Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per l'Economia e l'Impresa

Dal college alla NBA: analisi delle determinanti nelle scelte al draft NBA

Relatore Prof. Adriano Paggiaro
Dipartimento di Scienze Statistiche

Laureando: Alessandro Vigolo
Matricola n. 1227359

Anno Accademico 2022/2023

Alla mia famiglia, per avermi
sempre supportato

A Pietro e Francesca, che porterò
sempre con me

Indice

Introduzione

Il *draft* è un evento annuale tramite il quale i giocatori entrano a far parte del massimo campionato americano di basket: la NBA (*National Basketball Association*). Lo scopo di questo elaborato è capire quali sono le statistiche di gioco che consentono ai giocatori di college di rientrare tra le prime scelte delle squadre. Per fare ciò si analizzeranno dati relativi a giocatori che hanno frequentato il college e sono stati selezionati al *draft* per entrare nella NBA negli ultimi dieci anni, utilizzando principalmente il software *Stata* ed *RStudio* per alcune analisi grafiche.

Nel primo capitolo verrà presentato l'argomento con alcuni cenni storici per contestualizzare il *draft* e la sua struttura nell'ambiente della NBA.

Nel secondo capitolo viene dettagliato il metodo di raccolta dei dati e costruzione del dataset, che sarà poi presentato ed esplorato attraverso delle analisi descrittive e delle correlazioni, per iniziare a comprendere se già queste possono darci informazioni riguardo al nostro problema di interesse.

Nel terzo capitolo si procede poi con le analisi su *Stata* attraverso una serie di regressioni, per capire ogni modello cosa fa emergere di diverso rispetto ad un altro.

Infine, nel quarto capitolo, vengono condotte delle analisi di approfondimento su alcuni aspetti più specifici come le differenze tra i vari ruoli di ogni giocatore, e analisi di robustezza per giustificare alcune scelte prese durante la costruzione del dataset e lo sviluppo della tesi.

Letteratura

Per l'ambiente sportivo europeo, e quindi anche per quello italiano, il *draft* è un meccanismo ancora sconosciuto e che non viene ancora utilizzato. Sicuramente questo è dovuto principalmente alla scarsa disponibilità di atleti, legata alle dimensioni ridotte della popolazione italiana rispetto a quella americana, ma anche al fatto che la NBA, come la maggior parte delle leghe americane, è un sistema chiuso che non prevede promozioni e retrocessioni e quindi il *draft* diventa un buon sistema per "riequilibrare" la lega, aiutando le squadre che stanno performando peggio a provare a competere di nuovo con le migliori.

Ne segue quindi che la grande maggioranza delle ricerche e degli studi sul *draft* provengano dall'ambiente americano. Una grande importanza per la ricerca e studi sulla NBA è da attribuire all'economista sportivo David Berri che ha più volte analizzato la NBA e l'ambiente dei college americani da un punto di vista economico ed econometrico basandosi sulle performance dei giocatori. Nei suoi articoli, Berri [1] [2]analizza una serie di dati relativi a performance di giocatori di college per capire come questi influenzino le scelte dei *general manager* e *talent scout* quando seguono i futuri giocatori NBA. Tra le sue conclusioni vediamo come, tra i fattori che un giocatore può controllare, i punti segnati hanno un peso molto significativo tra i giocatori scelti più in alto in un *draft*. L'importanza di segnare molti punti per un giocatore per farsi notare dai talent scout era già stata confermata nel 2007 [3] nel suo lavoro con Brook e Schmidt in cui gli autori sottolineavano che concentrandosi su tale statistica, i giocatori NBA potevano aumentare le loro probabilità di essere scelti al *draft* e quindi di conseguenza anche di ottenere uno stipendio più alto nella loro carriera futura. Inoltre, seguendo questa osservazione, affermano che i "*decisions maker*" (ovvero gli scout e i manager NBA) non seguono un ragionamento completamente razionale nel valutare un giocatore perché non si basano solo su un'analisi di tutte le statistiche di gioco, ma si concentrano, come afferma Berri [3], sull'evento che più si nota nel guardare in una partita di basket, ovvero i canestri segnati.

Altri studi sulle performance dei giocatori al college, come quello di Dennis Coates [4], affermano che la produttività di un giocatore nella sua carriera collegiale è un buon indicatore per prevedere la sua posizione al *draft* e lo stipendio che potrà ottenere nella NBA. Ancora sull'argomento, Evans [5]conclude che tra le statistiche di gioco c'è una correlazione fra punti segnati, assist, palle rubate e stoppate e l'essere scelti ad una posizione elevata al *draft.* Tra gli altri fattori non collegati direttamente alle performance sul campo, Evans nota come la scelta al *draft* sia influenzata anche dalla squadra di provenienza del giocatore di college e dall'esperienza dell'allenatore che lo ha guidato. Allo stesso modo anche chi ha frequentato meno anni di college sembra avere più probabilità di essere scelto prima rispetto a chi gioca più stagioni al college. Sempre riguardo al numero di partite giocate, Groothius [6]conclude che i giocatori che passano più tempo al college vengono scelti più in fondo, mentre i talenti che mostrano subito un gran potenziale dopo la *high school,* tendono a giocare meno anni al college, a conferma di quanto già detto prima da Evans [5]. Le squadre NBA inoltre sono disposte a rischiare di scegliere un giocatore con poca esperienza ma con molto potenziale per poterne trarre più guadagno economico in un futuro.
È chiara, quindi, l'importanza che il *draft* ricopre nel mondo della NBA, ovviamente per le società da un punto di vista economico, ma anche per i giocatori come rampa di lancio verso il mondo del basket professionistico.

Data l'assenza di studi italiani ed il mio grande interesse nel mondo del basket, sarà interessante analizzare questi argomenti concentrandosi maggiormente sugli aspetti statistici attraverso le nozioni che ho appreso in questi anni di studi.

Inoltre, sarà interessante analizzare classi di *draft* più recenti rispetto a quelle trattate in letteratura, dato che il gioco si evolve nel tempo e le priorità delle scelte delle varie squadre al *draft* potrebbero essere diverse rispetto agli anni passati.

Capitolo 1 – Background

1.1 Il draft

Nel mondo dello sport statunitense e in generale nell'ambiente nord-americano esiste un sistema di scelta di atleti chiamato *draft* [7]che mira a dare possibilità a giovani atleti americani e internazionali di competere nelle migliori leghe sportive americane, tra cui le più importanti la NFL nel football, la MLB nel baseball e la NBA nel basket. Allo stesso tempo garantisce la possibilità per ogni squadra di acquisire nuovi giocatori e ampliare la propria rosa nonché di ringiovanirla, in quanto gli atleti scelti provengono per lo più dai college e quindi sono giovani che, si spera, possano avere del potenziale da esprimere. Data la grande disponibilità di giovani atleti che l'America possiede, ogni anno, in ognuna delle leghe prima citate, una serie di giocatori si rende disponibile ad essere scelta soprattutto tra gli atleti facenti parte del grande bacino collegiale americano. In numero sicuramente minore, anche atleti internazionali al di fuori degli Stati Uniti possono dichiararsi eleggibili per il *draft*. L'ordine di scelta nei *draft* è inversamente proporzionale ai risultati ottenuti durante l'ultima stagione[1]: questo sistema punta a dare, dopo ogni stagione, una possibilità di migliorare direttamente per il prossimo anno o per porre le basi di una ricostruzione della squadra. Ne segue quindi che il *draft* è una risorsa importante per le squadre che vi partecipano: scegliere un buon giocatore tra quelli disponibili, potrebbe portare la squadra a ottenere buoni risultati ma anche a creare interesse tra i tifosi nel vedere la propria squadra scegliere un potenziale campione.

[1] Nei campionati sopra citati la stagione si divide sempre in due fasi, la stagione regolare e i playoff. Per decidere l'ordine del *draft* si tiene conto dei risultati ottenuti nella stagione regolare

1.2 Il draft NBA

In particolare, nella NBA [8], il *draft* è un evento annuale istituito a partire dall'anno 1947 con molto seguito negli USA e non solo, in cui le 30 squadre NBA scelgono a turno atleti provenienti da college o da squadre internazionali che hanno dichiarato la loro volontà ad essere scelti o che rientrano automaticamente nei criteri di eleggibilità.

Il *draft* NBA così come lo vediamo noi adesso è frutto di una serie di modifiche e aggiustamenti avvenuti nel corso degli anni. [9] Dalla sua creazione nel 1947 fino al 1974 prevedeva una serie di scelte lunghissime senza limiti, in cui le squadre sceglievano i giocatori ad oltranza senza limite di round. Dato il poco interesse e la noia che un evento del genere poteva suscitare, vennero introdotti, nel 1974, dei primi limiti di round portando il *draft* ad avere un massimo di 10 round. In seguito ad un'altra serie di altri *draft* lunghissimi, dopo i quali la maggior parte dei giocatori selezionati dopo le 60-70esime scelte non trovava spazio nella NBA, si passò prima a 3 round nel 1988 e infine alla forma attuale di 2 round nel 1989.

1.2.1 La struttura del draft NBA

Il *draft* NBA si svolge nella off-season, solitamente a giugno, ovvero nel periodo di pausa del campionato in cui le società hanno la possibilità di sistemare la propria squadra attraverso gli scambi o l'acquisizione di giocatori senza contratto e appunto tramite il *draft*. [8] L'evento si divide in due turni da 30 scelte ciascuno, per un totale di 60 giocatori selezionati ogni anno. L'ordine di scelta viene definito in base al posizionamento ottenuto nella stagione regolare appena conclusa (senza contare quindi le partite giocate ai *playoff* [2]): le squadre col peggior record, tra le 14 non qualificate ai *playoff*, hanno maggiori possibilità di ottenere le prime scelte; probabilità che cala man mano che ci si avvicina alle squadre con record migliori.

[2] I *playoff* sono la seconda fase del campionato NBA in cui, al termine del campionato, le squadre meglio classificate si sfidano per il titolo

Per stabilire l'ordine delle prime 14 scelte si svolge ogni anno, a partire dal 1985, la *draft lottery* [10], ovvero una vera e propria estrazione per definire l'ordine di scelta prima del *draft*. Dal 2019 la regola vigente prevede che solo le prime 4 squadre tra quelle con il record peggiore vengano estratte dalla lotteria mentre l'ordine delle restanti viene definito a scalare seguendo il criterio peggiore-migliore. Alle 3 peggiori squadre viene assegnata la stessa probabilità di ottenere la prima scelta, questo per scoraggiare il fenomeno del *tanking* [11], ovvero la decisione di una società di perdere partite di proposito durante la stagione regolare per avere un record negativo e di conseguenza più possibilità di ottenere la prima scelta. Il meccanismo di estrazione si basa su un sistema di probabilità e combinazioni in cui 14 palline numerate vengono inserite in una macchina, dalla quale verranno poi estratte a combinazioni di 4. La squadra che vede estratta la sua combinazione, precedentemente assegnatagli, ottiene la prima scelta, e dopo aver reinserito le palline, si procede all'estrazione delle seguenti tre scelte. L'ordine dei restanti slot viene quindi stabilito in automatico rispetto al record di stagione.

1.2.2 I criteri di eleggibilità

L'eleggibilità di un atleta al *draft* è attualmente regolamentata da diversi criteri frutto di una serie di aggiustamenti e di revisioni delle precedenti regole in vigore. [12] Per i giocatori internazionali al compimento dei 22 anni, mentre per gli atleti americani che hanno almeno 19 anni e che hanno completato 4 anni di college, scatta l'eleggibilità automatica per il *draft* ed entrano quindi a far parte dei giocatori selezionabili dalle squadre NBA.

