Stockholms universitet

Att med multinomial logistisk regression förklara sannolikheter i fotbollsmatcher

Sebastian Rosengren

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2012:6
Matematisk statistik
September 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2012:6**,
http://www.math.su.se/matstat

Att med multinomial logistisk regression förklara sannolikheter i fotbollsmatcher

Sebastian Rosengren*

September 2012

Sammanfattning

I denna uppsats studerar vi möjligheten att genom multinomial logistisk regressionsmodeller förklara sannolikheterna för de olika utfallen i fotbollsmatcher. Dessa sannolikheter antas förklaras, genom två logitfunktioner, av en uppsättning förklarande variabler. Vi undersöker sedan hur väl dessa modeller presterar i prediktionssammanhang. Fotbollsmatcherna som analyseras i denna uppsats är matcher ifrån Svenska Spels Stryktipset, då det till varje sådan match finns en uppsjö av tänkbara förklarande variabler. Flera tillvägagångssätt används för att producera modeller, och den slutgiltiga modellen beror på två förklarande variabler och förklarar sannolikheterna på ett sätt som överensstämmer väl med data.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: sebastian.rosengren@hotmail.se. Handledare:Mikael Petersson.

Förord och tack

Detta är en kandidatuppsats som leder till en kandidatexamen i matematisk statistik våren 2012. Jag skulle vilja tacka min handledare Mikael Pettersson samt Anders Björkström för värdefull hjälp och rådgivning.

Abstract

In this report we will study the possibillity that through multinomial logistic regression explain the probabilities for the different outcomes in a footboll game. These probabilities are assumed to be explained by a set of explanatory variables through two logit functions.

We use multiple approaches to produce our models, and the final model depends on two explanatory variables and explains the probabilities in a way that is coherent with the data.

Sammanfattning

I denna uppsats studerar vi möjligheten att genom multinomial logistisk regressionsmodeller förklara sannolikheterna för de olika utfallen i fotbollsmatcher. Dessa sannolikheter antas förklaras, genom två logitfunktioner, av en uppsättning förklarande variabler. Vi undersöker sedan hur väl dessa modeller presterar i prediktionssammanhang.

Fotbollsmatcherna som analyseras i denna uppsats är matcher ifrån Svenska Spels Stryktipset, då det till varje sådan match finns en uppsjö av tänkbara förklarande variabler.

Flera tillvägagångssätt används för att producera modeller, och den slutgiltiga modellen beror på två förklarande variabler och förklarar sannolikheterna på ett sätt som överensstämmer väl med data.

Innehåll

1 Inledning

Stryktipset är ett populärt spelformat av Svenska Spel. Det går ut på att man ska tippa så många rätt som möjligt i 13 fotbollsmatcher, som främst spelas i den engelska ligan. Själva spelformatet är dock inte intressant för vårat ändamål. Då Stryktipset är så populärt så finns det extremt mycket information till varje match på kupongen, se beskrivning av datamaterial. Idén är att det ska finnas predikterande kraft i denna information, som vi sedan kan använda för att beskriva sannolikheterna för varje utfall i en match. För att beskriva utfallens sannolikheter används en multinomial logistisk regressionsmodell då en sådan modell ofta lämpar sig för modellering av kategorisk data.

Målet med uppsatsen är att undersöka om man med en multinomial logistiskt regressionsmodell kan förklara sannolikheterna för utfallen i en fotbollsmatch på ett lämpligt sätt.

2 Teori

2.1 Multinomial logistisk regression

Antag att vi har en diskret responsvariabel Y som kan anta ett av tre värden: 1, X, eller 2. Till denna responsvariabel tillhör en uppsättning förklarande variabler $x_1, x_2, ..., x_k$. För att utveckla den multinomiala modellen väljs nu en kategori som referenskategori, som de andra kategorierna ska jämföras mot. Kutym är att den vanligast förekommande kategorin (1, X, 2) tas som referens. Med kategori 1 som referenskategori är följande den multinomiala logistiska regressionsmodellen

$$f_1(x) = \log \frac{P(Y = 2|x)}{P(Y = 1|x)} = \beta_{10} + \beta_{11} \cdot x_1 + \beta_{12} \cdot x_2 + ... + \beta_{1k} \cdot x_k \qquad (1)$$

$$f_2(x) = \log \frac{P(Y = X|x)}{P(Y = 1|x)} = \beta_{20} + \beta_{21} \cdot x_1 + \beta_{22} \cdot x_2 + ... + \beta_{2k} \cdot x_k \qquad (2)$$

Ovanstående tillsammans med villkoret att $P(Y = 1|x) + P(Y = X|x) + P(Y = 2|x) = 1$ ger följande ekvivalenta modell

$$P(Y = 1|x) = \frac{1}{1 + e^{f_1(x)} + e^{f_2(x)}}$$

$$P(Y = 2|x) = \frac{e^{f_1(x)}}{1 + e^{f_1(x)} + e^{f_2(x)}}$$

$$P(Y = X|x) = \frac{e^{f_2(x)}}{1 + e^{f_1(x)} + e^{f_2(x)}}$$

Modellen är ofta lämplig i situationer då responsvariabeln är kategorisk dels därför att funktionen $\frac{e^{f(x)}}{1+e^{g(x)}+e^{h(x)}}$ rent matematisk sätt är mycket flexibel, dels för att relationen mellan parametrar och oddskvoter blir mycket enkel. Det gäller nämligen att

$$\Omega = \frac{P(Y = j|x_1, ..., x_i + 1, ..., x_k)/P(Y = 1|x_1, ..., x_i + 1, ..., x_k)}{P(Y = j|x_1, ..., x_i, ..., x_k)/P(Y = 1|x_1, ..., x_i, ..., x_k)}$$

$$= e^{\beta_{ji}}$$

Detta samband är den största anledningen till varför logistiska regressionsmodeller är ett så kraftfullt analytiskt verktyg.

Givet en uppsättning data $(y_i, x_{1i}, x_{2i}, ..., x_{pi}),\ i = 1, 2, ..., n$ så skattas modellens $2 \cdot (p+1)$ parametrar med maximum likelihood-metoden. Om vi låter $y_{1i} = 1$ om $Y_i = 1$ och $y_{xi} = 1$ om $Y_i = X$, samt $y_{2i} = 1$ om $Y_i = 2$ så blir likelihoodfunktionen för data

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} P(Y_i = 1|x)^{y_{1i}} \cdot P(Y_i = X|x)^{y_{xi}} \cdot P(Y_i = 2|x)^{y_{2i}}$$

Denna funktion maximeras numeriskt och maxarg(L($\boldsymbol{\beta}$)) tas som parameterskattningar.
Modellen kan självklart utvidgas till kategorivariabler med fler än tre kategorier, men ovanstående teori räcker för att tillgodogöra sig resultaten i denna rapport. För mer om multinomial logistisk regression se [3].

