Analys av NFL drafting och faktorers inverkan

med oddsvärdering

Författare:
Pawandeep Dhanoa
dhanoa@kth.se

Handledare:
Jimmy Olsson

Maj 2014
Examensarbete inom farkostteknik, grundnivå, 15 hp, SA105X
Institutionen för Matematik, inriktning Matematisk Statistik
Kungliga Tekniska Högskolan

SAMMANFATTNING

Amerikansk fotboll är en välkänd sport i Amerika. Förutom utformningen och reglerna av spelet och underlaget det spelas på, finns även en andra skillnader som inte är direkt synliga. En skillnad som påpekas i denna studie är hur lagen inom sporten skaffar an sig spelare. Varje lag har en styrelse som tillsammans med lagets ledarstab avgör vilka spelare som är mest passande för laget. Dessa spelare rekryteras till laget efter att ha tagit examen från ett amerikanskt college. Det gör de möjligt till att värva in nya och unga spelare till truppen. Om laget vill ha andra typer av nya spelare finns det även en möjlighet att byta ut någon eller några spelare. Detta ses som ett handelssystem, där spelare byts ut mot andra spelare lagen emellan. Dessa två möjligheter till att värva nya spelare till laget kallas för *drafting*, vilket är en process för att rekrytera spelare. Den här studien fokuserar på rekrytering av spelare från college. I Amerika är det ett system som används till att värva både nya och gamla spelare till respektive lag inom *National Football League, NFL*.

Denna studie utvärderar och analyser de faktorer som grundar sig till hur varje lag inom NFL, rekryterar spelare från den amerikanska college ligan (NCAA). Faktorerna som analyserades var varje individuell spelares prestationer under college ligan. Med hjälp av logistisk analys kunde dessa faktorer sedan bestämmas. Detta gjordes genom att analysera oddset och dess förändring för varje spelares prestationer. Denna studie kom fram till att på varje sätt en positions spelare presterar under college ligan, så finns det viktiga faktorer som bidrar till att dessa spelare rekryteras till NFL.

Abstract

American football is a well-known sport in America. In addition to the design and rules of the game and the surface it is played on, there are also other differences that are not directly visible. One difference noted in this study is how the teams in the game provide players. Each team has a board of directors together with the team's leadership and they will determine which players are most suitable for the team. These players are recruited to the team after graduating from a U.S. college. This makes it possible to recruit new and young players to the squad. If the team wants other types of new players, there is also an opportunity to replace one or more players. This is seen as a trading, where players are replaced with other players between the two teams. These two opportunities to recruit new players to the team are called *drafting*, which is a process of recruiting players. This study focuses on the recruitment of players from college. In America, it is a system used to recruit both new and old players to their respective teams in the National Football League, NFL.

This study evaluates and analyzes the factors based on how each team in the NFL recruits players from the U.S. college league (NCAA). The factors analyzed were for instance each individual player’s performance during the college league, such as age, passing touchdowns and tackles. By using logistic analysis, these factors could then be determined by analyzing the odds and its change for each player's performance based on their positions in the team. This study shows that in every way a positional player performs during the college league, there will be important factors that contribute to recruitment to the NFL.

Förord

Jag vill tacka min handledare under detta kandidatexamensarbete, Jimmy Olsson; Universitetslektor inom KTH Matematik. För stort tålamod och förståelse kring min arbetssituation med kandidatuppsatsen.

Dessutom vill jag tacka Erik Castillo för tips och hänvisningar till områden inom detta ämne, som har möjliggjort det till en enklare undersökning.

Jag vill även påpeka att denna analys inte resulterar modeller då det endast granskas oddsförändringar på varje verkande faktor.

Pawandeep Dhanoa
Stockholm Maj 2014

Innehållsförteckning

1 Inledning

Det är viktigt att till en början få en förståelse till vad detta arbete kommer att handla om. Därför syftar det första kapitlet till att ge en bakgrund till området och sporten som denna analys kommer att fokusera på.

1.1 Bakgrund

Inom exempelvis klassisk europeisk fotboll som i amerikansk folkmun kallas för *"soccer"* finns det fotbollslag som kan värva respektive sälja en spelare genom att köpa eller sälja spelare per kontrakt. På andra sidan Atlanten, dvs. den Nord Amerikanska kontinenten är fotboll något som i vårt samhälle refereras till som amerikansk fotboll. Skillnaden mellan sporterna är väldigt stor då amerikansk fotboll påminner väldigt mycket om rugby. Förutom att sporterna skiljer åt sig väldigt mycket, så är också kontrakten och reglerna kring spelarna och hur dessa införskaffas till de olika lagen väldigt olika över kontinenterna. Detta arbete fokuserar på det amerikanska så kallade *"drafting" - rekryteringssystemet* kring spelarna. Arbetet kommer att grunda sig till vilka faktorer som påverkar *oddset* för att en ny spelare rekryteras till NFL från amerikanska universitet, s.k. college och detta kommer att beskrivas genom att analysera vilka faktorer som ökar en spelares odds att bli draftad.

1.2 NFL: National Football League

Inom amerikansk fotboll kallas den högsta elitligan för NFL som är en förkortning av National Football League. Man får inte förväxla amerikansk fotboll med den europeiska då amerikansk fotboll är snarlik rugbyn men varje enskild spelare har hjälm och diverse skyddsutrustning. Förutom dessa skillnader så är även regelverken olika mellan amerikansk fotboll och rugby. Det amerikanska drafting-systemet fungerar på så sätt att de nya enskilda spelarna som nu finns tillgängliga för NFL draftas mellan lagen. Med detta menas att exempelvis ifall en ny spelare tar steget från NCAA, National Collegiate Athletic Association, football till NFL, så draftas denna spelare av ett NFL lag. Vad innebär då att "draftas"? Drafting innebär att varje NFL-lag, får värva en spelare utan ett NFL kontrakt, direkt från college. Varje NFL-lag har rätt att välja 6 stycken spelare från NCAA football att drafta till sitt lag. Drafting-turordningen bestäms på så sätt att det lag som har presterat bäst i ligan under förgående säsong hamnar på sista plats i turordningen och det sämst presterande laget på första plats i turordningen. Varje lag väljer en spelare åt gången under 6 omgångar.

1.3 Avgränsningar

Undersökningen i detta examensarbete kommer att avgränsas till de *vanligaste draftade spelarna* baserade på positioner och vilka faktorer som står till grund för detta urval.

Detta görs för att hålla arbetet till en kandidatuppsats samt att tillgängliga databaser endast ger information inom specifika områden.