Chi intende dichiararsi eleggibile prima (nel gergo del *draft* "*early entrants*") deve comunicarlo entro 60 giorni al *draft* e deve avere almeno 19 anni. Inoltre, solamente per i giocatori non internazionali, l'atleta deve aver frequentato almeno un anno di college dopo l'uscita dalla *high school*. Questa regola, chiamata nel mondo del basket americano "*one and one rule*" [13] venne introdotta nel 2005 dalla lega NBA per incoraggiare gli atleti ad essere più maturi al momento dell'ingresso nel campionato, ma anche a seguito di alcuni casi particolari e controversie accadute nei precedenti *draft*. Infatti, alcuni giocatori vi accedevano direttamente appena finita la *high*

school, per citarne alcuni, da Kobe Bryant nel 1996 a LeBron James e Dwight Howard rispettivamente nel 2003 e 2004.

Il commissario NBA dell'epoca, David Stern, era contrario alla visione di retribuzione e sicurezza finanziaria che la NBA stava creando nei giovani americani e alle troppe pressioni a cui venivano sottoposti da general managers e scout già alla loro giovane età. Ovviamente questa regola ha generato anche pareri contrastanti, in particolare nell'ottica di usare il college solo come "momento di passaggio" e non per motivi di istruzione e formazione dello studente. Inoltre, ha contribuito a creare dei giri d'affari in cui i college si contendono di nascosto i talenti uscenti dalle varie *high school* promettendo in cambio denaro alle famiglie degli atleti, andando contro al regolamento per cui la scelta del college deve essere libera e senza alcun vincolo contrattuale o compenso monetario.

Capitolo 2 – Il dataset

2.1 Raccolta dei dati

Per questa tesi si utilizzerà un dataset contenente dati relativi alle performance di giocatori di college durante la loro carriera pre-NBA, costruito attraverso siti specializzati in basket collegiale americano. In particolare, dal sito Sports Reference[3] sono state prese le principali statistiche di gioco, mentre alcuni approfondimenti e la lista dei giocatori scelti sono stati reperiti dalla pagina web di Bart Torvik[4], creatore del sito T-Rank[5]. La scelta di utilizzare soltanto i dati relativi ai college, e non anche ai giocatori internazionali, è stata fatta in linea con i lavori di Berri sulla NBA [1] e la WNBA (*Women* NBA) [2] per rendere le statistiche di gioco più confrontabili tra loro. Tra i campionati da cui provengono le scelte internazionali, infatti, c'è molta differenza sia nei confronti del college che all'interno dei campionati stessi per via dello stile di gioco, della competitività di ogni lega e dal diverso livello di tecnica dei giocatori. In questo modo, tra i 600 atleti scelti in totale (30 per turno, 60 in totale per ogni *draft*) negli ultimi 10 anni nei vari *draft* NBA (dal 2022 al 2013), ne verranno considerati 456, ovvero coloro che hanno frequentato almeno un anno di college (circa il 76% sul totale delle scelte). Il numero di atleti provenienti dal college negli anni considerati, come vediamo in Figura 1, è sempre maggiore rispetto ai giocatori internazionali ed è sufficientemente grande da permetterci di avere un buon dataset su cui lavorare.

[3] Indirizzo pagina web: https://www.sports-reference.com/cbb/
[4] Indirizzo pagina web: https://barttorvik.com/trank.php#
[5] T-Rank è un sito web creato da Bart Torvik per analizzare il basket collegiale americano da un punto di vista statistico. In questo elaborato viene utilizzato unicamente per reperire la lista aggiornata dei giocatori frequentanti il college scelti al *draft*.

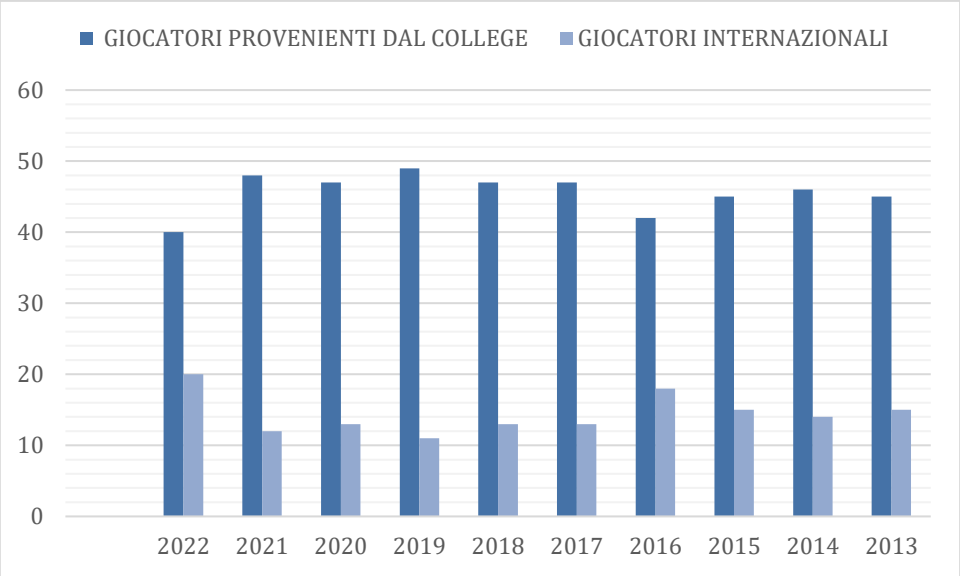

Figura 1: grafico della distribuzione della provenienza delle scelte al draft negli ultimi 10 anni

A seguito di questa decisione la posizione di ogni giocatore è stata "aggiustata" in modo da considerare le scelte come se i giocatori provenissero tutti dai college. In questo modo nel dataset la variabile PICK indicherà la posizione effettiva in cui un atleta è stato selezionato quell'anno, mentre ADJPICK è la variabile aggiustata ottenuta eliminando i giocatori internazionali, creando quindi una sorta di classifica per ogni anno tra i giocatori del college. Ad esempio, se un giocatore ha come valore di ADJPICK "10", vuol dire che quell'anno, tra i giocatori di college disponibili, è stato la decima scelta.

2.2 Presentazione del dataset

Tra le variabili descrittive presenti nel dataset troviamo YEAR, che indica l'anno di appartenenza di un giocatore[6] ed include i seguenti valori: *freshman* (FR) se è al primo anno di college, *sophomore* (SO) se è al secondo, *junior* (JR) e *senior* (SR) rispettivamente al terzo e quarto anno. In Figura 2 vediamo come si distribuiscono i giocatori presenti nel dataset secondo le categorie sopra citate.

[6] Classificazione degli studenti americani nei college in base agli anni trascorsi nell'università [19]

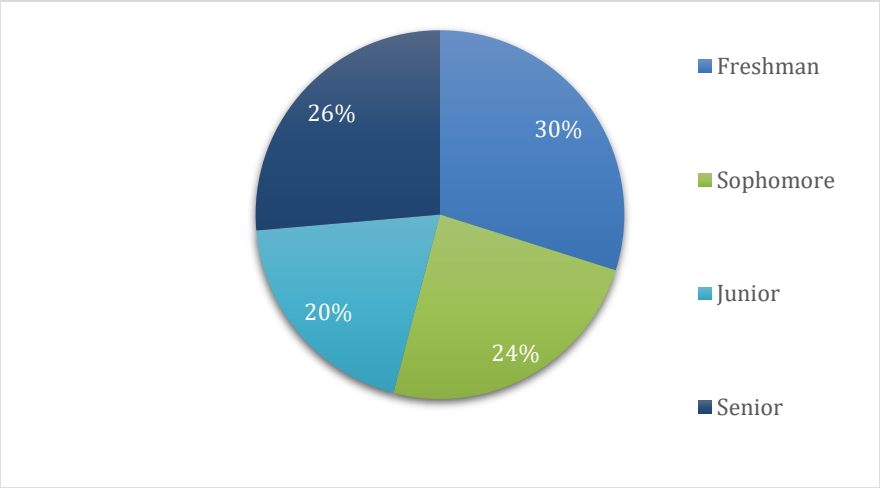

Figura 2: composizione del dataset secondo l'anno di appartenenza

Viene poi indicato il nome e cognome del giocatore con NAME e la sua altezza con HEIGHT. TEAM e CONFERENCE indicano invece rispettivamente la squadra di appartenenza e quello che in termini italiani potremmo definire il girone di appartenenza[7]. ROLE indica il ruolo del giocatore opportunamente generalizzato in 5 grandi classiche categorie: playmaker (1), guardia (2), ala piccola (3), ala grande (4), centro (5). Soprattutto nel basket americano, con l'evoluzione del gioco sono stati introdotti una serie di ruoli specifici per le diverse caratteristiche di ogni giocatore, ma per comodità e per corrispondenza al basket italiano in questo elaborato ci si riferirà a quelli che sono i classici "macro ruoli". Il *playmaker* ha fondamentalmente il ruolo di regista della squadra e quindi una spiccata dote nella gestione della palla e nell'abilità di far segnare i propri compagni, mentre la *guardia* si contraddistingue per le abilità al tiro e la capacità di realizzare punti. Le *ali* invece sono in genere dotate di grande atletismo e fisicità che vengono sfruttate *dall'ala piccola* soprattutto in attacco nelle penetrazioni a canestro e dall'*ala grande* in difesa e nel farsi valere sotto canestro. Il *centro* infine è in genere il giocatore più alto e fisico del quintetto che avrà il compito di catturare il maggior numero possibile di rimbalzi e di lottare sotto canestro. Ovviamente questa breve descrizione dei ruoli del basket è dettata dai compiti che ogni giocatore dovrebbe avere in campo ed è

[7] Il campionato collegiale americano è gestito dalla NCAA (*National Collegiate Athletic Association)* ed è composto da tre divisioni. La prima divisione, suddivisa in conference in base allo stato, è quella più importante e dalla quale provengono la maggior parte dei giocatori scelti al *draft* [18]

molto generale come classificazione, dato che a seconda del modo di giocare di ogni squadra e delle caratteristiche di ogni giocatore, i "confini" tra i ruoli sono spesso molto sottili. Vediamo nella Figura 3 la distribuzione dei giocatori del dataset a seconda del ruolo giocato.

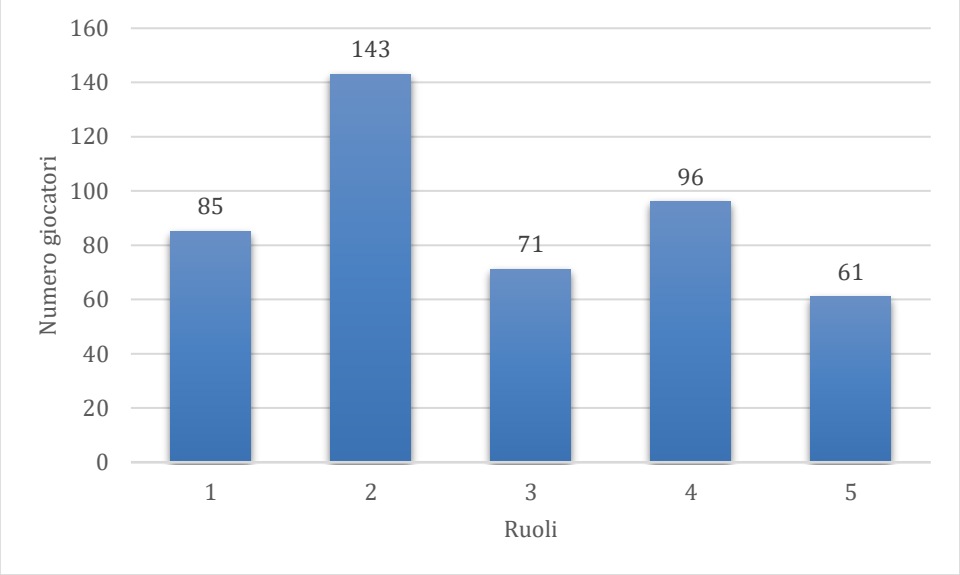

Figura 3: numero di giocatori per ogni ruolo

Ancora GP indica le partite giocate in totale e GS il numero di partite giocate da titolari, mentre MP indica i minuti giocati a partita.