2.2 χ^2-test

När man vill testa en hypotes H_0 mot en alternativ hypotes H_A där $H_0 \bigcup H_A$ beror på r fria parametrar, och H_0 tilldelar värden till k av dessa parametrar, så gäller det att

$$-2\log(\Lambda) \approx \chi^2(k), \quad \Lambda = \frac{\sup_{\boldsymbol{\beta} \in H_0} L(\boldsymbol{\beta})}{\sup_{\boldsymbol{\beta} \in H_0 \bigcup H_A} L(\boldsymbol{\beta})}$$

Denna statistika kan alltså användas för att testa H_0 mot H_A genom att förkasta H_0 om $-2\log(\Lambda) > \chi^2_\alpha(k)$ där α är testets signifikansnivå.

Detta test används för att testa om vissa parametrar eller delmängder av parametrar är lika med noll i multinomiala logistiska regressionsmodeller, se [2] för mer om detta.

2.3 AIC

Aikaikes informationskriterium, AIC, är ett relativt mått på hur bra modellen passar data. Vanligtvis gäller det att AIC $= 2k - 2\log(L(\hat{\boldsymbol{\beta}}))$ där k är antal parametrar i modellen. Enligt en viss statistisk teori, se [1], så ska man ur en uppsättning möjliga modeller välja den modell som har lägst AIC-värde. En modells AIC-värde säger alltså ingenting på egen hand utan ska bara användas för att jämföra mot andra modeller. I denna rapport används AIC i stor utsträckning för att jämföra olika modeller.

3 Beskrivning av datamaterial

Datamaterialet består av tre dataset, som används i olika steg av analysen. Alla dataset består av matcher, med tillhörande förklarande variabler, från stryktipset. Responsvariabeln är i samtliga fall självklart matchens utfall 1, X, eller 2. I nedanstående tabell så sammanfattas de förklarande variablerna.

1	X1	Andel av svenska folkets pengar som ligger på hemmavinst
2	XX	Andel av svenska folkets pengar som ligger på kryss
3	X2	Andel av svenska folkets pengar som ligger på bortavinst
4	Y1	Svenska Spels odds för hemmavinst, precis innan matchstart
5	YX	Svenska Spels odds för kryss, precis innan matchstart
6	Y2	Svenska Spels odds för bortavinst, precis innan matchstart
7	T1	Antal av tio tidningar som tror på hemmavinst
8	TX	Antal av tio tidningar som tror på kryss
9	T2	Antal av tio tidningar som tror på bortavinst
10	H1	Antal av de fem senaste matcherna som hemmalaget vunnit
11	HX	Antal av de fem senaste matcherna som hemmalaget kryssat
12	H2	Antal av de fem senaste matcherna som hemmalaget förlorat
13	B1	Antal av de fem senaste matcherna som bortalaget vunnit
14	BX	Antal av de fem senaste matcherna som bortalaget kryssat
15	B2	Antal av de fem senaste matcherna som bortalaget förlorat
16	VH	Andel matcher som hemmalaget hittills vunnit på hemmaplan
17	OH	Andel matcher som hemmalaget hittills kryssat på hemmaplan
18	PH	Genomsnittlig antal poäng hemmalaget hittills tagit per match
19	VB	Andel matcher som bortalaget hittills vunnit på hemmaplan
20	OB	Andel matcher som bortalaget hittills kryssat på hemmaplan
21	PB	Genomsnittlig antal poäng bortalaget hittills tagit per match

3.1 Allmänt

Matcher där det saknas data eller då lagen inte har spelat fem matcher än är borttagna ifrån dataseten. Data är hämtad från sajten www.warnsater.se, se [4]. All analys och simulering är gjord i R och Matlab.

3.2 Dataset 1

Dataset 1 består av 358 matcher ifrån stryktipset säsongen 2010, till varje match finns det värden på samtliga förklarande variabler. Den inledande analysen görs på detta dataset.

3.3 Dataset 2

Dataset 2 är en utökning av Dataset 1, och består av 756 matcher ifrån stryktipset säsongerna 2010-2011. Till varje match finns det nu, som konsekvens av den inledande analysen, färre förklarande variabler.

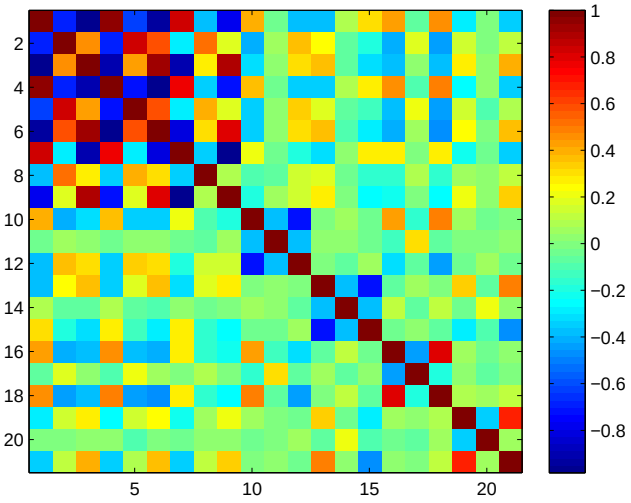
3.4 Dataset 3

Dataset 3 består av 495 matcher från stryktipset säsongen 2009, där det till varje match tillhör samma förklarande variabler som figurerar i dataset 2. Dataset 3 används för att testa olika modellers prediktionskraft. I rapporten används ett mått som vi kallar för prediktionsmått, och med det menar vi andelen korrekt tippade matcher i dataset 3, givet att vi tippar på den match som modellen säger har störst sannolikhet att gå in.

4 Statistisk modellering och analys av data

4.1 Inledande undersökning

Följande analys sker på data ifrån dataset 1. Vi börjar med att konstatera att många av de förklarande variablerna är starkt korrelerade, då vissa är ett mått på samma sak. Till exempel så avspeglar folkets proportioner, svenska spels odds, och tio tidningars tips alla troligheten för ett visst utfall i en match. Figur 1 ämnar att illustrera korrelationen i data, där mörka färger är stark positiv eller stark negativt korrelation. En konsekvens av att data är så korrelerat är att det blir svårt att veta om en viss variabel ska ingå i en modell eller inte. Eftersom det blir vanskligt att dra slutsatser ifrån signifikanstest, då variabler som bör ingå i modellen kan bli ej signifikant skilda ifrån noll som konsekvens av korrelationerna.



