Stockholms universitet

# Vilka faktorer leder ett basketlag till framgång? En studie med logistisk regression.

Louise Böhme

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2020:7
Matematisk statistik
Juni 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2020:7**
http://www.math.su.se/matstat

# Vilka faktorer leder ett basketlag till framgång? En studie med logistisk regression.

Louise Böhme*

Juni 2020

**Sammanfattning**

Syftet med detta arbete är att ta reda på vad som kan få ett basketlag att nå framgång. För att finna svaret på den frågan kommer jag undersöka vilka faktorer som under en basketmatch har störst påverkan och betydelse för matchresulatet och vilket lag som vinner. Till mitt förfogande har jag nyttjat ett datamaterial hämtat från NBA (National Basketball Association) officiella hemsida bestående av alla basketmatcher som spelades under NBA-säsongen 2018/2019. Med hjälp av logistisk regression analyserar jag den binära responsvariabeln vinst/förlust och hur den påverkas av ett antal förklarande variabler. Det visar sig att det inte nödvändigvis behövs särskilt många variabler för att tillsammans kunna förklara utfallet i en basketmatch, vilket ger upphov till många delade åsikter och teorier i frågan. Men en rödtråd påvisar att ett basketlag bör fokusera på att framförallt skjuta bra från den vanligaste distansen, se till att komma till avslut ofta, ta många offensiva returer samt spela ett aggressivt försvarspel för att öka sina chanser till vinst och därmed framgång.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: louise_bohme@hotmail.com. Handledare: Felix Wahl, Mathias Millberg Lindholm, Kristofer Lindensjö .

# Innehåll


**1 Inledning** **4**
1.1 Syfte & frågeställning . . . . . 4
1.2 Tidigare forskning och undersökningar . . . . . 5

**2 Data** **5**
2.1 Träningsmängd & testmängd . . . . . 6
2.2 Förlängning . . . . . 6
2.3 Hemmaplan/Bortaplan . . . . . 7
2.4 Beskrivning av data . . . . . 7
2.4.1 The Four Factors . . . . . 9

**3 Teori** **10**
3.1 Binär variabel . . . . . 10
3.2 Multipel logistisk regression . . . . . 10
3.2.1 Estimering av koefficienter . . . . . 12
3.3 Wald-test . . . . . 13
3.4 Oddskvot . . . . . 14
3.5 Klassifikationstabell . . . . . 14
3.6 Utvärdering av prediktiv förmåga . . . . . 16
3.6.1 Log Likelihood Loss (Log-loss) . . . . . 16
3.6.2 Klassificeringsförmåga . . . . . 16
3.6.3 ROC-curve & AUC . . . . . 17
3.7 Analysering av variablernas effekt . . . . . 19
3.7.1 Konfidensintervall . . . . . 19
3.7.2 Wald-baserat konfidensintervall . . . . . 20
3.7.3 Likelihood kvot-baserat konfidensintervall . . . . . 20
3.8 Kollinearitet . . . . . 20
3.8.1 Variansinflationsfaktorn (VIF) . . . . . 21
3.9 Akaike's Information Criterion (AIC) . . . . . 21
3.10 Stegvis variabelselektion . . . . . 21

**4 Statistisk analys** **22**
4.1 Metod . . . . . 23
4.1.1 Undersökning av eventuella korrelationssamband . . . . . 23
4.1.2 Grundmodellerna . . . . . 24
4.1.3 Variabelselektion . . . . . 25
4.2 Modellutvärdering & resultat . . . . . 27
4.2.1 Tolkning av resultat . . . . . 32

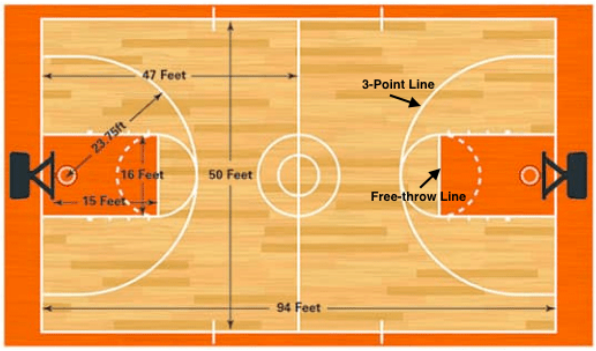
**5 Slutsats** **35**

**6 Diskussion** **36**

**A Bilaga** **41**
A.1 Korrelationsmatris . . . . . 41
A.2 Parvisa plottar . . . . . 42
A.3 Modell 1 - The four factors . . . . . 43
A.4 Oddskvoter och Konfidensintervall . . . . . 44
A.5 Predikterade värden . . . . . 45

# 1 Inledning

Basket är idag en av de mest populära sporterna i världen och uppskattas ha ca 825 miljoner följare runt om i världen [13]. Men störst är kanske sporten i USA, hemlandet till den högst rankade basketligan NBA (National Basketball Association).
Under en match är det 5 spelare från varje lag som möts på planen samtidigt där det lag som lyckas göra flest poäng vinner. Poängsystemet är baserat på vart på planen du gör mål. 2 poäng utdelas då du gör mål inom det så kallade korgområdet, medans distansskott (längre än 6,75m från korgen) ger 3 poäng. Blåser domaren straff utdelas 1,2 eller 3 fria kast från straffkastlinjen där vardera ger 1 poäng. Antalet straffkast som utdelas beror på spelsituationen.
En match är indelad i 4 perioder om 12 minuter (effektiv tid), dvs 48 minuter totalt. Om matchen inte är avgjord inom ordinarie speltid spelas det en förlängning på 5 minuter. Skulle matchen fortfarande inte vara avgjord spelas det ytterligare 5 minuter förlängningstid tills dess att ett lag kan koras som vinnare. Grundserien i NBA består av 30 lag som är uppdelade i två konferenser, den Östra konferensen och den Västra konferensen. Varje lag spelar total 82 matcher, 41 hemmamatcher och 41 bortamatcher, mellan oktober och april där de 8 bästa lagen i varje konferens sedan går vidare till slutspel.




Figur 1: En avspegling av en basketplan där området innanför 3-poängslinjen är det s.k korgområdet som ger 2 poäng per satt skott.

## 1.1 Syfte & frågeställning

Syftet med mitt arbete är att framställa en modell som kan redogöra för vilka faktorer som har störst inverkan på utfallet i en basketmatch och vilka faktorer som tycks spela en mindre roll. Jag utvärderar min modells förmåga genom att undersöka hur väl den kan prediktera utfallet av en match med hjälp av de observerade värdena på de förklarande variablerna. Förhoppningsvis ska detta kunna leda till ett svar på frågan vad som krävs för att ett lag ska nå framgång.

## 1.2 Tidigare forskning och undersökningar

Som en konsekvens av sportens populäritet, samt den noggranna och avancerade statistiken som förs under samtliga NBA matcher, finns det en hel del olika analyser och tidigare studier inom området. Men komplexiteten i sporten ger upphov till en mängd olika åsikter, synvinklar och perspektiv och det verkar inte finnas någon studie som fullt enar världens basketfantaster. Vissa påstår att ett lags defensiva handlingar är de som väger tyngst då ett bra försvar ger upphov till fler egna offensiva anfall (possessions) samt att det inte spelar någon större roll om ett lag är bra på att göra poäng om de inte kan stoppa motståndarna från att göra detsamma. Andra menar att ett lag med starka individuella kvalitéer och skickligheter i det offensiva är att föredra. En av de mest uppmärksammade personerna när det kommer till jakten på att beskriva framgångar inom basket är Dean Oliver, som i sin bok "Basketball on Paper" 2002 redovisade sin filosofi om "the four factors" i ett försök att identifiera hur fyra väsentliga strategier relaterar till framgångsrika resultat. De fyra faktorerna som Oliver bedömer är de viktigaste och mest inflytelserika redovisas nedan [10]

1. Effektiv skottprocent (40%)
2. Turnover rate (25%)
3. Offensiv returtagning (20%)
4. Straffkast rate (15%)

Dessa fyra parametrar syftar på poängeffektivitet, vikten av att ta hand om bollen i anfall, ta så många returer som möjligt samt att komma till straffkastlinjen så ofta som möjligt. Oliver menar på att andra prestationer som ett lag åstadkommer under en match inte är lika förekommande och sker inte lika frekvent som de fyra nyss nämda bedrifterna, och därför är dessa heller inte lika betydande för resultatet. Procentsatserna innanför parenteserna är de viktade värdena som representerar den relativa betydelsen för utfallet i en match för varje faktor (Se avsnitt 2.3.1 för en vidare förklaring av dessa faktorer). Trots att denna filosofi fått stor uppmärksamhet så är det många som är skeptiska kring den relativa fördelningen mellan faktorerna och anser att procentsatserna bör se något annorlunda ut [2].

# 2 Data

I mitt arbete har jag använt mig av data hämtad från NBA:s officiella websida från grundserien för säsongen 2018/2019 [11]. Datan ges i tabellform där varje rad representerar ett enskilds lags prestation under en match och kolumnerna de olika faktorerna som noggrant noteras under varje match.

## 2.1 Träningsmängd & testmängd

För att minska risken för överanpassning i min modell har jag delat in min data i två olika mängder, en träningsmängd och en testmängd. Träningsmängden används i själva framtagandet av min modell då det är denna mängd som min modell kommer anpassas på. Testmängden är den del av datan som min framtagna modell sedan appliceras på för undersöka hur bra modellen tycks förklara utfallet. Det är alltså denna mängd som modellens prediktiva förmåga utvärderas på samt en bedömning av de förklarande variablerna görs. Det ideala är att antalet vinster och förluster i träningsmängden är ekvivalenta, då en skillnad kan komma att hämma modellens prediktiva förmåga. Jag har därför delat upp mitt datamaterial så att träningsmängden består utav 614 stycken observationer, där hälften av matcherna är vinster och hälften är förluster, och detsamma för testmängden på 616 stycken observationer. Observationerna i träningsmängden representerar även de 614 första matcherna på säsongen, och observationerna i testmängden de 616 sista matcherna på säsongen. Detta för att jag tycker att det känns mer logiskt att ta fram en modell som baseras på "historisk data" för att sedan analysera om denna kan förklara de matcher som spelas i "framtiden", än tvärtom eller slumpvis fördelning av observationer.

## 2.2 Förlängning

Som tidigare nämnt så är matcherna i NBA upplagda på sådant vis att en match aldrig kan bli oavgjord, utan förlängningsminuter tillkommer tills dess att ett lag slutligen kan koras som vinnare. Om det spelas förlängning ger ju detta självklart upphov till att mer skott hinner skjutas, fler returer kommer att tas, antalet steals, turnovers och blockar har chans att öka jämfört med de som gjordes under den ordinarie speltiden. De förklarande variablerna beror alltså väldigt starkt på antalet minuter spelade, och resultatet från de matcher som avgjorts på förlängning blir alltså rätt så vilseledande jämt emot de matcher som avgjorts på ordinarie speltid. Det finns ett par tillvägagångssätt man kan tillämpa för kringgå denna "förlängningsproblematik". Det första, och kanske enklaste, alternativet är att utesluta alla matcher som avgjorts på förlängning från datamatrialet. Detta gör dock modellen begränsad till viss del då den därmed, rent teoretisk, inte kan säga någonting om matcher som inte avslutats på ordinarie speltid. En begränsad modell vill man i regel helst undivka så långt det går. Av alla NBA-matcher som spelades under säsongen 2018-2019 var det dock endast 5% av matcherna som gick till förlängning, så hade jag valt att följa denna metod hade min modell troligtvis inte påverkats särskillt mycket ändå.
Det andra alternativet är att anpassa datamaterialet efter antalet spelade matchminuter genom att dividera alla parametrar med antalet spelade matchminuter. På så sätt blir inte modellen begränsad på samma sätt som i alternativet innan. Att anpassa datamaterialet efter spelade matchminuter ger upphov till värden på en annan "skala" än tidigare, då parametrarna nu är representerade per spelad minut och inte per match, vilket kan göra att resultatet kan uppfattas som något mer svårtolkat. Detta var något jag själv upplevde när jag provade att

tillämpa denna metod. För att få tillbaka mina värden på "samma skala" som tidigare dividerade jag mina parametrar en gång till, denna gång med 48 som är antalet spelade minuter under ordinarie matchtid. På så sätt får jag mina parametrar givna per match igen utan att ha begränsat modellen.

## 2.3 Hemmaplan/Bortaplan

Av alla matcher som slutade i vinst under säsongen 2018-2019 visade sig 66.3% av matcherna vara spelade på det vinnande lagets hemmaplan.
På liknande vis som med matcher spelade på övertid, så är det även här svårt att utesluta multikollineart beteende bland variablerna (*multikollinearitet* beskrivs närmare i avsnitt 3.8) och det faktum att resterande förklarande variabler skulle kunna bero på om matchen är spelad på hemmaplan eller inte. Man kan nämligen lätt tänka sig att ett lag på sin hemmaplan får ett psykisk övertag jämt emot motståndarna då det oftast möter ett större publikstöd, och att de där med har en tendens att prestera bättre. Jag har därför valt att analysera matchresultat (det vill säga vinst eller förlust) utifrån hemmalagets perspektiv. I det här arbetet anser jag dock att båda lagens prestationer under matchens gång är av intresse för att kunna förklara matchresultatet. Jag har därför valt att, för varje rad i min hämtade data från NBAs officiella hemsida, addera motståndarnas (det vill säga bortalagets) värden för motsvarande parametrar för att få ett sånt bra helhets intryck av matchbilden som möjligt.

Efter denna anpassning är det viktigt att poängtera att den modell som konstrueras i det här arbetet nu kommer beskriva utfallet **för hemmalaget** i en basketmatch.

