

# Logistisk regression

Aalborg Intelligence

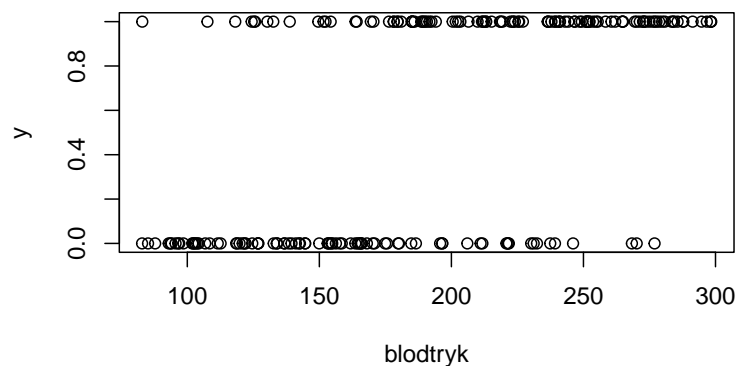
## Logistisk regression

I det følgende ses på logistisk regression. Måske har du allerede hørt begrebet logistisk regression i gymnasieundervisningen i forbindelse med logistisk vækst. Dette er en anden type af logistisk regression end den som behandles i denne note. Senere vil vi dog se eksempler hvor der alligevel kan laves en sammenhæng med logistisk vækst.

Logistisk regression, i den betydning vi kigger på i denne note, bruges, når man gerne vil modellere hvordan en sandsynlighed afhænger af en anden variabel. Som et eksempel forestiller vi os, at vi er interesserede i at undersøge sammenhængen mellem (systolisk) blodtryk og risikoen for hjerte-kar-sygdom.

Vi vil kigge på et datasæt bestående af 2000 mennesker, som har fået målt deres blodtryk. Desuden har de fået undersøgt, om de lider af hjertekarsygdom. Datasættet er fiktivt, men det er lavet til at ligne virkelige data<sup>1</sup>. Vi kalder blodtrykket for  $x$ , mens vi lader  $y$  være en variabel, der er 1 hvis personen lider af hjertekarsygdom og 0 ellers. På figuren nedenfor har vi tegnet samhörrende  $x$ - og  $y$ -værdier ind i et koordinatsystem for de første 200 personer i datasættet.

<sup>1</sup> De fleste mennesker har et systolisk blodtryk mellem 100 og 180 mmHg. I datasættet har vi genereret en masse mennesker med meget højt eller lavt blodtryk for illustrationens skyld.



Hvordan skal man beskrive sammenhængen mellem  $x$  og  $y$ ? Det ser ud til at der blandt folk med meget lavt blodtryk er flest raske, mens der for folk med meget højt blodtryk er flest syge, men ved de fleste blodtryksværdier ser der ud til at være både syge og raske. I stedet for at se direkte på sammenhængen mellem  $x$  og  $y$  vil vi derfor se på hvordan *sandsynligheden* for hjerte-kar-sygdom ( $y = 1$ ) afhænger af blodtrykket. Vi vil betragte denne sandsynlighed som en funktion  $p(x)$  som afhænger af blodtrykket  $x$ . Vi skal se på hvordan man kan bestemme denne funktion.

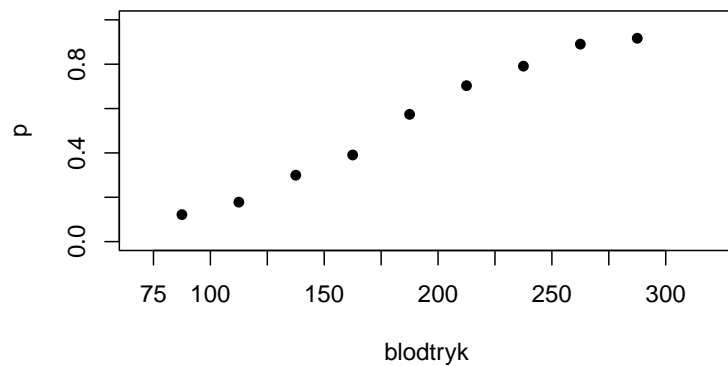
For at få en idé om, hvordan  $p(x)$  kunne se ud, kigger vi datasættet fra før. Vi inddeler nu blodtrykket i intervaller af længde 25.

|      | Blodtryk | 75,100] | 100,125] | 125,150] | 150,175] | 175,200] | 200,225] | 225,250] | 250,275] | 275,300] |
|------|----------|---------|----------|----------|----------|----------|----------|----------|----------|----------|
| Rask | 173      | 194     | 145      | 156      | 98       | 57       | 47       | 25       | 19       |          |
| Syg  | 24       | 42      | 62       | 100      | 132      | 135      | 178      | 203      | 210      |          |

For hvert interval beregnes andelen af patienterne, der lider af hjerte-kar-sygdom som estimat for sandsynligheden for hjerte-kar-sygdom i den gruppe. F.eks. er der 236 personer med et blodtryk i intervallet  $]100,125]$ . Heraf er 42 syge. Andelen af syge i denne gruppe er derfor

$$\frac{42}{236} \approx 0.178.$$

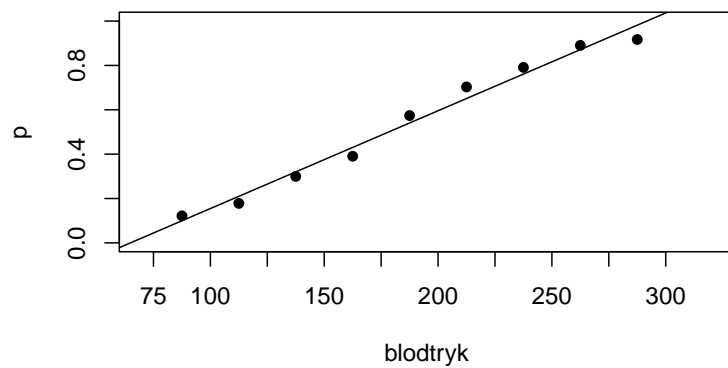
Der laves et plot hvor der for hver blodtryksinterval laves et punkt hvor  $x$ -værdien er midtpunktet for blodtryks- intervallet og  $y$  er den estimeret sandsynlighed ved intervallet.