De vanligaste draftade spelare är:

FB – Fullback

QB – Quarterback

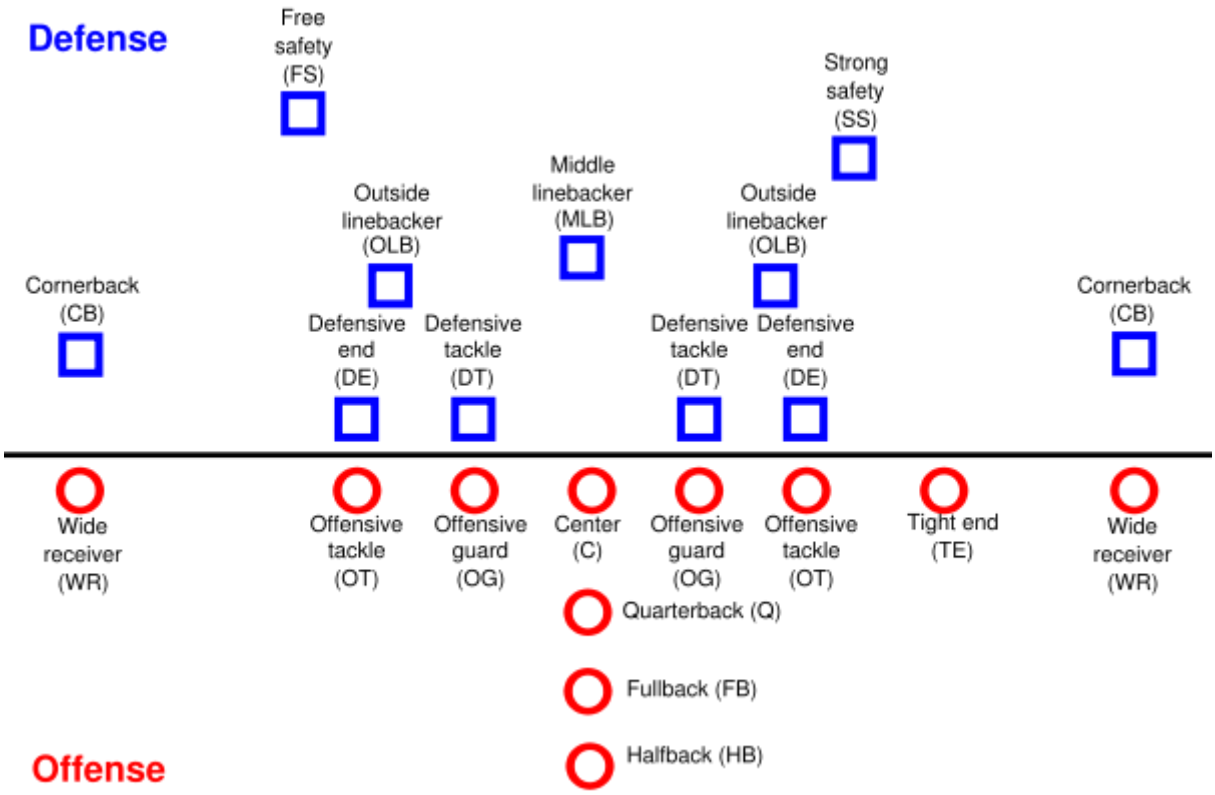
RB – Runningsbacks, omfattande såväl FB och Halfback (HB)

TE – Tight End

WR – Wide Receiver

DB – Defensive backs, som består av CB:s, FS:s, SS:s.

Samtliga spelares positioner finns utritade i *Figur 1.*



Figur 1 – Alla positioner inom amerikansk fotboll.

Arbetet utförs genom att analysera och utvärdera vilka faktorer som har störst oddsförändring till *oddset* för att en spelare blir rekryterad från universitetsnivå, till amerikansk fotboll på elitnivå.

2 Matematisk teori

Följande matematiska begrepp och termer kommer att spela en viktig roll i den kommande statistiska analysen.

2.1 Terminologi

Under denna del presenteras termer och begrepp som används inom denna rapport.

2.1.1 Oberoende variabler

Inom statistiken används väldigt ofta *oberoende variabler*, även kallade kovariater, som t e x $\boldsymbol{X}, X_i \; eller \; x_i$**.** Dessa variabler beskriver i sig andra variabler, vilket man oftast är intresserad av.

2.1.2 Beroende variabler

Med oberoende variabler beskriver man de sökta variablerna, som då är *beroende variabler.* Dessa betecknas som $\boldsymbol{Y}, Y_i \; eller \; y_i$. *En beroende variabel* beskrivs ofta inom statistiken med flera oberoende variabler, vilket då kallas för en multipel.

2.1.3 Koefficienter

För att kunna beskriva en variabels inverkan används konstanter som varje variabel multipliceras med, konstanterna kallas för *koefficienter* och betecknas med β_i. Skärningen (intercept) med $y - axeln$ betecknas med β_0.

2.1.4 Felterm

Feltermen betecknas med ϵ_i och kallas även för *residual* eller *slumpterm*. Det är ett mått på skillnaden mellan de observerade och beräknade värdet på den beroende variabeln. Det är även en term som inte kan förklaras rent generellt.

2.1.5 Signifikans

Inom statistiken används olika signifikansnivåer för att bestämma om variabler skall inkluderas eller inte. Det kan tänkas som ”ett område” med en viss *tillförlitlighet* man utför *tester* i och väl i detta måste man följa vissa matematiska regler. Det vanligaste använda ”området” är den med en signifikansnivå på 5 %, där tillförlitligheten då är på 95 %.

2.1.6 Dummyvariabel

Är en oberoende variabel som endast kan anta två värden, dvs. 0 och 1. Denna variabel inkluderas eftersom att det finns faktorer som behöver elimineras då dessa faktorer inte har någon inverkan på den beroende variabeln.

2.1.7 Strukturtolkning

Detta kretsar kring denna rubrik. En *strukturtolkning* utförs då man tittar på hur varje oberoende variabel påverkar den beroende variabeln. Det görs genom att titta på varje koefficient dessa oberoende variabler multipliceras med och på så sätt får man en uppfattning av varje faktors inverkan.

2.2 Linjär multipel regressionsanalys

Under denna del presenteras ett av det matematiska området som denna analys kommer att grunda sig kring. Därefter presenteras påbyggnaden av regressionsanalysen i ett senare avsnitt.

2.2.1 Multipla regressionsmodellen

Den vanligaste multipla regressionsmodellen beskrivs som:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ \ldots + \beta_n x_{ni} + \epsilon_i, \tag{2.1}$$

där y_i är beroende av variabeln x_{ni}. Koefficienterna $\beta_1, \beta_2, \ldots, \beta_n$ är konstanter framför varje respektive variabel som behöver skattas, kalibreras, från data för att modellen skall kunna användas praktiskt. Den sista termen ϵ_i är feltermen som inte kan förklaras och är avvikelsen från det observerade värdet och den skattade regressionslinjen.