Per quanto riguarda le statistiche di gioco, sono state utilizzate le medie per partita della carriera collegiale di ogni giocatore in merito a punti, rimbalzi, assist, palle rubate, stoppate e palle perse. Per chiarire a cosa si riferiscono queste statistiche viene riportata qui sotto una breve descrizione di ognuna, conforme con il manuale ufficiale per le statistiche NCAA[8] (National Collegiate Athletic Association) [14].

- Punti segnati (PTS): i punti realizzati sono contati ogni qualvolta un giocatore realizza un canestro che può essere da due punti o tre punti a seconda che sia avvenuto da dentro o fuori l'arco dei tre punti. I canestri invece realizzati dal

[8] La National Collegiate Athletic Association è un'associazione senza scopo di lucro che gestisce gli atleti-studenti del mondo universitario americano [18]

tiro libero, in seguito a un fallo subito durante l'azione di tiro, valgono un punto;

- Rimbalzi totali (TRB): vengono registrati nel momento in cui un giocatore riesce a catturare il pallone dopo un tiro sbagliato avversario (rimbalzo difensivo, DRB) o della propria squadra (rimbalzo offensivo, ORB);
- Assist (AST): viene attribuito ad un giocatore quando effettua un passaggio che porta un altro compagno a realizzare un canestro;
- Palle rubate (STL): viene registrata nel momento in cui un giocatore "ruba" il pallone ad un avversario dalle mani o intercetta un passaggio, guadagnando così il possesso di gioco;
- Palle perse (TOV): le palle perse sono in genere collegate alle palle rubate: infatti ogni volta che un giocatore perde il controllo del possesso di gioco, viene registrata una palla persa. È chiaro quindi che le palle perse sono una statistica negativa per un giocatore, ma non sempre è indice di una prestazione totalmente negativa come vedremo nelle analisi;
- Stoppate (BLK): vengono registrate quando un giocatore che sta difendendo devia un tiro avversario in modo legale, senza commettere falli.

Descrizione	Variabile	Osservazioni	Media	Std.Dev	Min	Max
Scelta	PICK	456	29.56	1.73	1	60
Scelta aggiustata	ADJPICK	456	23.37	1.33	1	49
Anno di appartenenza	YEAR	0				
Altezza	HEIGHT	456	6.51	0.32	5.1	7.1
Nome e cognome	NAME	0				
Squadra	TEAM	0				
Conference	CONF	0				
Ruolo	ROLE	456	2.79	1.33	1	5
Partite giocate	GP	456	78.88	3.86	15	149
Partite da titolare	GS	456	61.56	3.26	0	145
Minuti giocati	MP	456	28.28	4.36	15.2	36.7
Rimbalzi offensivi	ORB	456	1.44	0.82	0.1	4
Rimbalzi difensivi	DRB	456	3.95	1.31	1.8	8.6
Rimbalzi totali	TRB	456	5.38	1.94	2.1	11.8
Assist	AST	456	2.24	1.48	0.2	8.7
Punti segnati	PTS	456	12.82	3.34	5.2	27.4
Palle rubate	STL	456	1.01	0.44	0.1	2.9
Stoppate	BLK	456	0.78	0.69	0	4.4
Palle perse	TOV	456	1.87	0.62	0.6	5.2
Anno di scelta	DRAFTYEAR	456	2017.5	2.84	2013	2022

Tabella 1: riassunto delle variabili presenti nel dataset

2.3 Analisi descrittive

Il problema di interesse della tesi sarà quello di capire la relazione tra le variabili, in particolare tra ADJPICK e le varie statistiche di gioco, per vedere come quest'ultime influenzino la scelta di un giocatore al *draft*. Dalla descrizione delle variabili nella presentazione del dataset possiamo già aspettarci qualcosa riguardo alle principali statistiche. Ci aspettiamo chiaramente che i punti segnati abbiano un peso molto importante per la scelta di un giocatore, così come gli assist e i rimbalzi perché sono le tre statistiche base del basket che saltano di più all'occhio nell'osservazione di un giocatore. Stesso discorso per le palle rubate e le stoppate che potrebbero però risentire del fatto che, riguardando fasi di gioco prettamente difensive, si notano meno in una partita. Mentre invece le palle perse dovrebbero influire negativamente nella valutazione di un giocatore. Ci aspettiamo inoltre che ci siano delle differenze sostanziali per quanto riguarda le medie di tali statistiche tra i diversi ruoli, come viene sottolineato nei seguenti grafici.

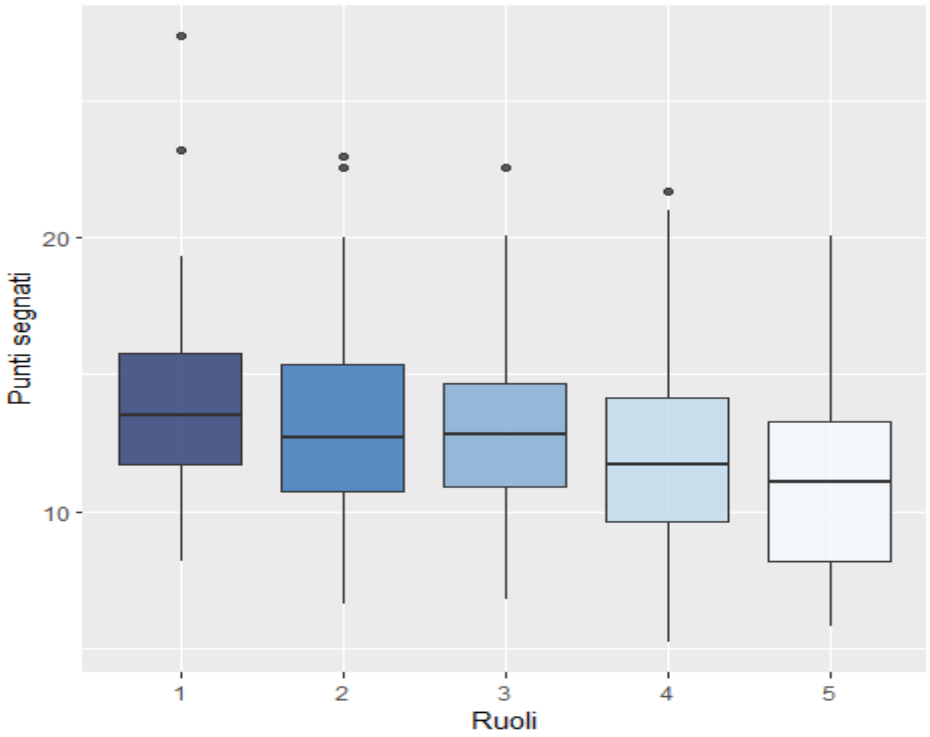

Figura 4: boxplot dei punti segnati per ruolo

Come possiamo notare dalla Figura 4, per quanto riguarda PTS non ci sono grandi differenze, infatti, la media si afferma attorno a quella generale, tranne per i centri (5), dove cala leggermente.

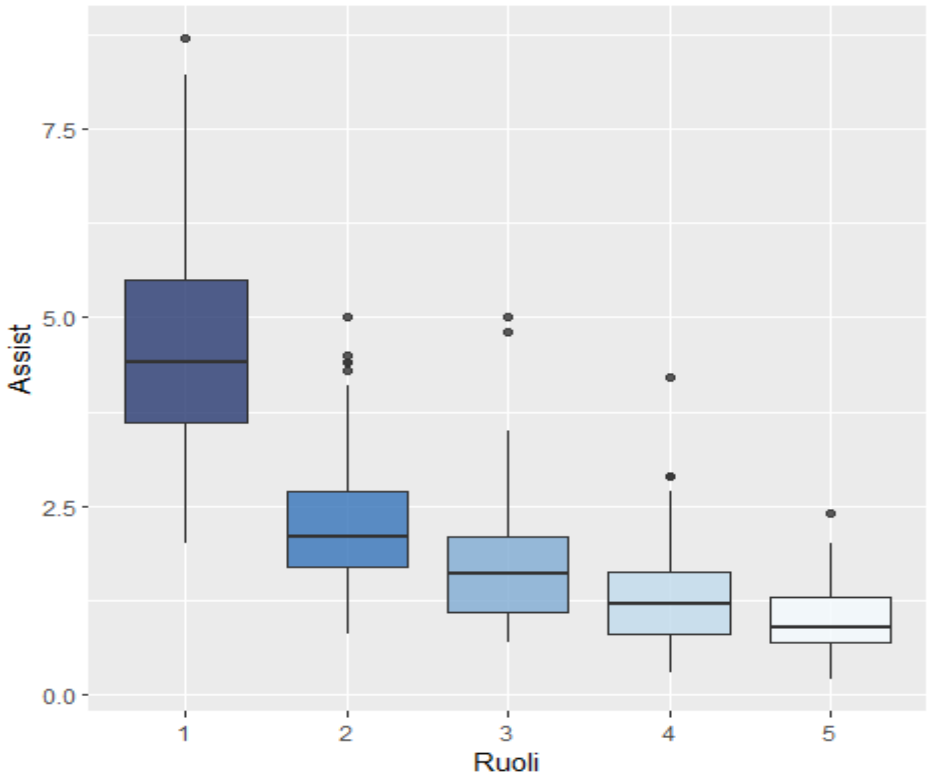

Figura 5: boxplot degli assist per ruolo

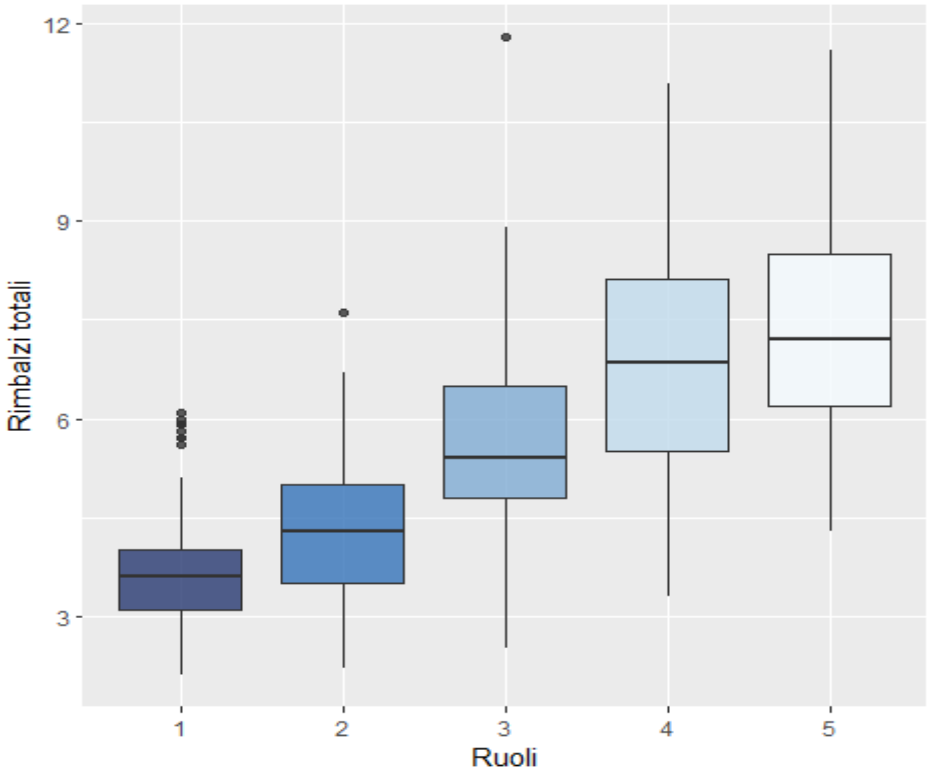

Figura 6: boxplot dei rimbalzi totali per ruolo

Per quanto riguarda gli assist e i rimbalzi invece notiamo delle differenze significative tra ruoli: soprattutto i playmaker (1), ma anche le guardie (2), hanno valori di AST molto elevati (come vediamo in Figura 5) e pochi rimbalzi (Figura 6). Mentre le ali (3 e 4) e i centri (5), al contrario, hanno un gran numero di rimbalzi (Figura 6) e pochi assist (Figura 5). Ancora più nello specifico notiamo che i playmaker e le guardie sono caratterizzati da un buon numero di palle rubate (Figura 7) ma anche da palle perse (Figura 9), mentre i centri e le ali grandi dalle stoppate (Figura 8).