Umiddelbart kunne det være fristende at lave lineær regression. Vi forestiller os altså en forskrift

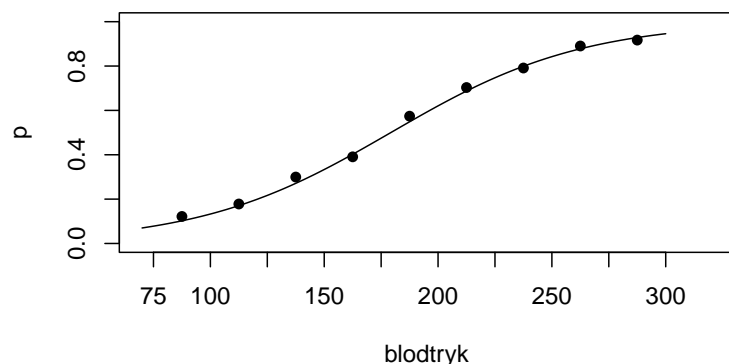
$$p(x) = ax + b.$$

Hvis vi indtegner den bedste rette linje i plottet, får vi nedenstående:



Der er et problem her: En sandsynlighed ligger altid mellem 0 og 1, men regressionslinjen ovenfor skærer  $x$ -aksen ved blodtryksværdier omkring 80. Det betyder, at sandsynligheden er negativ for blodtryk under 80. Tilsvarende får vi sandsynligheder, der er større end 1 ved blodtryk over 300. Det giver selvfølgelig ikke mening.

Hvis vi kigger på figur (ref) igen, ser sammenhængen da heller ikke lineær ud, men snarere S-formet. Vi kunne altså forestille os at grafen for  $p$  ser ud som indtegnet nedenfor:



## Odds

For at komme nærmere hvordan funktionsforskriften for  $p$  skal se ud, kigger vi på *oddsene* for sygdom i stedet for sandsynligheden<sup>2</sup>. Oddsene  $O$  for en hændelse er defineret som sandsynligheden for hændelsen  $p$  divideret med sandsynligheden for komplementærhændelsen, som er  $1 - p$ . Altså

$$O = \frac{p}{1 - p}$$

Odds måler således, hvor mange gange mere sandsynlig hændelsen er i forhold til komplementærhændelsen. Hvis fx sandsynligheden for hjerte-kar-sygdom er  $p = \frac{4}{5}$  så er odds for sygdom

$$O = \frac{\frac{4}{5}}{\frac{1}{5}} = 4.$$

<sup>2</sup> Du kender måske begrebet odds fra sportsgambling. Det er dog en anden betydning af ordet end det vi bruger her. Inden for gambling angiver odds, hvor mange gange man får pengene igen, hvis en bestemt hændelse indtræffer (fx. et bestemt hold vinder). Gambling odds er naturligvis udregnet ud fra de sandsynlighedsteoretiske odds for hændelsen, men er altid justeret for at sikre at bookmakeren vinder i det lange løb.

Det er altså fire gange så sandsynligt at være syg som at være rask.

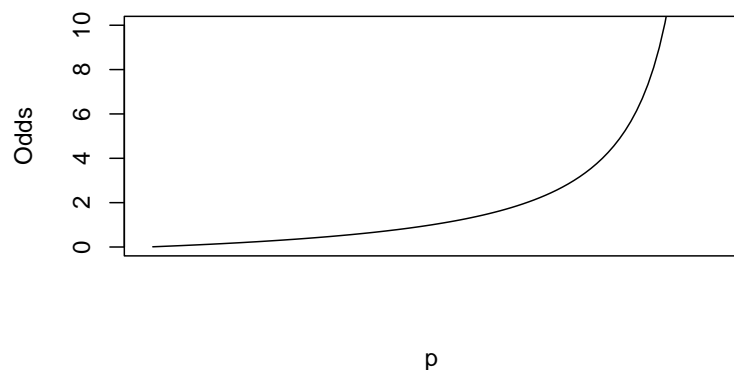
For at få en lidt bedre fornemmelse for, hvordan odds fungerer, kan vi lave en tabel, der viser odds svarende til forskellige værdier af  $p$ :

| $p$ | $\frac{1}{5}$ | $\frac{1}{4}$ | $\frac{1}{3}$ | $\frac{1}{2}$ | $\frac{2}{3}$ | $\frac{3}{4}$ | $\frac{4}{5}$ |
|-----|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| $O$ | $\frac{1}{4}$ | $\frac{1}{3}$ | $\frac{2}{3}$ | 1             | 2             | 3             | 4             |

Funktionen, der omdanner sandsynligheder til odds har forskriften

$$f(p) = \frac{p}{1-p},$$

hvor  $p$  kan ligge mellem 0 og 1. Grafen for  $f$  er tegnet nedenfor.

