diversi modelli Pitagorici

Come spiegato in precedenza, la caratteristica chiave del modello Pitagorico è la sua capacità di stimare la percentuale di vittorie utilizzando una sola variabile come fattore esplicativo. In particolare, il primo adattamento eseguito è il seguente: il modello Pitagorico viene stimato per ciascuno dei dieci indici di forza selezionati nel capitolo 2, l'adattamento genera i valori che hanno assunto i parametri $\alpha$ relativi, essi sono elencati nella tabella 4.1. Successivamente, per ciascuno dei vari modelli è stato calcolato il coefficiente di determinazione $R^2$. Questo calcolo aiuta a definire quale modello, e quindi quale statistica espressa in termine di indice di forza, si adatta meglio al dataset.

Dalla Tabella 4.1, possiamo ottenere informazioni importanti sui valori dei parametri $\alpha$ per ciascun modello e sui coefficienti di determinazione $R^2$ relativi ai vari modelli adattati. L'obbiettivo è definire quale sia la statistica che genera la miglior stima del numero di partite vinte di ciascun giocatore in un determinato periodo di tempo.

La cosa interessante è che i risultati che otteniamo in questa tabella sono coerenti con quanto specificato nello studio di Kovalchik (2016). Guardando la tabella, possiamo notare che la statistica che riguarda i break point vinti ottiene i risultati migliori in termini di $R^2$. Questo valore ci dice quanto bene il modello spiega le differenze nella percentuale di vittorie dei giocatori.

Tabella 4.1: Valori Parametro e Indice di Determinazione dei Modelli Pitagorici per il periodo 2004-2014

| Variabile esplicativa del modello | $\hat{\alpha}$ | $R^2$ (%) |
|---|---|---|
| Aces | 0.430 | 13.4 |
| Punti vinti in risposta alla prima di servizio | 2.012 | 49.2 |
| Punti vinti in risposta alla seconda di servizio | 1.851 | 33.9 |
| Opportunità di Break point | 2.284 | 80.9 |
| Punti vinti nel tiebreak | -2.429 | 56.2 |
| Punti vinti con la prima di servizio | 0.152 | 0.1 |
| Percentuale di doppi falli | -0.306 | 2.6 |
| Percentuale di punti vinti con la prima di servizio | 7.163 | 66.2 |
| Percentuale di punti vinti con la seconda di servizio | 5.532 | 57.3 |
| Break point vinti | 1.722 | 90.8 |

Un altra variabile che ottiene risultati significativi in termini di $R^2$ è il modello che usa l'indice di forza realtivo alle opportunità di break point come variabile esplicativa, infatti assieme alla variabile riguardante i break point vinti sono le uniche che usate al interno del modello Pitagorico, garantiscono almeno un valore di 75% per il coefficente di determinazione.

In pratica, questi risultati ci suggeriscono che ci potrebbe essere un legame importante tra il numero di break point vinti e il risultato finale delle partite. L'alto valore di $R^2$ ci dice che il modello Pitagorico che prende in considerazione questa statistica è in grado di spiegare perché alcuni giocatori vincono di più rispetto ad altri, ponendo l'attenzione sulla rilevanza dei break point vinti nell'andamento delle partite di tennis.

## 4.2 Adattamento Modello Lasso

Nel capitolo precedente è stato illustrato il concetto e il funzionamento del modello Lasso, una tecnica di regressione che permette la selezione automatica delle variabili più significative.

In questa sezione, esploreremo i risultati ottenuti mediante l'applicazione della procedura di selezione delle variabili e analizzeremo il modello di regressione lineare multivariato che ne consegue.

Nella Tabella 4.2 sono riportati i valori dei coefficienti stimati dal modello Lasso realtivi agli indici di forza delle statistiche. È interessante notare che la procedura ha scelto otto variabili, riducendo a zero l'influenza di altre due. Tra queste variabili, si evidenziano coefficienti più elevati associati al rapporto dei punti vinti giocati con la prima di servizio e al tasso di successo dei break point, suggerendo una maggiore importanza