## 2.4 Beskrivning av data

Under en match förs det noggrann och avancerad matchstatistik, både på lagnivå och individuell nivå, där jag i det här arbetet valt att fokusera på den statistik som sammanfattar *lagens prestation*. I tabell 1 redovisas all den matchdata som registreras *per lag* under en match.
Jag kommer dock inte använda mig utav alla de redovisade faktorerna i tabell 1 vid framtagandet av mina modeller, utan enbart de mest relevanta för mitt syfte. Skottstatistiken för de tre olika poänggivande distanserna kommer jag låta sammanfattas av andelen satta skott per distans. Lagens tagna returer kommer representeras av två variabler, offensiva returer och defensiva returerna. Variabeln för det totala antalet tagna returer kommer jag bortse ifrån helt då denna beskrivs av de två andra nyss nämna parametrarna. Jag kommer även bortse från lagens totala antal gjorda poäng, detta eftersom att denna variabel uppenbarligen avslöjar slutresultatet av matchen, vilket är just det jag vill att min modell ska kunna beskriva utifrån resterande faktorer.
Med hjälp utav statistiken i tabell 1 kan man beräkna Dean Olivers "the four factors". Dessa presenteras i tabell 2, och hur dessa variabler beräknas redovisas i avsnitt 2.3.1.

| Matchdata avseenede ett lag under en match | | |
|---|---|---|
| Variabelnamn | Översättning | Förklaring |
| TM | Team | Lag |
| PTS | Points | Gjorda poäng |
| FGM | Field goals made | Antal satta skott inom korgområdet |
| FGA | Field Goals Attempted | Antal skottförsök inom korgområdet |
| FG% | Field Goal % | Andelen satta skott inom korgområdet |
| 3PM | 3 Point Field Goals Made | Antal satta 3-poängsskott |
| 3PA | 3 Point Field Goals Attempted | Antal 3-poängs skottförsök |
| 3P% | 3 Point Field Goal % | Andel satta 3-poängsskott |
| FTM | Free Throws Made | Antal satta straffkast |
| FTA | Free Throws Attempted | Antal skjutna straffkast |
| FT% | Free Throws % | Andel satta straffkast |
| OREB | Offensive Rebounds | Antal offensiva returer |
| DREB | Defensive Rebounds | Antal deffensiva returer |
| REB | Rebounds | Totala antalet returer |
| AST | Assists | Antal målgivande passningar |
| TOV | Turnovers | Antal avbrutna bollinnehav |
| STL | Steals | Antal snodda bollinnehav (från motståndarna) |
| BLK | Blocks | Antal blockade skottförsök på motståndarna |
| PF | Personal Fouls | Totala antalet gjorde fouls |

Tabell 1: En överblick över den matchdata som registreras per lag efter varje spelad match. De gråmarkerade raderna representerar de mest relevanta variablerna för det här arbetet. Värden från både hemmalag och bortalag för samtliga relevanta variabler används vid modellframtagandet. Det är även underförstått att dessa variabler är normaliserade per spelade match på ordinarie speltid som beskrivs i avsnitt 2.2.

| The Four Factors | | |
|---|---|---|
| Variabelnamn | Översättning | Förklaring |
| eFG% | Effective Field Goal % | Effektiv skottprocent |
| TOV% | Turnover Rate | Andel offensiva anfall utan ett skottförsök |
| OREB% | Offensive Rebounding % | Andel returer på offensiv planhalva jämtemot motståndarlaget |
| FT-rate | Free Throw Rate | Andelen gånger ett offensivt anfall leder till straffkast |

Tabell 2: *The Four Factors* som Dean Oliver anser är de variabler som avgör utfallet i en match.

### 2.4.1 The Four Factors

Här nedan följer en kort beskrivning och motivering av "the four factors", samt hur dessa parametrar beräknas. [10]

- **Effective Field Goal Percentage**

$$eFG\% = \frac{FGM + 0,5 * 3PM}{FGA} \tag{1}$$

  eFG% är ett korrigerat mått för att identifiera ett lags effektiva skottprocent på distans, vilket innefattar alla skott som sker på planen förutom sttaffkast. Detta är alltså ett gemensamt mått för ett lags distansskytte, och till skillnad från FG% tar eFG% även hänsyn till 3-poängsskyttet och det faktum att dessa skott är värda 50% mer än ett på närmare distans.

- **Turnover Rate**

$$TOV\% = \frac{TOV}{FGA + 0,44 * FTA + TOV} \tag{2}$$

  Ju fler offensiva anfall ett lag har jämfört med sin motståndare, desto fler möjligheter att komma till avslut som kan leda till poäng har laget jämför med sin motståndare. Därför anser Oliver att det är av yttersta vikt att minimera antalet gånger ett offensivt anfall avbryts utan att ett skottförsök har gjorts. Detta mäts av Turnover rate.

- **Offensive Rebound Percentage**

$$OREB\% = \frac{OREB}{OREB + MOT.DREB} \tag{3}$$

  Offensive rebounding percentage beräknar andelen returer tagna på den offensiv planhalva jämtemot motståndarlaget. Att ta offensiva returer är också ett sätt att få fler offensiva anfall. Många offensiva returer som tas är dessutom i en position relativt nära korgen där skottprocenten är väldigt hög och där man har en stor chans att bli "foulad" för att på så sätt få komma till straffkastlinjen. Variabeln *MOT.DREB* syftar på antalet tagna defensiva returer av motståndarlaget.

- **Free Throw Rate**

$$FTrate = \frac{FTM}{FGA} \tag{4}$$

  Free throw rate mäter hur ofta ett lag tar sig till straffkastlinjen och hur många gånger de sätter ett straffkast. Straffkast har den högsta skottprocenten av alla skott som en spelare kan få och kallas ibland för "gratispoäng" då spelaren får skjuta ostört från ett väldigt bekvämt avstånd.

# 3 Teori

I det här avsnittet kommer jag att redovisa teorin bakom de många begrepp och koncept som ligger till grund för mitt arbete. I stora delar, och kanske främst avsnitt 3.2 om multipel logistisk regression, utgår jag ifrån litteratur av Alan Agresti "An Introduction to Categorical Data Analysis" [1]. Då andra källor används så hänvisas det till i texten.

## 3.1 Binär variabel

Mitt mål är, som tidigare nämnt, att ta fram en modell vars avsikt är att säga något om resultatet i en basketmatch och vilka faktorer det är som påverkar det. Med tanke på att vi vet att en NBA-match aldrig slutar oavgjort, kändes informationen om en match har slutat i vinst eller förlust för hemmalaget som det mest naturliga valet av respons.
Variablen vinst/förlust i vårt dataset är en kategorisk variabel som dessutom är binär. En variabel kallas binär då utfallet enbart antar två värden som kan definieras av två klasser. Antingen så sker en händelse (klass 1) eller så sker den inte (klass 0). I vårt fall så är utfallen alltså att man antingen vinner, eller så gör man det inte (förlust).

När en responsvariabel är binär så är relationen mellan de förklarande variabelerna och responsen inte nödvändigtvis linjär och feltermerna kan inte förväntas vara normalfördelade, vilket innebär att linjär regression inte är en lämplig metod. Istället kan logistisk regression användas för att få fram en modell som beskriver den valda responsen. Den logistiska regressionsmodellen skattar sannolikheten att vår responsvariabel tillhör klass 1 (det vill säga att en händelse inträffar) givet ett antal förklarande variabler. Precis som i linjär regression är de förklarande variablerna en linjärkombination av variabler som kan vara både kategoriska och kontinuerliga [15].

## 3.2 Multipel logistisk regression

En logistisk regressionsmodell är ett specialfall av en så kallad generaliserad linjär modell. En generaliserad linjär modell brukar man säga är uppbyggd utav 3 komponenter; en komponent som syftar på responsvariabeln $Y$ och dess fördelning, de "systematiska komponenterna" som syftar på den linjära kombinationen av förklarande variabler i modellen samt en "länk-funktion" som är en funktion $g(.)$ som relaterar väntevärdet av vår responsvariabel, $\mu$, till de systematiska komponenterna.

$$g(\mu) = \alpha + \beta_1 x_1 + ... + \beta_k x_k \tag{5}$$

Ekvation 5 speglar en generlaiserad linjär modell där $g(\mu)$ är den så kallade länk-funktionen, $\alpha$ interceptet, $X = (x_1, ..., x_k)$ är de förklarande variablerna

och $\beta = (\beta_1, ..., \beta_k)$ dess koefficienter.

En binär responsvariabel $Y$ kan vi anta följa en Bernoulli-fördelning där sannolikheten för att en händelse inträffar är $P(Y = 1|X = x) = p(x)$ och sannolikheten att händelsen inte inträffar är $P(Y = 0|X = x) = 1 - p(x)$. Väntevärdet av en Bernoulli-fördelad variabel vet vi i sin tur definieras som $E(Y = 1|X = x) = p(x)$.

Skulle vi naivt välja att modellera en binär respons med linjär regression hade vi kunnat ställa upp följande ekvation

$$p(x) = \alpha + \beta_1 x_1 + ... + \beta_k x_k \tag{6}$$

Det finns dock en hel del brister kring den ovanstående ekvationen som gör att denna inte kan betraktas som en optimal modell för en respons av det binära slaget.

Dels så är det viktigt att poängtera att relationen mellan $p(x)$ och $x$ oftast inte är linjär för en binär respons, utan dess samspel brukar beskrivas av en S-formad kurva där $p(x)$ förändras kontinuerligt i $x$ (se figur 2).

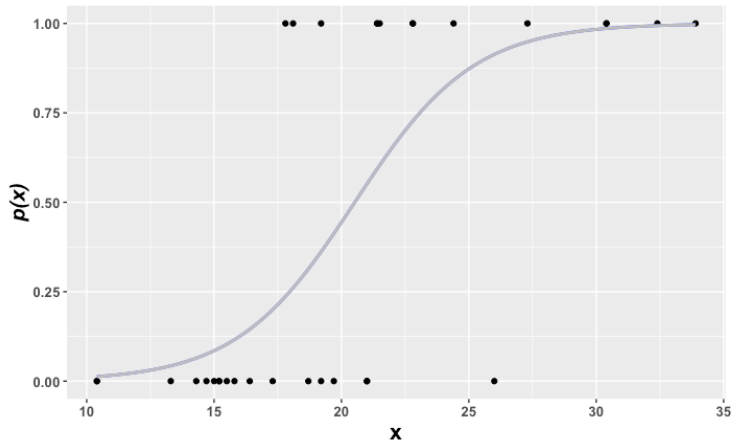
Vi ser dessutom att $VL$ och $HL$ är begränsade och bundna inom olika intervall vilket även det gör ekvation 6 problematisk. Vi vet att $p(x) \in [0, 1]$ medans $\beta x_i \in [-\infty, \infty]$.

För att komma runt denna problemtik är det därför ytterst tacksamt att tillämpa odds och log-odds inom logistisk regression. Odds är definierat som $\frac{p(x)}{1-p(x)}$ och beskriver hur troligt det är att en händelse ska inträffa. Log-odds är helt enkelt den naturliga logaritmen av odds. Genom att nyttja log-odds som länk-funktion i den logistiska regressionsmodellen kan ett linjärt förhållande mellan denna och de förklarande variablerna $x_i$ antas [15]. Dessutom så är $\log(\frac{p(x)}{1-p(x)}) \in [-\infty, \infty]$ vilket eliminerar skillnaderna i begränsningsantagandena.
En logistisk regressionsmodell kan därför följaktligen beskrivas av

$$g(E(Y)) = \log(\frac{p(x)}{1 - p(x)}) = \alpha + \beta_1 x_1 + ... + \beta_k x_k. \tag{7}$$

Löser vi ut $p(x)$ ur ekvation 7 erhåller vi den så kallade *logistiska funktionen* som genererar ett värde mellan 0 och 1 för den skattade sannolikheten att en händelse inträffar

$$p(x) = P(Y = 1|X = x) = \frac{\exp(\alpha + \beta_1 x_1 + ... + \beta_k x_k.)}{1 + \exp(\alpha + \beta_1 x_1 + ... + \beta_k x_k.)} \tag{8}$$




Figur 2: En logistisk regressionsmodell anpassad på ett dataset. De svarta punkterna representerar de binära utfallen (0 eller 1) för den givna datan, och kurvan den logistiska regressionsmodellens skattade sannolikheter för att utfall 1 ska ske. På y-axeln har vi sannolikheten för utfall 1 givet en förklarande variabel x, och på x-axeln en förklarande variabel x

### 3.2.1 Estimering av koefficienter

Interceptet $\alpha$ och de förklarande variablernas koefficienter $\beta = (\beta_1 ... \beta_k)$ är okända parametrar som vi, utifrån vår träningsmängd, skattar fram med hjälp utav *maximum likelihood metoden*. Härledningen av denna metod redovisas nedan och presenteras i Hosmer, Lemeshow och Sturdivants bok "Applied Logistic Regression" [7] .
För en binär variabel som är *Bernoulli fördelad* ges dess likelihood funktion av

$$L(\alpha, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \tag{9}$$

där $i = 1, 2...n$ motsvarar antalet observationer i träningsmängden och $y_i$ antar värdena 1 eller 0. Maximum likelihood metoden går ut på att hitta det $\alpha$ och $\beta$ som maximerar likelihooden i ekvation 9. För att underlätta de matematiska beräkningarna logaritmerar man oftast det ovanstående uttrycket enligt ekvation 10 som kallas för *log-likelihooden*

$$l(\alpha, \beta) = ln[L(\alpha, \beta)] = \sum_{i=1}^{n} y_i ln[p(x_i)] + (1 - y_i) ln[(1 - p(x_i))] \tag{10}$$

Därefter deriverar vi log-likelihooden med avseende på $\alpha$ och $\beta$ och sätter uttrycken lika med noll. Genom att lösa ekvationerna erhåller vi de estimerade parametrarna som maximerar likeliehood funktionen. De estimerade parametrarna tillsammans med ekvation 8 används sedan för att prediktera vår responsvariabel $Y$.

## 3.3 Wald-test

Ett wald-test kan användas för att utvärdera den statistiska signifikansen för varje estimerad koefficient i modellen. Tanken med testet är att pröva hypotesen att koefficienten för en oberoende variabel i modellen är lika med noll, $H_0 : \beta_0 = 0$, mot hypotesen att koefficienten skiljer sig från noll, $H_1 : \beta_0 \neq 0$. Om testet misslyckas med att avvisa nollhypotesen tyder detta på att en eliminering av variabeln inte skulle påverka modellen på något väsentligt sätt. I det fallet anses därför inte variabel som nödvändig bland modellens förklarande variabler.
När $H_0 : \beta = 0$ är sann så är, enligt Agresti [1], den ML estimerade koefficienten dividerat med dess standardfel approximativt standard normalfördelad (se ekvation 11). Ett test som använder denna typ utav test-statistika kallas för ett Z-test.

$$z = \frac{\hat{\beta} - \beta_0}{se(\hat{\beta})} = \frac{\hat{\beta}}{se(\hat{\beta})} \sim N(0,1) \tag{11}$$

I ett wald-test används istället den så kallade Wald-statistikan som är starkt kopplad till ovanstående ekvationen. Wald-statistikan beräknas nämligen genom att z-värdet kvadreras. Under antagandet att nollypotesen är sann följer denna statistika således en $\chi^2(1)$ fördelning då detta är resultatet av en standard normalfördelning i kvadrat, se ekvation 12.

$$W = z^2 = \frac{(\hat{\beta})^2}{Var(\hat{\beta})} \sim \chi^2(1) \tag{12}$$

Med hjälp utav Wald-statistikan kan sedan ett p-värde beräknas som avslöjar koefficientens signifikans, vilket fastställer om nollhypotesen ska förskastas eller ej.