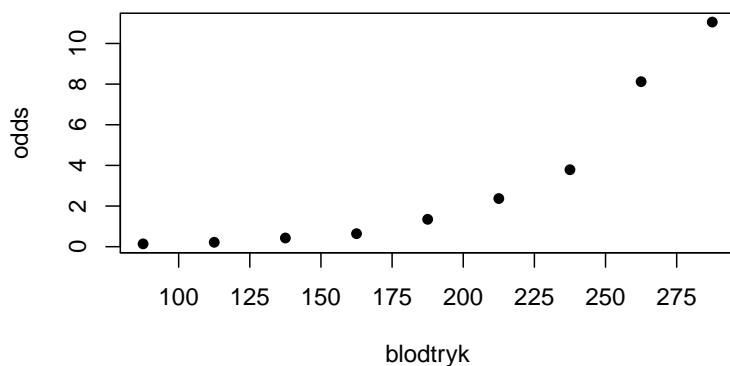


Vi ser at  $f$  altid er positiv, da både tæller og nævner er positive. Når  $p$  nærmer sig 0, nærmer tælleren sig 0 mens nævneren nærmer sig 1, så  $f(p)$  går mod 0. Når  $p$  nærmer sig 1, nærmer tælleren sig 1 og nævneren nærmer sig 0, så hele brøken  $f(p)$  går mod uendelig. Værdimængden for  $f$  består derfor af alle de positive reelle tal.

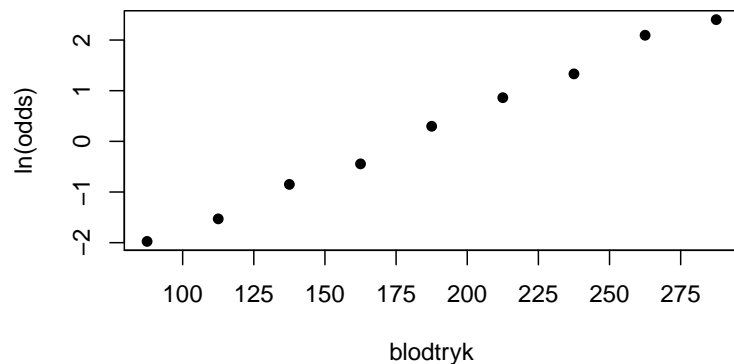
I vores dataeksempel, hvor sandsynligheden for hjerte-kar-sygdom er en funktion  $p(x)$ , bliver oddsene for hjerte-kar-sygdom også en funktion af  $x$

$$O(x) = \frac{p(x)}{1 - p(x)}.$$

Nedenfor vises dataeksemplet fra før, men hvor vi nu har oddsene  $O(x)$  på  $y$ -aksen.



Vi ser at oddsene for sygdom stiger med blodtrykket. Kigger vi på grafen, ser tendensen ikke lineær ud. Det kunne derimod ligne en eksponentiel vækst. For at bekræfte dette, laver vi samme plot, men nu med den naturlige logaritme til oddsene  $\ln(O(x))$  på  $y$ -aksen.



Der ser nu ud til at være en lineær sammenhæng! Det kunne altså tyde på, at log-oddsene afhænger lineært af  $x$ . Det leder os frem til følgende model for log-oddsene:

$$\ln(O(x)) = ax + b.$$

Denne model kaldes *den logistiske regressionsmodel*. Virkelige data følger ofte en logistisk regressionsmodel.

Med denne model og ved blot at se på grafen kunne man fristes til at benytte lineær regression til at finde  $a$  og  $b$ . Her er det dog vigtigt at bemærke at et punkt fra grafen egentligt er kombinationen af data fra flere observationer som ikke engang havde samme  $x$ -værdi. Antal af observationer som hvert punkt er lavet ud fra er desuden heller ikke ens.

## Logit-funktionen og den logistiske funktion

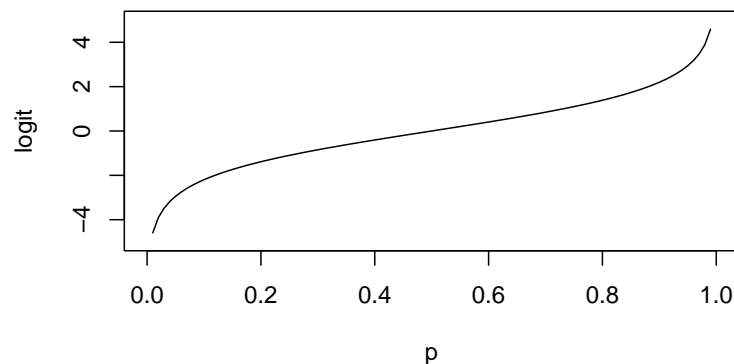
Når vi tager den naturlige logaritme til oddsene får vi

$$\ln(O) = \ln\left(\frac{p}{1-p}\right).$$

Funktionen på højresiden kaldes logit, altså

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right).$$

Definitionsmængden for logit-funktionen er det åbne interval  $]0, 1[$ . Vi fandt tidligere, at værdimængden for oddsene består af alle de positive reelle tal. Dette er netop definitionsmængden for  $\ln$ . Værdimængden for logit bliver derfor den samme som for  $\ln$ , nemlig alle de reelle tal. Grafen for logit er vist nedenfor.



Vi vil nu finde den inverse funktion til logit. Antag at

$$y = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right).$$

Vi vil prøve at udtrykke  $p$  som funktion af  $y$ . Vi tager først eksponentialfunktionen på begge sider af udtrykket og får

$$e^y = \frac{p}{1-p}$$

Så ganger vi med  $(1-p)$  på begge sider.

$$e^y(1-p) = p$$

Hvis parentesen ophæves får vi

$$e^y - p \cdot e^y = p$$

Vi kan så lægge  $p \cdot e^y$  til på begge sider og sætte  $p$  uden for parantes

$$e^y = p \cdot e^y + pe^y = p \cdot (e^y + 1)$$



Endelig kan vi isolere  $p$  og få

$$\frac{e^y}{e^y + 1} = p$$

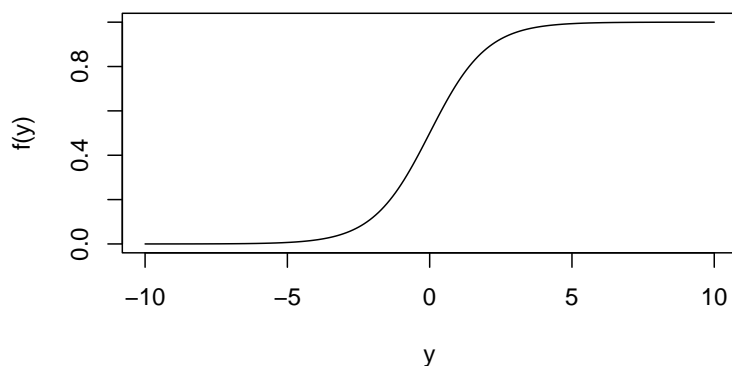
Her er  $p$  egentlig isoleret men vi kan forkorte brøken med  $e^y$  for at få et andet udtryk for  $p$ .