Figur 1: Korrelationskarta

Första åtgärden blir att sålla bort de variabler som inte verkar förklara sannolikheterna för de olika utfallen. Detta görs genom att en enkel multinomial logistisk regressionsmodell anpassas för varje enskild förklarande variabel. En variabel bedöms som överflödig och tas bort från fortsatt analys om P-värdet vid ett χ^2-test om variabeln ingår signifikant i modellen ($H_0 : \beta_{11} = \beta_{21} = 0$) är större än 0.25. Talet 0.25 är framtaget av Hosmer och Lemeshow, [3]. Att det inte är det sedvanliga 0.05 beror på att man misstänker att en variabel på egenhand inte förklarar tillräckligt men att den kan ingå i en delmängd av variabler som tillsammans förklarar sannolikheterna för utfallen bra. Denna gräns blir rimlig i vårat fall då variablerna H1 till PB (nummer 10-21) kan tänkas ha ovanstående egenskap. Då data

är så korrelerat så blir även denna metod för variabelreduktion lämplig. I nedanstående tabell återfinns P-värdena för de olika modellerna.

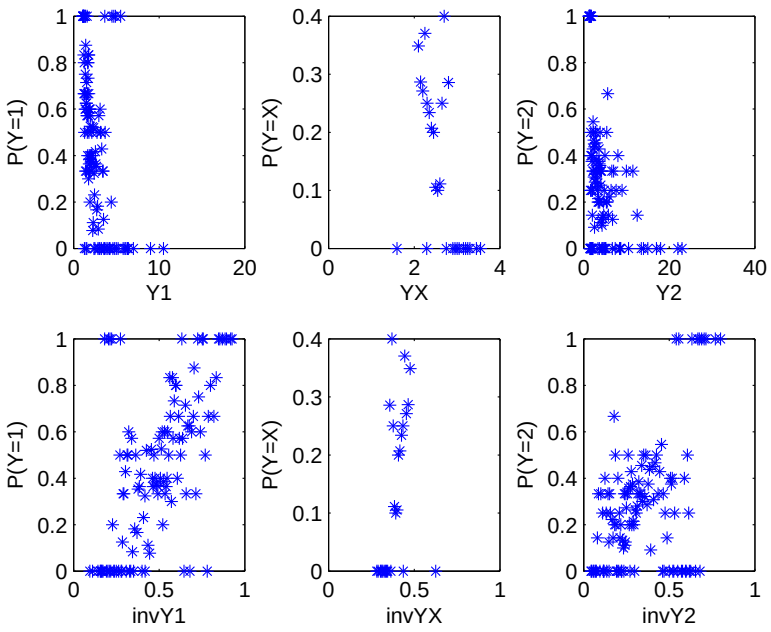
Var	P-värde
X1	2.31e-9
XX	2.76e-5
X2	1.17e-8
Y1	3.36e-10
YX	4.22e-6
Y2	6.88e-11
T1	1.84e-6
TX	0.07592
T2	5.69e-6
H1	0.00241
HX	0.53250
H2	0.00030
B1	0.07430
BX	0.26300
B2	0.04181
VH	4.61e-5
OH	0.84009
PH	4.65e-6
VB	0.75890
OB	0.25200
PB	0.61610

4.2 Inledande åtgärder

Följande variabler klarade inte signifikanskravet och utesluts nu ur den fortsatta analysen: HX, BX, OH, VB, OB, PB.

Med ett reducerat antal förklarande variabler finns det här möjlighet att utöka dataset 1, vilket görs och modelleringen sker nu på dataset 2. Att dataset 1 kan utökas med fler observationer faller oss i smaken, dels därför att det är bättre med fler observationer, dels därför att det är en standardåtgärd för data som är starkt korrelerat.

Då oddset för ett utfall ska vara ungefär ett genom sannolikheten för det utfallet, så undersöker vi om denna transformation är värd att göra. I Figur 2 är oddset samt det inverterade oddset plottat mot skattade sannolikheter. Då det finns ett mer linjärt samband mellan de inverterade oddsen och de skattade sannolikheter så transformerar vi oddsen och använder de inverterade oddsen i den fortsatta analysen. Att ett linjärt samband är att föredra beror på hur modellen är beskaffad, den beskriver helt enkelt linjära samband bättre än inverterade. Så Y1, YX, Y2 betecknar nu de inverterade oddsen.



Figur 2: Odds och inverterade odds plottade mot skattade sannolikheter

Vi använder tre olika tillvägagångssätt för att bygga vår modell. Det första sättet är ett mer analytiskt, där vi utför statistka test och tittar på en mängd olika mått för att bedöma modellämplighet. Det andra och tredje är mindre analytiska, där vi med ren datorkraft producerar en modell med ett så högt prediktionsmått som möjligt samt en med ett så lågt AIC-värde som möjligt.

4.3 Tillvägagångssätt 1

Vi delar upp mängden förklarande variabler i två kategorier, Informationskälla 1 och 2, I1 och I2. Anledningen är att variablerna X1 till T2 är prediktioner gjorda av andra och är därför ett mer direkt mått på sannolikheterna för de olika utfallen, medan H1 till B2 är mått på lagets form och hur bra laget varit hittills i säsongen. Analysen och modellbyggandet fortsätter nu med att två modeller anpassas, en baserad på I1 och en baserad på I2. Vi undersöker sedan om dessa två modeller kan kombineras till en bättre modell. Anledningen till detta tillvägagångssätt är att modellbyggandet blir mer hanterbart, då data från I1 är starkt korrelerade och bör därför få extra tillsyn utan andra störande faktorer. Samtidigt får man fram två lämpliga modeller från vardera informationskälla som tillsammans kan producera en bättre modell.

4.3.1 Modeller baserade på I1

Korrelationen i informationskälla 1 illustreras i nedanstående korrelationsmatris. Vi ser att variablerna som är ett mått på troligheten för samma utfall, till exempel X1, Y1, T1, är särskilt korrelerade. Detta föreslår att det är olämpligt att ha med mer än en av dessa trolighetsmått för varje utfall, då dessa skulle förklara mycket av varandra och därmed inte ingå med signifikanta effekter i modellen.