På matrisform kan vi utrycka den multipla regressionsmodellen som:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon} \tag{2.2}$$

där

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in R^n$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1i} \\ 1 & x_{21} & x_{22} & \ldots & x_{2i} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{ni} \end{pmatrix} \in n \times (i+1)$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_i \end{pmatrix} \in (i+1) \times 1, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_i \end{pmatrix} \in i + \times 1 \, .$$

2.2.2 Ordinary Least Squares, OLS

Ett verktyg för att estimera koefficienterna $\beta_0, \beta_1, \dots, \beta_i$, med $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_i$, där symbolen över varje koefficient betecknar estimerat värde, kallas för OLS. Detta verktyg minimerar summan av de kvadrerade feltermerna. Det vill säga, om en ekvation som

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ \dots + \hat{\beta}_k x_k \tag{2.3}$$

erhålls, så minimerar OLS

$$\sum_{k=1}^{n} \epsilon_k^2 \tag{2.4}$$

vilket med $\epsilon_k = (y_k - \hat{y}_k)^2$ ger

$$\sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \tag{2.5}$$

Det kan tolkas som skillnaden mellan det observerade och estimerade värdet i kvadrat. Ju mindre skillnaden är, desto bättre är koefficienterna estimerade.

2.2.3 Antaganden

För att regressionsmodellen ska gälla, så bör följande kriterier vara uppfyllda.

1. Den beroende variabeln y_i ska kunna skrivas som en linjärkombination av de oberoende variablerna, med koefficienter β_i framför respektive x_{ni} och ytterligare en koefficient β_i. Slumptermen ska adderas till denna linjärkombination, se (2.1).

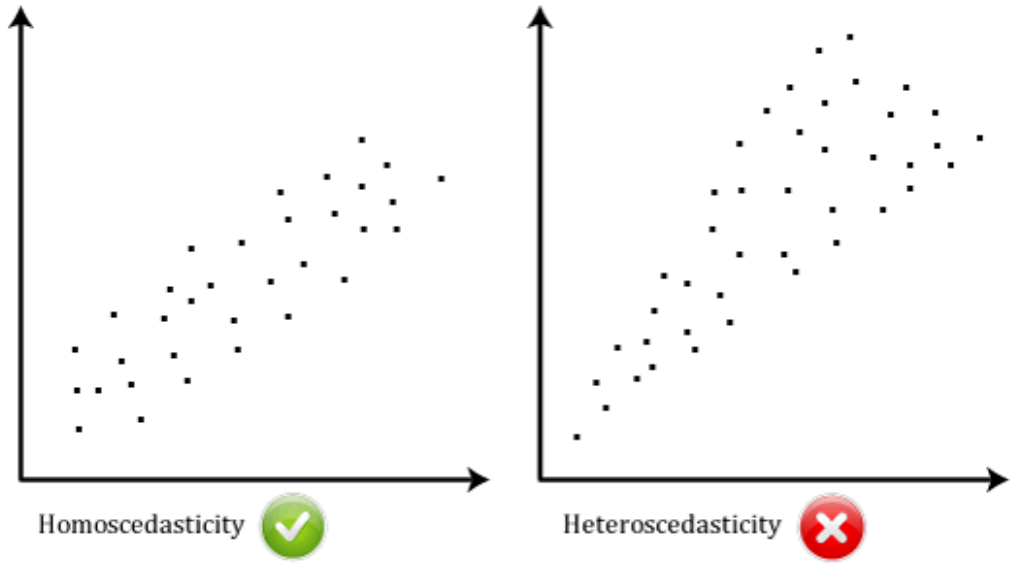
2. Det förväntade värdet på varje felterm ϵ_i är lika med noll, dvs.
$$E(\epsilon_i) = 0. \tag{2.6}$$
3. Alla feltermer ϵ_i har samma varians σ^2, (homoskedasticitet) dvs.
$$V(\epsilon_i) = \sigma^2. \tag{2.7}$$
4. Feltermerna är oberoende av varandra, dvs.
$$C(\epsilon_i, \epsilon_k) = 0. \tag{2.8}$$
5. Feltermerna ϵ_i är normalfördelade. Med kriterierna 2, 3, och 4 ovan fås att
$$\epsilon_i \sim N(0, \sigma^2). \tag{2.9}$$
6. De oberoende variablerna (kovariaterna) ska vara oberoende av varandra. Det ska inte gå att skriva dessa som linjärkombinationer av varandra (multikollinearitet).

2.3 Problem vid regressionsanalys

En regressionsanalys är aldrig felfri. Det kan uppstå fel i modellen p.g.a. att vissa antaganden inte är uppfyllda. Då finns det *tre* typer av fall som orsakar problem och dessa presenteras under detta avsnitt.

2.3.1 Heteroskedasticitet

Ifall feltermerna ϵ_i inte har konstant varians, dvs. om 3:e kriteriet från *antaganden* inte uppfylls så råder det *heteroskedasticitet*. Feltermerna har i detta fall olika varianser för alla observationer vilket påverkar standardavvikelsen i de sökta koefficienterna. Det kan upptäckas genom att plotta feltermerna mot de oberoende variablerna.



Figur 2 – Skillnaden mellan homoskedasticitet (eng. homoscedasticity) och *heteroskedasticitet* (eng. heteroscedasticity).

2.3.2 Endogenitet

Om en eller flera kovariater är korrelerade med feltermen uppkommer det *endogenitet*. Det leder till att OLS inte kan estimera korrekta koefficienter som tillhör respektive kovariater.

2.3.3 Multikollinearitet

Multikollinearitet innebär att det går att skriva en eller flera kovariater, som linjärkombinationer av varandra. Det finns två typer av fall för multikollinearitet, *perfekt* och *nästan perfekt (imperfekt) multikollinearitet.*

Vid *perfekt multikollinearitet* går det att skriva t ex en kovariat som en linjärkombination av en annan. Medan vid *imperfekt multikollinearitet* tillkommer det ytterligare en term som inte är en kovariat, men kan vara t ex en felterm.

När det kommer till OLS så går det inte att skatta koefficienterna när det råder *perfekt* multikollinearitet, men det går när fallet är *imperfekt* till en viss grad och även om vissa koefficienter skattas dåligt. Därför kan en modell bli effektivare ifall korrelerade variabler utelämnas.

2.4 Logistisk regression

En vanlig enkel multipel regressionsmodell kan inte beskriva denna undersökning eftersom att det sökta värdena, dvs. det den matematiska modellen ska beskriva, kan anta oändliga fler värden än bara två. Därför är det mer användbart att ta steget efter multipel regression vilket är multipel *logistisk regression.* Denna typ av metod är olinjär, men kan linjäriseras. Det är lämpligare att få ett resultat på den beroende variabeln som ligger på intervallet [0,1], för att sedan på ett finurligt sätt omvandla det till *odds* och *oddsförändringar.* Detta görs genom att tillämpa klassisk regression med logistisk regression.