$$p = \frac{\frac{e^y}{e^y}}{\frac{e^y}{e^y} + 1} = \frac{\frac{e^y}{e^y}}{\frac{e^y}{e^y} + \frac{1}{e^y}} = \frac{1}{1 + e^{-y}}.$$

Sammenlagt har vi vist, at den inverse funktion til logit er *den standard logistiske funktion* (også nogle gange kaldet *sigmoide-funktionen*) NOTE: eller sigmoid?

$$f(y) = \frac{1}{1 + e^{-y}}.$$

Grafen for den standard logistiske funktion er indtegnet nedenfor. Vi ser at grafen har en karakteristisk S-form, som vokser fra 0 mod 1.



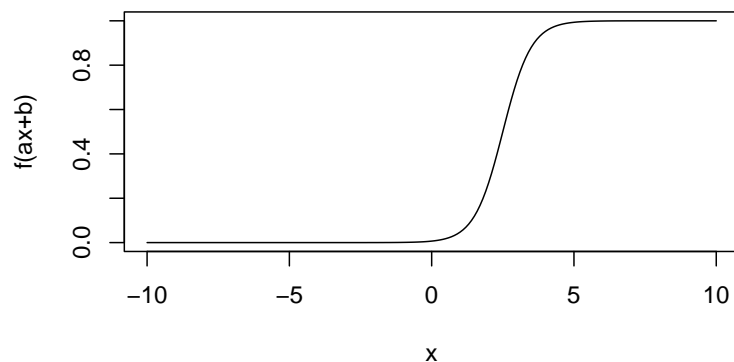
Hvis man har en funktion  $f$  kan væksthastigheden øges med en faktor  $a$  ved at se på  $f(ax)$ . Hvis man derimod ønsker at forskyde grafen med  $k$  i  $x$ -aksens retning kan det gøres ved at se på  $f(x - k)$ . Hvis man kombinere disse, og øger væksthastigheden med en faktor  $a$ , og derefter forskyder grafen med  $k = \frac{-b}{a}$  i  $x$ -aksens retning fås istedet funktionen

$$f\left(a \cdot \left(x - \left(\frac{-b}{a}\right)\right)\right) = f(ax + b).$$

Hvis man gør dette for den standard logistiske funktion, får man den generelle logistiske funktion

$$f(ax + b) = \frac{e^{ax+b}}{1 + e^{ax+b}} = \frac{1}{1 + e^{-(ax+b)}}.$$

Sammenlignet med den standard logistiske funktion får man altså en *S*-formet kurve, der vokser  $a$  gange så hurtigt og er forskudt med  $\frac{-b}{a}$ . Fx får man for  $a = 2$  og  $b = -5$ :



Det var en logistisk funktion af denne type, der blev brugt til at lave den S-formede kurve i Figur ?.

Her laves et datasæt hvor du selv kan forsøge at tilpasse modellen. Vigtigt at hjælpe læseren med lidt info, med at det jo ikke er meningen at grafen skal gå gennem punkterne, men istedet skal indikere hvor stor sandsynlighed der er for ..... lige ved det sted. Desuden bør man kunne få vist den bedste model.

## Den logistiske regressionsmodel

Vi vender nu tilbage til den logistiske regressionsmodel. Husk på, at da  $\ln(O(x)) = \text{logit}(p(x))$ , siger modellen

$$\ln(O(x)) = \text{logit}(p(x)) = ax + b.$$

Da værdimængden for logit var alle de reelle tal, giver det mening at modellere  $\text{logit}(p(x))$  med en lineær funktion.

Hvis vi gerne vil have et udtryk for oddsene, kan vi tage eksponentialfunktionen på begge sider

$$O(x) = e^{ax+b} = e^b \cdot e^{ax} = e^b \cdot (e^a)^x.$$

Hvis  $e^b$  kaldes  $b_{ny}$  og  $e^a$  kaldes  $a_{ny}$ , ses at  $O(x) = b_{ny} \cdot a_{ny}^x$  er en eksponentiel funktion med fremskrivningsfaktor  $a_{ny} = e^a$ . Derved vil %-delen som Odds for sygdom stiger med være  $r = e^a - 1$  hver gang blodtrykket stiger med 1 mmHg.

I forbindelse med logistisk regression kaldes  $e^a$  også for *odds-ratioen*. For at forstå hvorfor, kan vi forestille os to patienter, en med blodtryk  $x_1$  og en med blodtryk  $x_2$ . De har dermed oddsene

$$O(x_1) = e^b \cdot e^{ax_1} O(x_2) = e^b \cdot e^{ax_2}.$$

Lad os se på forholdet (ratioen) mellem de to personers odds:

$$\frac{O(x_1)}{O(x_2)} = \frac{e^b \cdot e^{ax_1}}{e^b \cdot e^{ax_2}} = \frac{e^{ax_1}}{e^{ax_2}} = e^{ax_1 - ax_2} = e^{a(x_1 - x_2)} = (e^a)^{x_1 - x_2}.$$

Forholdet mellem oddsene afhænger altså kun af forskellen  $x_1 - x_2$  mellem de to personers blodtryk. Hvis person 1 er har et blodtryk, der er 1mmHg højere end person 2, bliver forholdet (ratioen) mellem oddsene lige præcis  $e^a$ . Sagt på en anden måde: Odds for sygdom stiger med en faktor  $e^a$  hver gang blodtrykket stiger med 1 mmHg.