	X1	XX	X2	Y1	YX	Y2	T1	TX	T2
X1	1.00	-0.70	-0.96	0.95	-0.63	-0.94	0 .83	-0.38	-0.78
XX	-0.74	1.00	0.48	-0.69	0.81	0.59	-0.30	0.50	0.18
X2	-0.96	0.48	1.00	-0.90	0.45	0.94	-0.91	0.27	0.89
Y1	0.95	-0.69	-0.90	1.00	-0.72	-0.98	0.76	-0.34	-0.71
YX	-0.63	0.81	0.45	-0.72	1.00	0.57	-0.30	0.37	0.21
Y2	-0.94	0.59	0.94	-0.98	0.57	1.00	-0.81	0.30	0.78
T1	0.83	-0.30	-0.91	0.76	-0.30	-0.81	1.00	-0.36	-0.96
TX	-0.38	0.50	0.27	-0.34	0.37	0.30	-0.36	1.00	0.10
T2	-0.78	0.18	0.89	-0.71	0.21	0.78	-0.96	0.10	1.00

Vidare gäller följande linjära samband:

$$\begin{aligned} \text{X1+XX+X2} &= 100 \\ \text{T1+TX+T2} &= 10 \\ \text{Y1+YX+Y2} &\approx 1.26 \end{aligned}$$

Att Y1+YX+Y2 ≈ 1.26 syns i datamaterialet och detta betyder att, under vissa antaganden, den förväntade andelen pengar Svenska Spel ger tillbaka till sina kunder är ungefär $\frac{1}{1.26} = 0.79$, se appendix för motivering till detta.

Vi kan alltså inte basera vår modell på dessa tre uppsättningar variabler, då matrisen som behöver inverteras för att hitta ML-skattningarna av parametrarna blir singulär. Vi har nu reducerat antalet modeller att jämföra till 24 stycken, ty du kan välja den första förklarande variabeln på tre olika sätt, samma för den andra och den tredje, men då måste dra bort ovanstående tre modeller så $3^3 - 3 = 24$. I nedanstående tabell hittas modellernas P-värde vid test av att alla parametrar är lika med noll, modellernas AIC-värde, samt andelen korrekta prediktioner av matcher från dataset 3.

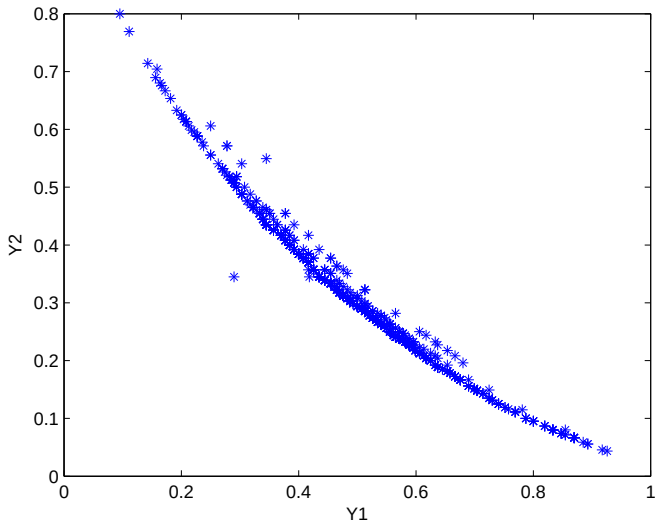
Model	P-värde	AIC	Prediktionsmått
(Y1, YX, X2)	<2.22e-16	1535.54	0.4636
(Y1, YX, T2)	<2.22e-16	1534.88	0.4575
(Y1, XX, Y2)	<2.22e-16	1532.03	0.4636
(Y1, XX, T2)	<2.22e-16	1535.46	0.4717
(Y1, XX, X2)	<2.22e-16	1533.13	0.4656
(Y1, TX, Y2)	<2.22e-16	1525.08	0.4575
(Y1, TX, T2)	<2.22e-16	1529.63	0.4535
(Y1, TX, X2)	<2.22e-16	1527.75	0.4595
(X1, YX, Y2)	<2.22e-16	1534.07	0.4372
(X1, YX, X2)	<2.22e-16	1539.16	0.4595
(X1, YX, T2)	<2.22e-16	1541.84	0.4474
(X1, XX, Y2)	<2.22e-16	1543.50	0.4676
(X1, XX, T2)	8.48e-16	1555.26	0.4717
(X1, TX, Y2)	<2.22e-16	1540.80	0.4575
(X1, TX, X2)	3.08e-16	1548.30	0.4656
(X1, TX, T2)	6.56e-16	1549.88	0.4636
(T1, TX, X2)	6.09e-16	1559.42	0.4575
(T1, TX, Y2)	<2.22e-16	1540.23	0.4555
(T1, XX, X2)	5.25e-16	1554.26	0.4534
(T1, XX, Y2)	<2.22e-16	1542.68	0.4656
(T1, XX, T2)	1.30e-13	1551.32	0.4615
(T1, YX, Y2)	<2.22e-16	1556.09	0.4534
(T1, YX, X2)	<2.22e-16	1541.66	0.4615
(T1, YX, T2)	<2.22e-16	1535.73	0.4555

Vi ser att modell (Y1, TX, Y2) har lägst AIC-värde, samt ett relativt högt prediktionsvärde. Modellerna (Y1, XX, T2) och (X1, XX, T2) har högst prediktionsvärden, men deras AIC-värde är såpass mycket högre än föregående modell och övervägs därför inte för fortsatt analys. Vi går alltså vidare i analysen med modellen (Y1, TX, Y2).

I Figur 3 hittas parameterskattningar för ovanstående modell. Görs ett χ^2-test av hypotesen $H_0 : \beta_{1i} = \beta_{2i} = 0$ så erhålles signifikanta P-värden för $i = 1, 2$ men ett P-värde på 0.0750 för $i = 3$, alltså det är troligt att Y2 inte ska ingå i modellen, se (1) och (2) för indexering. Plottas Y1 mot Y2, se Figur 4, så ser vi att dessa variabler förklarar mycket av varandra, detta tillsammans med ovanstående test gör att vi utesluter Y2 ur modellen.

```
Coefficients :
               Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  11.145757   3.665841  3.0404 0.0023624 **
x:(intercept)   6.062126   3.216415  1.8847 0.0594641 .
2:Y1          -16.483205   4.419743 -3.7294 0.0001919 ***
x:Y1           -9.285455   3.779848 -2.4566 0.0140271 *
2:TX           -0.323198   0.105857 -3.0531 0.0022646 **
x:TX           -0.093158   0.095405 -0.9764 0.3288445
2:Y2          -10.668316   4.761656 -2.2405 0.0250608 *
x:Y2           -5.723247   4.299146 -1.3313 0.1831062
```

Figur 3: (Y1, TX, Y2) - parameterskattningar



Figur 4: Y1 plottat mot Y2

Utökas modellen med en samspels effekt visar det sig att det inte är troligt

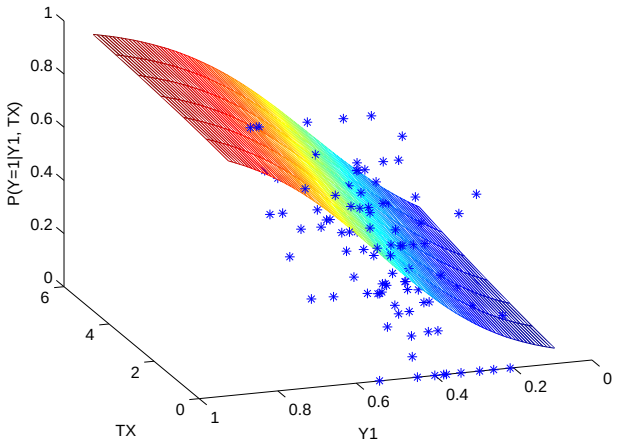
att en sådan effekt existerar, P-värdet vid motsvarande χ^2-test blir 0.1670.