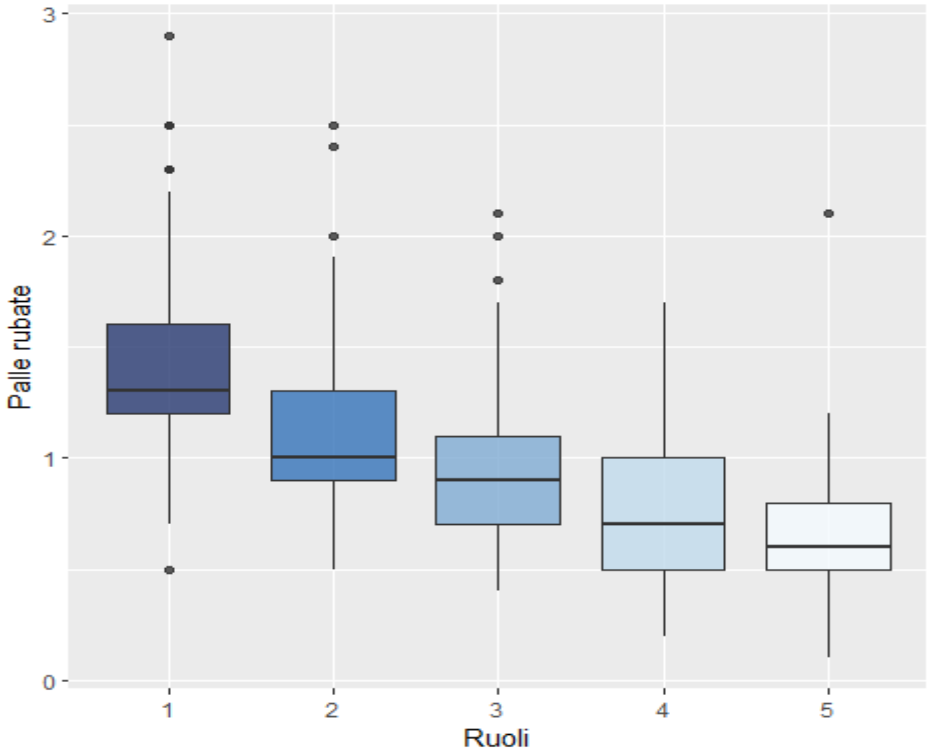

Figura 7: boxplot delle palle rubate per ruolo

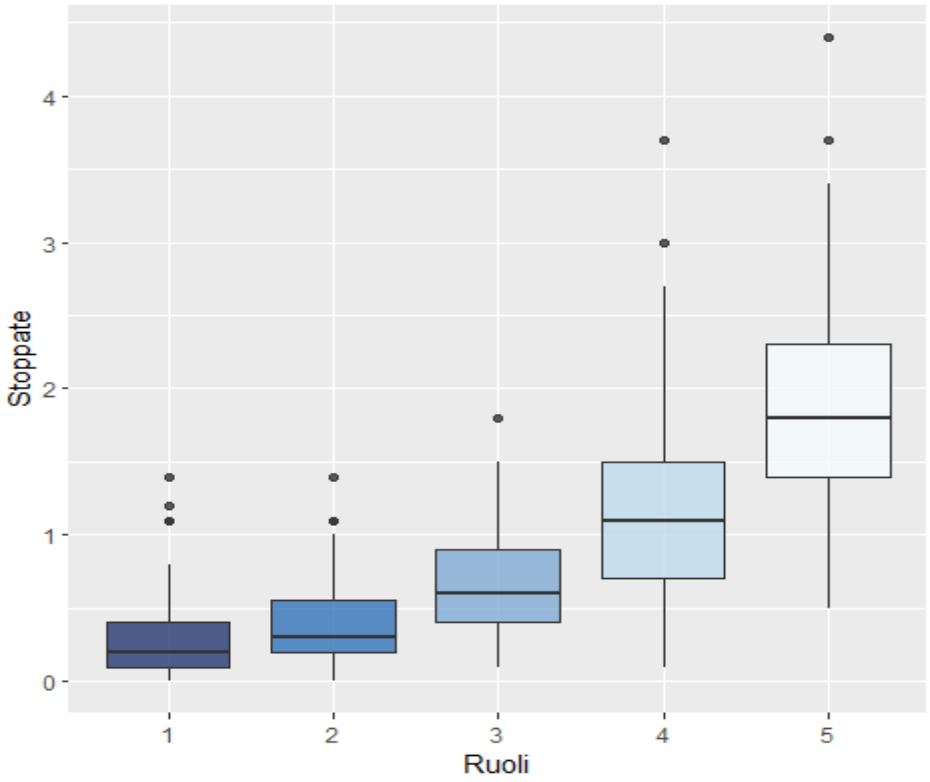

Figura 8: boxplot delle stoppate per ruolo

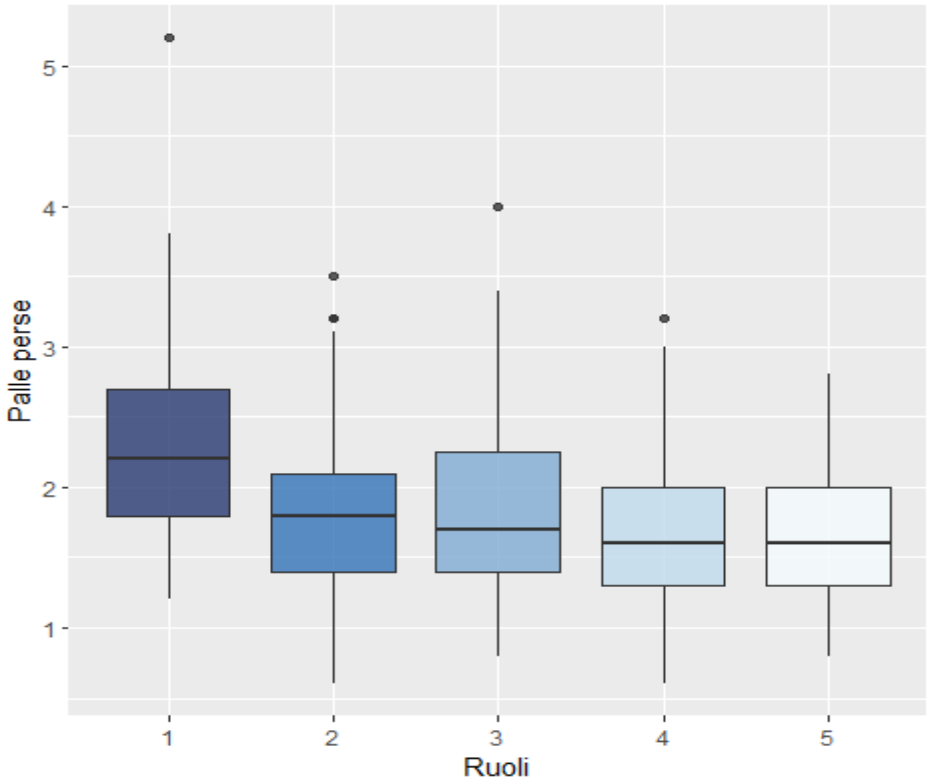

Figura 9: boxplot delle palle perse per ruolo

Da queste prime analisi descrittive emergono quindi le caratteristiche principali di ogni ruolo che erano state elencate nella presentazione del dataset e che vengono largamente confermate.

Inoltre, osserviamo anche in Tabella 2 come, sempre a seconda del ruolo, la posizione media con cui un giocatore è stato scelto al *draft* cambi molto: le ali piccole e i centri sono in media scelti ad una posizione migliore rispetto agli altri ruoli e i playmaker hanno la scelta media più alta.

RUOLO	MEDIA	STD.DEV.	FREQUENZA
1	25.20	13.56	85
2	23.36	12.30	143
3	21.05	14.31	71
4	24.68	13.50	96
5	21.51	13.43	61
Totale	23.37	13.30	456

Tabella 2: media della scelta a seconda del ruolo

Un altro aspetto interessante riguarda i minuti giocati: la media di minuti giocati a partita (MP) nel nostro dataset è di circa 28,27 minuti (Tabella 1). Contando che ogni partita di basket al college dura 40 minuti, è un buon numero. Ciò non deve però sorprendere molto, infatti bisogna tenere in considerazione che tutti i giocatori presenti nel dataset sono stati scelti per entrare nella NBA e sono un numero ristretto rispetto al totale dei giocatori di college disponibili, come vediamo in Tabella 3. È molto probabile quindi che ogni giocatore, nella sua carriera collegiale, avesse un ruolo importante per la propria squadra e di conseguenza un minutaggio elevato.

GIOCATORI NCAA COLLEGE	GIOCATORI ELEGGIBILI AL DRAFT (APPROSSIMATIVI)	SCELTE POSSIBILI AL DRAFT
18816	4181	60

Tabella 3: esempio relativo al 2020 del numero di atleti dichiaratosi eleggibili al draft

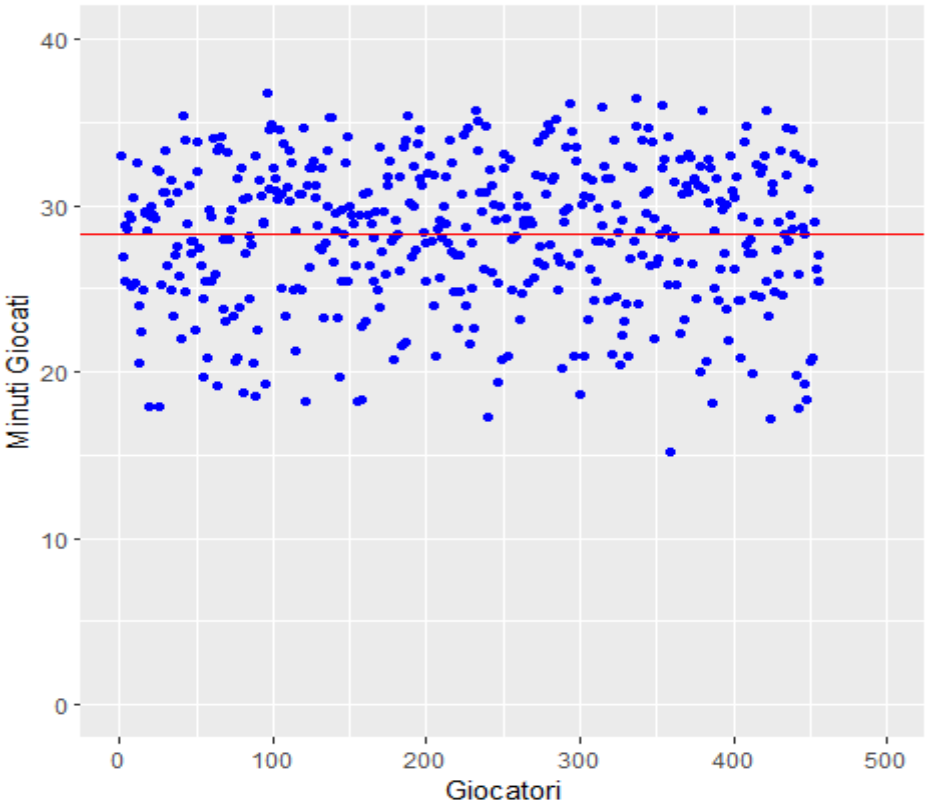

Figura 10: grafico di dispersione dei minuti giocati

2.3.1 Correlazioni

Vediamo ora attraverso una analisi delle correlazioni la relazione che esiste tra le variabili. Il comando *pwcorr* in *stata* calcola le correlazioni tra le variabili coppia per coppia, indicando il segno e la forza delle relazioni tra due variabili. Per ogni coppia il comando produce un coefficiente che varia tra 0 e 1: un valore di 0 indicherà una completa assenza di relazione tra le due variabili, 1 e -1 una relazione perfetta rispettivamente con direzione positiva o negativa. Ci aspettiamo quindi che i segni della correlazione siano negativi per tutte le variabili tranne che per le Palle perse poiché, come già anticipato, dovrebbe essere una statistica negativa per un giocatore e quindi correlata positivamente con la scelta

Vediamo quindi come interagiscono tra loro le variabili calcolando le correlazioni.