Tabella 4.2: Stime dei parametri del modello Lasso per le partite del periodo 2004-2014

| Variabili esplicative del modello | $\hat{\alpha}$ |
|---|---|
| Aces | – |
| Punti vinti in risposta alla prima di servizio | 0.320 |
| Punti vinti in risposta alla seconda di servizio | 0.120 |
| Break point vinti | 1.395 |
| Opportunità di Break point | – |
| Punti vinti nel tiebreak | 0.727 |
| Punti vinti con la prima di servizio | 0.011 |
| Percentuale di doppi falli | 0.088 |
| Percentuale di punti vinti con la prima di servizio | 1.587 |
| Percentuale di punti vinti con la seconda di servizio | 0.513 |

di queste due variabili nella determinazione dei valori della risposta. Diversamente, le statistiche sul rapporto di aces e sul rapporto delle opportunità di break point avute mostrano un coefficiente pari a 0, indicando la loro scarsa utilità nelle stime. Questo potrebbe essere causato dalla loro alta correlazione con altre variabili, come ad esempio accade per il numero di break point convertiti, che come evidenziato nel grafico 2.1, risulta essere altamente correlata con i break point conseguiti.

Il modello stimato a seguito della procedura Lasso assume la forma dell'Equazione 4.1, in cui $p = 8$ rappresenta il numero di variabili selezionate e $\alpha_i$ denota i relativi coefficienti. Tale equazione è una rappresentazione logistica dell'aspettativa di vittoria in funzione delle variabili selezionate.

$$E[\text{logit}(P(\text{Win}))] = \alpha_1 \ln\left(\frac{X_1}{Y_1}\right) + \ldots + \alpha_p \ln\left(\frac{X_p}{Y_p}\right) \tag{4.1}$$

L'adattamento del modello di regressione lineare normale utilizzando i parametri stimati dal modello Lasso produce un valore di $R^2 = 0.919$. Nonostante una leggera miglioria rispetto al modello Pitagorico, è importante notare che l'utilizzo di un modello di regressione lineare con 8 parametri introduce una complessità significativa senza un miglioramento sostanziale delle prestazioni.

Quindi possiamo affermare che l'adattamento del modello Lasso offre un metodo di selezione delle variabili significative e permette quindi la costruzione di un modello di regressione lineare, il quale offre degli ottimi risultati.

Tuttavia, l'incremento di complessità del modello con un modesto miglioramento delle prestazioni richiede un'attenta valutazione e considerazione dell'equilibrio tra complessità e precisione delle stime nel contesto specifico dell'analisi. Oltretutto dovrà in seguito essere confrontato in termini di stima fuori dal campione di adattamento,

quest'ultimo aspetto verrà valutato nel paragrafo seguente.

## 4.3 Validazione esterna 2015

Come appena anticipato, in questo paragrafo verrà esaminata l'effettiva funzionalità dei modelli utilizzati per la stima della percentuale di vittorie nell'intera stagione. La procedura di validazione esterna si basa sulla registrazione dei risultati della prima metà della stagione e sul confronto di diverse metodologie al fine di identificare quale si dimostri più efficace nella stima dei risultati dell'intera stagione. Il confronto sarà basato sulla radice dell'errore quadratico medio leave-one-out cross-validation (RMSE LOOCV).

La procedura di validazione esterna comporta la creazione di un dataset che presenta la stessa struttura di quello utilizzato per le stime precedenti, quindi il dataset includerà i giocatori che sono entrati nella top 100 almeno una volta durante i 6 mesi considerati, per i quali sono stati creati i vari indici di forza relativi alle variabili selezionate nel capitolo 2.

Successivamente, verranno confrontate quattro tipologie di modelli:

1. Modello pitagorico con un coefficiente $\alpha$ fissato a 2.

   Il primo modello produce delle stime delle percentuali di vittorie utilizzando la formula 1.2 ma non andando a stimare il coefficente $\alpha$ sui dati, bensì verrà attribuito invece il valore costante di 2 in linea con la definizione iniziale di "Teorema Pitagorico".

2. Modello Pitagorico con coefficente $\alpha$ stimato

   Il secondo modello produce delle stime delle percentuali di vittorie utilizzando la formula 1.2 stimando il modello sui dati.

3. Modello normale con variabili selezionate con la procedura lasso

   Il terzo modello produce delle stime delle percentuali di vittoria utilizzando la formula 3.1 stimando il modello sui dati e selezionando le variabili attraverso la procedura LASSO.

4. Modello basato sulla percentuale di vittoria

   Il quarto modello utilizzato è un metodo estremamente intuitivo che estende le percentuali di vittoria dei primi sei mesi all'intera stagione, assumendo che le prestazioni dei giocatori rimangano costanti nel tempo.