Modellen (Y1, TX) har ett AIC-värde på 1526.27 alltså större än (Y1, TX, Y2)-modellen, men prediktionsvärdet på modellen är 0.4636 som är ett bättre värde än (Y1, TX, Y2)-modellens. Dessutom ser vi att parameterskattningarna i (Y1, TX)-modellen är mer stabila sett till deras standardavvikelser, alltså modellen är mer robust och accepteras som den slutgiltiga ifrån uppsättningen modeller ifrån I1.

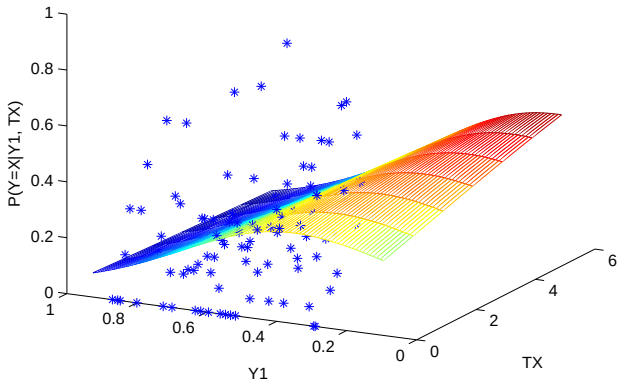
```
Coefficients :
                Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  3.106344   0.425409  7.3020 2.836e-13 ***
x:(intercept)  1.955643   0.415707  4.7044 2.546e-06 ***
2:Y1          -7.002033   0.812168 -8.6214 < 2.2e-16 ***
x:Y1          -4.622303   0.759108 -6.0891 1.135e-09 ***
2:TX          -0.276789   0.103219 -2.6816  0.007328 **
x:TX          -0.083419   0.094886 -0.8791  0.379321
---
```

Figur 5: (Y1, TX) - parameterskattningar

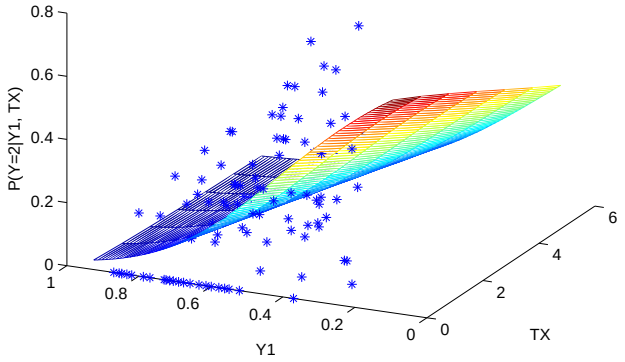
I Figur 6, 7, och 8 så plottas de tre sannolikhetsfunktionerna tillsammans med de skattade sannolikheterna. De skattade sannolikheterna baseras på medelvärdet av indikatorvariabler som har samma Y1 och TX värde, och för att graferna ska bli mer tydliga är de sannolikheter som är baserade på färre än tre observationer borttagna. Vi ser att modellen och de skattade sannolikheterna överensstämmer.



Figur 6: (Y1, TX) plottade mot P(Y=1|Y1,TX)



Figur 7: (Y1, TX) plottade mot P(Y=X|Y1,TX)



Figur 8: (Y1, TX) plottade mot P(Y=2|Y1,TX)

För att få en bättre bild av hur väl modellen förklarar sannolikheterna för de olika utfallen gör vi en tabell. I vänstra kolumnen hittas modellsannolikheterna för utfallen och i de högra så hittas de observerade andelarna matcher som gick in i den kategorin, samt hur många matcher som ingick i kategorin.

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	0.11	9	0.00	4	0.15	52
10-20	0.21	48	0.04	71	0.24	161
20-30	0.23	94	0.23	306	0.28	274
30-40	0.34	148	0.36	371	0.33	165
40-50	0.42	180	0.00	2	0.49	55
50-60	0.57	159	-	0	0.24	38
60-70	0.65	57	-	0	0.11	9
70-80	0.69	32	-	0	-	0
80-90	0.93	27	-	0	-	0

Sett till Figur 6, 7, och 8 samt till ovanstående tabell så slår vi fast att modellen lämpar sig för att förklara sannolikheterna för de olika utfallen i fotbollsmatcher. Vi undersöker nu mer djupgående hur väl modellen predikterar matcher ifrån dataset 3. Detta görs på samma sätt som i ovanstående tabell.

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	0.15	13	-	0	0.03	36
10-20	0.28	29	0.24	63	0.24	135
20-30	0.27	49	0.28	159	0.25	168
30-40	0.42	98	0.31	272	0.30	86
40-50	0.40	124	-	20	0.38	34
50-60	0.47	94	-	0	0.43	23
60-70	0.60	32	-	0	0.83	12
70-80	0.70	43	-	0	-	0
80-90	0.75	12	-	0	-	0

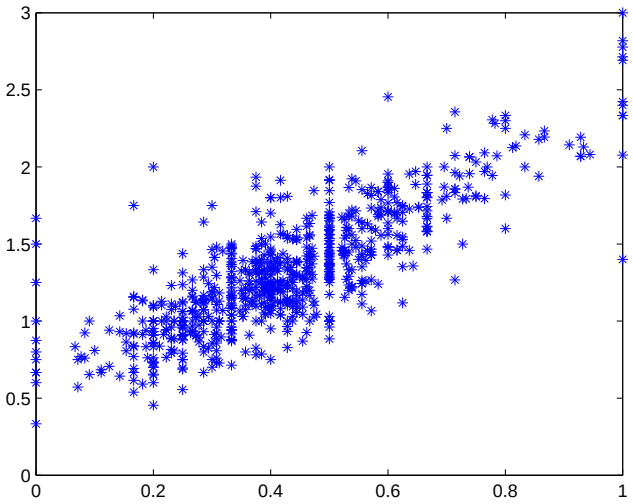
Modellen fungerar även för prediktion, men har en tendens att överskatta sannolikheten för hemmavinst.