2.4.1 Odds

Inom logistisk regression definieras *odds* eller *oddskvoten* som kvoten mellan sannolikheten att en händelse sker och inte sker. Alltså om p är sannolikheten att en händelse sker och $1 - p$ är sannolikheten att händelsen inte sker, fås

$$odds = \frac{p}{1-p}. \tag{2.10}$$

2.4.2 Oddsförändring

Oddsförändring definieras som kvoten mellan det nya och gamla oddset, dvs. om det gamla oddset är $odds_{11}$ och det nya $odds_{12}$ fås kvoten som

$$oddsförändring = \frac{odds_{12}}{odds_{11}}. \tag{2.11}$$

Oddsförändring beskriver hur mycket en analyserad variabel ökar (i procent) ifall de andra variablerna hålls konstanta. Det är ett viktigt begrepp inom detta arbete då analysen kommer att värdera oddsförändringar på respektive prestation.

2.4.3 Logiten

Sambandet mellan klassiska och logistiska regressionsmodellen kallas för *logiten av Y.* Den definieras som den naturliga logaritmen av oddsekvationen (2.10), dvs.

$$Y = ln\ (odds) = ln\left(\frac{p}{1-p}\right). \tag{2.12}$$

2.4.4 Sannolikheten, p

För att kunna bestämma oddsen av en händelse måste man först kunna bestämma sannolikheten för just den händelsen. Denna sannolikhet, p, kan lösas ur (2.12) som en funktion av den klassiska linjära regressionsmodellen.

Från tidigare är det givet att den beroende variabeln y kan uttryckas som en linjärkombination av ett antal oberoende variabler. Detta utnyttjas nu för att få den sökta funktionen.

Låt Y vara utfall, X_i vara kovariaterna (oberoende variabler) där $i = 1, 2, \dots, n$, och med $X_0 = 1$ samt feltermen ϵ. Då fås att

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots + \beta_n X_n + \epsilon \tag{2.13}$$

med (2.10) kan skrivas på formen,

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots + \beta_n X_n + \epsilon. \tag{2.14}$$

Med exponenten e^x skrivs (2.12) som,

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} \tag{2.15}$$

vart därefter sannolikheten p kan lösas ut.

Förenkling av parentesen,

$$p = (1-p)e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} \tag{2.16}$$

$$p = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} - pe^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} \tag{2.17}$$

omflyttning av termer

$$p + pe^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} \tag{2.18}$$

utbrytning av p

$$p\ (1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon}) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon} \tag{2.19}$$

$$p = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+ \dots+\beta_n X_n+\epsilon}}. \tag{2.20}$$

2.4.5 Maximum Likelihood Estimering (av β_i)

För att estimera de viktiga koefficienterna β_i inom logistisk regression, används *ML – Maximum Likelihood* metoden (eng. MLE- Maximum Likelihood Estimation). Metoden går ut på att estimera de sökta koefficienterna så att dessa maximerar det sökta utfallet.

Anta att det finns givna datavärden $x_1, x_2, \dots, x_n$ som är utfall av stokastiska variabler $X_1, X_2, \dots, X_n$. Dessa är oberoende samt lika fördelade och observationerna kommer från en okänd fördelning $f_X(x)$. Då finns det en parameter (koefficient) $\hat{\beta}$ som maximerar sannolikheten att erhålla de givna datavärdena, förutsatt att $\hat{\beta}$ ligger inom samma utfallsrum som $x_1, x_2, \dots, x_n$. Det går alltså att skatta en parameter så att sannolikheten att erhålla det sökta värdet maximeras.

För att illustrera detta hämtas exempel, 11.12 (s.256), ur *"Blom m.fl. Sannolikhetslära och statistikteori med tillämpningar, Studentlitteratur".*

Anta följande: x är en observation av en stokastisk variabel X, där $X \in Bin(n, p)$ och att sannolikheten p ligger i samma utfallsrum som kan anta värden $0 \leq p \leq 1$. Där även n antal observationer registrerats. Sannolikheten p skall skattas som $\hat{p}_{ML}$ och med *Maximum Likelihood* fås,

$$L(p) = p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{2.21}$$

som med logaritmering ger

$$\ln L(p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p). \tag{2.22}$$

Maximum ges där derivatan är noll. Genom derivering med avseende på p fås,

$$\frac{d \ln L(p)}{dp} = \frac{x}{p} - \frac{(n-x)}{1-p} = 0 \tag{2.23}$$

som resulterar

$$\hat{p}_{ML} = \frac{x}{n}. \tag{2.24}$$

2.5 Tester

Då koefficienter är estimerade, är det lämpligt att testa dessa med hjälp av statiska testverktyg. I detta avsnitt kommer det att presenteras *två* tester, som prövar signifikansen hos respektive koefficient och dessa två tester är sammankopplade med varandra samt även med klassisk regression.

2.5.1 Z test

Ett ***Z test*** är ett testverktyg på koefficienten β, och utförs på det med hjälp av den estimerade standardavvikelsen SE, enligt

$$Z = \frac{\hat{\beta}}{SE} \tag{2.25}$$

där Z motsvarar kvoten, som sedan jämförs med kvantilen på den signifikansnivå man väljer att utföra testet på. Nollhypotesen H_0 sätts alltid (oavsett signifikansnivå α) till att koefficienten $\beta = 0$.

T e x. Ifall $\hat{\beta}$ testas med ett *tvåsidigt* test på signifikansnivå $\alpha = 5\ \% = 0.05$,

fås kvantilen

$$\lambda_{0.025} = 1.96.$$

Nollhypotesen är som alltid

$$H_0\colon\ \hat{\beta} = 0.$$

Om beloppet av det erhållna Z värdet är mindre än kvantilen, dvs. om $|Z| < \lambda_{0.025}$, så förkastas inte nollhypotesen på signifikansnivå $\alpha = 5\ \%$, utan koefficienten kan inkluderas i analysen, varvid normalfördelat.

2.5.2 Wald– test

Ett ***Wald – test*** är väldigt lik Z – test, det är sammankopplat på sådant sätt att Z – värdet endast behöver kvadreras för att erhålla *Wald* – värdet. Alltså

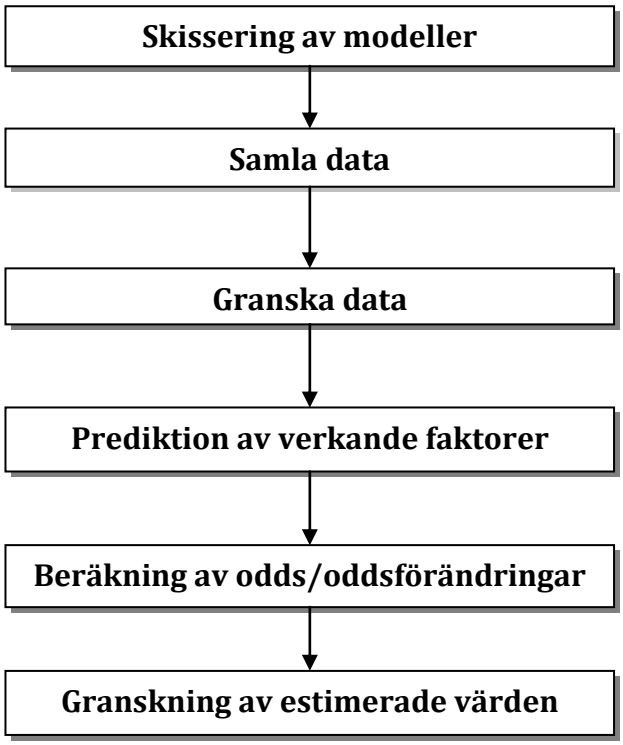
$$Wald = Z^2 = \left(\frac{\hat{\beta}}{SE}\right)^2 \tag{2.26}$$

varpå även kvantilen λ kvadreras (oavsett signifikansnivå α). Testet utförs även på samma vis om Z testet. Men ifall nollhypotesen är sann, så tillhör den testade koefficienten en χ^2 *(chi2)* – fördelning till skillnad från *Z test*, där koefficienten istället tillhör en normalfördelning.