Poiché l'obiettivo principale di questo studio è osservare le stime delle probabilità di vittoria e, di conseguenza, il numero di partite vinte durante l'intera stagione. L'approccio preferenziale prevede l'utilizzo di una metodologia che punti a stimare i risultati al di fuori del campione già osservato. In particolare, questa stima sarà applicata alla seconda metà della stagione, poiché la prima metà è stata oggetto di osservazione e costituisce parte integrante del set di adattament dei nostri modelli.

Il confronto dei modelli elencati in precedenza verrà effettuato utilizzando l'indice RMSE, ma i calcoli verranno svolti applicando una trasformazione $logit^{-1}$ alla variabile risposta in modo da esprimere i risultati in termini di probabilità. Questo riscalamento dei valori della radice dell'errore quadratico medio permette di avere un interpretazione più semplice per i risultati ottenuti. Tale confronto sarà utile per valutare l'efficacia dei vari modelli nel predire i risultati delle partite dell'intera stagione, basandosi sui dati e quindi sulle variabili osservate solo nella prima parte.

Attraverso quindi l'analisi comparativa di tali modelli, saremo in grado di determinare l'efficacia e l'utilità del modello Pitagorico, sarà dunque importante considerare i risultati ottenuti e valutare l'equilibrio tra complessità del modello e precisione predittiva, al fine di selezionare il modello più adatto per le future analisi e previsioni.

Nella tabella 4.3, vediamo i seguenti risultati di RMSE calcolati tramite LOOCV per il campione out-of-sample:

Tabella 4.3: RMSE da LOOCV per il campione out-of-sample

| Modello | RMSE LOOCV |
|---|---|
| Pitagorico $\alpha$=2 | 0.164 |
| Pitagorico $\alpha$ stimato | 0.082 |
| Modello lasso | 0.171 |
| Modello Percentuali primo semestre | 0.167 |

Dall'analisi, possiamo notare che il modello Pitagorico con $\alpha$ stimato ha i migliori risultati. Questo evidenzia l'efficacia del modello stesso. Infatti, il valore di 0.082 indica che, considerando una stagione di 50 partite, la nostra stima devierà in media di circa $\pm 4$ incontri per giocatore.

Nonostante l'esistenza di modelli più complessi che offrono risultati leggermente migliori in alcune situazioni, il modello pitagorico dimostra di essere efficace nella previsione dei risultati delle partite. Questo risultato è coerente con la letteratura scientifica che ha evidenziato l'utilità del modello pitagorico nel contesto sportivo. La sua capacità di fornire stime ragionevolmente accurate con un solo parametro sottolinea la sua validità e praticità nell'analisi dei risultati sportivi.

## 4.4 Applicazione a dati 2010-2020

In questo paragrafo verranno applicati nuovamente i modelli descriti nel capitolo precedente al set di dati che comprende le partite degli anni 2010-2020 selezionate secondo le regole descritte in precedenza, i risultati che otteniamo sono i seguenti:

Tabella 4.4: Stime dei parametri $\alpha$ nei diversi modelli Pitagorici, applicati nei due istanti temporali

| | Valori $\hat{\alpha}$ | |
|---|---|---|
| Variabile esplicativa del modello | (2004-2014) | (2010-2020) |
| Aces | 0.430 | 0.465 |
| Punti vinti in risposta alla prima di servizio | 2.012 | 2.092 |
| Punti vinti in risposta alla seconda di servizio | 1.851 | 1.990 |
| Opportunità di Break point | 2.284 | 2.103 |
| Punti vinti nel tiebreak | -2.429 | -2.016 |
| Punti vinti con la prima di servizio | 0.152 | 0.077 |
| Percentuale di doppi falli | -0.306 | -0.198 |
| Percentuale di punti vinti con la prima di servizio | 7.163 | 6.925 |
| Percentuale di punti vinti con la seconda di servizio | 5.532 | 5.737 |
| Break point vinti | 1.722 | 1.695 |

Nella tabella 4.4 è evidente la presenza di divergenze nei parametri stimati nei due differenti periodi temporali. Tuttavia, queste differenze possono essere attribuite alle variazioni nei set di dati di adattamento. Ciò che riveste particolare interesse è l'indice di determinazione, poiché ci permette di valutare se vi siano variazioni nell'importanza di alcune statistiche per la determinazione di quale effettivamente sia la statistica che permette di determinare all'interno del modello Pitagorico il miglior adattamento ai dati.