Intutionen bakom Wald-testet (och wald-statistikan) är att undersöka hur stor skillnaden är mellan den ML estimerade koefficient $\hat{\beta}$ och koefficienten under nollhypotesen $\beta_0$. Är skillnaden stor resulterar det i att värdet på statistikan kan bli högt (detta eftersom att täljaren kvadreras). En wald-statistika som antar ett högt värde långt ifrån noll är därmed troligtvis inte $\chi^2$-fördelad då denna fördelning, med 1 frihetsgrad, är lika med noll eller väldigt nära noll. I ett sådant scenario kan vi därför förkasta att nollhypotesen skulle vara sann. Om den estimerade koefficienten skulle ha stor varians så skulle dock en koefficient under nollhypotesen kunna ge liknande likelihood som den ML estimerade koefficienten, trots att de skiljer sig åt. Därav $Var(\hat{\beta})$ i nämnaren.

Eftersom att Wald-statistikan är så starkt förankrad till ekvation 11 brukar man i generella drag uttrycka denna likt z-statistikan och beräkna p-värdet baserat på en standard normalfördelning. De z-värden och 2-sidiga p-värden som kommer presenteras i resultatdelen av det här arbetet kan därför kopplas till ett wald-test likt det som redovisas i det här avsnittet. [7]
Viktigt att poängtera är att det finns flera olika test att tillämpa för att undersöka signifikansen hos en modells koefficienter, och en vanlig motsvarighet

till wald-test i logistisk regression är likelihood kvot-testet. Detta test kommer jag beskriva kort i avsnitt 3.7.3 .

## 3.4 Oddskvot

I ekvation 7 representerar varje koefficient $\beta_i$ den förväntade skillnaden i log-odds om $x_i = k$ ökar med en enhet till $x_i = k+1$ när resterande förklarande variabler hålls konstanta. Dock så är en förändring i log-odds ett ganska svårtolkat mått. För en lättare association av koefficienternas betydelse används istället ofta *oddskvot* som mått. Oddskvoten är som det låter kvoten mellan två odds och förklarar hur mycket troligare eller otroligare (i termer av odds) det är att en händelse ska inträffa när $x_i = k$ ökar med en enhet till $x_i = k + 1$. Hosmer, Lemeshow och Sturdivant definierar i sin bok "Applied Logistisc Regression" [7] ett odds av att en händelse inträffar enligt ekvationen nedan

$$\frac{p(x)}{1 - p(x)} \tag{13}$$

Om vi antar att $\beta$ är skillnaden i log-odds när dess tillhörande förklarande variabel ökar med 1, så är relationen mellan samma regressionskoefficient och dess oddskvot $\exp(\beta)$, vilket härleds nedan enligt [7]

$$\beta = \log(\frac{p(x = k + 1)}{1 - p(x = k + 1)}) - \log(\frac{p(x = k)}{1 - p(x = k)}) \tag{14}$$

$$\begin{aligned} \exp(\beta) &= \exp(\log(\frac{p(x = k + 1)}{1 - p(x = k + 1)}) - \log(\frac{p(x = k)}{1 - p(x = k)})) \\ &= \frac{\exp(\log(\frac{p(x=k+1)}{1-p(x=k+1)}))}{\exp(\log(\frac{p(x=k)}{1-p(x=k)}))} \\ &= \frac{\frac{p(x=k+1)}{1-p(x=k+1)}}{\frac{p(x=k)}{1-p(x=k)}} \end{aligned} \tag{15}$$

När vi ökar $x_i$ med 1 så ökar alltså oddsen, $\frac{p(x)}{1-p(x)}$, med faktorn $\exp(\beta_i)$.

För en tydligare förståelse kan vi återkoppla nollhypotesen från avsnitt 3.3. Om nollhypotesen är sann, det vill säga $\beta_i = 0$, ger det oddskvoten $\exp(\beta_i) = 1$. Oddsen multipliceras då med 1 och förhållandet förblir oförändrat. Detta betyder att $x_i$ inte har något statistiskt samband med $Y$, vilket är samma slutsats som beskrivs i avsnitt 3.3 när nollhypotesen inte förkastas.

## 3.5 Klassifikationstabell

Precis som vid de flesta modellframtagningar är det av stort intresse att utvärdera en logistisk regressionsmodells prestanda och pricksäkerhet i att faktiskt kunna utföra korrekta uppskattningar om sanna observerade utfall. För att kunna

bedömma en modells "duglighet" vill vi ofta kunna svara på frågor som "*när det observerade värdet är ja, hur ofta lyckas vår modell prediktera ett ja?*" och "*när det observerade värdet är nej, hur ofta lyckas vår modell prediktera ett nej?*". Dessa frågeställningar svarar för modellens *Sensitivitet* respektive *Specificitet*.

För en binär variabel som endast antar två värden finns det två möjliga felaktigheter som kan uppstå vid prediktion. Om vi låter 0 definiera ett positivt utfall och 1 definiera ett negativt utfall så kallas den första felaktigheten, Typ I fel, för "false positive (FP)" och uppstår då ett utfall är predikterat att tillhöra klass 0 men egentligen tillhör klass 1. Den andra felaktigheten, Typ II fel, kallas för "false negative (FN)" och uppstår då ett utfall är predikterat att tillhöra klass 1 men egentligen tillhör klass 0. Detta kan illustreras i en klassifikationstabell likt den i figur 3. En klassifikationstabell illustrerar inte bara de felaktigt predikterade utfallen, utan även de utfall som faktistkt blivit korrekt predikterade. Dessa observationer kallas " true positiv (TP)" och " true negativ (TN)" [8]. Med hjälp utav värdena i klassifikationstabellen kan vi nu beräkna modellens

| | | True Condition | |
|---|---|---|---|
| | | 0 | 1 |
| Predicted | 0 | TP | FP |
| Condition | 1 | FN | TN |
| | **Total** | P | N |

Figur 3: Klassifikationstabell

Sensitivitet och Specificitet som:

$$\text{Sensitivitett} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{Specificitet} = \frac{TN}{TN + FP} \tag{17}$$

Andra mått som är vanligt förekommande vid bedömningen en modells duglighet är *true positive rate (TPR)* och *false positive rate(FPR)*. Om vi predikterat ett utfall att vara positivt så svarar TPR mot sannolikheten att denna upskattning är korrekt och FPR mot sannolikheten att den är felaktig. True positive rate är förknippad med modellens sensitivitet och False positive rate är relaterad till modellen specificitet enligt ekvationerna nedan [7]:

$$\text{True positive rate} = Sensitivitet = \frac{TP}{TP + FN} \tag{18}$$

$$\text{False positive rate} = 1 - Specificitet = \frac{FP}{FP + TN} \tag{19}$$

## 3.6 Utvärdering av prediktiv förmåga

För att kunna avgöra om de förklarande variablerna i mina framtagna modeller faktiskt kan användas för att beskriva utfallet i en basketmatch väljer jag att undersöka modellernas prediktiva förmåga. Det finns flera olika tillvägagångssätt och metoder ämnade för det här syftet, och här nedan presenteras de metoder som jag har valt att tillämpa i min analys.

### 3.6.1 Log Likelihood Loss (Log-loss)

Uttrycket för en så kallad loss-funktion kan se olika ut beroende på vilken situation den är ämnad för, men gemensamt är dess syfte att utvärdera prestandan hos en algoritm. I vårt fall vill vi hitta en loss-funktion som kan beskriva hur nära vårt predikterade utfall $\hat{y}$ är det sanna observerade värdet $y$. Detta kan bli uppfyllt genom att minimera den så kallade binära "Cross-entropy"-funktionen som är definierad enligt ekvation 20 [4]. För en binär variabel kallas även denna funktion för "Log Likelihood Loss"-funktionen (Log-loss"), detta eftersom att uttrycket nedan i stortsätt minimerar den log likelihood funktion för en binär variabel som är definierad i ekvation 10.

$$H = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i))) \tag{20}$$

I ekvation 20 utgör $y_i$ de sanna observationerna och $p(y_i)$ är den predikterade sannolikheten att $y_i = 1$.
Daniel Godoy redovisar i sin artikel för "towardsdatascience.com" [5] att Log-loss funktionen hjälper oss att avgöra hur bra/dålig vår modell är på att prediktera sannolikheter genom att returnera höga värden för dåliga prediktioner och låga värden för bra prediktioner. Det optimala är alltså att få ett log-loss värde lika med noll som menar på att våra predikterade värden inte avviker alls från de sanna observerade värdena och förlusten där emellan är alltså noll. Omvänt gäller det om vi får ett högt log-loss värde, då är förlusten mellan våra predikterade värden och de sanna värdena stort.

### 3.6.2 Klassificeringsförmåga

Som redan visat så är det väldigt tacksamt att sammanställa en modells prediktiva kapacitet enligt klassificationstabellen i figur 3 för att skapa sig en tydlig bild över hur bra en modell faktiskt kan antas vara och inte.
Utöver att beräkna en modells sensitivitet och specificitet kan man enkelt använda klassificationstabellen till att beräkna proportionen av felaktigt skattade händelser från modellen, det så kallade *felklassificeringsvärdet*, som självklart önskas vara så lågt som möjligt.

$$Felklassificeringsvärdet = P(\text{Felaktig prediktion}) = \frac{FP + FN}{FP + FN + TP + TN} \tag{21}$$

Även den totala andelen korrekt skattade händelser kan såklart beräknas. Detta är det mått jag valt att använda vid mina bedömningar i det här arbetet och kommer benämnas som modellens *klassificeringsförmåga* i den fortsatta analysen. Agresti definierar klassificeringsförmågan enligt ekvation 23. [1]

$$Klassificeringsförmåga = P(\text{Korrekt prediktion}) = \frac{TP + TN}{FP + FN + TP + TN} \quad (22)$$

Man kan såklart beräkna modellens felklassificeringsvärde och klassificeringsförmåga utan att nödvändigtvis behöva ställa upp en klassifikationstabell, detta är något som James, Witten, Hastie och Tibshirani redovisar i sin bok "An introduction to statistical learning" [8].
För utfallen $y_i$ där $i = 1 \ldots N$, definieras felklassificeringsvärdet som

$$\frac{1}{N} \sum_{i=1}^{N} I(y_i \neq \hat{y}_i) \quad (23)$$

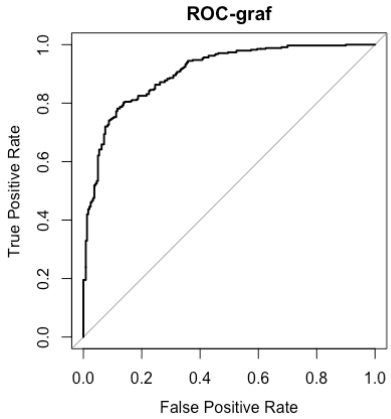
där $\hat{y}_i$ är den skattade sannolikheten för vinst och $I(y_i \neq \hat{y}_i)$ en indikatorvariabel som returnerar 1 om händelsen är felaktigt skattad och 0 om händelsen är korrekt skattad [8]. Modellens klassificeringsförmåga beräknas sedan som

$$\text{Klassificeringsförmåga} = 1 - \text{Felklassificeringsvärdet} \quad (24)$$

### 3.6.3 ROC-curve & AUC

*Receiver Operator Characteristic (ROC)-kurvan* är också ett vanligt förekommande hjälpmedel vid utvärdering av en statistisk modells noggrannhet, där informationen från klassifikationstabellen i avsnitt 3.5, baserat på olika tröskelvärden, används. Ett tröskelvärde är den gräns som avgör över vilken estimerad sannolikhet vi predikterar att händelsen har inträffat och under vilken vi predikterar att händelsen inte har inträffat. Om tröskelvärdet är satt till 0.5 så skulle alla sannolikheter $\geq 0.5$ i vårt fall räknas som en vinst och alla sannolikheter $< 0.5$ som en förlust. ROC-kurvan presenteras i en tvådimensionell graf där modellens *sensitivitet*, True Positive rate, plottas mot *1-specificiteten*, False positive rate, för alla tröskelvärden. Ett exempel på en ROC-graf illustreras i figur 4. Punkten (1,1) representerar det fall där tröskelvärdet är satt så pass lågt att alla utfall kommer predikteras som positiva, motsatt scenario är i punkten (0,0) där tröskelvärdet är så pass högt att inga utfall kommer predikteras som positiva. Ett högt tröskelvärde ger hög specificitet men låg sensitivitet och det motsatta för ett lågt tröskelvärde. Den optimala punkten i grafen är (0,1) där vår modell predikterar alla utfall korrekt. Den sträckade diagonala linjen i grafen visar vart True positive rate är lika med False positive rate . För eventuella punkter på denna linje innebär det att andelen korrekta prediktioner är densamma som andelen felaktiga prediktioner.
Agresti [1] hävdar att en ROC-graf kan anses vara mer informativ än en klassificationstabell av den anledningen att den sammanfattar modellens prediktiva förmåga för alla möjliga tröskelvärden. Men för att kunna använda ROC-kurvan

som ett mått för en modells noggranhet behöver man beräkna *arean under ROC-kurvan (AUC)*. Värdet på AUC ligger mellan 0 och 1 och syftar alltså på den area som uppstår under ROC-kurvan. Ett värde nära 1 indikerar på att modellen har mycket god precision i sina prediktioner och det motsatta för ett värde nära 0. Om AUC är lika med 0.5 betyder det att ROC-kurvan sammanfaller med den streckade diagonala linjen i figur 4. Detta är som tidigare nämnt inte ett önskevärt scenario då det betyder att vår modell inte har någon direkt prediktionskapacitet alls eftersom att andelen korrekta prediktioner är densamma som andelen felaktiga prediktioner. Man brukar därför tala om att ett AUC-värde mindre eller lika med 0.5 tyder på en modell med dålig precisionsförmåga. [8]




Figur 4: Ett illustrerat exempel på en ROC-kurva där Sensitiviteten (True positive rate) är plottat mot 1-Specificiteten (False positive rate) för alla möjliga tröskelvärden. Arean som uppstår under kurvan är den s.k AUC.