3. Metod

Det är inte så lätt att undersöka och analysera viktiga faktorer inom det amerikanska drafting systemet enbart multipel logistisk regressionsanalys, p.g.a. av att matematiken inte kan tillämpas till hundra procent. Ur vissa aspekter dyker det upp situationer inom drafting-systemet som är rimligt intuitivt, men inte överensstämmer med den statistiken och dess tester. Därför är det viktigt att granska felen som uppstår inom statistiken väldigt noga och kombinera intuitionen med logiken. För att avgöra ifall vissa aspekter som går emot resultaten från de statiska testverktygen överensstämmer med verkligheten, även om matematiken inte kan beskriva den till hundra procent. Man bör alltså vara väl medveten om felen som kan uppstå vid studier och analyser. Det är en av de viktiga orsakarna till varför detta examensarbete endast kommer att syfta till vilken faktor som ger upphov till *oddsförändring*.

För att göra arbetet smidigare så följs ett blockschema på hur analysen skall gå tillväga:



3.1 Skissering av modeller

För att kunna ha något att grunda sitt arbete kring så är det viktigt att ha en utgångspunkt att utgå från, det kan vara t ex en frågeställning, undersökning eller en problemställning. I detta arbete syftar grunden till en matematisk *logistisk regressionsmodell*, som skall beskriva de största och viktiga faktorerna över en spelares odds att bli draftad (värvad) av ett NFL lag, direkt från college ligan (NCAA). Men att ta fram en modell är inte relevant i detta arbete då de granskade variablerna kan beskrivas med enbart *oddsförändring.* Dock måste det läggas en grund, för att kunna ha nått att utgå ifrån.

En vanlig enkel multipel regressionsmodell, kan inte beskriva denna undersökning eftersom att den sökta utfallet, dvs. det den matematiska modellen ska beskriva, kan anta oändliga fler värden än bara två. Därför är det mer användbart att ta steget efter och titta på oddset av att en händelse sker.

En sådan modell ser ut på formen:

$$ln(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots + \beta_n x_n \qquad (3.1)$$

där kovariaterna $x_1, x_2 \ldots x_n$ är variabler som utfallet (oddset) är beroende av och betakoefficienterna $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ är de sökta koefficienterna som beskriver varje variabels inverkan på utfallet.

Vid denna utgångspunkt har det därför skisserats fram modeller intuitivt för respektive spelarpositioner, där alla granskade faktorer finns inom samtliga modeller.

3.2 Samla data

Inom statistiken är det viktigt att ha en bra mängd data eftersom att hela analysen bygger på dessa data. Skulle datavärdena vara inkonsekventa skulle hela modellen (som bygger på datavärden) bli instabil och inte hålla för eventuella analyser.

En bra källa för att hämta data ifrån är en amerikansk hemsida som har statistik lagrade för flera amerikanska sportligor (*se källförteckning)*. Statistiken beskriver varje individuell spelares prestationer, ålder, positioner etc. Under förgående säsong i NCAA och deras drafting positioner. Eftersom att dessa värden endast fanns på *html* format, så behövde allting föras över till en Microsoft Excel fil, så att det sedan kunde implementeras i **IBM SPSS Statistics 22**. Även om det finns datavärden som går 50 – 60 år bakåt hämtades det ändå bara datavärden från 2013 och 20 år bakåt. Eftersom under dessa årtionden skedde det en hel del förändringar i NFL branschen. Bl.a. utökades antal lag i ligan och det resulterade antal utökade värvningar med åren. Det var viktigt att ta hänsyn till detta, därför att annars leder till stora problem med analysen då multikollinearitet är ett problem som uppträder väldigt ofta inom logistisk regressions.

3.3 Granskning av data

Vid detta block av diagrammet börjar själva analysen. För att kunna utvärdera och analysera drafting oddsen för en spelare, baserad på en spelare position och så är det lämpligt att kontrollera efter vilka variabler som skall representera vad och hur detta ska gå ihop med de matematiska teorierna detta arbete speglar kring.

Eftersom de insamlade datavärdena inte är komplett för vissa spelares statistik, så begränsas därför modellerna till de allra vanligaste valda spelarna, baserat på deras positioner. T ex så finns det mer värden (statistik) för en Quarterback, än en Offensive guard (OG). Med hjälp av dummy variabler och filtrering så tas värden på de andra spelarpositionerna bort, eftersom att de inte har någon inverkan på analysen av detta arbete. Ifall detta in tas till hänsyn finns det risk för att analysen inte kommer att resultera de sökta faktorerna. Relevanta problem som tas upp i avsnitt 2.3 kan uppstå och därför skall det till högsta pris undvikas att stöta på problem som finns listade under detta avsnitt.

3.4 Prediktion av verkande faktorer

Då data är granskad och filtrerad är det viktigt att använda logiken till att fundera över vilka koefficienter som söks och deras inverkan på dess beroende variabel (utfallet).

Även här är det bra att skissera över vilka faktor som kan förväntas sig ha en inverkan på respektive positions spelare. Med hjälp av kunskaper och lära om sporten går det att få en ungefärlig uppfattning över vilka koefficienter som bör förväntas. Detta kan underlätta för en då koefficienterna kan bli lättare att tolka och utvärdera, det sista som en vill erhålla är resultat som är helt bortom rimliga gränsvärden. T e x så vet man att för en för en Wide Receiver (WR), har faktorer som *rushing yards* (antal sprungna yards med boll) och *receiving yards* (antal sprungna yards utan boll), en stor inverkan och alltså är det rimligt att koefficienterna framför dessa variabler är snäppet större än de som har en mindre inverkan. Det är just kring dessa banor som tankesättet bör spegla så att analysen inte far mot värden som inte kan förklaras med varken ord eller matematik.