Tabella 4.5: Indice di determinazione percentuale dei modelli Pitagorici nei diversi periodi

| | Valori $R^2$ (%) | |
|---|---|---|
| Variabile esplicativa del modello | (2004-2014) | (2010-2020) |
| Aces | 13.4 | 15.9 |
| Punti vinti in risposta alla prima di servizio | 49.2 | 55.2 |
| Punti vinti in risposta alla seconda di servizio | 33.9 | 36.0 |
| Opportunità di break point | 80.9 | 80.3 |
| Punti vinti nel tiebreak | 56.2 | 42.5 |
| Punti vinti con la prima di servizio | 0.1 | 0.2 |
| Percentuale di doppi falli | 0.26 | 0.1 |
| Percentuale di punti vinti con la prima di servizio | 66.2 | 67.0 |
| Percentuale di punti vinti con la seconda di servizio | 57.3 | 59.2 |
| Break point vinti | 90.8 | 91.0 |

Osservando la tabella 4.5 notiamo che la statistica relativa ai break point vinti continua a fornire i risultati migliori, migliorando dello 0.2% i risultati precedenti, per quanto riguarda le altre statistiche osserviamo un aumento significativo per la percentuale di punti vinti in risposta e per la percentuale di punti vinti al servizio, mentre una diminuzione per i punti vinti nel tiebreak e la percentuale di doppi falli.

Presentiamo in seguito anche i valori del modello Lasso:

TABELLA 4.6: Stime dei parametri del modello Lasso per le partite del periodo 2010-2020

| Variabile | Coefficiente |
|---|---|
| Aces | – |
| Punti vinti in risposta alla prima di servizio | – |
| Punti vinti in risposta alla seconda di servizio | 0.040 |
| Break point vinti | 1.544 |
| Opportunità di Break point | 0.063 |
| Punti vinti nel tiebreak | 0.419 |
| Punti vinti con la prima di servizio | 0.107 |
| Percentuale di doppi falli | 0.088 |
| Percentuale di punti vinti con la prima di servizio | 0.011 |
| Percentuale di punti vinti con la seconda di servizi | 0.264 |

Possiamo evidenziare dalla tabella 4.6, che ci sono stati dei cambiamenti dei valori dei parametri. Come osservato in precedenza, anche attraverso il modello lasso si può notare una tendenza generale dei coefficenti ad essere contratti verso lo zero, contrariamente un aumento per il parametro dei break point vinti, questo potrebbe indicare una maggiore influenza della variabile in questione nella determinazione dei valori delle stime, in linea con l'aumento del $R^2$.

### 4.4.1 Validazione esterna 2021

Nella fase successiva di validazione esterna, i risultati ottenuti si rivelano differenti da quelli ottenuti nel 2015.

Nel dettaglio, quando estendiamo la nostra stima al di fuori dei dati di addestramento (modello out-of-sample), il modello Pitagorico continua a emergere come il più efficace, come è evidenziato nella tabella 4.7. Questo conferma l'eccezionale capacità del modello Pitagorico nel prevedere il numero di partite vinte da un singolo giocatore nel corso di un anno.

È degno di nota che il modello Pitagorico non si distacca dagli altri modelli in maniera sostanziale come accadeva nel periodo temporale precedente. Quello che osserviamo

è un peggioramento generale delle prestazioni predittive. Ma nonostante questo peggioramento questo livello di precisione sottolinea l'efficacia del modello nella cattura delle dinamiche delle partite tennistiche e nella formulazione di previsioni attendibili.

Tabella 4.7: RMSE calcolato sulle stime LOOCV out-of-sample seconda metà del 2021

| Modello | RMSE LOOCV |
|---|---|
| Pitagorico $\alpha$=2 | 0.221 |
| Pitagorico $\alpha$ stimato | 0.210 |
| Modello lasso | 0.220 |
| Modello Percentuali primo semestre | 0.212 |

## 4.5 Differenze tra le diverse superfici

Le superfici dei campi da tennis giocano un ruolo determinante nel plasmare le dinamiche e le strategie di gioco. Un esempio significativo di questo impatto è evidente nelle diverse proprietà delle superfici come la terra battuta, l'erba e il cemento.