## 3.7 Analysering av variablernas effekt

Utöver att ta fram en modell med avsikt att kunna skatta sannolikheten att ett hemmalag vinner, är det största syftet med arbetet att utvärdera de enskilda förklarande variablernas effekt på utfallet för att på så sätt kunna säga något om vilka faktorer som är av större vikt och vilka som spelar en mindre roll för resultatet. Som nämnt i 3.3 är oddskvoten ett användbart verktyg i den här frågan då denna förklarar hur oddsen av en vinst för hemmalaget påverkas av en ökning i de enskilda variablerna. Jag kommer därför att beräkna oddskvoten samt ett tillhörande konfidensintervall för varje variabel i den slutliga modellen för att granska deras effekter. Denna metod beskrivs i både "An Introduction to Categorical Data Analysis" av Alan Agresti [1] och i "Applied Logistic Regression" av Hosmer, Lemeshow och Sturdivant [7], vilket är den litteratur som jag valt att utgå ifrån i följande avsnitt.

### 3.7.1 Konfidensintervall

Inledningsvis vill jag bara börja med att kort förklara vad ett konfidensintervall är och vad syftet med dessa är. För att hantera den osäkerhet som kan uppstå när vi skattar fram en parameter kan ett konfidensintervall skapas, som innebär att om vi skulle upprepa vår skattning av samma parameter flera gånger så skulle intervallet *med viss säkerhet* kunna antas täcka in det sanna värdet. Säkerheten kallas för *konfidensgrad* där de vanligaste valen av konfidensgrad brukar vara 90, 95 eller 99%. Jag kommer konstruera 95% konfidensintervall i den här undersökningen.
Tanken i det här arbetet är att konstruera konfidensintervall över de förklarande variablernas skattade koefficienterna. Vi får sedan enligt Agresti ett konfidensintervall för den multiplikativa effekten på oddsen, det vill säga oddskvoten, genom att exponentiera ändpunkterna i konfidensintervallen. Ändpunkterna i ett tvåsidigt $100(1-\alpha)\%$ konfidensintervall för koefficienten $\beta$ i den logistiska regressionsmodellen vet vi kan beskrivas som

$$\hat{\beta} \pm z_{1-\frac{\alpha}{2}} se(\hat{\beta}) \tag{25}$$

där $\alpha$ är den valda signifikansnivån, $z_{1-\frac{\alpha}{2}}$ normalkvantilen för den valda signifikansnivå och $se(\hat{\beta})$ standardfelet för den estimerade koefficienten. Ett konfidensintervall för oddskvoten $exp(\beta)$ och dess ändpunkter kan därmed beskrivas enligt

$$[\exp(\hat{\beta} - z_{1-\frac{\alpha}{2}} se(\hat{\beta})), \exp(\hat{\beta} + z_{1-\frac{\alpha}{2}} se(\hat{\beta}))] \tag{26}$$

Om intervallet $[1.2, 1.5]$ definierar ett konfidensintervall över en koefficients oddskvot så betyder det att oddsen för en vinst ökar som minst med 20% och som mest med 50% när den förklarande variabeln ökar med 1. Det finns olika metoder att tillämpa för att utforma ett konfidensintervall och nedan presenterar jag kort de två kanske vanligaste konfidensintervallen för just logistisk regression.

### 3.7.2 Wald-baserat konfidensintervall

Hosmer, Lemeshow och Sturdivant skriver att konfidensintervall för de skattade koefficienterna oftast baseras utifrån deras respektive wald-test, vilket är ett test jag beskriver i avsnitt 3.3. Ett sådant konfidensintervall kallas således för ett Wald-baserat konfidensintervall (eller bara Wald-konfidensintervall). Här utgår man ifrån att Wald-statistikan följer en standard normalfördelning och $z_{1-\frac{\alpha}{2}}$ i uttrycket ovan är lika med kvantilen för en sådan fördelning givet signifikansnivån $\alpha$.

### 3.7.3 Likelihood kvot-baserat konfidensintervall

En annan metod är att konstuera konfidensintervall baserat på ett likelihood kvot-test istället. Detta hypotestest går egentligen ut på att jämföra likelihooden av två modeller genom att beräkna kvoten mellan dem, för att på så sätt avgöra om koefficienten för en viss variabel tycks vara statistisk signifikant för utfallet eller ej. Test-statistikan i det här fallet kan enligt Agresti beskrivas som uttrycket nedan

$$-2log(\frac{L_0}{L_1}) = -2(l_0 - l_1) \sim \chi^2(1) \tag{27}$$

där $l_0$ är den maximerade log-likelihood funktionen under nollhypotesen och $l_1$ den maximerade log-likelihood funktionen över den modell som inkluderar den koefficient som man vill testa signifikansen på. Till skillnad från wald-statistikan så kommer här $z_{1-\frac{\alpha}{2}}$ i uttrycket ovan vara kvantilen för en $\chi^2(1)$ fördelning.

Jag kommer i det här arbetet utgå från wald-baserde konfidensintervall i mina analyser. Vilken metod som anses vara den bästa kan dock diskuteras, vilket jag pratar mer om i diskussionen i avsnitt 6.

## 3.8 Kollinearitet

Kollinearitet uppstår när de oberoende variablerna korrelerar i stor grad med varandra, vilket är ett vanligt fenomen när man handskas med flera förklarande variabler. Problemet som uppstår vid detta fenomen är att det kan bli svårt att avgöra vilken av de förklarande variablerna som faktiskt påverkar vår valda responsvariabel. När variabler korrelerar starkt med varandra uppstår det variation som delas mellan flera variabler, och eftersom att regressionskoefficienter oftast skattas med hjälp av en variabels unika variation, leder detta till att en del variation går till spillo och våra koefficientskattningar blir mindre precisa (standardfelen ökar)[12]. För att kontrollera eventuell kollinearitet mellan variabler kan man bland annat beräkna korrelationen mellan dem med hjälp utav en korrelationsmatris samt plotta variablerna mot varandra. Tyvärr kan inte all kollinearitetsproblematik upptäckas genom dessa åtaganden. Det är nämligen möjligt för kollinearitet att existera mellan tre eller flera variabler, även fast inga variabelpar visar på särskilt stark korrelation. En sådan situation kallas för multikollinearitet. I stället för att inspektera korrelationsmatrisen, är det bättre

att bedöma multikollinearitet genom att beräkna variansinflationsfaktorn (VIF) [8].

### 3.8.1 Variansinflationsfaktorn (VIF)

VIF-faktorn beräknas enligt James, Witten, Hastie och Tibshirani i boken "An introduction to statistical learning" som

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \tag{28}$$

där $R^2_{X_j|X_{-j}}$ står för den förklaringsgrad som anger hur mycket av variationen i variabeln $x_j$ som förklaras utav de andra variablerna. Om $R^2_{X_j|X_{-j}}$ är nära 1 förekommer det mulltikollinearitet och VIF-faktorn kommer vara stor. Det minsta möjliga värdet på VIF är 1, som indikerar att det inte existerar någon som helst mulltikollinearitet. I praktiken förekommer det dock oftast en liten mängd av kollinearitet mellan variablerna av naturliga skäl, därför brukar man i regel säga att ett VIF-värde större än 10 är problematiskt [8].

Utöver de åtaganden som är nämda ovan, skriver Rolf Sundberg i sitt kompendium "Lineära statistiska modeller" [14] att en känsla för den tillgängliga datan ibland kan vara tillräcklig för att upptäcka kollinaritetsproblem. Avsnitt 2.2 och 2.3 beskriver sådanna fall där jag redan innan min analys misstänker att variabler kan bidra till ovanstående problematik, och hur jag även väljer att hantera dem.

## 3.9 Akaike's Information Criterion (AIC)

AIC är ett bra mått om man vill granska en modells förmåga att beskriva viss data utan att riskera att den är för överanpassad, detta då det belönar modeller med hög passform men straffar dem samtidigt om de blir alltför komplexa [3]. AIC-värdet i sig säger dock inte så mycket eftersom att det inte finns någon bestämmelse kring vad ett "bra" värde är. AIC är därför bara ett användbart mått när flera modeller kan jämföras med varandra. Formulan för AIC definieras i "Applied statistic inference" av Held och Bové [6] som:

$$AIC = -2l(\hat{\theta}_{ML}) + 2p \tag{29}$$

där den maximala log-likelihooden straffas av antalet parametrar p i modellen. Kriteriet är negativt orienterat vilket betyder att, bland flertalet olika modeller, är det den modell med lägst AIC som anses vara bäst.

## 3.10 Stegvis variabelselektion

Stegvis variabelselektion är en användbar metod för att ur en stor uppsättning tänkbara förklarande variabler välja ut en "bästa delmängd" för den avsedda modellen. Det går ut på att man successivt antingen inkluderar eller eliminerar

en variabel åt gången i modellen till dess att ett visst optimalt kriterium är uppfyllt. I vårt fall är ett lägsta AIC-värde det optimala kriteriet, vilket beskrivs i avsnittet ovan.

I Rolf Sundbergs kompendium [14] förklarar han att det finns olika varianter av stegvis variabelselektion, där *"forward selection"* syftar på att inkludera och utvidga modellen med en variabel i taget medans *"backward elimination"* eliminerar den variabel som är mest "icke-signifikant" ur modellen tills dess att alla variabler är statistiskt signifikanta. *"Stepwise regression"* kan man säga är en blandning mellan de två nyss nämna typerna och kan anses som en mer avancerad version av variabelselektion. Denna metod tar hänsyn till att ett tillägg av en variabel till modellen kan minska signifikansen hos andra redan befintliga variabler. Av denna anledning så kontrolleras signifikansen hos alla kandiderande variabler i modellen efter det att en ny variabel har adderats. Skulle en variabel då visa sig vara obetydlig för modellen tas den bort. Det är denna metod som används i den statistiska analysen av det här arbetet.

# 4 Statistisk analys

Dean Olivers teori om vilka faktorer som under en basketmatch leder till framgång för ett lag fångade verkligen min uppmärksamhet. Därför kommer jag dels undersöka en modell som enbart är baserad på "the four factors" för att se om jag kommer fram till liknande resultat som Oliver med de metoder som presenteras i det här arbetet, samt göra en bedömning kring hur bra modellen faktiskt verkar kunna förklara utfallet i en basketmatch. Sedan kommer jag att konstruera två andra modeller, där den ena modellen inte bara baseras på "the four factors" utan inkluderar även andra variabler som eventuellt skulle kunna förklara matchresultatet medans den andra exkluderar "the four factors" helt bland de förklarande variablerna.
Min analys kommer således gå ut på att, efter statistiska implementeringar, jämföra och utvärdera dessa 3 modeller med varandra för att avslutningsvis sedan göra en bedömning kring vilka faktorer som tycks ha störst inflytande för ett framgångsrikt resultat i en basketmatch.

Modellerna anpassas med hjälp av funktionerna glm och step från R. Tillsammans med funktionen summary presenterar glm-funktionen bland annat de ML-skattade koefficienterna med tillhörande p-värden samt AIC värden för given modell medans step-funktionen utför så kallad stegvis variabelselektion, vilket är centrala delar i min modellframtagning. De p-värden som presenteras utav glm-funktionen kan härledas från wald-testet som nämns i avsnitt 3.3

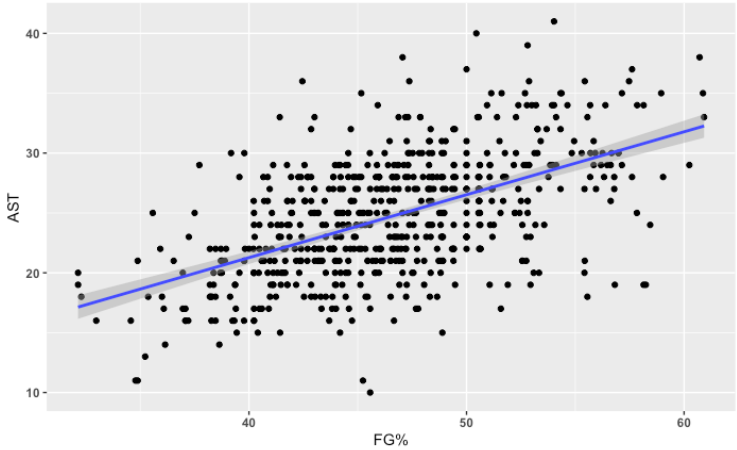
## 4.1 Metod

I det här avsnittet presenteras det tillvägagångssätt som tillämpats vid modellframtagningen. Till mitt förfogande kommer jag här att nyttja träningsmängden, vilket är den del av datamaterialet som modellerna kommer utformas på.

### 4.1.1 Undersökning av eventuella korrelationssamband

En förutsättning för logistisk regression är att observationerna ska vara oberoende varandra samt att de förklarande variablerna inte ska ha något starkt samband mellan sig, detta för att minimera de risker som kan uppstå vid multikolinearitet som jag diskuterat i avsnitt 3.8. Därför inleder jag min analys med att studera mitt datamaterial för att finna eventuella korrelationssamband. I de fall där stark korrelation verkar mellan två variabler utesluts en utav dem. Vilken variabel som utesluts låter jag bestämmas helt utav AIC-värdet på den resulterade modellen, den variabel som ingår i den modell med lägst AIC är den som behålls kvar.
För att hitta eventuella korrelationssamband har jag valt att börja med att beräkna korrelationen mellan samtliga relevanta variabler och även plottat de variabler som tyder på ett visst samspel mot varandra. I figur 8 i bilagan presenteras den korrelationsplott som jag utgått ifrån. Gränsen för vilket värde som räknas som "stark korrelation" är godtycklig. Jag har i det här arbetet valt att titta närmare på de variabel som visar på korrelation mer än $\pm 55\%$, testat modeller med och utan dessa variabler och därefter tagit ett beslut om variabler ska uteslutas eller ej baserat på AIC. I mina slutliga modeller har jag sedan valt att beräkna VIF-faktorn för varje variabel för att kontrollera att det inte existerar mulltikollinearitet.

Den givna datan innehåller bland annat *antalet målgivande passningar (AST)* och *andelen satta skott inom korgområdet (FG%)*. Dessa variabler är exempel på korrelationssamband som påträffades i datamaterialet. I figur 5 är antalet målgivande passningar *(AST)* plottade mot andelen satta skott inom korgområdet *(FG%)* där vi kan se ett tydligt linjärt samband yttra sig. Korrelationen mellan variablerna beräknas till 61%. För en utav de framtagna modellerna erhöll modellen ett AIC på 237.22 då parametern $AST$ plockats bort, och ett AIC på 309.2 då $FG\%$ istället tagits bort. Här valde jag alltså att behålla variablen för andelen satta skott inom korgområdet i modellen och uteslöt antal målgivande passningar, vilket rent intiutivt inte är ett helt oväntat beslut. Jag hanterade de resterande korrelationssambanden som påträffades på motsvarande sätt. Parvisa plottar för dessa variabler redovisas i figur 9 och 10 i bilagan.




Figur 5: En parvis plott mellan parametrarna $AST$ och $FG\%$ där observationerna tydligt följer ett linjärt mönster. Den blå linjen representerar den linjära regressionslinjen mellan de två variablerna.