Odds kan være svære at forstå intuitivt. Vi vil derfor prøve at finde tilbage til et udtryk for sandsynligheden for sygdom  $p(x)$ . Vi havde modellen

$$\text{logit}(p(x)) = ax + b.$$

Bruger vi den inverse funktion (som vi fandt var den logistiske funktion) på begge sider, får vi

$$p(x) = \frac{e^{ax+b}}{1 + e^{ax+b}}.$$

eller

$$p(x) = \frac{1}{1 + e^{-(ax+b)}}.$$

Vi får altså et logistisk funktionsudtryk for  $p$ , deraf navnet “den logistiske regressionsmodel”. Nu skal vi bare finde ud af hvordan  $a$  og  $b$  bestemmes ud fra et datasæt, dette er dog ikke helt simpelt.

Tidligere definerede vi egentlig  $O(x)$  ud fra  $p(x)$ , men ud fra formlerne

$$O(x) = e^{ax+b} \quad \text{og} \quad p(x) = \frac{e^{ax+b}}{1 + e^{ax+b}}$$

ses også let hvordan  $p(x)$  findes ud fra  $O(x)$  ved

$$p(x) = \frac{O(x)}{1 + O(x)}.$$

Lad os se på hvordan man selv kan få en ide om estimater for  $a$  og  $b$  ud fra de oprindelige grafer.

Vi bruger her at

$$ax + b = \text{logit}(p) = \ln\left(\frac{p}{1-p}\right).$$

og se på hvor midtpunktet er og hvor hurtigt funktionen vokser. Ved midtpunktet er  $p = 50\% = 0,5$ . Derved er

$$ax + b = \ln\left(\frac{0.5}{1-0.5}\right) = \ln(1) = 0.$$

Og ved at isolere  $b$  fås at  $x = \frac{-b}{a}$ . Hvis vi ønsker at se på hvor meget  $x$ -værdien skal for  $p$  ændres fra f.eks. 20% til 80% (note med det er symmetrisk sandsynlighed) kan det gøres på følgende måde. Lad først  $x_{20\%}$  være  $x$ -værdien svarende til  $p = 20\% = 0,2$  og tilsvarende ved  $p = 80\% = 0,8$ . Vi ønsker nu at finde

$$a \cdot x_{80\%} + b - (a \cdot x_{20\%} + b) = \ln\left(\frac{0.8}{1-0.8}\right) - \ln\left(\frac{0.2}{1-0.2}\right).$$

Hvis man ser på de to brøkker er de egentlig ens blot tæller og nævner er byttet om. Ved at benytte at  $\ln(a) - \ln(b) = \ln(\frac{a}{b})$  kan man vise at  $\ln(\frac{a}{b}) = -\ln(\frac{b}{a})$ . Ved at benytte dette og omskrive venstresiden fås

$$a \cdot (x_{80\%} - x_{20\%}) = \ln\left(\frac{0.8}{1-0.8}\right) - \left(-\ln\left(\frac{0.8}{1-0.8}\right)\right) = 2 \cdot \ln\left(\frac{0.8}{1-0.8}\right).$$

Derved bliver

$$x_{80\%} - x_{20\%} = \frac{2 \cdot \ln\left(\frac{0.8}{1-0.8}\right)}{a} = \frac{2 \cdot \text{logit}(0.8)}{a}.$$

Her kan man selvfølgelig også bestemme  $a$  hvis man har en ide om hvor hurtig man går fra en sandsynlighed fra 20 % til 80 % ved

$$a = \frac{2 \cdot \ln\left(\frac{0.8}{1-0.8}\right)}{x_{80\%} - x_{20\%}} = \frac{2 \cdot \text{logit}(0.8)}{x_{80\%} - x_{20\%}}.$$

## Maksimum likelihood estimation

I den logistiske regressionsmodel indgår to ukendte parametre  $a$  og  $b$ . Hvis vi har et datasæt, hvordan finder vi så de værdier af  $a$  og  $b$  der passer bedst til vores data? Som regel bruges maksimum likelihood metoden, som er en teknik, der stammer fra statistikken. Kort fortalt er idéen at vælge de parametre, der gør vores data så sandsynligt som muligt.

Lad os kalde punkterne i vores datasæt  $(x_i, y_i)$ , hvor  $i = 1, \dots, n$  er en nummerering af datapunkterne. Her angiver  $x_i$  blodtrykket hos den  $i$ 'te person, og  $y_i$  er en variabel, der antager værdien 1 hvis  $i$ 'te person har hjerte-kar-sygdom og er 0 ellers, dvs.

$$y_i = \begin{cases} 1, & \text{hvis } i\text{'te person har hjerte-kar-sygdom,} \\ 0, & \text{hvis } i\text{'te person ikke har hjerte-kar-sygdom.} \end{cases}$$

For hvert par  $(x_i, y_i)$  kan vi nu forsøge at beregne sandsynligheden  $p_i$  for at  $i$ 'te person faktisk har den sygsomsstatus  $y_i$  som vi observerer når vi ved at blodtrykket er  $x_i$ . Hvis den  $i$ 'te person er syg, dvs.  $y_i = 1$ , er

$$p_i = p(x_i) = \frac{e^{ax_i+b}}{1 + e^{ax_i+b}}.$$

Hvis patienten er rask, altså  $y_i = 0$ , er

$$p_i = 1 - p(x_i) = 1 - \frac{e^{ax_i+b}}{1 + e^{ax_i+b}} = \frac{1}{1 + e^{ax_i+b}}.$$

Bemærk at  $p_i$  afhænger af de ukendte parametre  $a$  og  $b$ . Vi kan altså opfatte  $p_i$  som en funktion  $p_i(a, b)$ .