La terra battuta è ben nota per la sua capacità di rallentare la velocità della palla e diminuire l'altezza del rimbalzo. Questa caratteristica è fondamentale nell'influenzare il gioco, portando a scambi più prolungati e richiedendo maggiore resistenza fisica da parte dei giocatori. La palla tende ad affondare leggermente nella superficie e a scivolare sulla terra battuta, contribuendo a una riduzione della sua velocità. Questo si traduce in scambi più estesi e in un maggior coinvolgimento tattico, con i giocatori che cercano di costruire il punto attraverso strategie a lungo termine invece che mirare a vincere il punto con un singolo colpo.

Al contrario, l'erba produce effetti opposti. La superficie erbosa consente alla palla di scivolare rapidamente, riducendo il tempo di reazione dei giocatori e richiedendo una notevole abilità nel gestire gli scambi rapidi. Il rimbalzo sulla superficie erbosa è generalmente più basso, portando a punti spesso più brevi, in cui la precisione e la capacità di adattamento a situazioni di gioco in continua evoluzione sono fondamentali.

Il cemento si colloca al centro di queste due estremità. La superficie in cemento offre un rimbalzo costante e uniforme, senza variazioni significative nell'altezza del rimbalzo. Ciò contribuisce a un gioco equilibrato tra potenza e controllo, consentendo ai giocatori di sfruttare sia la velocità che la precisione nei loro colpi.

In conclusione, la scelta della superficie del campo da tennis gioca un ruolo cruciale nello sviluppo delle tattiche e degli stili di gioco dei tennisti. Ogni tipo di superficie presenta sfide uniche, richiedendo adattamenti specifici e abilità particolari da parte dei

giocatori per ottenere prestazioni ottimali. In questo lavoro si intende dimostrare se la differenza tra le superfici influisce in modo significativo sulle strategie di gioco, tanto da evidenziare se esiste un'altra statistica, che utilizzata nel modello Pitagorico, possa determinare la probabilità di vittoria di un singolo giocatore nel corso di una stagione. In particolare, si analizzano gli aspetti che potrebbero indicare come i punti del break point vinti diventino meno influenti rispetto ad altre statistiche.

Di seguito sono riportati i valori dei parametri e i diversi indici di determinazione dei vari modelli Pitagorici.

TABELLA 4.8: Stime dei parametri $\alpha$ per i diversi modelli Pitagorici, considerando le partite del periodo 2010-2020 nelle diverse superfici

| | Valori $\hat{\alpha}$ | | |
|---|---|---|---|
| Variabile del modello | Cemento | Terra | Erba |
| Aces | 0.500 | 0.425 | 0.642 |
| Punti vinti in risposta alla prima di servizio | 1.986 | 2.272 | 2.049 |
| Punti vinti in risposta alla seconda di servizio | 2.092 | 1.650 | 1.979 |
| Opportunità di Break point | 1.768 | 1.954 | 1.578 |
| Punti vinti nel tiebreak | -1.974 | -1.464 | -1.825 |
| Punti vinti con la prima di servizio | 0.034 | 0.331 | 1.059 |
| Percentuale di doppi falli | -0.363 | -0.359 | -0.409 |
| Percentuale di punti vinti con la prima di servizio | 6.862 | 6.821 | 7.020 |
| Percentuale di punti vinti con la seconda di servizio | 5.044 | 4.561 | 3.872 |
| Break point vinti | 1.514 | 1.585 | 1.391 |

TABELLA 4.9: Indici di determinazione percentuali dei modelli pitagorici stimati nelle diverse superfici per le partite del periodo 2010-2020

| | Valori $R^2$ (%) | | |
|---|---|---|---|
| Modello | Cemento | Terra | Erba |
| Aces | 14.6 | 10.1 | 19.1 |
| Punti in risposta alla prima di servizio | 51.7 | 52.7 | 54.8 |
| Punti in risposta alla seconda di servizio | 36.6 | 31.5 | 34.1 |
| Opportunità di Break point | 65.9 | 68.5 | 61.1 |
| Punti nel tiebreak | 49.6 | 33.4 | 49.0 |
| Punti con la prima di servizio | 0.01 | 0.04 | 0.02 |
| Percentuale di doppi falli | 0.3 | 0.4 | 0.5 |
| Percentuale punti con la prima di servizio | 65.0 | 66.8 | 64.1 |
| Percentuale punti con la seconda di servizio | 51.7 | 51.8 | 33.5 |
| Break point vinti | 84.7 | 83.4 | 80.7 |

I risultati presentati nelle tabelle offrono spunti di riflessione rilevanti. I valori del $R2\%$ associate ai break point vinti emergono come i valori maggiori indipendentemente dalla superficie considerata, nonostante mostri una notevole diminuzione nella loro

media. Un aspetto di particolare interesse è l'accentuato declino dei break point vinti sulla superficie erbosa, in contrasto con l'andamento dei punti guadagnati in risposta. Tale evidenza potrebbe suggerire che in situazioni in cui la velocità di gioco è elevata, come avviene sulla superficie erbosa, la risposta al servizio acquisisce un'importanza maggiore nella determinazione del andamento dell'incontro, a discapito dell'impatto dei break point vinti rispetto ad altre superfici di gioco.