### 4.1.2 Grundmodellerna

Efter granskning av mitt datamaterial kan de olika grundmodellerna konstrueras vars slutliga modeller är de som sedan kommer att jämföras och bedömmas. Som nämnt i inledningen av det här avsnittet så är det träningsmängden som används i kommande steg.

- Den första modellen är som redan nämnt baserad på Dean Olivers "the four factors" vilka presenteras i tabell 2. Denna modell redovisas även i tabell 12 i bilagan.

- Den andra grundmodellen kommer bestå utav alla relevanta variabler av intresse från den matchdata som presenteras i tabell 1, både för hemmalaget och bortalaget. Dock med viss justering av de korrelationssamband som upptäckts i avsnittet ovan. Denna grundmodell redovisas här nedan i tabell 3.

- Den tredje grundmodellen baseras på en kombination av förklarande variabler från modell 1 och 2. Modellen utgår ifrån "the four factors", men adderar även de variabler från tabell 1 som inte visar på något starkt korrelationssamband till de fyra givna faktorerna. Även denna grundmodell redovisas nedan i tabell 4 .

**Grundmodell 2**

| Variabel | Estimering | Std.Fel | z värde | Pr( > \|z\|) |
|---|---|---|---|---|
| FG% | 0.66 | 0.10 | 6.84 | $8.16 \cdot 10^{-12}$ |
| 3P% | 0.22 | 0.04 | 6.46 | $1.04 \cdot 10^{-10}$ |
| FT% | 0.13 | 0.02 | 5.52 | $3.42 \cdot 10^{-8}$ |
| OREB | 0.50 | 0.08 | 6.57 | $5.12 \cdot 10^{-11}$ |
| DREB | 0.08148 | 0.06 | 1.44 | 0.15 |
| TOV | -0.42 | 0.07 | -5.77 | $7.86 \cdot 10^{-9}$ |
| BLK | 0.01 | 0.08 | 0.19 | 0.85 |
| PF | -0.28 | 0.05 | -5.19 | $2.13 \cdot 10^{-7}$ |
| STL | 0.56 | 0.09 | 6.39 | $1.71 \cdot 10^{-10}$ |
| MOT.FG% | -0.71 | 0.10 | -7.21 | $5.42 \cdot 10^{-13}$ |
| MOT.3P% | -0.14 | 0.03 | -5.08 | $3.70 \cdot 10^{-7}$ |
| MOT.FT% | -0.10 | 0.02 | -4.56 | $5.25 \cdot 10^{-6}$ |
| MOT.OREB | -0.47 | 0.08 | -6.18 | $6.25 \cdot 10^{-10}$ |
| MOT.DREB | -0.11 | 0.06 | -1.82 | 0.07 |
| MOT.BLK | -0.04 | 0.08 | -0.55 | 0.59 |
| MOT.PF | 0.34 | 0.06 | 6.06 | $1.34 \cdot 10^{-9}$ |

Tabell 3: De variabler som ingår i grundmodell 2, dess skattade koefficienter, standardfel, z-värde och p-värde. Se tabell 1 för beskrivning av variabelnamn.

**Grundmodell 3**

| Variabel | Estimering | Std.Fel | z värde | Pr( > \|z\|) |
|---|---|---|---|---|
| eFG% | 0.72 | 0.082 | 8.80 | $< 2 \cdot 10^{-16}$ |
| TOV% | -0.45 | 0.08 | -5.71 | $1.10 \cdot 10^{-8}$ |
| OREB% | 0.27 | 0.04 | 6.45 | $1.13 \cdot 10^{-10}$ |
| FTrate% | 0.20 | 0.03 | 5.95 | $2.71 \cdot 10^{-9}$ |
| DREB | 0.03 | 0.05 | 0.51 | 0.61 |
| BLK | -0.045 | 0.07 | -0.61 | 0.54 |
| PF | -0.21 | 0.05 | -4.72 | $2.36 \cdot 10^{-6}$ |
| STL | 0.51 | 0.08 | 6.28 | $3.34 \cdot 10^{-10}$ |
| MOT.FG% | -0.68 | 0.09 | -7.75 | $9.37 \cdot 10^{-15}$ |
| MOT.3P% | -0.13 | 0.02 | -5.23 | $1.67 \cdot 10^{-7}$ |
| MOT.FT% | -0.10 | 0.02 | -4.63 | $3.71 \cdot 10^{-6}$ |
| MOT.OREB | -0.44 | 0.07 | -6.31 | $2.86 \cdot 10^{-10}$ |
| MOT.DREB | 0.06 | 0.07 | 0.81 | 0.42 |
| MOT.BLK | 0.02 | 0.07 | 0.35 | 0.73 |

Tabell 4: De variabler som ingår i grundmodell 3, dess skattade koefficienter, standardfel, z-värde och p-värde. Se tabell 1 för beskrivning av variabelnamn.

#### 4.1.3 Variabelselektion

Nästa steg är att utföra *Stepwise regression* med hjälp utav funktionen *step* i R på grundmodellerna för att få fram de modeller som genererar lägst AIC. Jag

kommer endast utföra det här steget på grundmodell 2 och 3. Detta eftersom att modellen baserat på Dean Olivers teori redan anses vara en *slutgiltig modell* som jag enbart kommer göra de slutliga modellutvärderingarna på och jämföra med de två resterande modellerna.
I tabell 5 och 6 presenteras de slutliga variablerna som ingår i modell 2 och 3 efter det att Stepwise regression utförts på grundmodellerna.

**Slutgiltig Modell 2**

| Variabel | Estimering | Std.Fel | z värde | Pr( > \|z\|) | VIF |
|---|---|---|---|---|---|
| FG% | 0.75 | 0.09 | 8.75 | $< 2{\cdot}10^{-16}$ | 4.84 |
| 3P% | 0.21 | 0.033 | 6.38 | $1.79{\cdot}10^{-10}$ | 1.81 |
| FT% | 0.14 | 0.02 | 6.16 | $7.50{\cdot}10^{-10}$ | 1.83 |
| OREB | 0.531 | 0.07 | 7.25 | $4.09{\cdot}10^{-13}$ | 2.66 |
| TOV | -0.36 | 0.06 | -5.77 | $7.88{\cdot}10^{-9}$ | 1.88 |
| PF | -0.30 | 0.053 | -5.78 | $7.35{\cdot}10^{-9}$ | 1.83 |
| STL | 0.51 | 0.08 | 6.48 | $8.93{\cdot}10^{-11}$ | 1.88 |
| MOT.FG% | -0.77 | 0.09 | -8.58 | $< 2{\cdot}10^{-16}$ | 5.03 |
| MOT.3P% | -0.14 | 0.03 | -5.10 | $3.46{\cdot}10^{-7}$ | 1.67 |
| MOT.FT% | -0.11 | 0.02 | -4.93 | $8.41{\cdot}10^{-7}$ | 1.44 |
| MOT.OREB | -0.50 | 0.07 | -6.82 | $9.02{\cdot}10^{-12}$ | 2.65 |
| MOT.PF | 0.36 | 0.06 | 6.49 | $8.68{\cdot}10^{-11}$ | 1.81 |

Tabell 5: De slutliga variabler som ingår i Modell 2, dess skattade koefficienter, standardfel, z-värde, p-värde och VIF-faktor. Se tabell 1 för beskrivning av variabelnamn.

**Slutgiltig Modell 3**

| Variabel | Estimering | Std.Fel | z värde | Pr( > \|z\|) | VIF |
|---|---|---|---|---|---|
| eFG% | 0.68 | 0.07 | 9.68 | $< 2{\cdot}10^{-16}$ | 4.46 |
| TOV% | -0.46 | 0.07 | -6.58 | $4.63{\cdot}10^{-11}$ | 1.74 |
| OREB% | 0.25 | 0.03 | 7.44 | $1.03{\cdot}10^{-13}$ | 1.90 |
| FT-rate | 0.19 | 0.03 | 6.40 | $1.59{\cdot}10^{-10}$ | 1.68 |
| PF | -0.21 | 0.04 | -4.74 | $2.09{\cdot}10^{-6}$ | 1.36 |
| STL | 0.49 | 0.07 | 6.57 | $4.95{\cdot}10^{-11}$ | 1.71 |
| MOT.FG% | -0.68 | 0.08 | -8.88 | $< 2{\cdot}10^{-16}$ | 4.12 |
| MOT.3P% | -0.13 | 0.02 | -5.38 | $7.55{\cdot}10^{-8}$ | 1.62 |
| MOT.FT% | -0.10 | 0.02 | -5.03 | $4.89{\cdot}10^{-7}$ | 1.37 |
| MOT.OREB% | -0.44 | 0.06 | -6.73 | $1.73{\cdot}10^{-11}$ | 2.17 |

Tabell 6: De variabler som ingår i Modell 3, dess skattade koefficienter, standardfel, z-värde, p-värde och VIF-faktor. Se tabell 1 för beskrivning av variabelnamn.

Vi noterar att samtliga förklarande variabler i tabellerna ovan har statistiskt

signifikanta p-värden, vilket tyder på att variablerna är av betydelse och modellen skulle komma att påverkas om någon utav parametrarna plockades bort. Som nämnt i teoridelen representerar de estimerade koefficienterna variablernas effekt på log-oddsen att ett hemmalag vinner, där en positiv koefficient tyder på att den förklarande variabeln har ett positivt samband till log-oddsen och omvänt för en negativ koefficient. Här kan vi självklart föreställa oss att de koefficienter med störst absolutbelopp är de variabler som har störst effekt på log-oddsen och därmed störst påverkan på utfallet. En mer intiutiv tolkning av koefficienterna får vi när vi beräknar oddskvoten av varje enskild variabel vilket kommer utföras i nästa avsnitt. Även VIF-faktorerna för samtliga variabler är låga som sig bör. Jag noterar dock att de parametrar som sticker ut med högst VIF är *andelen satta skott inom korgområdet* för båda hemmalag och bortalag (FG% och MOT.FG%)

En för mig ganska oväntat iakttagelse som jag reagerar på är att antalet defensiva returer (DREB) inte anses vara en statistisk signifikant variabel i någon utav modellerna. Jag hade nog förväntat mig att denna parameter skulle vara given bland de som kan förklara utfallet av en match då jag tänker mig "att alla returer är bra returer". Tar hemmalaget en defensiv retur betyder det att de får möjligheten att gå till anfall, samtidigt som de hindrar bortalaget från att ta en offensiv retur. Det verkar ju onekligen dock som att det finns andra defensiva kvalitèr som är mer betydelsefulla för ett lag än just dessa returer. Bland annat att störa motståndarnas skytte så gott det går eller att stjäla så många anfall från motståndarna som möjligt (STL) för att hindra dem från att komma till avslut överhuvudtaget.

## 4.2 Modellutvärdering & resultat

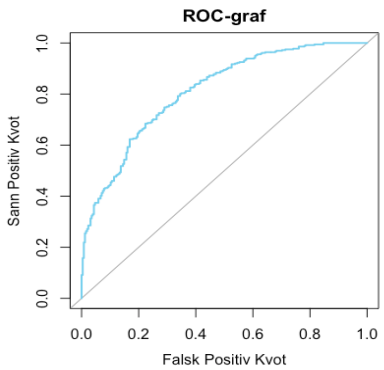
När de slutliga modellerna har tagits fram utvärderas dereas prediktiva förmåga för att se hur väl modellerna faktiskt klarar av att förklara utfallet. Detta görs genom att nu applicera de framtagna modellerna på *testmängden*, och därefter utförs de metoder som redovisas i avsnitt 3.6. Samtliga modellers AIC, Log-loss värde, klassificeringsförmåga och AUC sammanfattas i tabell 7 och i figur 6 presenteras de ROC-grafer som tillhör respektive modell.

**Utvärdering av prediktiv förmåga**

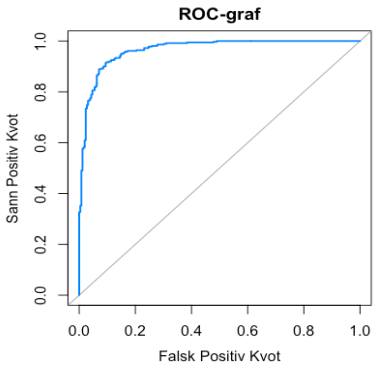
| Modell | AIC | Log-loss | Klass.förmåga | AUC |
|---|---|---|---|---|
| Modell 1 | 634.33 | 0.52 | 0.741 | 0.811 |
| Modell 2 | 233.10 | 0.25 | 0.905 | 0.966 |
| Modell 3 | 239.69 | 0.21 | 0.907 | 0.975 |

Tabell 7: Log-loss, klassificeringsförmåga och AUC för respektive modell.

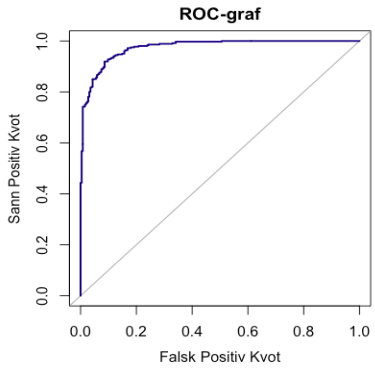
Den kanske mest intressanta iaktagelsen utifrån dessa resultat är att samtliga modeller tycks ha en ytterst god prediktiv förmåga med märkbart höga klassificeringsvärden och AUC. Modell 2 och 3 utmärker sig särskilt mycket, då




(a) Modell 1




(b) Modell 2

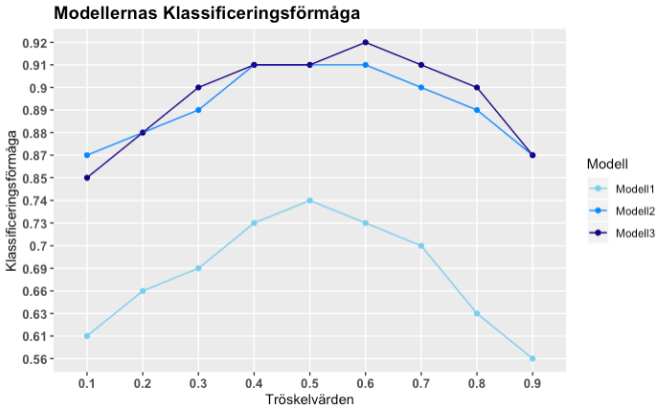



(c) Modell 3

Figur 6: ROC-grafer för samtliga 3 modeller.

båda modellerna uppvisar en klassificeringsförmåga på nästan 91% och AUC-värden större än 0.95. Detta betyder alltså att dessa modeller lyckas prediktera rätt utfall för hemmalaget över 90% av fallen, och som vi ser i figur 6 är dessa modellerns ROC-kurvor ytterst nära grafens optimala punkt $(1, 1)$ som beskrevs i avsnitt 3.6.3. Utifrån AIC så är det modell 2 som erhåller det lägsta värdet och därför anses vara "den bättre modellen". Men precis som i Log loss, klassificeringsförmåga och AUC så är skillnaden mellan de två modellernas AIC inte särskillt stor. Jag skulle därför vilja påstå att modellerna är relativt likvärdiga baserat på dessa resultat.
Trots att Modell 1, som består utav "the four factors", erhåller minst eftertraktade värden på samtliga poster skulle jag inte vilja påstå att modellen är dålig och inte kan användas i det tänkta syftet. Jag skulle snarare vilja hävda att modellen är fullt användbar och beskriver utfallet förvånadsvärt bra för att bara bestå utav fyra förklarande variabler då AUC är hela 0.81 (vilket kan anses som högt).