Nu kigger vi på den samlede sandsynlighed for at observere netop de værdier  $y_1, \dots, y_n$ , som vi faktisk har observeret, når vi ved at patienternes blodtryk er givet ved  $x_1, \dots, x_n$ . Til det formål antager vi, at personerne i datasættet er udvalgt uafhængigt af hinanden. (Afhængigheder kan fx opstå hvis mange af personerne er i familie med hinanden, går i klasse sammen eller bor i samme by. I så fald kan de have noget tilfælles, der gør at deres  $y$ -værdier er mere ens end ellers. Familiemedlemmer kan fx have samme arvelige tendens til hjerte-kar-sygdom. Som regel forsøger man at undgå sådanne afhængigheder når man indsamler data.)

For at komme videre, er vi nødt til at vide lidt om *uafhængighed af hændelser*: Husk på at to hændelser  $A$  og  $B$  er uafhængige hvis man kan finde sandsynligheden for *fælleshændelsen*  $A \cap B$  (at  $A$  og  $B$  indtræffer på en gang) ved at gange de enkelte sandsynligheder sammen:

$$P(A \cap B) = P(A) \cdot P(B).$$

Uafhængighed af  $n$  hændelser  $A_1, \dots, A_n$  betyder tilsvarende at sandsynligheden for at alle  $n$  hændelser indtræffer på samme tid  $A_1 \cap A_2 \cap \dots \cap A_n$  kan findes som et produkt af sandsynlighederne for de enkelte hændelser:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_n).$$

(Desuden skal der gælde, at hver gang vi udtager  $m$  ud af de  $n$  hændelser, skal sandsynligheden for de  $m$  hændelser indtræffer samtidig kunne findes ved en tilsvarende produktformel, men dette skal vi ikke bruge i det følgende).

For at vende tilbage til vores data så lader vi  $A_1$  være hændelsen at første patient har sygdomsstatus  $y_1$ ,  $A_2$  være hændelsen at anden patient har sygdomsstatus  $y_2$  osv. Bemærk at  $P(A_i)$  er

det samme som det vi tidligere kaldte  $p_i(a, b)$ . Hændelsen at vi observerer alle de  $y_1, \dots, y_n$  som vi faktisk observerer på samme tid er fælleshændelsen  $A_1 \cap A_2 \cap \dots \cap A_n$ . Da vi antog at de  $n$  personer er udvalgt uafhængigt af hinanden, kan vi bruge produktformlen:

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) \cdots P(A_n) = p_1(a, b) \cdot p_2(a, b) \cdots p_n(a, b).$$

Bemærk, at sandsynligheden for vores udfald  $y_1, \dots, y_n$  afhænger af  $a$  og  $b$ . Den kan derfor betragtes som en funktion af to variable

$$L(a, b) = p_1(a, b) \cdot p_2(a, b) \cdot \dots \cdot p_n(a, b).$$

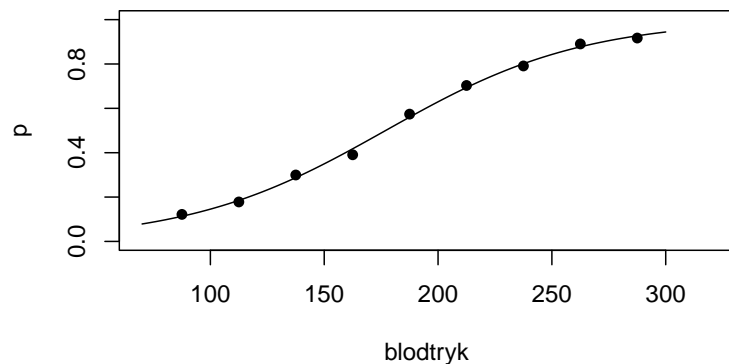
Denne funktion kaldes likelihoodfunktionen. Idéen med maksimum likelihood metoden er at vælge de værdier af  $a$  og  $b$ , der gør sandsynligheden  $L(a, b)$  for det vi har observeret så stor som muligt. Vi søger altså de  $a$  og  $b$ , der maksimerer funktionen  $L(a, b)$ . Dette maksimum kan ikke udregnes eksakt. I stedet kan man bruge numeriske metoder, fx gradient descent, som I kan læse mere om her... **link til note**. Disse optimeringsmetoder vil gøre brug af de partielle afledte, og sidst i dokumentet gennemgås flere detaljer til dette.

Finder man  $a$  og  $b$  ved hjælp af maksimum likelihood metoden i vores dataeksempel, fås funktionen

$$p(x) = e^{0.0230x - 4.07}.$$

$$T_{jekoppomdeter0,023eller0,03}$$

Grafen for  $p$  er vist nedenfor. Vi har en odds-ratio på  $e^{0.023} \approx 1,0233$ . Odds for hjerte-kar-sygdom stiger derfor med en faktor 1,0233 (altså 2,33%) for hver gang blodtrykket stiger med 1 mmHg.



## Prædiktion

Lad os nu sige, at vi har estimeret  $a$  og  $b$  ud fra et datasæt. Vi har altså fundet en model for risikoen for sygdom

$$p(x) = \frac{e^{ax+b}}{1 + e^{ax+b}}.$$

Hvis vi så får en ny patient med blodtryk  $x$ , kan vi prøve at bruge modellen til at forudsige (prædiktere) om patienten er syg eller ej. Vi kan beregne sandsynligheden  $p(x)$  for at patienten er syg ud fra vores model. Den mest oplagte prædiktionsregel er at prædiktere det mest sandsynlige:

- Hvis  $p(x) > 1/2$ : Patienten er syg.
- Hvis  $p(x) \leq 1/2$ : Patienten er rask.