Osservando con attenzione la tabella dei valori dei parametri (tabella 4.8), si evidenzia una considerevole riduzione dei valori dei parametri stessi. Questa marcata deviazione dai valori ipotetici del teorema Pitagorico potrebbe essere attribuita sia alle dimensioni significativamente ridotte del sottocampione considerato, sia alla natura specifica della superficie in esame. Il tipo di superficie su cui si disputano le partite potrebbe influire in modo sostanziale sulla dinamica di gioco, portando ad adattamenti differenti dei modelli pitagorici. Una fase successiva di indagine ha coinvolto una comparazione dettagliata dei parametri stimati sulla superficie in cemento con quelli rilevati sulle altre due superfici. Questa analisi è stata condotta attraverso una procedura ipotetica appropriata, volta a determinare se le discrepanze nei valori dei parametri fossero statisticamente rilevanti. Come si può osservare dai risultati espressi nella tabella 4.10 hanno rivelato che i parametri associati alle superfici in cemento e terra battuta non presentavano divergenze statisticamente significative, suggerendo un certo grado di coerenza nelle performance dei giocatori su queste due superfici. D'altro canto, il parametro associato alla superficie erbosa ha manifestato una distinzione statisticamente significativa rispetto agli altri due. Questi risultati mettono in luce l'importanza delle

Tabella 4.10: Risultati dei test d'ipotesi

| Test | Valore t | p-value |
|---|---|---|
| Terra | 2.345 | 0.971 |
| Erba | 3.456 | 0.0005 |

singole superfici nel modellare le sfaccettature del gioco. La marcata deviazione nel parametro riferito alla superficie erbosa enfatizza l'effetto notevole che tale superficie può esercitare sulle prestazioni globali dei giocatori. In altre parole, la peculiarità intrinseca della superficie erbosa sembra influenzare in modo particolare il processo di stima del modello pitagorico nell'effettuare le previsioni, comportando una minore precisione nella determinazione della percentuale di vittorie di un giocatore da parte della statistica riguardante il numero di break point vinti a favore invece dei punti vinti con la seconda di servizio.

In definitiva, i risultati sottolineano l'importanza dell'adattamento dei modelli analitici alle variazioni delle condizioni di gioco, evidenziando come il contesto possa influire in modo significativo sulle prestazioni e sui parametri stimati. Le analisi condotte forniscono ulteriori elementi di riflessione nel valutare l'applicabilità e l'interpretazione dei modelli pitagorici in diverse situazioni di gioco.

# Capitolo 5

# Conclusioni

Nell'ambito delle analisi condotte, è stata valutata l'efficacia del modello Pitagorico nella previsione dei risultati nel tennis. In particolare, è stato dimostrato che il modello Pitagorico che utilizzava come variabile concomitante l'indice di forza relativo alla statistica "numero di break point vinti dal giocatore" aveva l'indice di determinazione $R^2$ con il valore più alto tra gli altri modelli che utilizzavano ulteriori indici di forza, in linea con lo studio di riferimento (Kovalchik, 2016). Inoltre, sono stati osservati i risultati della validazione esterna del 2015 (out-of-sample), il modello Pitagorico con $\alpha$ stimato ha dimostrato di essere il migliore nel predire la percentuale di partite vinte da un singolo giocatore nel corso di un anno e, di conseguenza, il numero di partite vinte nel corso dell'anno.

Successivamente, sono state applicate le stesse analisi ai dati del periodo 2010-2020 con lo scopo di analizzare e in caso scoprire possibili cambiamenti e tendenze del gioco. Sono state osservate alcune differenze nei valori dei parametri e nei valori degli indici di determinazione relativi tra i due periodi temporali. I modelli Pitagorici, sia con $\alpha$ fisso che stimato che con $\alpha$ fissato a 2, hanno dimostrato buone prestazioni in entrambi i periodi, confermando l'efficacia di questo modello nella previsione dei risultati nel tennis.