4.3.2 Modeller baserade på I2

Vi börjar modellbyggandet med att ansätta en modell som innehåller alla förklarande variabler ifrån I2, och i Figur 9 sammanfattas modellen.

```
Coefficients :
                Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  1.8923982  0.6751198  2.8031 0.0050621 **
x:(intercept)  0.6181424  0.6404345  0.9652 0.3344486
2:H1           0.0345083  0.1150819  0.2999 0.7642847
x:H1          -0.0406185  0.1089413 -0.3728 0.7092617
2:H2           0.0804896  0.1069222  0.7528 0.4515780
x:H2           0.0412585  0.1024818  0.4026 0.6872471
2:B1          -0.0072889  0.1086202 -0.0671 0.9464986
x:B1           0.0042735  0.1068691  0.0400 0.9681028
2:B2          -0.2880369  0.1099267 -2.6203 0.0087862 **
x:B2          -0.0601342  0.1052201 -0.5715 0.5676551
2:VH           0.1097113  0.9011327  0.1217 0.9030984
x:VH          -1.2270539  0.8454826 -1.4513 0.1466947
2:PH          -1.6033612  0.4607823 -3.4796 0.0005021 ***
x:PH          -0.3098387  0.4193898 -0.7388 0.4600379
```

Figur 9: (H1, H2, B1, B2, VH, PH) - parameterskattningar



Figur 10: VH plottad mot PH

I Figur 10 så plottas VH mot PH, och vi ser att dessa förklarar varandra väl. Med ovanstående modellsammanfattning och Figur 10 som stöd så kan vi med gott samvete utesluta VH ur modellen. I appendix hittas modellsammanfattningen för (H1, H2, B1, B2, PH)-modellen.

Ingen av variablerna H1, H2, och B1 ingår med signifikanta effekter i någon kategori, och ett χ^2-test utförs därför på nollhypotesen $H_0 : \beta_{11} = \beta_{21} = \beta_{12} = \beta_{22} = \beta_{13} = \beta_{23} = 0$. P-värdet för detta test blir 0.9749 och nollhypotesen accepteras. Parameterskattningar för reviderad modell sammanställs i Figur 11.

```
Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  2.157318   0.396100  5.4464 5.140e-08 ***
x:(intercept)  0.890677   0.370564  2.4036 0.0162358 *
2:B2          -0.277272   0.082189 -3.3736 0.0007420 ***
x:B2          -0.066146   0.077332 -0.8554 0.3923510
2:PH          -1.624197   0.264190 -6.1478 7.855e-10 ***
x:PH          -0.884951   0.234792 -3.7691 0.0001638 ***
```

Figur 11: (B2, PH) - parameterskattningar

Modellen utökas med en samspelseffekt, då det är tänkbart att PH inte påverkar sannolikheterna med samma effekt för olika värden på B2. I figur 12 sammanställs denna modells parameterskattningar.

```
Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  2.356549   0.721402  3.2666 0.0010884 **
x:(intercept)  2.221205   0.719473  3.0873 0.0020201 **
2:B2          -0.340017   0.317344 -1.0714 0.2839686
x:B2          -0.685929   0.293526 -2.3369 0.0194464 *
2:PH          -1.735410   0.526067 -3.2988 0.0009709 ***
x:PH          -1.871306   0.515206 -3.6321 0.0002811 ***
2:B2:PH        0.026898   0.240088  0.1120 0.9107955
x:B2:PH        0.461680   0.209102  2.2079 0.0272501 *
```

Figur 12: (B2, PH, B2·PH) - parameterskattningar

I nedanstående tabell sammanfattas modellernas AIC-värde, deras P-värde vid test om alla variabler saknar effekt, samt deras prediktionsmått.

Modell	P-värde	AIC	Prediktionsmått
(H1, H2, B1, B2, VH, PH)	4.17e-8	1586.91	0.4271
(H1, H2, B1, B2, PH)	2.17e-8	1585.57	0.4291
(B2, PH)	3.91e-11	1574.83	0.4271
(B2, PH, B2·PH)	4.52e-11	1573.45	0.4291

Ett χ^2-test utförs på hypotesen $H_0 : \beta_{13} = \beta_{23} = 0$ i modellen (B2·PH) och vi erhåller ett P-värde på <0.05, och H_0 förkastas. Samspelsmodellen har även lägst AIC-värde och högst prediktionsmått och accepteras därför som den slutgiltiga modellen ifrån informationskälla 2, vi kallar modellen för

modell 2. Nedanstående tabeller ämnar att visa hur väl modellen förklarar sannolikheterna för de olika utfallen samt hur modellen presterar rent prediktionsmässigt. Det är samma typ av tabeller som i avsnitt 4.3.1.

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	-	0	0	2	0.22	37
10-20	0.40	15	0.12	34	0.24	142
20-30	0.28	72	0.27	411	0.32	265
30-40	0.34	199	0.29	307	0.29	235
40-50	0.42	233	-	0	0.27	74
50-60	0.53	159	-	0	0	1
60-70	0.76	59	-	0	-	0
70-80	0.69	13	-	0	-	0
80-90	1.00	4	-	0	-	0

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	0.00	1	0	3	0.26	35
10-20	0.28	18	0.21	46	0.24	105
20-30	0.36	75	0.29	246	0.20	133
30-40	0.37	111	0.32	196	0.35	146
40-50	0.44	104	0.33	3	0.35	55
50-60	0.48	119	-	0	0.29	7
60-70	0.55	55	-	0	0.62	13
70-80	0.75	8	-	0	-	0
80-90	1.00	3	-	0	-	0

Modellen kan ej förklara och prediktera sannolikheterna för bortavinst på ett lämpligt sätt.

4.3.3 Modeller baserade på I1 och I2

Kombineras de två modellerna från I1 och I2, så erhålles följande parameterskattningar.

```
Coefficients :
              Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  3.508517   0.796569  4.4045 1.060e-05 ***
x:(intercept)  3.093746   0.788812  3.9220 8.780e-05 ***
2:Y1          -6.025926   0.949811 -6.3443 2.234e-10 ***
x:Y1          -4.709794   0.908427 -5.1846 2.165e-07 ***
2:TX          -0.291046   0.103523 -2.8114  0.004932 **
x:TX          -0.062544   0.095831 -0.6527  0.513980
2:B2           0.028543   0.338632  0.0843  0.932828
x:B2          -0.490381   0.305334 -1.6061  0.108263
2:PH          -0.500275   0.582807 -0.8584  0.390678
x:PH          -0.966549   0.559599 -1.7272  0.084129 .
2:B2:PH       -0.116400   0.255935 -0.4548  0.649251
x:B2:PH        0.437653   0.217501  2.0122  0.044200 *
```

Figur 13: (Y1, TX, B2·PH) - parameterskattningar

Modellen har ett AIC-värde på 1526.68 och ett prediktionsmått på 0.4534.