I praksis kan der dog være et problem, hvis man gerne vil kunne forudsige en meget sjælden sygdom. I dette tilfælde vil det ofte være sådan at  $p(x) \leq 1/2$  for alle patienter. Ingen ville blive diagnosticeret med sygdommen på denne måde - og så er prædiktionsalgoritmen jo ikke meget værd. Derfor vælger man ofte et lavere delepunkt end  $p(x) = 1/2$ . Dermed kommer man til at fejldiagnosticere en hel del patienter. Til gengæld får man fanget flere af dem, der faktisk er syge.



Her på siden har vi flere eksempler på algoritmer, som vil kunne bruges til at prædiktere om patienter er syge eller raske, fx neurale netværk og Bayes klassifikation (flere?)<sup>3</sup>. Fordelen ved at bruge logistisk regression er, at man ikke bare får en prædiction, men også en model for, hvordan sandsynligheden  $p(x)$  afhænger af variabelen  $x$ . Dermed opnår man en indsigt i, hvordan sammenhængen mellem fx blodtryk og hjerte-kar-sygdom er. Ved hjælp af odds-ratioer kan vi endda sætte tal på hvordan odds for sygdom ændrer sig når blodtrykket vokser. Dette er i modsætning til mange andre prædiktionsalgoritmer, der blot giver en prædiction, uden at brugeren af algoritmen ved hvor den kommer fra. Inden for medicin er det ofte vigtigt at kende baggrunden for en given prædiction, så man kan forholde sig kritisk til resultatet og rådgive patienten om hvordan man sænker risikoen for sygdom (fx med blodtrykssænkende medicin). Til gengæld har de mere avancerede algoritmer mulighed for at give en mere præcis prædiction.

<sup>3</sup> Logistisk regression er i øvrigt et meget simpelt eksempel på et neuralt netværk.

## Multipel logistisk regression

I praksis er der selvfølgelig flere faktorer end blodtryk, der afgør ens risiko for hjerte-kar-sygdom. Fx stiger risikoen med alderen, ligesom rygning øger risikoen. Vi kan opstille en model, der inddrager alle tre variable på en gang. Lader vi  $x_1$  betegne blodtryk,  $x_2$  betegne alder, og  $x_3$  betegne antal cigaretter, man ryger pr. dag, kan man tilføje dem til formelen for log-odds ved

$$\ln O(x_1, x_2, x_3) = a_1x_1 + a_2x_2 + a_3x_3 + b,$$

dvs.

$$O(x_1, x_2, x_3) = e^{a_1x_1 + a_2x_2 + a_3x_3 + b}$$

og

$$p(x_1, x_2, x_3) = \frac{e^{a_1x_1 + a_2x_2 + a_3x_3 + b}}{1 + e^{a_1x_1 + a_2x_2 + a_3x_3 + b}}.$$

Hvordan skal vi forstå denne model? Jo, lad os forestille os en patient med alder  $x_1$  og blodtryk  $x_2$ , som ryger  $x_3$  cigaretter om dagen. Hvis vedkommende begynder at ryge 1 cigaret mere

om dagen (og vi forestiller os at alder og blodtryk er uændret) så vil odds-ratioen være

$$\frac{O(x_1, x_2, x_3 + 1)}{O(x_1, x_2, x_3)} = \frac{e^{a_1 x_1 + a_2 x_2 + a_3 (x_3 + 1) + b}}{e^{a_1 x_1 + a_2 x_2 + a_3 x_3 + b}} = e^{a_3}.$$

Den ekstra daglige cigaret vil altså øge odds for sygdom med en faktor  $e^{a_3}$ . Tilsvarende har  $e^{a_1}$  og  $e^{a_2}$  fortolkninger som odds-ratioer, når hhv. blodtryk og alder stiger med 1 mens alle andre variable fastholdes. Selv om modellen tager alle tre variable i betragtning, får vi altså mål for den individuelle effekt af hver af de tre variable.

Maximum likelihood metoden kan igen benyttes til at estimere parametrene  $a_1, a_2, a_3$  og  $b$ . Framingham ?????????? datasættet som vi så på tidligere indeholder alle de tre nævnte variable. Estimerer man parametrene, får man følgende

$$O(x_1, x_2, x_3) = e^{0.02x_1 + 0.06x_2 + 0.02x_3 - 6.77}.$$

Odds for hjerte-kar-sygdom stiger således med en faktor  $e^{0.02} \approx 1,02$  (altså med 2%) for hver cigaret man ryger om dagen.

## anvendelser af logistisk regression til andet

Predektion af sygdom ud fra måling af stof, hvor det er naturligt at man har stof i krop, men højt indhold vil indikere en sygdom. Evt. blodsukker og diabetes. Undersøgelse af om en person tænkes at fortage køb i forretning ud fra viden om alder og indkomst.

## Sammenhæng mellem logistisk regression og logistisk vækst

Det er selvfølgelig oplagt at tænke på om der er en sammenhæng mellem logistisk regression og logistisk vækst, både ud fra navnet men også ud fra grafernes forløb.

Ved logistisk vækst har vi en funktion af typen

$$f(x) = \frac{M}{1 + c \cdot e^{-M \cdot a \cdot x}}$$

Hvor  $c$  normalt er et positivt tal og  $M$  er en øvre grænse for  $f(x)$  og hvor

$$\lim_{x \rightarrow \infty} f(x) = M.$$

Funktionen er desuden løsning til differentialligningen

$$y' = a \cdot y \cdot (M - y)$$

hvor  $a > 0$ .

Lad os se på et eksempel hvor man ønsker at se på hvor hurtigt det går før en befolkning får immunitet mod en sygdom, og hvor vi med tiden regner med at alle bliver immune. For at beskrive dette ved logistik vækst vil man typisk lave et datasæt hvor man til givne tidspunkter ved hvor mange der er immune. Egentlig kræver det stort arbejde at få et sådan datasæt da det kræver information fra alle personer.