Infine, è stata valutata l'influenza delle diverse superfici di gioco (cemento, terra battuta ed erba) sulle prestazioni e sui risultati. In particolare, ci si aspetta che nelle superfici come la terra e il cemento, dove gli scambi per ogni punto sono composti mediamente da più colpi e implicano un maggiore sforzo fisico, i giocatori focalizzino le loro energie più nella gestione dei punti importanti che negli altri punti del game, quindi nella conversione dei break point. Al contrario, nella superficie come nell'erba, essendo gli scambi molto più veloci, quindi meno dispendiosi a livello di energie fisiche e mentali, i giocatori preferiscono distribuire le loro energie più equamente in tutti i punti, valorizzando in particolare la risposta in modo da indirizzare immediatamente

l'andamento dello scambio; questo comporta anche nel modello Pitagorico relativo a questa statistica un miglioramento in termini di precisione dei risultati. Tuttavia non è stata individuata alcuna statistica che, indipendentemente dalla superficie, permetta di ottenere valori dell'indice di determinazione maggiori rispetto a quelli ottenuti dai break point convertiti. Questo supporta l'ipotesi che i giocatori che gestiscono meglio i punti fondamentali del match sono quelli che vincono il maggior numero di partite, indipendentemente dalla superficie del campo.

Complessivamente, il modello Pitagorico si è dimostrato efficace nella previsione dei risultati nel tennis, offrendo stime ragionevolmente accurate con un solo parametro. Nonostante l'esistenza di modelli più complessi, il modello Pitagorico rimane una scelta valida per l'analisi e la previsione dei risultati nel tennis.

## 5.1 Future ricerche

Le ricerche future potrebbero concentrarsi sulla previsione della vittoria nella singola partita, confrontando le probabilità date dai modelli Pitagorici applicati sui singoli giocatori calcolate su un periodo precedentemente definito, ad esempio 9 o 6 mesi, per valutare quale dia i migliori risultati. Inoltre, l'utilizzo delle nuove tecnologie potrebbe fornire ulteriori statistiche riguardo al gioco del tennis, consentendo di scoprire nuove tendenze e di migliorare la previsione dei risultati.

# Appendice

Tabella .1: nomi nei grafici

| Rapporti di forza delle seguenti statistiche | nome nei grafici |
|---|---|
| Aces | v1 |
| Punti vinti in risposta alla prima di servizio | v2 |
| Punti vinti in risposta alla seconda di servizio | v3 |
| Break point vinti | v4 |
| Opportunità di Break point | v5 |
| Punti vinti nel tiebreak | v6 |
| Punti vinti con la prima di servizio | v7 |
| Percentuale di doppi falli | v8 |
| Percentuale di punti vinti con la prima di servizio | v9 |
| Percentuale di punti vinti con la seconda di servizio | v10 |

# Bibliografia


Braunstein, A. (2010). Consistency and pythagoras. *Journal of Quantitative Analysis in Sports* **6**, 1–16.

Caro, C. A. & Machtmes, R. (2013). Testing the utility of the pythagorean expectation formula on division one college football: An examination and comparison to the morey model. *Journal of Business & Economics Research* **11**, 537–542.

Cha, D. U., Glatt, D. P. & Sommers, P. M. (2007). An empirical test of bill james's pythagorean formula. *Journal of Recreational Mathematics* **35**, 117–130.

Cochran, J. J. & Blackstock, R. (2009). Pythagoras and the national hockey league. *Journal of Quantitative Analysis in Sports* **5**, 1–13.

Davenport, C. & Woolner, K. (1999). Revisiting the pythagorean theorem: Putting bill james' pythagorean theorem to the test. *The Baseball Prospectus* .

Hamilton, H. H. (2011). An extension of the pythagorean expectation for association football. *Journal of Quantitative Analysis in Sports* **7**, 1–18.

James, B. (1981). Baseball abstract. *Self-published* .

Kovalchik, S. (2016). Is there any pythagoream theorem in tennis? *De Gruyter* .

Rosenfeld, J. W., Fisher, J. I., Adler, D. & Morris, C. (2010). Predicting overtime with the pythagorean formula. *Journal of Quantitative Analysis in Sports* **6**, 1–19.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* , 267–288.

Vollmayr-Lee, B. (2002). More than you probably ever wanted to know about the "pythagorean" method.