3.5 Beräkning av odds/oddsförändringar

Efter att ha predicerat fram koefficienter så återstod det att beräkna fram de med hjälp av *Multinomial Logistic Regression* i SPSS. Eftersom denna funktion även har en inbyggd modellanpassare så underlättar det för användaren att utvärdera sina koefficienter. Det tar ett tag att sätt sig in i hur programmet arbetar men det finns instruktioner att följa på bl.a. tillverkarnas hemsida.

Eftersom detta arbete kretsade kring att oddsvärdera fem stycken spelarpositioners inverkande faktorer så har det vid denna del av blockdiagrammet, tagits fram just det i SPSS. Innan det utförs analyser på de fem spelarpositionerna måste den beroende variabeln y kodas om i SPSS. Det görs genom att sätta värdet *1* på den beroende variabeln och där *"1"* betyder att *spelaren blir draftad.* Då syftet är att analysera de inverkande faktorer som påverkar en positionsspelares chans att bli draftad, d.v.s. vilka faktorer som ger bidrag till att spelaren blir draftad. Därefter implementeras detta för varje spelare.

För erhålla de sökta oddsförändringarna så måste först sannolikheten beräknas fram enligt (2.20) där koefficienterna skattas med SPSS (enligt *Maximum Likelihood*). De β-koefficienter som finns listade under resultat för varje spelarposition är däremot de logaritmerade oddsförändringar. Då detta utförs i SPSS måste det väljas en referenskategori. Denna referenskategori valdes till DB, vartefter oddsförändringarna beräknas givet att DB:s redan är draftade. Varför just DB:s väljs som referenskategori är för att dessa spelare ingår i ett lags defensiv och ett lag byggs alltid från bakåt till framåt. Dessutom minimeras problem som ger upphov till multikollinearitet.

Då sannolikheterna är kända för varje inverkande faktor är nästa steg att beräkna oddsen och dess förändringar, ifall en liten variation sker i någon av de oberoende och granskade faktorerna. Detta uppnås genom att utnyttja (2.10) och (2.11), vilket SPSS då resulterar fram i tabellformat.

För att undersöka om varje oberoende variabel med en estimerad koefficient var lämplig att inkludera i analysen utfördes det därefter Wald – test på varje estimerad koefficient.

Nollhypotesen sattes alltid (oavsett värdet på den estimerade koefficienten) till att $\beta = 0$ på signifikansnivå $\alpha = 0.05$. Därefter utfördes det tester som det beskrivs i avsnitt 2.5.

3.6 Granskning av koefficienter

Det sista stadiet i denna studie än den allra viktigaste. Då alla tester är gjorda är det dags att utvärdera vilka faktorer som har den största inverkan på varje spelares positioner. Att en koefficient har passerat ett matematiskt test och anses vara lämplig, betyder det inte alltid att den är ”rimlig”. Värdet på koefficient kan t e x vara skyhög i förhållande till andra värden trots att den har testats som lämplig. Därför är det bra att ha en grund att gå på. Den grunden är i detta arbeta att läsa på och studera hur varje person, på dessa positioner spelar. Utifrån det kan man som sista steg, dra slutsats om en variabel har den inverkan på oddset såsom spelaren agerar i verkligheten. Därför går en stor del av arbetet till att granska och tolka det estimerade koefficienterna med beräknade oddsförändringar och deras bidrag till studien.

För att lämna ut icke relevanta inverkande faktorer på respektive spelare användes först och främst resultaten från Wald – testen. Alla variabler vars koefficient gav $P - värde$ under signifikansnivå α ansågs som signifikanta och dokumenterades. De faktorer vars koefficienter *inte* gav upphov till oddsförändring och de som inte ansågs vara signifikanta ($P - värde$ över α) utelämnades ur analysen.

4 Resultat

Under detta avsnitt finns resultat tabellerade för respektive spelarpositioner, med hjälp av SPSS. Antal observations data, dvs. nollskilda värden var 398 (av 5012). Resterande data bestod av nollor eller värden tillhörande andra spelare. I SPSS användes den inbyggda dummy kodningsfunktionen för varje kategori och så att utfallet bestod positioner med värdet 1.

De variabler som hade mest inverkan på respektive spelare togs med och dess koefficienter som statistisk lämpliga. Lämpligheten testades med *Wald – test* och de som var statistisk signifikanta, d.v.s. de som gav ett $P - värde$ (p – value) under signifikansnivån $\alpha = 0.05$ dokumenterades. Samtliga resultat presenteras med tre decimalers noggrannhet.

4.1 Variabler

Nedan följer varje variabelbeteckning och dess representation.

V1	*Positions*	*Spelarpositioner (FB, QB, RB, TE och WR)*
V2	*Age*	*Ålder*
V3	*Starts*	*Startade matcher*
V4	*Passes completed*	*Lyckade passningar*
V5	*Passes attempted*	*Passnings försök*
V6	*Yards gained by passning*	*Erhållna passnings yards*
V7	*Passing touchdowns*	*Passningar ledda till touchdowns*
V8	*Interceptions throwns*	*Passningar ledda till motståndare*
V9	*Rushing attempts*	*Försök till språng*
V10	*Rushing yards gained*	*Erhållna yards vid språng*
V11	*Rushing touchdowns*	*Språng ledda till touchdowns*
V12	*Receptions*	*Mottagningar*
V13	*Receivning yards*	*Erhållna yards innan mottagning*
V14	*Receivning touchdowns*	*Mottagningar ledda till touchdowns*
V15	*Tackles*	*Tacklingar*
V16	*Sacks*	*Kontringar på motståndarnas speluppbyggnad*
V17	*College/University*	*Universitet*

Tabell 4.1 – Variabler som användes i analysen.

4.2 Resultat för varje *spelarposition*

Med koefficienter (logaritmerade oddsförändringar) β, standardavvikelse $S.A$ och anti-logaritmerade värden e^{β} fås resultaten presenterade från *tabell 4.2 – tabell 4.8.* β-koefficienterna utgör de logaritmerade värdena i den logistiska regressionen. De säger inget mer än att positiva värden ger *ökat* odds ifall en liten variation sker på respektive variabel. Däremot anger de anti-logaritmerade värdena e^{β} hur mycket oddset ökar i procent, då det är en kvot mellan det nya samt det gamla odds värdet, ifall en liten

ändrig sker på respektive variabel. T.ex. om $e^{\beta} = 1{,}612$ så betyder det att oddset på dess variabel (i detta fall V11). ökar med 61,2 % om det sker en liten variation på variabeln.

Här näst följer samtliga resultat för varje spelarposition och dessa faktorer som hade störst inverkan på oddsen.

4.2.1 Resultat för en Fullback

FB	β	$S.A.$	$Wald$	$P-$ värde	e^{β}
V2	1,523	9,095	,028	,037	4,585
V3	2,269	11,862	,037	,038	9,669
V11	,478	2,021	,056	,013	1,612
V17	1,527	53,913	,001	,041	4,604

Tabell 4.2 – Resultat för en Fullback och dess faktorer.