Vi börjar med att testa om samspelseffekterna verkligen behöver ingå i modellen, då det verkar som att effekterna av B2 och PH inte bör göra det. Tolkningen av en samspelseffekten B2·PH är att PH påverkar sannolikheterna med olika effekt på olika nivåer av B2, men om det är troligt att B2 och PH saknar effekt så blir det vanskligt att ha med en samspelsterm i modellen. Därför utförs ett χ^2-test av nollhypotesen $H_0 : \beta_{15} = \beta_{25} = 0$ och ett P-värde på 0.05 erhålles. Vi ska enligt statistiska principer förkasta H_0, men vi accepterar istället hypotesen då vi föredrar en modell utan samspelseffekter dels därför effekterna av B2 och PH inte ingår signifikant i denna modell dels därför en modell utan samspelseffekter beror mindre på datamaterialet vilket är en fördel i prediktionssammanhang.

Parameterskattningar för den reviderade modellen (Y1, TX, B2, PH) hittas i appendix. Det verkar som att variablerna från I1 förklarar mer av sannolikheterna än variablerna ifrån I2. Ett nytt χ^2-test utförs därför på hypotesen $H_0 : \beta_{13} = \beta_{23} = \beta_{14} = \beta_{24} = 0$ och ett P-värde på 0.1082 erhålles. Vi accepterar därmed H_0.

Modellen (Y1, TX) föredras alltså över (Y1, TX, B2, PH) då den är mer generell och robust, har ett lägre AIC-värde, och bättre prediktionskraft. Denna modell accepteras som en av de slutgiltiga modellerna. Vi kallar modellen för Modell 1. Analys av denna modell hittas i avsnitt 4.3.1.

4.4 Tillvägagångssätt 2

Totalt sett finns det 15 tillgängliga variabler att basera en modell på. Tar vi inte hänsyn till samspelseffekter så finns det $\sum_{i=1}^{15} \binom{15}{i} = 32767$ olika modeller att beräkna prediktionsmått för. Resultaten för de tre modellerna med högst prediktionsmått presenteras i nedanstående tabell.

Modell	P-värde	AIC	Prediktionsmått
(Y1, XX)	<2.22e-16	1532.29	0.4919
(Y2, XX, H1, H2)	5.26e-16	1547.70	0.4899
(Y2, H2)	<2.22e-16	1541.75	0.4879

Modellen (Y1, XX) har lägst AIC-värde och högst prediktionsvärde av de tre presenterade modellerna och väljs därför som den slutgiltiga modellen ifrån denna analysmetod. Vi kallar modellen för Modell 3, och parameterskattningar för denna modell hittas i Figur 14.

```
Coefficients :
               Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  3.2109072  0.8887776  3.6127  0.000303 ***
x:(intercept)  1.4354154  0.9200443  1.5602  0.118722
2:Y1          -6.6460823  0.9553065 -6.9570 3.476e-12 ***
x:Y1          -4.1238929  0.9731055 -4.2379 2.257e-05 ***
2:XX          -0.0176014  0.0205755 -0.8555  0.392299
x:XX           0.0086519  0.0203839  0.4244  0.671238
```

Figur 14: (Y1, XX) - parameterskattningar

Nedanstående tabeller ämnar att illustrerar hur väl modellen förklarar sannolikheterna samt hur vida modellen lämpar sig för prediktion.

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	0.00	4	0.00	5	0.15	47
10-20	0.20	40	0.05	80	0.22	156
20-30	0.25	101	0.27	269	0.29	283
30-40	0.34	170	0.33	400	0.37	175
40-50	0.43	175	-	0	0.32	60
50-60	0.57	142	-	0	0.26	23
60-70	0.63	63	-	0	0.25	8
70-80	0.69	32	-	0	0.00	2
80-90	0.93	27	-	0	-	0

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	0.20	10	0.00	2	0.03	30
10-20	0.32	28	0.24	63	0.20	101
20-30	0.29	69	0.29	139	0.24	168
30-40	0.37	131	0.31	290	0.32	125
40-50	0.43	104	-	0	0.47	35
50-60	0.50	56	-	0	0.33	21
60-70	0.57	42	-	0	0.82	11
70-80	0.70	43	-	0	0.50	2
80-90	0.73	11	-	0	-	0

Sett till dessa tabeller så är modellen lämplig båda för att förklara sannolikheterna för utfallen, samt för att predikterar framtida matcher.

4.5 Tillvägagångssätt 3

Vi tar inte heller i denna metod hänsyn till samspelseffekter, då detta blir för beräkningstungt. Utav av de 32767 möjliga modeller så presenteras den med lägst AIC-värde i nedanstående tabell.

Modell	P-värde	AIC	Prediktionsmått
(Y1, TX, Y2)	<2.22e-16	1525.08	0.4575

Vi kallar modellen för modell 4, och parameterskattningar för denna modell hittas i avsnitt 4.3.1. Nedanstående tabeller är de sedvanliga som används för modellanalys.

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	0	0	0	18	0.16	74
10-20	0.20	46	0.05	55	0.27	135
20-30	0.22	119	0.21	265	0.27	246
30-40	0.39	166	0.35	416	0.32	180
40-50	0.43	178	-	0	0.43	79
50-60	0.58	125	-	0	0.23	39
60-70	0.61	46	-	0	0	1
70-80	0.76	33	-	0	-	0
80-90	0.80	35	-	0	-	0
90-100	1.00	5	-	0	-	0

Modellsannolikhet	Hemmavinst	Antal	Kryss	Antal	Bortavinst	Antal
0-10	-	0	0	8	0.11	65
10-20	0.26	35	0.27	52	0.26	102
20-30	0.26	57	0.31	133	0.25	158
30-40	0.43	117	0.30	301	0.28	101
40-50	0.38	107	-	0	0.40	40
50-60	0.47	78	-	0	0.61	28
60-70	0.54	37	-	0	-	0
70-80	0.59	34	-	0	-	0
80-90	0.78	27	-	0	-	0
90-100	1.00	2	-	0	-	0

Modellen verkar förklara sannolikheterna väl, den presterar dock inte rent prediktionsmässigt. En tydlig indikation på att modellen beror för mycket på datamaterialet den är baserad på.

4.6 Slutgiltig modell

De fyra modeller som har analyserats fram presenteras i nedanstående tabell.

	Modell	AIC	Prediktionsmått
Modell 1	(Y1, TX)	1526.27	0.4636
Modell 2	(B2, PH, B2·PH)	1573.45	0.4291
Modell 3	(Y1, XX)	1532.29	0.4914
Modell 4	(Y1, TX, Y2)	1525.08	0.4575

Sett till all ovanstående analys så står det mellan (Y1, TX) och (Y1, XX)-modellen då dessa har presterat bäst. Då målet med arbetet var att hitta en modell som på lämpligt sätt förklarar sannolikheterna för de olika utfallen i en fotbollsmatch så väljs (Y1, TX)-modellen som den slutgiltiga, då den gör detta bättre, för dataset 2, än (Y1, XX).