Klassificeringsförmågan i tabell 7 är baserad på ett tröskelvärde lika med 0.5, vilket är det tröskelvärde som oftast används i liknande undersökningen men som nödvändigtvis inte behöver vara det som ger det mest optimala resultatet. I figur 7 har jag beräknat klassificeringsförmågan för de tre modellerna givet olika tröskelvärden, detta för att kontrollera vilken gräns som tycks vara den ideala för varje enskild modell. Där ser vi att modell 1 och modell 2 båda pikar kring ett tröskelvärde på 0.5, medans modell 3 tycks få absolut bäst resultat med ett tröskelvärde på 0.6. Dock så är denna ökning minimal och klassificeringsförmågan förbättras endast med 1 procentenhet.




Figur 7: Modellernas klassificeringsförmåga beräknat med olika tröskelvärden.

Min naturliga reaktion och känsla efter denna modellutvärdering var att "detta är för bra för att vara sant". Hur kan mina modeller lyckas skatta utfallen så pass bra som de gör? Är detta över huvud taget rimligt?
En förklaring kan vara det faktum att jag låter mina modeller förklara utfallet för just hemmalaget. Precis som jag nämnde i avsnitt 2.3 så spelades majoritetet av alla vinster under säsongen på det vinnande lagets hemmaplan. Och då den logistiska regressionsmodellen beskriver sannolikheten att en händelse ska inträffa, vilket i det här fallet syftar på sannolikheten att hemmalaget vinner, skulle kanske den naturliga majoriteten utav vinster i datamaterial kunna vara en bidragande faktor som gör min data mer "lätt-predikterad" för denna typ av modell.
En annan bidragande faktor till att modell 2 och modell 3 lyckades utmärka en sådan särskilt god prediktiv förmåga kan självklart ha och göra med att de förklarande variablerna i dessa modeller representerar parametrar från båda lagen. Förmodligen så kan dessa förklarande variabler tillsammans ge en så pass övergripande och målande bild utav händelseförloppet att utfallet blir lätt att uppskatta, jämfört med när parametrarna endast representerar ena laget. Detta skulle kunna förklara varför modell 2 och 3 exempelvis har en klassificeringsförmåga på över 90% medans modell 1, vars förklarande variabler näst intill

enbart utgår ifrån hemmalagets parametrar, visar på en klassificeringsförmåga betydligt lägre än de två andra i tabell 7.
Jag har dock försökt att undvika denna problematik så gott det går genom att medvetet exkludera de mest uppenbara parametrarna i mitt datamaterial i början av modellframtagningen. Som jag tidigare förklarat så använder jag mig inte av *antalet gjorde poäng per lag* då vem som helst kan prediktera ett korrekt utfall efter att enbart ha observerat dessa två parametrar. Av samma anledning har jag valt att inte använda vairabler så som *FGM,3PM* och *FTM* (se tabell 1 för beskrivning av variablerna) i mina modeller eftersom att dessa beskriver antalet **satta skott** från de olika distanserna. Att observera både hemmalaget och bortalagets satta distansskott från en match ger en ganska klar bild om vilka som bör ha vunnit (utan att veta den totala poängen per lag), så en modell baserat på delvis dessa parametrar förväntas också ha en näst intill perfekt prediktionskapacitet.

Så trots mina medvetna exkluderingar av de mest uppenbara parametrarna, tycks det alltså ändå finnas variabler i mitt datamaterial som tillsammans kan förklara resultatet i en match på ett utomordentligt bra sätt, men som är mindre uppenbara. Hur bra modellerna lyckas prediktera resultaten under NBA-säsongen är dock inte mitt primära syfte i den här rapporten. Detta gör jag enbart för att se så att de förklarande variablerna i mina framtagna modeller faktiskt *kan* användas för att beskriva utfallet i en match, vilket de uppenbarligen kan enligt resultaten från tabell 7. Jag tänker därför inte lägga någon större vikt i att göra någon vidare analys kring modellernas prediktionskapacitet, men ämnet vidrörs en aning i diskussionen som kommer senare i rapporten.

Det huvudsakliga syftet med det här arbetet är att bedömma vilka faktorer som ett basketlag bör satsa på för att nå framgång och alltså försöka kartlägga de faktorer som tycks ha störst inverkan på resultatet. Hittils har vi redan sållat bort de parametrar som inte är statistiskt signifikanta och som därmed inte verkar har någon betydelse alls för det slutliga utfallet. Antal blockerade skott (BLK) och tagna defensiva returer (DREB) verkar till exempel vara genomgående variabler i både modell 2 och 3 som inte tycks ha några effekter på utfallet alls då ingen utav dessa variabler finns kvar i de slutliga modellerna. De förklarande variabler som istället ingår i våra slutliga modeller har en betydande effekt på resultatet, frågan är bara *hur* betydelsefull denna effekt är.
Vi hade kunnat granska storleken på de skattade koefficienterna i de slutliga modellerna och därefter rangordna de förklarande variablerna för att få en uppfattning om vilka faktorer som verkar spela en större roll än andra för matchresultatet. Men som tidigare nämnt är koefficienterna, i de slutliga modellerna, i relation till log-oddsen att ett hemmalag vinner. Av den anledningen har jag valt att göra en vidare analys av de förklarande variablerna för att få en bättre överblick av parametrarna samt för att skapa mig en ännu bättre uppfattning i frågan. Detta har jag valt att göra enligt metoden som redovisas i avsnitt 3.7, det vill säga beräkna varje enskild variabels oddskvot med tllhörande konfidensintervall i varje modell. Dessa resultat sammanfattas i tabell 8, 9 och 10.

I bilagan, figur 11, 12, 13, redovisas även mer visuella illustrationer av konfidenintervallen. Som tidigare nämnt så har jag valt att utforma "wald-baserade" konfidensintervall som jag tar fram med den hjälp av den inbyggda funktionen *confint* i R. I denna funktion kan man välja att nämna "wald" som ett av argumenten för att konfidensintervallen ska bli wald-baserade.

**Modell 1**

| Variabel | Oddskvot | Konfidensintervall |
|---|---|---|
| eFG% | 1.24 | [ 1.19,1.30 ] |
| OREB% | 1.08 | [ 1.05,1.11 ] |
| FT-rate | 1.06 | [ 1.03,1.09 ] |
| TOV% | 0.90 | [ 0.85,0.96 ] |

Tabell 8: Beräknade oddskvoter och tillhörande konfidensintervall för varje förklarande variabel i modell 1.

**Modell 2**

| Variabel | Oddskvot | Konfidensintervall |
|---|---|---|
| FG% | 2.13 | [1.80,2.51] |
| 3P% | 1.24 | [1.16,1.32] |
| FT% | 1.15 | [1.10,1.21] |
| OREB | 1.69 | [1.47,1.95] |
| TOV | 0.69 | [0.61,0.79] |
| PF | 0.74 | [0.67,0.82] |
| STL | 1.66 | [1.42,1.94] |
| MOT.FG% | 0.46 | [0.39,0.55] |
| MOT.3P% | 0.87 | [0.82,0.92] |
| MOT.FT% | 0.90 | [0.86,0.94] |
| MOT.OREB | 0.60 | [0.52,0.70] |
| MOT.PF | 1.43 | [1.28,1.59] |

Tabell 9: Beräknade oddskvoter och tillhörande konfidensintervall för varje förklarande variabel i modell 2.

**Modell 3**

| Variabel | Oddskvot | Konfindensintervall |
|---|---|---|
| eFG% | 1.98 | [1.72, 2.27] |
| TOV% | 0.63 | [0.55, 0.72] |
| OREB% | 1.29 | [1.20, 1.38] |
| FT-rate | 1.21 | [1.14, 1.28] |
| PF | 0.81 | [0.75, 0.89] |
| STL | 1.63 | [1.41, 1.89] |
| MOT.FG% | 0.51 | [0.44, 0.59] |
| MOT.3P% | 0.88 | [0.84, 0.92] |
| MOT.FT% | 0.90 | [0.87, 0.94] |
| MOT.OREB | 0.65 | [0.57, 0.73] |

Tabell 10: Beräknade oddskvoter och tillhörande konfidensintervall för varje förklarande variabel i modell 3.

Fördelen med att betrakta värden på en oddskvot-skala är som tidigare nämnt att resultaten bli mer intiutiva och lätt-tolkade då oddskvoten beskriver den mulltiplikativa effekten på oddsen. Detta implicerar att oddskvoten kan anses motsvara den procentuella förändringen på oddsen när den förklarande variabeln ökar med 1. En oddskvot nära 1 tyder på att variabeln har en liten effekt på oddsen och omvänt för en oddskvot längre ifrån 1. Agresti skriver i paragraf 4.2.2 " *Confidence intervals for effects* "[1] att konfidenintervallets undre- och övregräns över en oddskvot kan löst tolkas som variablens minsta troliga effekt på oddsen respektive maximalt troliga effekt. Men, koefficientskattningen och därmed oddskvoten är fortfarande vår bästa gissning av vad effekten är. Ett brett konfidensintervall tyder på att oddskvoten har en stor standardavvikelse, och vår osäkerhet i oddskvotens riktiga effekten blir stor.

### 4.2.1 Tolkning av resultat

Om vi börjar med att granska resultatet för "the four factors" i modell 1, kan vi konstatera att ordningen av variablernas betydelse för utfallet ter sig likt Dean Olivers teori. Effective field goal percentage (eFG%) har en oddskvot på 1.24, vilket är den oddskvot som skiljer sig mest ifrån 1 och variabeln visar därför på störst inverkan på oddsen att hemmalaget vinner. En oddskvot lika med cirka 1.24 betyder att om eFG% ökar med 1, så ökar oddsen för vinst med 24%. På samma sätt erhåller Turnover rate (TOV%) en oddskvot på 0.90 som alltså betyder att oddsen för vinst sjunker med cirka 10% när TOV% ökar med 1, vilket är den oddskvot som näst efter eFG% skiljer sig mest från 1. Offensive rebound percentage (OREB%) och turnover rate (FT-rate) visar sig öka oddsen med 8.4% respektive 6.1% enligt oddskvoten och är, precis som Oliver hävdar, de variabler som har minst effekt på utfallet av de fyra faktorerna.

De främsta likheterna mellan samtliga modeller är att effektiv skottprocent

(eFG%) och skottprocent inom korgområdet (FG%) utmärker sig för att vara de viktigaste och mest betydelsefulla faktorerna för utfallet i en basketmatch. För att ett lag ska öka sina chanser till att vinna måste laget alltså först och främst kunna skjuta bra. Det är nog heller ingen slump att just FG% tycks vara den variabel som ligger i framkant bland skottdistanserna med tanke på att skott inom korgområdet utan tvekan utgör det vanligaste sättet att göra mål på. 3-poängsskott må generera fler poäng per satt skott, men frekvensen av dessa är betydligt lägre under en match i jämförelse, vilket avspeglas i 3P%-variabels oddskvot jämt emot de andra variablerna (se tabell 9). Även de parametrar som relaterar till offensiv returtagning (OREB) tillhör de variabler med högst oddskvot i samtliga modeller. Detta har antagligen att göra med det faktum att offensiva returer gör så att laget får möjlighet till ett ytterligare anfall eller leder till lätta poäng nära korgen alternativt straffkast. Modell 2 och 3 visar även att motsvarande parametrar för bortalaget är de som har störst negativ effekt på oddsen för att hemmalaget ska vinna, vilket inte är helt oväntat eftersom att dessa parametrar då generera poäng till bortalagets fördel. Tittar vi till exempel närmare på bortalagets andel satta skott inom korgområdet (MOT.FG%) ser vi att dess konfidensintervall avslöjar att variabeln skulle kunna minska oddsen att hemmalaget vinner med hela 60% . Parametrar relaterade till antal avbrutna anfall (TOV) visar även de tillhöra de variabler som tycks ha störst inverkan på utfallet i samtliga modeller. Dessa parametrar (TOV och TOV%) visar på en negativ effekt på oddsen att hemmalaget ska vinna, vilket inte heller är förvånande då många avbrutna anfall ger laget färre möjligheter att komma till avslut, som istället ger upphov till att motståndarlaget får chansen att generera poäng.

Andra noterbara likheter mellan modell 2 och 3 är att båda modellerna antyder att hemmalagets antal gjorda fouls (PF) och antal stulna bollinnehav (STL) är signifikanta faktorer. I de flesta fall leder en foul till att motståndarlaget får skjuta straffkast. Det finns även en regel som säger att en spelare måste avbryta matchen om hen själv har orsakat 5 stycken folus, detta kallas för att spelaren blir "ut-foulad". Det finns alltså tydliga fördelar för motståndarlaget om ett lag orsakar många fouls, men det är samtidigt inget som tycks väga särskillt tung för ett matchresultat eftersom att variabeln PF inte står för de mest inflytelserika oddskvoterna i någon utav modellerna. Mer intressant är istället STL som visar sig ha en rätt hög oddskvot i båda modellerna som tyder på en betydligt större effekt på oddsen än många andra parametrar. Enligt konfidensintervallen för STL ser vi att denna variabel i bästa fall nästan kan fördubbla oddsen att hemmalaget vinner om den ökar med 1 (se tabell 12 och 13). Detta kan kanske tyckas låta mycket, men man får heller inte glömma att de fördelar som kommer av att ta en STL är både att det ofta leder till att man får möjlighet att göra ett så kallat "fast break" ,som syftar på ett snabbt anfall där försvararna inte hunnit positionera sig och sannolikheten för poäng därför stor, men också den mentala boost som ett lag får av att stjäla bollen från motståndarna vilket bidrar till ett psykiskt övertag och momentum i matchen.