4.2.2 Resultat för en Quarterback

QB	β	$S.A.$	$Wald$	$P-$ värde	e^{β}
V2	,754	3,120	,058	,009	2,126
V4	,042	,154	,076	,043	1,043
V5	,007	,084	,006	,037	1,007
V7	,057	,863	,004	,047	1,059
V8	,089	,959	,009	,026	1,094
V17	,776	29,116	,027	,049	2,173

Tabell 4.3 – Resultat för en Quarterback och dess faktor.

4.2.3 Resultat för en Runningback

RB	β	$S.A.$	$Wald$	$P-$ värde	e^{β}
V2	,827	3,108	,071	,019	2,287
V3	,306	2,922	,011	,034	1,358
V9	,314	,583	,289	,046	1,368
V10	,434	1,191	,133	,017	1,543
V12	,568	1,060	,287	,008	1,770
V15	,173	,718	,058	,009	1,188
V17	1,143	41,663	,001	,028	3,135

Tabell 4.4 – Resultat för en Runningback och dess faktorer.

4.2.4 Resultat för en Tight End

TE	β	$S.A.$	$Wald$	$P-$värde	e^{β}
V2	,773	2,017	,156	,040	2,161
V11	,148	1,074	,019	,019	1,160
V12	,136	,211	,416	,039	1,146
V15	,173	1,276	,018	,032	1,188
V17	,347	2,997	,011	,048	1,414

Tabell 4.5 – Resultat för en Tight End och dess faktorer.

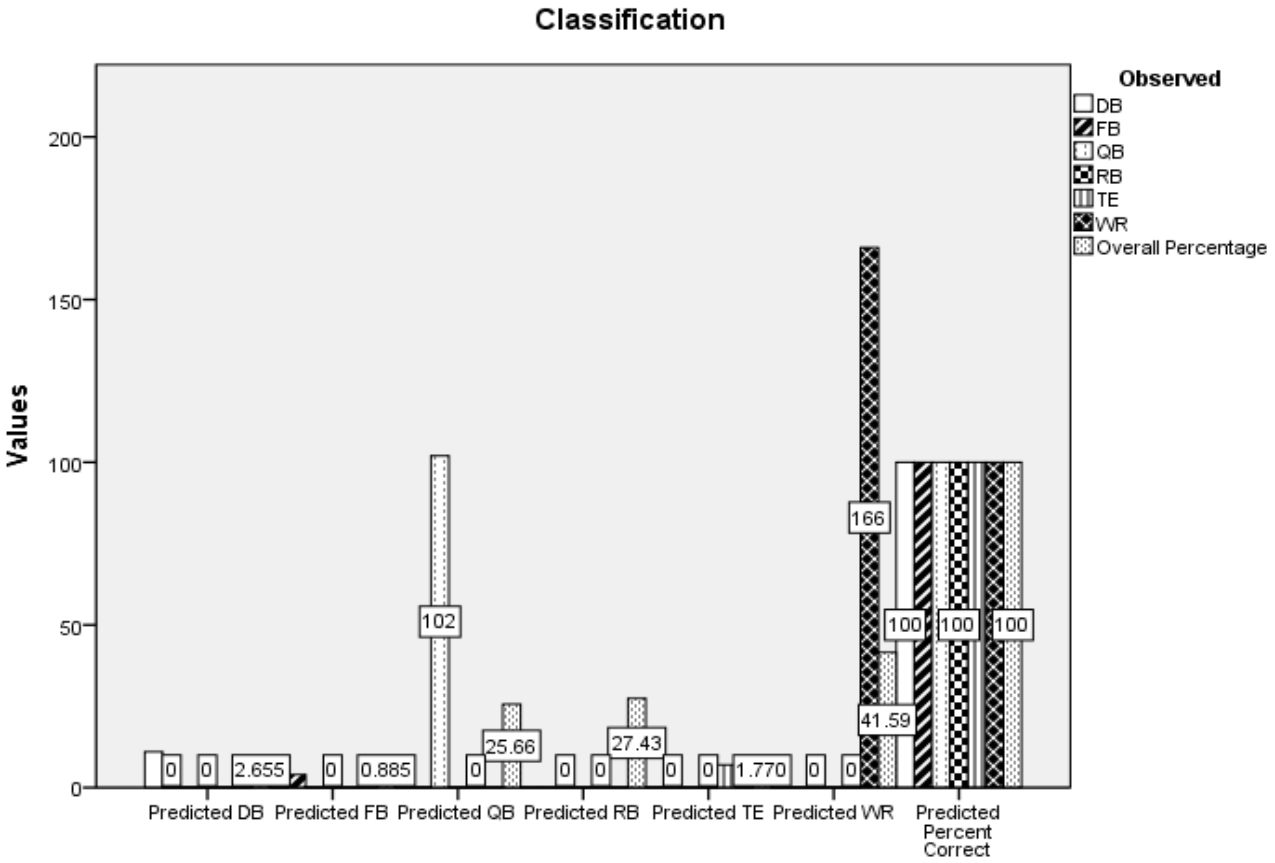
4.2.5 Resultat för en Wide Receiver

WR	β	$S.A.$	$Wald$	$P-$värde	e^{β}
V2	,856	3,065	,078	,043	2,355
V10	,848	27,649	,000	,037	2,334
V11	,375	37,969	,000	,028	1,455
V12	,104	25,567	,000	,050	1,109
V14	2,310	48,168	,002	,029	10,373
V17	,208	41,026	,003	,036	1,232

Tabell 4.6 – Resultat för en Wide Receiver.

Noterbart är det att V14 ger stort upphov till en odds ökning hos en Wide receiver. Det kan tolkas som lämpligt eftersom en Wide receivers främsta uppgift är att ta emot passningar från sina medspelare. För att sedan med hjälp av löpningar erhålla (touchdown) poäng åt laget.

Här näst visas det i *figur 3* hur bra SPSS anpassade sina estimerade koefficienter med samt hur väl dessa koefficienter kan förutse odds.



Figur 3 – Anpassning av de skattade koefficienter och dess oddsvärden.

5 Diskussion

Under detta avsnitt diskuteras arbetet och resultatet över studien.

5.1 Diskussion om *studien*

Denna studie kunde ha utförts på ett smidigare sätt ifall man redan var bekant området innan arbetet påbörjades. En del andra teorier (t ex logistisk regression) behövde läsas in och förstå för att kunna tillämpa de på denna analys.

Det upptäcktes multikollinearitet i början av analysen. Vilket anses vara väldigt vanligt inom logistisk regression. Detta resulterade i att värdena på standardavvikelserna var orimliga i förhållande till de andra avvikelserna.

5.2 Diskussion om *Metod*

Då problemställning var satt från början så behövdes det hämtas material och data över sporten, matematiken och programmet SPSS. Då det från början inte räckte med att titta på dessa sökta faktorer med enbart klassisk regressionsanalys, vilket gjorde att det ägnades tid åt att få insikt på logistisk regression och förstå det grundligt. Detta resulterade i att modellerna kunde skisseras i början av blockschemat.