5 Resultat

Det verkar som att modell 1 förklarar sannolikheterna för de olika utfallen väl, sett till Figur 6, 7, 8 samt tabellen på sid. 19. Det verkar alltså vara möjligt att genom multinomial logistisk regression förklara sannolikheterna för de olika utfallen i en fotbollsmatch. Vi konstaterar också att för vårat ändamål så är det bättre att basera sin modell på mått ifrån experter (variablerna nr. 1-9) istället för att basera modellen på andra mått (variablerna nr. 10-21).

6 Diskussion

Det är ett svårt problem att hitta modeller som förklarar sannolikheterna i fotbollsmatcher, då dessa uppenbarligen beror på ett stort antal variabler. Att sannolikheterna beror på många olika faktorer syns i vår modell också, ty det inverterade hemmaoddset ingår i den slutgiltiga modellen, och man får anta att Svenska Spel sätter oddsen med hänsyn till en mängd olika faktorer. En nackdel med slutmodellen är dock att den innehåller det inverterade hemmaoddset som förklarande variabel. Detta medför att man inte kan använda modellen till speciellt mycket. Ett tänkbart användningsområde hade varit att använda modellen för att hitta spelbara fotbollsmatcher, alltså matcher där något av oddsen är för högt satt. Dessa matcher blir svåra att identifiera med vår slutgiltiga modell då oddset ingår i modellen, så när man väl stöter på en match med ett överodds så leder det till att modellen producerar en överskattad sannolikhet, och matchen kanske då inte identifieras som spelbar. En annan nackdel med modellen är att om man fixerar oddset och ökar antalet tidningar som tror på oavgjort så ökar inte bara den skattade sannolikheten för oavgjort utan också den för hemmavinst. Helst skulle man vilja att en ökning i antal tidningar som tror på oavgjort medför en ökning av den skattade sannolikheten för oavgjort och en minskning av de andra sannolikheterna.

7 Appendix

7.1 Figurer

```
Coefficients :
               Estimate Std. Error t-value  Pr(>|t|)
2:(intercept)  3.621605   0.498163  7.2699 3.597e-13 ***
x:(intercept)  1.870306   0.470872  3.9720 7.127e-05 ***
2:Y1          -5.952819   0.944856 -6.3002 2.972e-10 ***
x:Y1          -4.780135   0.904759 -5.2833 1.269e-07 ***
2:TX          -0.281126   0.103185 -2.7245   0.00644 **
x:TX          -0.075859   0.095345 -0.7956   0.42625
2:B2          -0.092753   0.090970 -1.0196   0.30792
x:B2           0.093131   0.085461  1.0897   0.27583
2:PH          -0.656104   0.313058 -2.0958   0.03610 *
x:PH          -0.020886   0.287328 -0.0727   0.94205
```

Figur 15: (Y1, TX, B2, PH) - parameterskattningar

```
Coefficients :
               Estimate Std. Error t-value Pr(>|t|)
2:(intercept)  1.8776930  0.6683570  2.8094 0.004963 **
x:(intercept)  0.7457053  0.6341500  1.1759 0.239630
2:H1           0.0343318  0.1145178  0.2998 0.764334
x:H1          -0.0544496  0.1086284 -0.5012 0.616198
2:H2           0.0812177  0.1061418  0.7652 0.444164
x:H2           0.0227659  0.1016357  0.2240 0.822761
2:B1          -0.0078851  0.1081759 -0.0729 0.941893
x:B1           0.0174926  0.1063198  0.1645 0.869315
2:B2          -0.2864404  0.1098111 -2.6085 0.009094 **
x:B2          -0.0544648  0.1050459 -0.5185 0.604120
2:PH          -1.5587211  0.2984318 -5.2230 1.76e-07 ***
x:PH          -0.7799739  0.2689879 -2.8997 0.003736 **
```

Figur 16: (H1, H2, B1, B2, PH) - parameterskattningar

7.2 Motivering till Svenska Spels utbetalningar

Låt $p_1,\ p_x,\ p_2$ vara sannolikheterna för utfallen i en fotbollsmatch. Antag att sedan att Svenska Spel skattar dessa sannolikheter med $\hat{p}_1,\ \hat{p}_x,\ \hat{p}_2$. För att kommande resonemang ska vara giltigt kräver vi att dessa summerar sig till 1. Ett rättvist odds för utfallet i ges av $\frac{1}{p_i}$, $i = 1, x, 2$, eftersom då är din förväntade vinst 0kr. Om nu Svenska Spel säljer odds av storleken $\frac{1}{\hat{p}_i}$ och tar in proportionen $\hat{p}_i$ av alla de spelade pengarna på alternativ i så blir andelen pengar som Svenska Spel förväntas ge tillbaka till spelarna 100% av alla intagna pengar, se nedan. Låt Z vara intagna pengar och O_i oddset för utfal i, då ges den förväntade andelen pengar som Svenska Spel ger tillbaka till spelarna av

$$\frac{Z \cdot \hat{p}_1 \cdot O_1 \cdot p_1 + Z \cdot \hat{p}_x \cdot O_x \cdot p_x + Z \cdot \hat{p}_2 \cdot O_2 \cdot p_2}{Z}$$

$$= \frac{\hat{p}_1 \cdot p_1}{\hat{p}_1} + \frac{\hat{p}_x \cdot p_x}{\hat{p}_x} + \frac{\hat{p}_2 \cdot p_2}{\hat{p}_2}$$
$$= p_1 + p_x + p_2 = 1$$

Om vi nu antar att Svenska Spel, för att gå med vinst, säljer odds av storleken $O_i = \frac{y}{\hat{p}_i}$ där $y < 1$ så ger samma räkningar som ovan en förväntat andel på y. Då vi har observerat att $\frac{1}{O_1} + \frac{1}{O_x} + \frac{1}{O_2} = \frac{\hat{p}_1}{y} + \frac{\hat{p}_x}{y} + \frac{\hat{p}_2}{y} \approx 1.26$ så ger detta att, under våra ovanstående antaganden, att andelen pengar Svenska Spel ger tillbaka till sina kunder är ungefär lika med $\frac{1}{1.26} = 0.79$.

Referenser

[1] Alan Agresti, *Categorical Data Analysis*, Wiley, *p. 216-217*

[2] Bernard W. Lindgren, *Statistical Theory*, Chapman & Hall/CRC, 4th Edition, *p. 298*

[3] David W. Hosmer & Stanley Lemeshow, *Applied Logistic Regression*, Wiley, *p. 91-142, 260-339*

[4] http://www.warnsater.com/Stryktipset/