	Scelta	Punti Segnati	Rimbalzi Totali	Assist	Palle Rubate	Stoppate	Palle Perse
Scelta	1.0000						
Punti Segnati	-0.2702	1.0000					
Rimbalzi Totali	-0.2057	0.1789	1.0000				
Assist	-0.0320	0.3609	-0.3356	1.0000			
Palle Rubate	-0.1453	0.2739	-0.1862	0.6342	1.0000		
Stoppate	-0.2025	-0.1241	0.6262	-0.4299	-0.2494	1.0000	
Palle Perse	-0.1522	0.5898	0.1201	0.6223	0.4477	-0.1039	1.0000

Tabella 4: output dell'analisi delle correlazioni

Innanzitutto, i segni dei coefficienti sono coerenti con il problema di interesse: una correlazione con direzione negativa rispetto alla scelta indicherà che quella

variabile è legata a una scelta in posizione bassa al *draft*. Tra tutte notiamo che i punti segnati e i rimbalzi totali hanno la correlazione più forte rispetto alle altre variabili e con segno negativo, come vediamo anche nei grafici in Figura 11 e Figura 12.

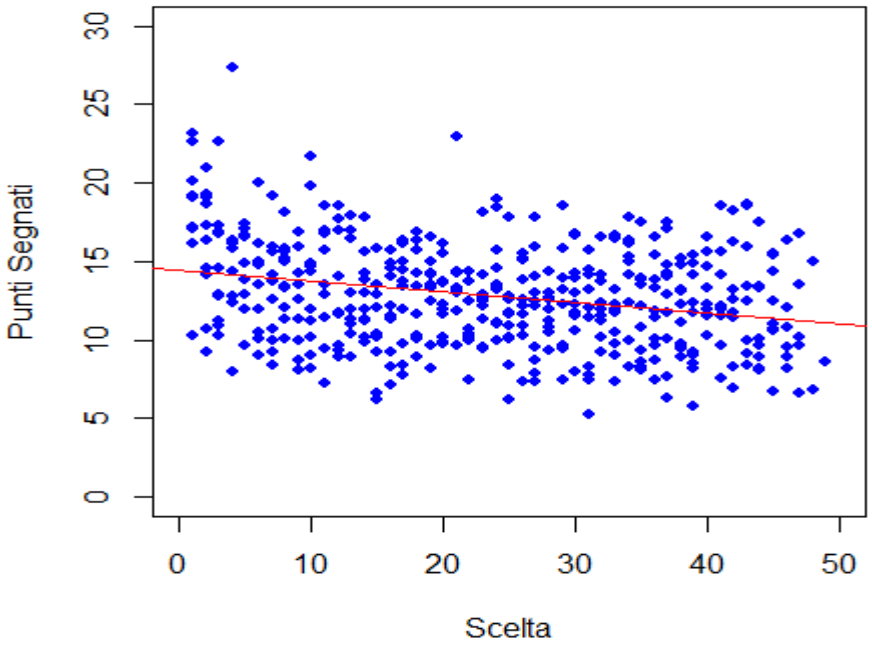

Figura 11: scatterplot scelta aggiustata in funzione dei punti segnati

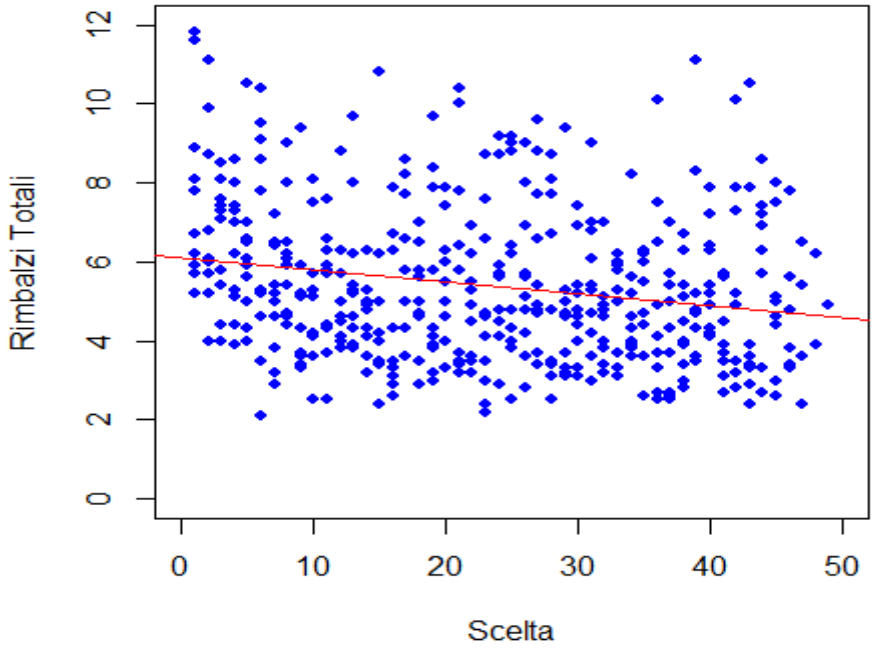

Figura 12: scatterplot scelta aggiustata in funzione dei rimbalzi totali

L'unica eccezione per le aspettative riguardo i segni delle correlazioni è per la variabile TOV. Sarebbe lecito aspettarsi qui un segno positivo dato che le palle perse sono un deficit per un giocatore. Cerchiamo di capire a cosa potrebbe essere dovuto questo segno positivo aggiungendo anche i minuti giocati a partita da ciascun giocatore (MP) e le partite giocate (GP).

	Scelta	Punti Segnati	Rimbalzi Totali	Assist	Palle Rubate	Stoppate	Palle Perse	Minuti Giocati	Partite Giocate
Scelta	1.000								
Punti Segnati	-0.270	1.000							
Rimbalzi Totali	-0.206	0.179	1.000						
Assist	-0.032	0.361	-0.335	1.000					
Palle Rubate	-0.145	0.274	-0.186	0.634	1.000				
Stoppate	-0.202	-0.124	0.626	-0.430	-0.249	1.000			
Palle Perse	-0.152	0.589	0.120	0.622	0.448	-0.104	1.000		
Minuti Giocati	-0.140	0.716	0.045	0.578	0.493	-0.247	0.553	1.000	
Partite Giocate	0.485	-0.239	-0.223	-0.004	-0.089	-0.191	-0.244	-0.135	1.000

Tabella 5: output dell'analisi delle correlazioni

Aggiungendo i minuti giocati si nota in Tabella 5 che questi sono correlati negativamente con la scelta, come ci si poteva aspettare, e si può provare a dare una prima spiegazione del segno della variabile palle perse: i minuti giocati sono fortemente correlati con segno positivo con i punti e con le palle perse, perché più un giocatore è in campo più ha possibilità di realizzare punti ma anche di generare più palle perse. Inoltre, nel basket moderno, il giocatore più forte della squadra tende ad avere più spesso la palla in mano e gli schemi sono incentrati su di lui, per cui ha più probabilità di perdere palla. A conferma di ciò vediamo come gli assist, che generalmente sono registrati dal ruolo del playmaker, siano correlati con segno positivo con le palle perse. Il playmaker, essendo il regista della squadra, avrà molto spesso il pallone tra le mani e quindi avrà anche più probabilità di perdere palla.

Per quanto riguarda le partite giocate notiamo che hanno correlazione positiva con la scelta e negativa con tutte le altre variabili, per cui sembra che i giocatori che sono rimasti per più tempo al college siano poi stati scelti in fondo al *draft* e abbiano registrato in media statistiche di gioco piuttosto basse. Al contrario quindi, chi ha le migliori performance sono quei giocatori che hanno meno partite giocate. Questo si

conferma essere, come già affermato da Groothius [6], il fenomeno degli *early entrants* ovvero di quei giocatori che sarebbero già pronti ad entrare nella NBA dopo la *high school* poiché mostrano un gran potenziale e sono fisicamente già formati per la NBA, ma sono obbligati a fare almeno un anno di college per la già citata *one and one rule*.

Potrebbe essere interessante anche considerare le statistiche *per minute,* ovvero una trasformazione delle variabili iniziali ottenuta dividendo le statistiche medie per partita con i minuti giocati a partita da ogni giocatore.

Ad esempio, per i punti giocati vale che:

$$PTSpm = PTS/MP$$

Equazione 1

	Scelta	Punti Segnati	Rimbalzi Totali	Assist	Palle Rubate	Stoppate	Palle Perse
Scelta	1.0000						
Punti Segnati	-0.2605	1.0000					
Rimbalzi Totali	-0.1252	0.1374	1.0000				
Assist	0.0028	-0.0326	-0.5411	1.0000			
Palle Rubate	-0.1137	-0.0918	-0.2844	0.5061	1.0000		
Stoppate	-0.1615	-0.0006	0.7068	-0.4807	-0.2190	1.0000	
Palle Perse	-0.0909	0.3126	0.0869	0.4231	0.2548	-0.0113	1.0000

Tabella 6: analisi delle correlazioni utlizzando le statistiche per minute ottenute in Equazione 1

Nelle correlazioni con le statistiche *per minute* notiamo che i coefficienti delle correlazioni sono leggermente diminuiti in valore assoluto, ma rimane ancora il segno negativo delle palle perse. Questo ci porta a pensare quindi che il segno negativo della correlazione palle perse-scelta sia dovuto in gran parte alla correlazione delle palle perse con i punti segnati (variabile più correlata negativamente con la scelta), come vediamo in Figura 13.

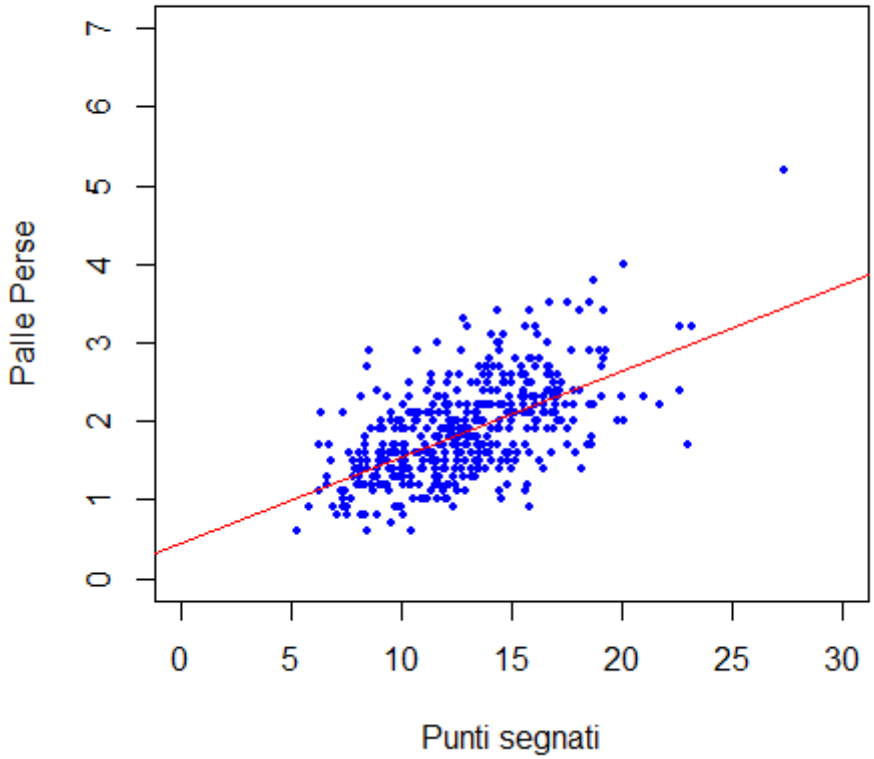

Figura 13: scatterplot punti segnati in funzione delle palle rubate

Capitolo 3 – Regressione lineare

3.1 Metodo

La regressione lineare è un modello statistico che consente di studiare la relazione tra due generiche variabili Y e X, permettendoci di capire come cambia Y al variare di X. Nel caso in cui l'interesse venga rivolto ad una singola variabile dipendente il modello è detto di regressione lineare semplice e si può indicare con la seguente equazione:

$$Y = \beta_0 + \beta_1 x + u$$

Equazione 2

La variabile Y è detta variabile dipendente o variabile risposta mentre X è la variabile indipendente. β_1 è la "pendenza", il coefficiente che indica come varia Y al variare di X e β_0 è l'intercetta del modello, ovvero il valore che Y assumerebbe se la variabile X fosse nulla. Infine, la variabile *u* indica il termine di errore del modello ovvero tutti i fattori che non siamo riusciti ad osservare ma che potrebbero influenzare la variabile risposta. Introducendo più variabili al modello siamo in grado di controllare più fattori che potrebbero influire sulla variabile dipendente, e si parlerà quindi di regressione lineare multipla:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + u$$