Detta är en utav de tydligaste skillnaderna mellan modell 1 och de förklarande

variablerna i modell 2 och 3. Resultaten från modell 2 och 3 visar på att STL har en relativt stor multiplikativ effekt på oddsen att hemmalaget vinner och därmed anses vara en väldigt betydelsefull variabel för utfallet, medans det är en parameter som Dean Oliver inte väljer att nämna alls i sina teorier. Istället låter Oliver *Free throw rate* (FT-rate) bygga en utav stenarna i the four factors, samtidigt som FT%,MOT.FT% och FT-rate (det vill säga alla parametrar som relaterar till straffkast på något sätt) tillhör de parametrar som visar på minst effekt på oddsen i modell 2 och 3.
Man kan argumentera för att det tillkommer många andra fördelar med att få skjuta straffkast, bland annat då detta är direkt kopplat till att motståndarna har orsakat en foul (PF). Att hamna i "foul-problem" med många fouls per spelare kan vara förödande för ett lag. Detta eftersom att det kan bidra till att spelare blir mer avvaktande i sitt försvarsspel för att undvika att få fler onödiga fouls, samt att man kan mista viktiga spelare som blir "ut-foulade".

Man kan alltså konstatera att det finns många likheter, men också ett par märkbara olikheter, mellan modellerna och dess förklarande variabler. Trots detta tycks ändå alla modeller kunna förklara utfallet i en basketmatch väl då samtliga modeller anses ha en god prediktiv förmåga. Detta gjorde mig nyfiken på att undersöka hur alternativa modeller till Dean Olivers "the four factors" skulle bete sig. Hur starka är egentligen Olivers påståenden? Finns det andra kombinationer utav fyra parametrar som skulle kunna ersätta the four factors minst lika bra? I tabell 11 har jag sammanställt resultaten från 4 olika alternativa modeller som består utav 4 parametrar vardera .

**Alternativa Modeller**

| | Förklarande Variabler | AIC | Log loss | Klass.förmåga | AUC |
|---|---|---|---|---|---|
| 1 | FG%,OREB,STL,TOV | 661.68 | 0.55 | 0.713 | 0.777 |
| 2 | eFG%,OREB%, TOV%,STL | 631.69 | 0.52 | 0.732 | 0.810 |
| 3 | MOT.FG,MOT.OREB, MOT.PF,MOT.FT | 686.33 | 0.55 | 0.711 | 0.784 |
| 4 | MOT.eFG%,MOT.OREB%, MOT.TOV%,MOT.FTrate | 649.59 | 0.51 | 0.742 | 0.818 |

Tabell 11: Beräknade värden för AIC, Log loss, klassificeringsförmåga och AUC för fyra modeller.

Baserat på dessa resultat tycks de modeller som ändå inkluderar eFG%,OREB% och TOV% som är ett par utav "the four factors" (rad 2 och rad 4) vara att föredra. Detta eftersom att dessa modeller har lägst AIC och Logloss samt högst klassificeringsvärde och AUC, vilket ändå bekräftar Olivers påståenden till viss mån. Men jämför vi dessa två alternativa modeller med resultatet från modell 1 i tabell 7 kan vi se att alla modellerna genererar så gott som identiska resultat. Dessutom är det modellen på rad 2 i tabell 11 bestående utav eFG%,OREB%,

TOV% och STL den modell som visar på lägst AIC och därför skulle kunna anses som den "bättre" modellen framför Dean Olivers four factors. Så uppenbarligen tycks det finnas andra alternativa kombinationer som kan förklara utfallet i en match i princip lika bra som "the four factors" . En intressant iaktagelse är att rad 4, där jag beräknat "the four factors" för bortalagets parametrar, genererar näst intill exakt samma värden på Log loss, klassificeringsförmåga och AUC som Dean Olivers motsvarighet i modell 1. Det enda som märkbart skiljer modellerna åt är deras AIC-värden, där modell 1 står för det betydligt lägre värdet.

Jag noterar det faktum att oddskvoterna och konfidensintervallen för "the four factors" skiljer sig åt väldigt mycket i modell 1 och modell 3 (se tabell 8 och 10). Detta beror ju delvis på att modell 3 innefattar fler variabler, som dessutom visar sig kunna förklara utfallet av en match ännu bättre, vilket påverkar de skattade koefficienterna i modellen som i sin tur verkar på oddskvoterna och dess utseenden. I diskussionen resonerar jag mer kring utformningen av själva konfidensintervallen, vilket dock skulle kunna vara en bidragande faktorer till den stora skillnaden i resultat som vi ser i det här avsnittet.

# 5 Slutsats

Så vilka är de mest inflytelserika faktorerna i en basketmatch? Vart bör ett hemmalag lägga sitt stora fokus för att vinna?
Ja, jag kan definitivt konstatera att den frågan visar sig vara lika komplex och mångsidig som den förutspåddes vara i inledningen. Olika kombinationer av förklarande variabler visar sig ha en förmåga att kunna förklara utfallet av en basket match väldigt väl, vilket gör att det inte riktigt finns något entydigt svar. Det blir svårt att säga vad som egentligen är rätt och fel när man handskas med en fråga med många svar, vilket antagligen är anledningen till att en hel basketvärld är oense.

Dean Oliver har tagit fram en bra modell med 4 faktorer som absolut kan anses förklara utfallet i en basketmatch. Det är dock uppenbart att fokuset i hans teori ligger i det offensiva anfallsspelet. Man kan väldigt tydligt se att modeller där man inkluderar fler parametrar har en ännu bättre förmåga att beskriva utfallet då de antagligen ger en bättre helhets bild av händelseförloppet, inte minst då man även inkluderar motsvarande parametrar för bortalaget. I dessa modeller framgår det att det *finns* defensiva kvaliteer i ett lag som *är* betydelsefulla för att ett lag ska ha större chans till att vinna, vilket inte framkommer alls i Olivers teori.

Jag kan dock tycka att Oliver har hittat ett bra sätt att sammanfatta några utav de viktigaste egenskaperna som ett lag bör lägga stor vikt i om de vill nå framgång, där bland *effektiv skottprovent (eFG%)*, *andel returer på offensiv planhalva (OREB%)* och *andel offensiva anfall som inte slutar med ett poängförsök (TOV%)*. Men jag skulle även vilja tillägga att ett bra, och framförallt

aggresivt, försvarspel som leder till många *stulna bollinnehav (STL)* också är av stor betydelse för att ett hemmalag ska öka sina chanser till vinst, och definitivt bör räknas med bland de mest inlfytelserika faktorerna.

Så sammanfattningsvis anser jag att ett hemmalag bör absolut fokusera på att kunna skjuta bra, ta många offensiva returer, vara sickliga med bollen i anfall så att de kan komma till avslut ofta samt spela ett aggresivt försvarspel genom att stjäla många bollinnehav från bortalaget, för att öka deras möjligheter att vinna och nå framgång.
Självklart bör även lag se till att sätta sina straffkast och inte orsaka många onödiga fouls eftersom att dessa parametrar också visar sig ha en viss betydelse för utfallet. Men baserat på dessa faktorers oddskvoter ser man att effekten på oddsen att hemmalaget vinner är mindre i jämförelse med andra, och jag skulle därför vilja påstå att det inte är utav dessa parametrar ett lag kommer vinna matcher på.

Ett dåligt försvar visar sig absolut kunna sänka ett lags chanser till att vinna. Framför allt om laget inte lyckas störa motståndarlagets avslut inom korgområdet med tanke på att MOT.FG% visar sig kunna sänka oddsen för vinst med när 60% (se undregräns i konfidensintervallen). Men överlag bland de variablerna som anses vara mest betydelsefulla, tycks ändå de offensiva kvaliteerna vara de som väger tyngst. Därför kan jag förstå varför Dean Oliver kanske valt att fokusera på just den offensiva kapaciteten i ett lag när han framställt sina "the four factors". Men som många andra är jag inte heller fullt överrens med allt som påstås i hans teorier, trots att "the four factors" faktistk visar sig ha en relativt god förmåga att kunna förklara utfallet i en basketmatch.
Som tidigare nämnt i avsnitt 1.2 menar Oliver att de parametrar som inte ingår i hans fyra valda faktorer sker så pass sällan under en match och kan därav anses som obetydliga för utfallet. Rent teoretisk sätt skulle ett lag kunna stjäla bollen utav motståndarna (STL) varje gång motståndarlaget anfaller (sen hur realistisk den tanken är kan självklart diskuteras). Poängen är att antal tagna STL är resultatet av ett aggresivt försvarsspel. Ett lag med få tagna STL per match kan i praktiken öka denna frekvens genom att ändra på sin strategi och lägga stort fokus och träning i att jobba upp ett aggresivare försvar. Jag hävdar alltså att detta är något som lagen faktiskt kan påverka till viss del, och om fler STL per match skulle kunna näst intill dubbla oddsen att vinna hävdar jag att det definitivt är något ett lag bör lägga fokus på om de vill vinna många matcher.

# 6 Diskussion

I avsnitt 3.7 skriver jag om konfidensintervall och förklarar att jag valt att utgå ifrån wald-konfidensintervall i det här arbetet men att jag likväl hade kunnat utforma andra konfidensintervall med samma syfte, till exempel likelihood-kvot baserade konfidensintervall. Det finns många argument som är till fördel för

likelihood-kvot baserade konfidensintervall. För det första så klarar ett likelihood-kvot test att testa hela modeller medans ett wald-test enbart testar enskilda koefficienter, vilket gör att likelihood-kvot testet får ett bredare tillämpningsområde i jämförelse. Enligt Agresti så ter sig båda testen relativt lika när vi handskas med stora datamaterial. Men om datamaterialet anses ha få observationer eller om de skattade sannolikheterna råkar vara väldigt nära 0 eller 1, är likelihood-kvot testet mer tillförlitligt än wald-testet och vi bör därför konstruera ett likelihood-kvot baserat konfidensintervall i dessa fall [1]. Även Hosmer, Lemeshow och Sturdivant påpekar detta och att det finns anledningar till att likelihood-kvot baserade konfidensintervall rekommenderas framför wald-konfidensintervall, så väl för skattade koefficienter som hela modeller. [7]
Men vad räknas då som ett stort datamaterial? Svaret på denna fråga är väldigt godtycklig och det beror mycket på hur datamaterialet man jobbar med ser ut. J.Scott Long rekommenderar i sin bok "Regression Models for Categorical and Limited Dependent Variables" [9] att inte utföra logistisk regression på ett datamaterial mindre än 100 observationer, och att data på över 500 observationer kan i breda drag anses som pålitligt. Han poängterar dock att det är datamaterialets struktur som egentligen avgör hur stort materialet behöver vara. [9]
Baserat på J.Scott Longs uttalanden anser jag att det datamaterial som används i det här arbetet kan räknas som stort eftersom att det total består utav 1230 stycken observationer, där träningsmängd och testmängd innehåller över 600 observationer var. Så av den anledningen ser jag inget fel i att använda wald-konfidensintervall. Men Agresti nämner ju även en problematik som kan uppstå då de skattade sannolikheterna är nära 0 och 1, och med tanke på att modell 2 och modell 3 visar på en sådan god prediktiv förmåga misstänkte jag att detta förmodligen var fallet i dessa två modeller. Detta är något jag får bekräftat när jag plottar de skattade värdena, se figur 14 och 15 i bilagan. Jag har därför anledningar att tro att konfidensintervallen som är framtagna i den här rapporten inte är fullt så pålitliga och precisa för ändamålet (gällande oddskvoterna i modell 2 och 3) och det hade nog varit ett bättre alternativ att utforma konfidensintervallen baserat på ett likelihood-kvot test. Min osäkerhet och okunskap i att generea ett pålitligt och korrekt likelihood-kvot baserat konfidensintervall i R var dock det största, och kanske enda, skälet till varför jag inte valde att gå på denna metod.

Med mer tid hade jag även kunna inkludera ännu fler observationer till min data. På NBAs officiella hemsida finns matchdata ända från säsongen 1996/1997 registrerad, och jag är väl medveten om att ju mer data desto bättre. Men tyvärr så var processen att hämta datan från hemsidan väldigt tidskrävande och jag fick därför nöja mig med statistik från säsongen 2018/2019. Min avvsikt var heller aldrig att använda data från alla matcher ända från 1996 eftersom att sporten förändrats och modifierats så pass mycket under dessa 20 år som gått. Nya regler har tillkommit och tempot är betydligt snabbare, vilket är reformer som man nog känner av i de flesta sporter. Det viktigaste för mig var därför att använda data som kan representera nutidens basketspel, vilket min undersökning ändå uppfyller trots något färre observationer.

En annan väg att gå om jag hade fått göra om arbetet är att möjligen bortse från all skottprocent inom korgområdet (FG%). Jag misstänker att det är denna parametern, som till störst del, är nyckeln till att mina modeller lyckas förklara utfallet så bra. Även om jag medvetet valt att enbart använda andelen satta skott för denna distans (för att göra situationen mindre uppenbar) så kan man nog ändå ganska väl tyda vilket lag som borde vunnit en match baserat på denna procent (givet att man har kännedom om båda lagens procent).Detta eftersom att det är övervägande flest avslut inom denna distans under en match. Att denna typ av skott är de absolut vanligaste är också anledningen till att det såklart kan anses vara självklart att denna faktor har en stor betydelse för att ett lag ska vinna och därför bör kunna tas förgivet. Vi noterade även att denna parameter för båda lagen erhåller störst VIF-faktor vilket också är en anledning till att jag överväger att behandla dessa parametrar annorlunda till nästa gång. Vissa påstår nämligen att en VIF-faktor över 5 är problematiskt, vilket är fallet för MOT.FG% i modell 2 (se tabell 5).

Jag vill slutligen nämna en sista reflektion som jag hade tagit hänsyn till om jag fått göra om arbetet. Tanken med att ha med en modell bestående av ”the four factors” var att utvärdera samt utmana Dean olivers teorier. Oliver nämner aldrig något om motståndarlagets prestationer, utan utgår enbart från ett lag, där av mitt val att låta modell 1 endast bestå utav fyra faktorer. Men det hade självklart varit intressant att se hur hur resultatet hade sett ut av en modell bestående av ”the four factors” för både hemmalaget och bortalaget. Med tanke på att modell 1 visar på relativt goda egenskaper när det kommer till att förklara resultatet i en match, och att vi efter denna undersökning dessutom kan konstatera att modellens prediktiva förmåga ökar när vi räknar med parametrar från motståndarlaget, tvivlar jag inte på att en sådan modell hade presterat väl. Så hade jag fått möjlighet att göra om arbetet hade jag nog valt att jämföra en fjärde modell likt denna.

# Referenser


[1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013.

[2] Mike Beuoy. *How NBA Games Are Won*. https://www.inpredictable.com/2015/02/how-nba-games-are-won.html, Jan 16, 2015.