Data värdena, som erhölls över NFL drafting, var inte lika begriplig i längden som den först ansågs vara i början av analysen. Trots att det fanns många värden, så saknades det även en del värden. Det var värden som var nollor och vissa element saknade även siffror. Som i sin tur gav problem för analysen som skedde i SPSS till en början. Dessutom så upptäcktes det att antalet lag hade utökats i NFL i början av 2000-talet och gav också upphov till multikollinearitet, eftersom vissa kategorier påverkade andra kategorier.

En annan stor tidsåtgång gick till att granska de erhållna datavärdena. Det kunde ha gjorts bättre eftersom de värden som plockades ut av alla erhållna var för lite. En tumregel inom statistiken är att alltid ha en bra uppsättning av data. Antalet värden som erhölls var inte tillräckliga för att utföra en ordentlig analys, vilket var önskemålet från början. Detta tyder på att hemsidan inte gav tillräckligt många datavärden från tidigare drafting säsonger.

Då problemen med logistisk regression åtgärdades, stämde logiken och resultaten överens med varandra, det vill säga de förväntade oddsfaktorerna på varje dominanta variabel för respektive position. Man kan på så sätt säga att de förväntade oddsfaktorerna stämde överens med verkligheten då man väl känner till varje spelare uppgift inom sporten. En WR:s främsta uppgift är att springa mot änden av motståndarens planhalva och ta emot en passning på vägen dit eller vid änden, från en medspelare, som oftast är QB:en. Därför är det helt logiskt att dominanta oddsfaktorerna för en WR bör vara *receptions, rushing yards, receivning touchdowns* etc.

Att använda SPSS som en nybörjare tar tid. Med tanke på alla inbyggda funktioner i varje kommando tar det ett bra tag innan man väl sätter sig i det. Till en början så orsakade det stora problem eftersom en annan stor del av analysen gick åt att förstå programmets kommandon och dessa inbyggda funktioner. Ifall detta inte gjordes i början, resulterade det att värden som beräknades fram på oddset var helt orimliga med logiken och teorin. Trots detta gav programmet inte några varningar om t e x avvikelserna hos respektive skattade koefficient.

Det allra sista stadiet var allra värt att ägna tiden åt, eftersom en person inte förlitar sig till siffror på samma sätt som en dator. Tiden kunde ha ägnats sig mer åt den sista delen av blockdiagrammet. Därför att det är det allra viktigaste då en studie av viktiga faktorer utförs. Därför kan det vara bra att välja ett område med en tydlig problemställning som kan analyseras med hjälp av de kunskaper en redan har tillgodo. På så sätt spar man tid på de mindre viktiga områden av arbetet och en stor del av tiden går då åt de allra viktigaste, vilket i denna studie är att analysera och tolka de värden som SPSS estimerar fram.

5.3 Diskussion om *Resultat*

Resultaten hos varje spelare ansågs vara ganska logiska och rimliga. Dock så kunde något mer korrektare värde ha beräknats fram ifall antal data värden hade varit fler. Med hjälp av noggranna metoder, erhölls ändå logiska korrekta resultat. De variabler som togs med i tabellerna, var det som hade störst inverkan på respektive spelare. De resterande gav inga förändringar på oddset och ansågs då utelämnas. Alla relevanta faktor gav en odds ökning hos varje spelare och den allra högsta fanns under WR, som hade en förändring på närmare 10 gånger det ursprungliga värdet på den representerade variabeln.

En annan faktor som kan påpekas är att college/universitet också har en inverkan på alla de analyserade positionerna. Med andra ord, vilket collegelag en spelare representerar har också en inverkan på oddset gällande alla de analyserade positionerna.

Trots att data kunde ha omstrukturerats ännu mer, t e x genom att dela in olika collegelag i kategorier, så estimerade ändå SPSS väldigt bra. Detta kan ses i *figur 3*.

6 Slutsats

Att kunna tillämpa matematiken inom en sportbransch, gjorde denna studie väldigt intressant. Kärnan med själva studien var att analysera de mest inverkande faktorer som ökar oddset till att en spelare rekryteras till Amerikansk fotboll på elitnivå från universitetsnivå. Det denna studie då kom fram till var att specifika spelare, baserat på deras positioner, har viktiga faktorer som resulterar förändringar på oddset till att värvas till NFL från NCAA. Det spelar alltså ingen roll vilken position man än spelar på, utan de största faktorerna är varje spelares uppgift på den amerikanska fotbollsplanen, d.v.s. hur väl de presterar i förhållande till sin uppgift. Oddsen för att värvas till elitnivå baseras på hur väl spelarna presterar i förhållande till sin uppgift.

6.1 Förslag till vidare studie

Det som kan vara intressant att titta på utifrån oddssättning av respektive verkande faktor, är att med hjälp av *logiten av y*, undersöka sannolikheten att en spelare på blir draftad med hänsyn till prestationer under collegeligan. Då kan det även vara bra att ha med andra verkande faktor, såsom skador och vikt på respektive spelare. Då detta ska göras är det viktigt att påpeka att ju fler faktorer studien involverar, desto mer datavärden behövs det. Detta är för att få en bra precision på de estimerade koefficienterna samt att undvika problem som alltid uppstår vid regressionsanalys även om det är klassisk linjär- eller logistisk regression. Dessutom skulle det även vara intressant att se vad de resterande positions spelare, som inte ingick i denna studie, har för verkande faktorer.

Källförteckning

Litteratur

Peter Kennedy – A Guide to Econometrics, 6th edition; Blackwell Publishing 2008

Harald Lang – Topics on Applied Mathematical Statistics, v. 0.97 Nov. 2013; KTH Teknikvetenskap

Damodar N. Gujarati – Basic Econometrics, 5th Rev edition; McGraw Hill Higher Education 2008

Blom m.fl. – Sannolikhetslära och statistikteori med tillämpningar, Upplaga 5:10; Studentlitteratur 2013
(Exempel 11.12 s.256)

Internet

Data över NFL drafting
http://www.pro-football-reference.com/draft
hämtad 28 mars 2014

Logistisk regression
http://www.jmg.gu.se/digitalAssets/1307/1307026_Nr_62_Logistisk_regression.pdf
hämtad 19 april 2014

Logistisk regression
http://userwww.sfsu.edu/efc/classes/biol710/logistic/logisticreg.htm
hämtad 19 april 2014

Logistisk regression
http://math.bu.edu/people/nkatenka/MA116/Week5Lecture3.pdf
hämtad 20 april 2014

Logistisk regression
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1065119
hämtad 20 april 2014

Figurer

Figur 1
http://www.peaux-rouges.fr/les-postes.html
hämtad 28 april 2014

Figur 2
http://stats.stackexchange.com/questions/76151/what-is-an-intuitive-explanation-of-why-we-want-homoskedasticity-in-a-regression
hämtad 28 april 2014