Equazione 3

Un modello di regressione lineare multipla può essere usato nel nostro caso per studiare la relazione tra ADJPICK, che nel nostro caso sarà la variabile dipendente, e tutte le altre variabili di interesse PTS, AST, TRB… che saranno le variabili indipendenti.

$$Adjpick = \beta_0 + \beta_1 * pts + \beta_2 * ast \ldots$$

(dove β_0 indica l'intercetta del modello)

Equazione 4

In questo modo potremo approfondire quanto già detto tramite le analisi descrittive e cercare di capire, attraverso diversi modelli di regressione, come le statistiche di gioco influiscano sulla scelta al *draft*. Il generico coefficiente β_i ci indicherà l'effetto di un aumento unitario della variabile x_i sulla variabile ADJPICK, "tenendo fermi" tutti gli altri fattori. Il nostro interesse sarà quindi capire quali variabili contribuiscono ad abbassare la scelta, quindi con coefficiente β_i negativi, e quali invece la fanno alzare, con coefficienti β_i positivi.

3.2 Stime OLS

Il metodo principale che verrà utilizzato per le analisi su *stata* si basa sulla stima dei parametri β dell'equazione di regressione attraverso il metodo dei minimi quadrati, anche detto *OLS* (dall'inglese *ordinary least squares*). Il metodo consiste nello scegliere le stime $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ che minimizzano la somma dei quadrati degli scarti tra Y e le X:

$$\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}\right)^2$$

Equazione 5

Il metodo OLS prevede alcune assunzioni fondamentali:

- Incorrelazione dei residui, ovvero tutto ciò che non ho osservato e potrebbe influire sulla var dipendente, deve essere incorrelato con le variabili esplicative:

$$E(u|x_1, x_2, \dots, x_k) = E(u) = 0$$

Equazione 6

- Assenza di collinearità tra le variabili esplicative: le X non devono essere costanti e perfettamente lineari tra loro;
- Il campione in analisi deve essere ottenuto in modo casuale dalla popolazione;
- I parametri β devono essere lineari.

A seguire vengono dettagliate le applicazioni dei modelli di regressione alle varie statistiche di gioco, andando ad integrare man mano le variabili presenti nel dataset e utilizzando diversi metodi per capire cosa emerge di diverso e di nuovo da ogni modello. Per ognuna di queste va ricordato che, come anticipato sopra, una stima negativa corrisponde a un abbassamento della posizione nella scelta al *draft.*

Procediamo quindi ad effettuare una regressione partendo dalle tre principali statistiche di gioco: Punti segnati, Rimbalzi totali, Assist. Usiamo come variabile risposta la scelta:

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-0.96	0.20	-4.74	0.000
Rimbalzi totali	-1.11	0.34	-3.22	0.001
Assist	0.01	0.48	0.01	0.989
Intercetta	41.66	2.76	15.09	0.000
Numerosità campionaria: 456				
R^2: 0.099				

Tabella 7: output regressione lineare

Di seguito un'analisi dei dati in Tabella 7.

- Punti segnati: coefficiente negativo e significativo (P>t < 0.05) questo ci dice che per ogni punto realizzato in più un giocatore guadagna quasi una posizione al *draft* (0.96 per la precisione), "tenendo fermi" tutti gli altri fattori
- Rimbalzi totali: coefficiente negativo e significativo, per ogni rimbalzo catturato in più un giocatore guadagna 1.11 posizioni al *draft*
- Assist: per gli assist invece il coefficiente, seppur molto piccolo, ha segno positivo e per niente significativo, con un *p-value* di 0.989

Questa prima regressione denota il segno negativo e la significatività dei punti segnati e dei rimbalzi totali, mentre per quanto riguarda gli assist, che dovrebbero essere una statistica che influisce in modo positivo sulla valutazione di un giocatore, risulta invece essere non significativa. Ma vediamo ora cosa succede aggiungendo le altre variabili del dataset per poter controllare più fattori:

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.16	0.22	-5.25	0.000
Rimbalzi totali	-0.18	0.43	-0.43	0.666
Assist	0.37	0.67	0.55	0.580
Palle rubate	-5.45	1.70	-3.21	0.001
Stoppate	-4.67	1.15	-4.08	0.000
Palle perse	1.15	1.47	0.79	0.432
Intercetta	45.49	2.83	16.06	0.000
Numerosità campionaria: 456				
R^2: 0.151				

Tabella 8:output regressione lineare

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.18	0.25	-4.66	0.000
Rimbalzi totali	-0.28	0.40	-0.70	0.482
Assist	-0.71	0.64	-1.10	0.272
Palle rubate	-4.78	1.58	-3.02	0.003
Stoppate	-3.06	1.06	-2.90	0.004
Palle perse	3.52	1.35	2.61	0.009
Minuti giocati	0.38	0.21	1.83	0.069
Partite giocate	0.14	0.01	9.82	0.000
Intercetta	20.26	4.42	4.58	0.000
Numerosità campionaria: 456				
R^2: 0.306				

Tabella 9:output regressione lineare

Considerando anche le altre statistiche principali e aggiungendo poi i minuti giocati e le partite giocate in Tabella 9, il modello sembra essere più vicino alle nostre

aspettative: il coefficiente di assist ora è negativo ma rimane poco significativo esattamente come i rimbalzi totali che perdono significatività rispetto al modello precedente. Saltano all'occhio i segni negativi molto forti e significativi di stoppate e palle rubate: per ogni stoppata e palla rubata in più, rispettivamente 3 e quasi 5 posizioni guadagnate al *draft*. I segni di queste due variabili rispettano le attese ma è sorprendente il forte effetto che sembrano avere sulla scelta. Per quanto riguarda le palle perse, il segno positivo e la grandezza del coefficiente smentiscono quanto trovato nelle correlazioni al Capitolo 2 e ci dicono che per ogni palla persa in più un giocatore perde 3.52 posizioni al *draft*, il che sembra rispettare la natura negativa di tale statistica. I minuti giocati e le partite giocate hanno invece entrambi segno positivo anche se i minuti non sono molto significativi.

Proviamo a capire la scarsa significatività di assist e rimbalzi totali da cosa potrebbe essere causata prendendo ad esempio la variabile relativa agli assist e analizzandola meglio:

Ruolo	Media	Std.Dev.	Frequenza
1	4.55	1.33	85
2	2.25	0.80	143
3	1.76	0.85	71
4	1.30	0.66	96
5	0.97	0.47	61
Totale	2.23	1.48	456

Tabella 10: media variabile assist a seconda del ruolo

Ruolo	Media	Std.Dev.	Frequenza
1	25.20	13.56	85
2	23.36	12.30	143
3	21.06	14.31	71
4	24.68	13.51	96
5	21.51	13.43	61
Totale	23.375	13.29966	456

Tabella 11: media variabile ADJPICK a seconda del ruolo

Dalla Tabella 10 notiamo che, nel nostro dataset, i playmaker registrano in media 4.55 assist (il numero più alto rispetto agli altri ruoli). Inoltre, un playmaker viene selezionato in media alla venticinquesima scelta, come vediamo in Tabella 11. La statistica degli assist potrebbe quindi essere troppo specifica per le caratteristiche del gioco di un solo ruolo (così come i rimbalzi per i centri), ed inoltre essere tipica di giocatori che vengono scelti in fondo al *draft.* Dobbiamo tenere conto, infatti, che in questi modelli stiamo confrontando giocatori all'interno del dataset che hanno ruoli diversi e quindi caratteristiche di gioco differenti, per cui è normale aspettarsi differenze significative tra alcune statistiche. Per provare a risolvere questo problema Berri [1] aveva proposto una sorta di standardizzazione delle variabili per renderle confrontabili tra loro e cercare di ridurre l'errore dovuto alla differenza di ruolo. Per fare ciò calcolava le medie di ogni statistica condizionate al ruolo e le medie di ogni statistica rispetto all'intero dataset senza distinzioni. A questo punto si sottrae alla variabile la sua media rispetto al ruolo del giocatore e poi sia aggiunge la media generale di quella variabile senza distinzione di ruolo, ottenendo le nuove variabili che vedremo nelle seguenti tabelle: PTS2, TRB2, AST2…

$$PTS2 = PTS - MediaPTSRuolo + MediaPTS$$

Equazione 7

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.06	0.23	-4.63	0.000
Rimbalzi totali	-0.49	0.53	-0.91	0.361
Assist	-0.37	0.86	-0.43	0.667
Palle rubate	-4.11	1.81	-2.27	0.024
Stoppate	-5.96	1.34	-4.44	0.000
Palle perse	1.54	1.47	1.05	0.294
Intercetta	46.36	2.86	16.23	0.000
Numerosità campionaria: 456				
R^2: 0.156				

Tabella 12: regressione con le variabili "standardizzate"

Aggiungiamo anche i minuti giocati e le partite giocate per vedere se è cambiato qualcosa rispetto al modello precedente con i dati "grezzi":

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.04	0.24	-4.40	0.000
Rimbalzi totali	-0.71	0.50	-1.44	0.150
Assist	-0.96	0.80	-1.21	0.228
Palle rubate	-3.25	1.65	-1.97	0.050
Stoppate	-4.54	1.22	-3.71	0.000
Palle perse	3.74	1.35	2.77	0.006
Minuti giocati	0.34	0.17	2.03	0.042
Partite giocate	0.14	0.01	9.67	0.000
Intercetta	21.87	3.93	5.57	0.000
Numerosità campionaria:456				
R^2: 0.314				

Tabella 13: regressione con le variabili "standardizzate" con i minuti giocati e le partite giocate

Il modello sembra già adattarsi meglio: come vediamo in Tabella 13, migliora leggermente da un punto di vista delle significatività dei rimbalzi totali e degli assist, mentre i coefficienti rimangono pressoché uguali al modello in Tabella 8 e Tabella 9.