[3] Sachine Date. *The Akaike Information Criterion*. https://towardsdatascience.com/the-akaike-information-criterion-c20c8fd832f2, Nov 9 2019.

[4] Sai Harshith Reddy Gaddam. *Log loss function match explained*. https://towardsdatascience.com/log-loss-function-math-explained-5b83cd8d9c83, Jan 6, 2019.

[5] Daniel Godoy. *Understanding binary cross-entropy / log loss: a visual explanation*. https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a, Nov 21, 2018.

[6] L. Held and D.S. Bové. *Applied Statistical Inference: Likelihood and Bayes*. Springer Berlin Heidelberg, 2013.

[7] D.W. Hosmer, S. Lemeshow, and R.X. Sturdivant. *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley, 2013.

[8] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2013.

[9] J.S. Long, J.S. Long, and J. Freese. *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications, 1997.

[10] Raymond Navarro. *The 4 Factors of winning basketball games by Raymond Navarro*. https://www.mensbasketballhoopscoop.com/winning-basketball-games/.

[11] NBA. *NBA official website*. https://stats.nba.com/teams/boxscores-traditional/?sort=gdatedir=1.

[12] Pär Nyman. *Föreläsning 10, Regressionsdiagnostik och experimentell design, Avsnitt 1.1 Multikollinearitet*. http://www.parnyman.com/files/lectures/140919$_n otes.pdf, sep$192014.

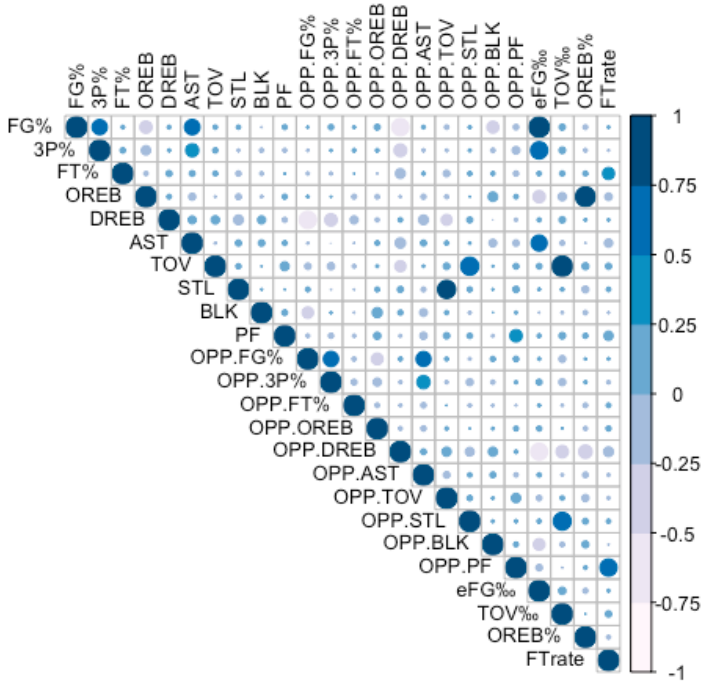
[13] Rosemary Powell. *De mest populära sporten i världen*. https://sv.ripleybelieves.com/most-popular-sports-in-world-39, Feb 2020.

[14] Rolf Sundberg. *Lineära statistiska modeller*. Kompendium, sida 70-72. Matematisk Statistik, Stockholms universitet, Oktober 2016.

[15] Fredrik Svensson. *Logistisk regression: genomförande, tolkning, odds ratio, multipel regression.* https://science.nu/lektion/logistisk-regression-genomforande-tolkning-odds-ratio-multipel-regression/, 2020.

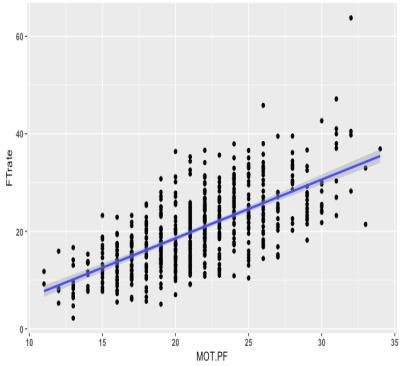
# A Bilaga

## A.1 Korrelationsmatris

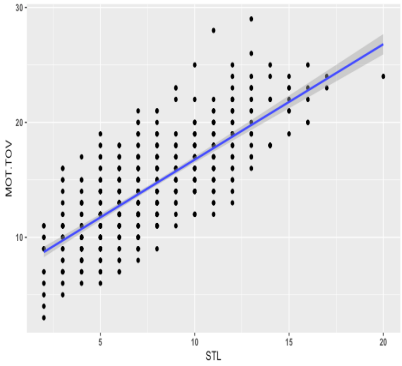



Figur 8: Korrelationsmatris

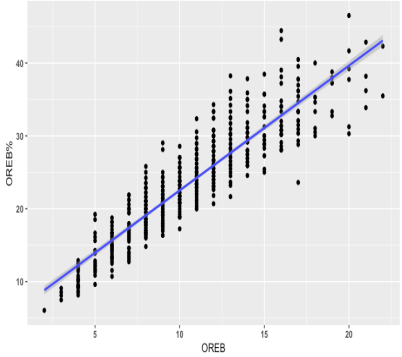
## A.2 Parvisa plottar




(a) FTrate - MOT.PF

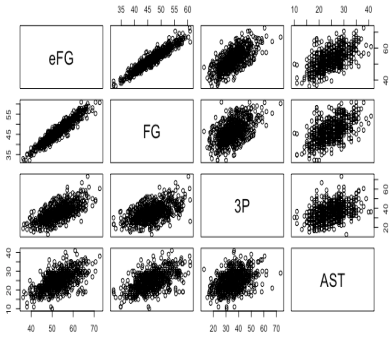



(b) MOT.TOV - STL

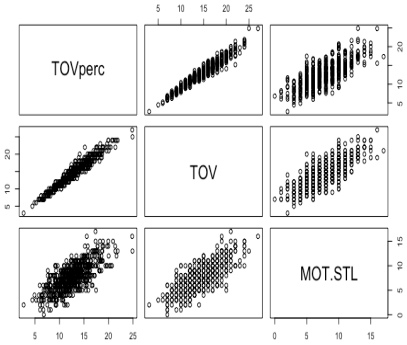



(c) OREB - OREB%

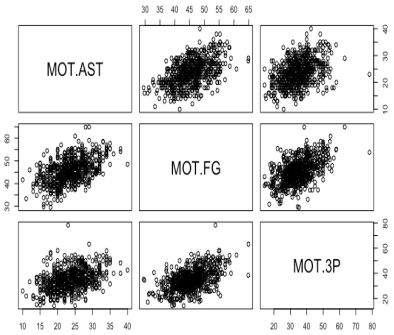
Figur 9: Parvisa plottar som påvisar linjära samband, där den blå linjen är den linjära regressionslinjen mellan variablerna. Se tabell 1 för beskrivning av parametrar.




(a) eFG% - FG% - 3P% - AST




(b) TOV% - TOV - MOT.STL




(c) MOT.AST - MOT.FG% - MOT.3P%

Figur 10: Parvisa plottar över variabler som påvisar linjära samband mellan fler än en variabel. Se tabell 1 för beskrivning av parametrar.

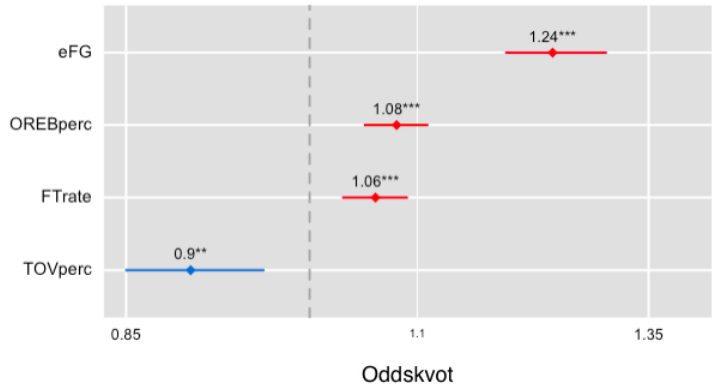
## A.3 Modell 1 - The four factors

**Modell 1**

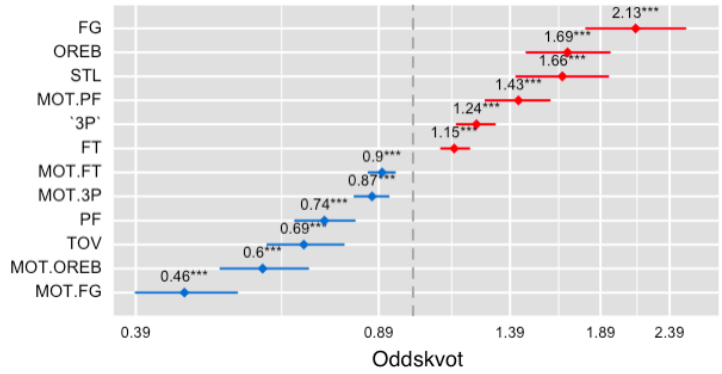
| Variabel | Estimering | Std.Fel | z värde | Pr( > \|z\|) | VIF |
|---|---|---|---|---|---|
| eFG% | 0.22 | 0.02 | 10.46 | $< 2{\cdot}10^{-16}$ | 1.12 |
| TOV% | -0.10 | 0.03 | -3.27 | 0.0010 | 1.07 |
| OREB% | 0.08 | 0.01 | 5.64 | $1.75{\cdot}10^{-8}$ | 1.06 |
| FT-rate | 0.06 | 0.01 | 4.44 | $9.14{\cdot}10^{-6}$ | 1.04 |

Tabell 12: De variabler som ingår i grundmodell 1, dess skattade koefficienter, standardfel, z-värde, p-värde och VIF-faktor. Se 2 för beskrivning av variabelnamn. Denna modell är baserad på Dean Olivers teorier.

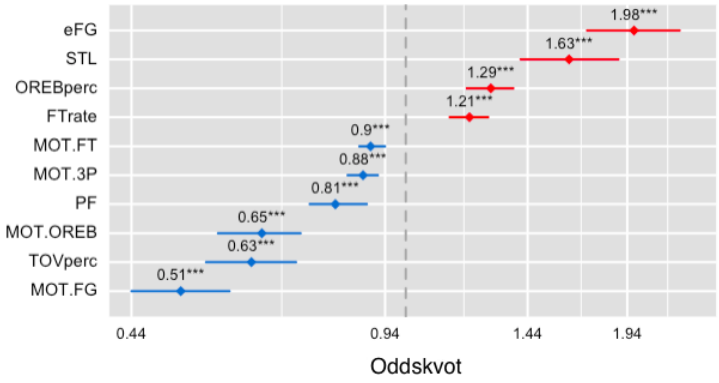
## A.4 Oddskvoter och Konfidensintervall




Figur 11: Visuallisering av de förklarande variablernas oddkvot och tillhörande konfidensintervall i modell 1.

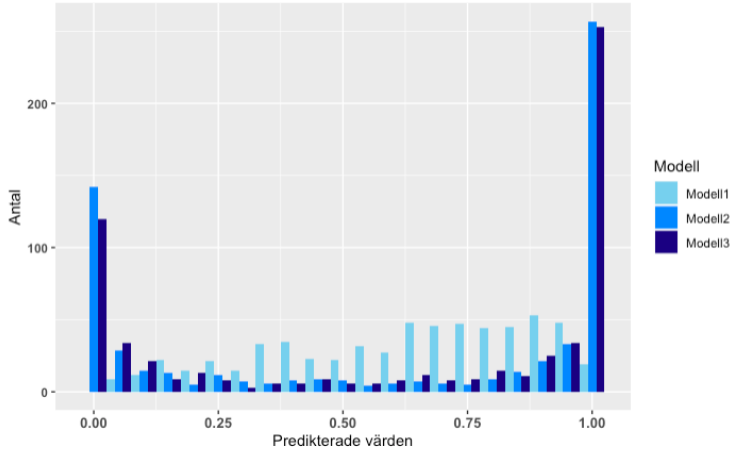



Figur 12: Visuallisering av de förklarande variablernas oddkvot och tillhörande konfidensintervall i modell 2.

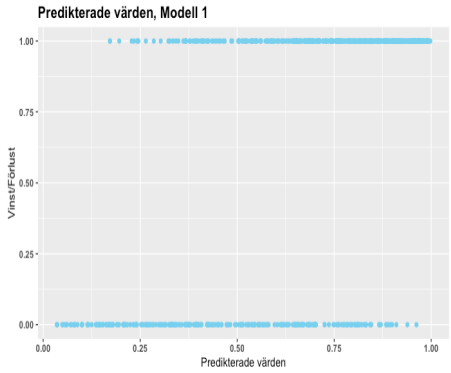



Figur 13: Visuallisering av de förklarande variablernas oddkvot och tillhörande konfidensintervall i modell 3.

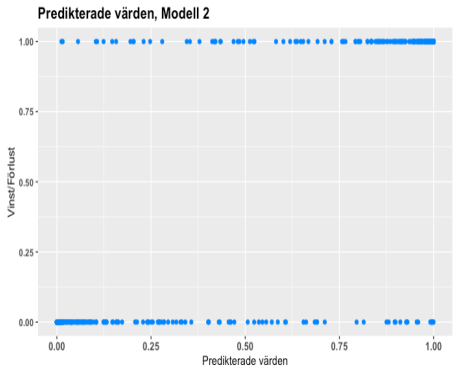
## A.5 Predikterade värden



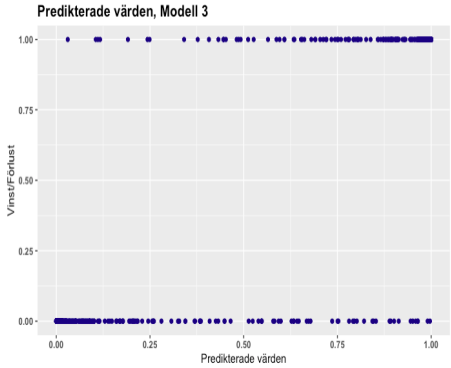

Figur 14: Modellernas predikterade värden. Vi ser att Modell 2 och Modell 3 har betydligt fler fall där värdena är nära 0 och 1 jämfört med Modell 1.




(a) Modell 1




(b) Modell 2




(c) Modell 3

Figur 15: Här plottas de predikterade värdena mot den binära responsvariabeln Vinst/Förlust. Vi kan se att Modell 2 och Modell 3 har högre koncentration av utfall i ändarna (nära noll och 1) på sina figurer jämfört med Modell 1 där utfallen är mer utspridda.