Questa sorta di standardizzazione proposta da Berri che si basa sugli scostamenti delle medie in base al ruolo, si poteva semplicemente ottenere considerando, all'interno del modello visto in Tabella 8 la variabile ROLE come una *dummy* che prende valore 1 per i playmaker, 2 per le guardie, 3 per le ali piccole, 4 per le ali grandi ed infine 5 per i centri. Si nota infatti che i risultati ottenuti sono identici:

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.06	0.23	-4.65	0.000
Rimbalzi totali	-0.49	0.53	-0.92	0.359
Assist	-0.37	0.86	-0.43	0.666
Palle rubate	-4.12	1.80	-2.28	0.023
Stoppate	-5.96	1.34	-4.46	0.000
Palle perse	1.55	1.47	1.05	0.293
RUOLO				
2	-3.00	2.42	-1.24	0.215
3	-4.02	3.16	-1.27	0.204
4	1.51	3.76	0.40	0.687
5	1.08	4.43	0.24	0.808
Intercetta	47.46	3.79	12.50	0.000
Numerosità campionaria				
R^2: 0.169				

Tabella 14: output regressione con variabile dummy ruolo

L'interpretazione della dummy ROLE è da leggere nel seguente modo: il software *stata* considera in automatico come base del modello il ruolo 1, ovvero il gruppo sulla base dei quali saranno calcolati i coefficienti degli altri gruppi. Per cui rispetto ai playmaker le guardie vengono scelte 3 posizione prima al *draft,* le ali piccole 4 e così via. Notiamo però che le significatività riguardanti i ruoli sono molto basse.

3.3 Regressione ad effetti fissi

L'idea proposta da Berri [1] di tenere conto del ruolo attraverso una "standardizzazione" che coglie lo scostamento di ogni osservazione dalla media del suo gruppo di appartenenza, è alla base della regressione ad effetti fissi. Un modello ad effetti fissi è un particolare modello di regressione che permette di analizzare un insieme di osservazioni, che possono essere caratterizzate da una divisione in gruppi (nel nostro caso i ruoli della pallacanestro), mantenendo costanti le medie di

ogni gruppo. In questo modo il modello permette di catturare la variabilità tra i diversi gruppi.

L'obiettivo degli effetti fissi è cercare di eliminare quei fattori non osservati che rimangono costanti tra i gruppi, indicati nella seguente equazione con a_i.

$$Y_{it} = \beta_1 x_{it} + a_i + u_{it}$$

Equazione 8

Adesso calcoliamo la media per ogni gruppo:

$$\bar{Y}_i = \beta_1 \bar{x}_i + a_i + \bar{u}_i$$

Equazione 9

Ora per eliminare a_i sottraiamo Equazione 9 a Equazione 8:

$$Y_{it} - \bar{Y}_i = \beta_1 (x_{it} - \bar{x}_i) + u_{it} - \bar{u}_i$$

Equazione 10

Il procedimento appena presentato non è altro che quello visto in Equazione 7, per cui possiamo verificare se i risultati coincidono, considerando un modello in cui la variabile ROLE è l'identificativo di gruppo, in questo caso i diversi ruoli. Osserviamo quindi in Tabella 15 risultati identici a quelli che avevamo ottenuto attraverso il modello con la variabile dummy in Tabella 14 e quello con le variabili standardizzate in Tabella 13.

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.06	0.23	-4.61	0.000
Rimbalzi totali	-0.49	0.53	-0.92	0.359
Assist	-0.37	0.86	-0.43	0.666
Palle rubate	-4.11	1.81	-2.28	0.023
Stoppate	-5.96	1.34	-4.46	0.000
Palle perse	1.55	1.47	1.05	0.293
Intercetta	46.36	2.85	16.28	0.000
Numerosità campionaria: 456				
R^2: within = 0.158, between = 0.095, totale = 0.144				
Rho: 0.039 (correlazione intraclasse)				
Test F (ipotesi u_i=0): F (4, 445) = 2.42 (Prob > F = 0.047)				

Tabella 15: output regressione ad effetti fissi

3.4 Statistiche per minute

Proviamo ora a considerare un modello con le statistiche calcolate sul minutaggio di ogni giocatore, per tenere conto che alcuni giocatori potrebbero giocare meno di altri ma avere comunque buone performance in rapporto al tempo di gioco. Si procede allo stesso modo della standardizzazione in Equazione 7 ma utilizzando, come già introdotto in Equazione 1, le statistiche *per minute* e moltiplicando infine tutto per 40, ovvero la durata di ogni partita di college, per rendere tutto confrontabile su una stessa base di minuti giocati:

$$PTS2pm = (PTSpm - MediaPTSRuolo + MediaPTS) \times 40$$

Equazione 11

Si noti che, per i nostri scopi, trasformare le statistiche moltiplicandole per 40 o lasciarle per il singolo minuto giocato, avrebbe portato agli stessi risultati dal punto di vista della regressione in quanto è una semplice trasformazione di scala. Ai fini di una migliore interpretazione si sceglie di utilizzare le statistiche *per 40 minutes* e, prima di procedere ad utilizzare le variabili in Equazione 11, si riporta anche un modello con le variabili *per minute* per poterli confrontare.

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.20	0.19	-6.25	0.000
Rimbalzi totali	-0.12	0.32	-0.38	0.705
Assist	-0.48	0.52	-0.92	0.357
Palle rubate	-4.47	1.26	-3.54	0.000
Stoppate	-2.52	0.75	-3.33	0.001
Palle perse	1.40	1.05	1.33	0.183
Intercetta	53.06	4.26	12.46	0.000
Numerosità campionaria: 456				
R^2: 0.130				

Tabella 16: output regressione con variabili per minute

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.18	0.19	-6.14	0.000
Rimbalzi totali	-0.58	0.41	-1.40	0.162
Assist	-1.24	0.67	-1.85	0.065
Palle rubate	-3.39	1.31	-2.58	0.010
Stoppate	-3.57	0.86	-4.13	0.000
Palle perse	1.77	1.05	1.68	0.093
Intercetta	57.35	4.44	12.92	0.000
Numerosità campionaria: 456				
R^2: 0.145				

Tabella 17: output regressione con variabili per minute e standardizzate

Considerando le statistiche *per minute,* rispetto al modello con le variabili standardizzate in Tabella 12, migliorano molto le significatività anche se rimbalzi totali, assist e palle perse rimangono ancora leggermente poco significative. I segni delle stime vengono confermati come nei precedenti modelli soprattutto i punti segnati confermando quanto visto nella letteratura nei lavori di Berri [1] [3]. Anche per quanto riguarda le palle rubate e le stoppate viene confermano il loro effetto molto forte negativo e significativo.

Come già visto per i modelli precedenti, potevamo ottenere gli stessi risultati attraverso un modello con la dummy ROLE o con una regressione ad effetti fissi.

Scelta	Stima	Std.err.	t	P>t
Punti segnati	-1.18	0.19	-6.16	0.000
Rimbalzi totali	-0.58	0.41	-1.41	0.160
Assist	-1.24	0.67	-1.85	0.065
Palle rubate	-3.39	1.31	-2.59	0.010
Stoppate	-3.57	0.86	-4.15	0.000
Palle perse	1.77	1.05	1.69	0.092
Ruolo				
2	-4.06	2.41	-1.68	0.093
3	-4.40	3.18	-1.38	0.168
4	2.42	3.90	0.62	0.534
5	2.38	4.69	0.51	0.612
Intercetta	58.48	5.03	11.61	0.000
				Numerosità campionaria: 456
				R^2: 0.160

Tabella 18: output regressione con variabile dummy ruolo, statistiche per minute

Scelta	Stima	Std.Err.	t	P>t
Punti segnati	-1.18	0.19	-6.16	0.000
Rimbalzi totali	-0.58	0.41	-1.41	0.160
Assist	-1.24	0.67	-1.85	0.065
Palle rubate	-3.39	1.31	-2.59	0.010
Stoppate	-3.57	0.86	-4.15	0.000
Palle perse	1.77	1.05	1.69	0.092
Intercetta	57.34	2.85	16.28	0.000
Numerosità campionaria: 456				
R^2: within = 0.146, between = 0.052, totale = 0.116				
Rho: 0.068 (correlazione intraclasse)				
Test F (ipotesi u_i=0): F(4, 445) = 3.73 (Prob > F = 0.054)				

Tabella 19: output regressione ad effetti fissi, statistiche per minute

Capitolo 4 – Ulteriori analisi

4.1 Analisi di robustezza

4.1.1 Variabile ADJPICK

La decisione di "aggiustare" la variabile PICK con la rimozione degli atleti internazionali da ogni *draft* è stata presa con il proposito di avere statistiche coerenti e provenienti da un unico database [15]. Infatti, dato che i giocatori non collegiali provengono da campionati a volte molto diversi tra loro e alcuni non molto conosciuti, sarebbe stato necessario confrontare diversi database di ogni campionato al di fuori degli USA. Su quelli però non abbiamo la certezza che siano precisi ed aggiornati e che seguano tutti lo stesso criterio di assegnazione delle statistiche, come invece l'abbiamo per i college USA. Per esempio, alcune delle statistiche, in particolare per gli assist, hanno un criterio di assegnazione che potrebbe variare a seconda del campionato e della federazione cestistica di ogni paese. Al contrario invece per i college possiamo fare affidamento su un regolamento effettivo che permette di avere almeno inizialmente un criterio comune per tutte le squadre di college, anche se poi la decisione di assegnare l'assist spetta allo statistico che segue la partita. [14]

Come si nota in Tabella 20 e Tabella 21, considerando PICK o ADJPICK i modelli sono molto simili soprattutto in termini di significatività, probabilmente anche per il numero di giocatori internazionali che è rimasto costante negli anni considerati, tranne che per il 2022 (40 college e 20 internazionali). I coefficienti, inoltre, sia per segno che per grandezza, confermano i risultati ottenuti nei modelli precedenti, portandoci a ritenere che la scelta di utilizzare ADJPICK non porta a commettere errori nelle costruzioni dei modelli.

Scelta	Stima	Std.Err.	t	P>t
Punti segnati	-1.59	0.25	-6.39	0.000
Rimbalzi totali	-0.75	0.54	-1.39	0.165
Assist	-1.70	0.87	-1.95	0.052
Palle rubate	-4.55	1.70	-2.67	0.008
Stoppate	-4.50	1.12	-4.01	0.000
Palle perse	2.80	1.36	2.05	0.041
Intercetta	73.68	5.76	12.79	0.000
Numerosità campionaria: 456				
R^2: 0.146				

Tabella 20: output regressione con la scelta non aggiustata (variabili standardizzate e per minute)

Scelta	Stima	Std.Err.	t	P>t
Punti segnati	-1.18	0.19	-6.14	0.000
Rimbalzi totali	-0.58	0.41	-1.40	0.162
Assist	-1.24	0.67	-1.85	0.065
Palle rubate	-3.39	1.31	-2.58	0.010
Stoppate	-3.57	0.86	-4.13	0.000
Palle perse	1.77	1.05	1.68	0.093
Intercetta	57.35	4.44	12.92	0.000
Numerosità campionaria: 456				
R^2: 0.145				

Tabella 21: output regressione con la scelta aggiustata (variabili standardizzate e per minute)

4.1.2 Regressioni per quinquennio

È interessante verificare quanto affermato nella Letteratura riguardo alle priorità nell'osservazione di un giocatore col passare del tempo, attraverso i dati con i nostri modelli, dato che il nostro dataset considera un periodo di dieci anni di scelte. Per fare ciò confrontiamo due modelli diversi: uno con le classi *draft* degli anni dal 2013 al 2017 e l'altro dal 2018 al 2022 (vengono usate le variabili *per minute* e standardizzate).

Scelta	Stima	Std.Err.	t	P>t
Punti segnati	-1.40	0.28	-5.05	0.000
Rimbalzi totali	-1.68	0.57	-2.94	0.004
Assist	-0.88	0.90	-0.97	0.331
Palle rubate	-3.76	1.77	-2.13	0.035
Stoppate	-2.03	1.19	-1.69	0.092
Palle perse	2.60	1.49	1.74	0.083
Intercetta	64.74	6.49	9.97	0.000
				Numerosità campionaria: 456
				R^2: 0.174

Tabella 22: output regressione quinquennio 2013-2017 (variabili standardizzate e per minute)

Scelta	Stima	Std.Err.	t	P>t
Punti segnati	-1.03	0.27	-3.85	0.000
Rimbalzi totali	0.66	0.61	1.08	0.281
Assist	-1.67	1.00	-1.67	0.096
Palle rubate	-2.06	1.96	-1.06	0.292
Stoppate	-5.99	1.26	-4.74	0.000
Palle perse	1.01	1.51	0.67	0.504
Intercetta	49.95	6.16	8.11	0.000
				Numerosità campionaria: 456
				R^2: 0.170

Tabella 23: output regressione quinquennio 2018-2022 (variabili standardizzate e per minute)

Notiamo che per quanto riguarda i punti segnati non vi sono grandi cambiamenti, i coefficienti sono simili e sempre molto significativi, segno che questa statistica è

rimasta negli anni sempre molto importante nella scelta di un giocatore. Per i rimbalzi nel primo quinquennio si ha un coefficiente negativo e significativo, al contrario del secondo dove addirittura il segno dei rimbalzi totali è positivo, mentre gli assist rimangono sempre poco significativi, anche se nel secondo quinquennio la significatività sembra aumentare leggermente. Le palle rubate e le stoppate confermano, come nei modelli considerati in precedenza, un forte effetto negativo, che per le stoppate aumenta di molto nel secondo quinquennio, più significativo che per le palle rubate. Le palle perse sembrano essere più significative nel primo quinquennio che nel secondo, ma sempre con segno positivo. Da queste analisi notiamo quindi qualche cambiamento anche se leggero: il dataset attraversa dieci anni di scelte al *draft,* un periodo forse non troppo ampio (2013-2022) per notare differenze significative o cambiamenti radicali nello stile di gioco. Da un altro punto di vista invece troviamo conferme sulla significatività di alcune variabili come i punti segnati e le stoppate.

4.2 Differenze per ruolo

Come verificato nel Capitolo 2, sono emerse delle differenze significative tra le statistiche considerate a seconda del ruolo. Anche nella valutazione di un giocatore, un osservatore sicuramente darà maggior peso, in base al tipo di giocatore che sta cercando, ad alcune statistiche rispetto ad altre. Nei precedenti modelli si è provato a tenere considerazione di ciò attraverso gli scostamenti delle medie prima in Equazione 7, poi introducendo la variabile *dummy* per il ruolo (Tabella 14) ed infine con una regressione ad effetti fissi. Proviamo ora invece a cogliere queste differenze considerando un modello di regressione diverso per ogni ruolo; in questo modo dovrebbe essere più facile fare confronti tra ruoli diversi e capire se qualche variabile è significativa per un ruolo rispetto ad un altro.

Riportiamo nelle tabelle sottostanti gli output prima della regressione generale senza distinzione di ruolo e poi delle regressioni con un modello diverso per ogni ruolo, per far emergere anche le differenze rispetto al modello completo.

Scelta	Stima	Std.Err.	t	P>t
Punti Segnati	-1.20	0.19	-6.25	0.000
Rimbalzi totali	-0.12	0.32	-0.38	0.705
Assist	-0.48	0.52	-0.92	0.357
Stoppate	-4.48	1.26	-3.54	0.000
Palle rubate	-2.52	0.76	-3.33	0.001
Palle perse	1.40	1.05	1.33	0.183
Intercetta	53.06	4.26	12.46	0.000
				Numerosità campionaria: 456
				R^2: 0.129

Tabella 24: output regressione con variabili per minute

	PLAYMAKER			GUARDIA		
Scelta	**Stima**	**Std.err.**	**P>t**	**Stima**	**Std.err.**	**P>t**
Punti Segnati	-0.84	0.47	0.077	-1.07	-2.98	0.003
Rimbalzi totali	-2.92	1.45	0.048	-1.99	-2.34	0.021
Assist	-1.64	1.14	0.154	0.02	0.01	0.991
Stoppate	-3.47	2.67	0.198	-0.34	-0.15	0.885
Palle rubate	-5.58	4.84	0.253	-3.99	-1.26	0.209
Palle perse	1.00	2.37	0.673	2.01	0.97	0.332
Intercetta	69.08	10.42	0.000	52.07	5.92	0.000
	Numerosità campionaria: 85			Numerosità campionaria: 143		
	R: 0.245			R: 0.100		

Tabella 25: output regressione divisa per i ruoli 1 e 2, statistiche per minute

	ALA PICCOLA			ALA GRANDE			CENTRO		
Scelta	**Stima**	**Std.err.**	**P>t**	**Stima**	**Std.err.**	**P>t**	**Stima**	**Std.err.**	**P>t**
Punti	-1.73	0.54	0.002	-1.38	0.42	0.002	-1.28	0.49	0.012
Rimbalzi	1.98	1.16	0.092	-0.01	0.71	0.979	-0.03	1.3	0.981
Assist	-2.38	1.92	0.220	-0.50	1.65	0.762	-3.62	2.89	0.217
Stoppate	-7.58	4.20	0.076	-3.22	3.28	0.329	-5.45	3.93	0.171
Palle rubate	-5.53	4.45	0.218	-3.09	1.32	0.021	-3.34	1.48	0.028
Palle perse	2.56	2.78	0.360	1.76	2.23	0.431	2.54	3.32	0.449
Intercetta	51.66	10.41	0.000	56.12	10.9	0.000	59.41	13.35	0.000
	Numerosità campionaria: 85			Numerosità campionaria: 96			Numerosità campionaria: 61		
	R: 0.231			R: 0.160			R: 0.271		

Tabella 26:output regressione divisa per i ruoli 1 e 2, statistiche per minute

Possiamo notare che la variabile punti segnati è l'unica che rimane significativa in tutti i modelli, confermando quindi che i punti segnati sono una priorità nell'osservazione di un prospetto collegiale. Notiamo poi che la variabile stoppate è significativa per le ali grandi e i centri, ma si nota che, per gli stessi ruoli, i rimbalzi totali non sono significativi al contrario di quanto ci si potesse aspettare, dato che è una statistica che contraddistingue questi due ruoli. Al contrario invece notiamo un aspetto interessante che riguarda playmaker e guardie: per questi ruoli i rimbalzi totali risultano significativi, probabilmente poiché per un playmaker o una guardia, oltre agli assist e ai punti segnati, un buon numero di rimbalzi potrebbe essere un valore aggiunto.

Un altro metodo per provare a cogliere meglio se esistono differenze significative tra i ruoli, è attraverso un modello definito *fully interacted model*, in cui le variabili vengono valutate per la loro interazione con la variabile ROLE, e dove l'intercetta e i coefficienti possono variare tra i vari gruppi. Il modello non verrà riportato in questo elaborato poiché risultava molto poco significativo, se non per la variabile dei rimbalzi per i ruoli 3 e 4, e non consentiva di trarre delle conclusioni adeguate. Probabilmente questo è dovuto alla numerosità del campione utilizzato che potrebbe essere non abbastanza elevata e presentare poca variabilità tra i ruoli per questo tipo di analisi.

Risultati e conclusioni

Dalle analisi condotte nella tesi sono quindi emersi diversi aspetti significativi dai vari metodi presi in considerazione: i vari modelli ci portano a concludere che i punti segnati sembrano essere la determinante che ha un maggiore impatto sulla scelta di un giocatore al *draft*, in linea con le precedenti ricerche sull'argomento, in particolare quelle di Berri [1] [2] e di Berri, Brook e Schimdt [3]. Un altro risultato più sorprendente riguarda invece l'influenza positiva che anche stoppate e palle rubate hanno sulla scelta, come già affermato in letteratura, ma che attraverso questa analisi si è visto avere un effetto ancora più forte rispetto ai risultati di Evans [5]. Da questi risultati emerge quindi un profilo di giocatore ideale che attraverso buoni doti offensive (punti segnati) e difensive (stoppate e palle rubate) potrebbe attirare l'attenzione delle squadre NBA in vista del *draft*. Per quanto riguarda le statistiche (assist, rimbalzi, palle perse e percentuali realizzative) che non hanno trovato conferme nei modelli di regressione, soprattutto in termini di significatività, non vuol dire che non sono importanti nella valutazione di un giocatore. Il dataset utilizzato infatti, potrebbe presentare dei limiti, ad esempio per quanto riguarda la numerosità campionaria: soprattutto per le analisi riguardanti i ruoli, è emersa poca variabilità tra ciascun gruppo dovuto probabilmente alla dimensione non molto elevata del campione. Inoltre, il numero di variabili coinvolte nelle analisi, chiaramente non rispecchia tutti i fattori che nella realtà potrebbero incidere nella scelta di un giocatore. Molti fattori, infatti, non sono controllabili nei modelli presentati nella tesi: basti pensare alle variabili legate alla bravura di ogni manager e talent scout nel saper scegliere il giocatore giusto in base alle necessità della propria squadra. Anche per quanto riguarda i giocatori, alcune variabili non possono essere controllate per migliorare la propria scelta al *draft* come, ad esempio, tutto ciò che riguarda la genetica, come l'altezza o le doti atletiche.

Quindi, i risultati emersi da questa tesi non possono essere una certezza riguardo alle statistiche su cui un giocatore di college dovrebbe concentrarsi per ottenere una scelta migliore, ma potrebbero essere un punto di partenza per ulteriori analisi future. Attraverso dataset più dettagliati e un controllo su più variabili in futuro si potrebbe arrivare a definire con precisione le determinanti che incidono sulla scelta

di un giocatore e ridurre al minimo la probabilità per le squadre di fare una cattiva scelta al *draft.*

Bibliografia

[1] D. Berri, «From college to the pros: predicting the NBA amateur player draft,» *Journal of Productivity Analysis,* pp. 25-35, 2011.

[2] J. Harris e D. J. Berri, «Predicting the WNBA draft: What matters most from college performance?,» *International Journal of Sport Finance,* p. 299, 2015.

[3] D. J. Berri, S. L. Brook e M. B. Schmidt, «Does one simply need to score to score,» *International Journal of Sport Finance,* pp. 190-205, 2007.

[4] D. Coates e B. Oguntimein, «The length and success of NBA careers: Does college production predict professional outcomes,» *International Journal of Sport Finance,* pp. 4-26, 2010.

[5] B. A. Evans e J. D. Pitts, «The Determinants of Draft Position for NBA Prospects,» *NEW YORK ECONOMIC REVIEW,* p. 22, 2017.

[6] P. A. Groothuis, J. R. Hill e T. J. Perri, «Early entry in the NBA draft: The influence of unraveling, human capital, and option value,» *Journal of Sports Economics,* pp. 223-243, 2007.

[7] «Draft» https://it.wikipedia.org/wiki/Draft.

[8] «Draft NBA»https://it.wikipedia.org/wiki/Draft_NBA.

[9] «DraftHistory»https://nbahoopsonline.com/Articles/History/Drafthistory.html.

[10] «Draft Lottery» https://en.wikipedia.org/wiki/NBA_draft_lottery.

[11] A. Bonfante, «Tanking, cos'è e come funziona la Draft Lottery?» 15 11 2016. https://www.nbareligion.com/2016/11/15/vocabolario-nba-cose-e-come-funziona-la-draft-lottery-parte-22/.

[12] «Criteri di eleggibilità» https://en.wikipedia.org/wiki/Eligibility_for_the_NBA_draft.

[13] «Draft History» http://www.thepostgame.com/blog/throwback/201507/remember-nba-david-stern-collective-bargaining-agreement-players-union.

[14] G. K. Johnson e J. Hamilton, «BASKETBALL STATICIANS' MANUAL» http://fs.ncaa.org/Docs/stats/Stats_Manuals/Basketball/2023EZ.pdf.

[15] «Basketball Reference College Stats» https://www.sports-reference.com/cbb/.

[16] «Probability of going pro» https://ncaaorg.s3.amazonaws.com/research/pro_beyond/2020RES_ProbabilityBeyondHSFiguresMethod.pdf.

[17] J. M. Wooldridge, Introductory econometrics: A modern approach, 2012.

[18] R. Fante, «Basketball NCAA» https://basketballncaa.com/come-funziona-il-campionato-di-basket-dellncaa/.

[19] A. Mauri, «Freshman, redshirt, ecc.: i giocatori Ncaa,» 2019. https://basketballncaa.com/freshman-redshirt-ecc-i-giocatori-ncaa/.