I stedet kunne man for hver dag der går tage en test af en (eller flere) tilfældig person og se om personen er immun eller ikke er immun. Derved får man et datasæt med punkter  $(x, y)$  hvor  $x$  er tid(dage) og  $y$  kunne være 0 svarende til ikke immun eller 1 svarende til immun. Dette datasæt er derved oplagt at undersøge med logistisk regression. Og derfra få en model af typen

$$p(x) = \frac{1}{1 + e^{-(a \cdot x + b)}} = \frac{1}{1 + e^{-a \cdot x - b}}.$$

Her ses selvfølgelig ikke på hvor mange der er immune men istedet sandsynligheden for at en tilfældig udvalgt er immun til tidspunktet  $x$ .

Det kan selvfølgelig være at man ønsker et bud på hvor mange der faktisk er immune blandt hele befolkningen med  $M$  personer. Hvis  $f(x)$  er antal immune efter  $x$  dage må det være oplagt at benytte modellen

$$f(x) = M \cdot p(x) = \frac{M}{1 + e^{-a \cdot x - b}}.$$

Ved  $a$  kan vi lige gange og dividere med  $M$  og kalde  $a/M$  for  $a_{ny}$ . Derved er

$$f(x) = \frac{M}{1 + e^{-M \cdot a_{ny} \cdot x - b}}.$$

Ved at benytte potensregnerregel får vi

$$f(x) = \frac{M}{1 + e^{-b} \cdot e^{-M \cdot a_{ny} \cdot x}}.$$

Hvis vi kalder  $e^{-b}$  for  $c$  får vi

$$f(x) = \frac{M}{1 + c \cdot e^{-M \cdot a_{ny} \cdot x}}.$$

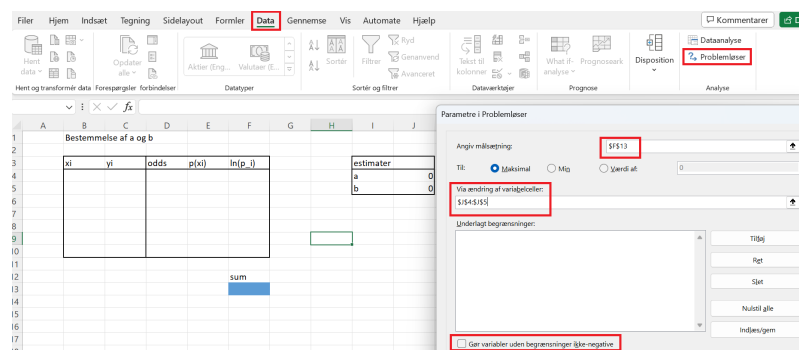
Som netop svarer til funktionstypen hørende til logistisk vækst.

Det er dog ikke ved alle eksempler at det er oplagt at lave denne kobling mellem de to emner, og det kræver desuden at  $M$  er kendt.

Andre eksempler hvor man kunne lave sammen kobling kunne f.eks. være undersøgelse af hvor hurtigt folk skifter til et nyt system som Mitid eller andet som alle med tiden skal bruge.

## Bestemmelse af $a$ og $b$ med excel-solve-værktøj

Man kan benytte excel til at finde estimat for  $a$  og  $b$ . Her skal man først og fremmest sørge for at man har aktiveret problemløser-værktøjet som gøres på følgende måde: Gå op under filer og vælg indstillinger. Derefter vælges tilføjelsesprogrammer. Nederst kan man vælge Excel-tilføjelsesprogrammer og trykke udfør. Til sidst kan man endelig tilføje problemløser-værktøjet fra en liste.



Figur 1: Illustration af excel ark til bestemmelse af  $a$  og  $b$  samt brug af problemløser

På billedet ses hvordan man kan lave et lille regneark hvor man kan beregne de vigtige størrelser. Desuden der der vist med rød hvor man finder problemløser og hvad der skal justeres.

Når man laver excel-arket er der lavet et par celler til de ukendte parametre  $a$  og  $b$ , som egentlig bare kan sættes til 0 fra starten. Ved beregning i cellerne med  $p(xi)$  er det selvfølgelig vigtigt at disse celler benyttes (og fastlåsning af referencere med vil være smart, hvor man har \$ foran både tal og bogstav ved reference).

Efter beregning af  $p(xi)$  vil det være oplagt at benytte formel ???, og fra noget af det efterfølgende teori følger det at

$$\ln(pi) = y_i \cdot \ln(p(x_i)) + (1 - y_i) \cdot \ln(1 - p(x_i)).$$

Til sidst skal man have fundet summen af alle  $l(p_i)$ , da det netop giver  $l(a, b)$ . Nu mangler man bare at benytte problemløseren til at sørge for at summen bliver maksimal ved at de to parametre  $a$  og  $b$  justeres.

## Ekstra detaljer til optimering

For at finde et maksimum eller minimum for en funktion af flere variable vil det være oplagt at se på de partielle afledte ligesom hvis man finder stationære punkter for en funktion af to variable.

Vi havde

$$L(a, b) = p_1(a, b) \cdot p_2(a, b) \cdot \dots \cdot p_n(a, b).$$

Tidligere indså vi også at

$$p_i(a, b) = \begin{cases} p(x_i), & \text{hvis } y_i = 1, \\ 1 - p(x_i), & \text{hvis } y_i = 0. \end{cases}$$

Denne kan egentlig skrives på følgende måde:

$$p_i(a, b) = p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i}.$$

For at se dette er det oplagt lige at se hvad der sker når  $y_i$  er henholdsvis 0 og 1 i udtrykket.

Dette udtryk benyttes når vi opskriver udtrykket for  $L(a, b)$  hvor vi desuden benytter notation produkt-notationen

$$L(a, b) = \prod_{i=1}^n (p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i}).$$

Istedet for at optimere ud fra  $L(a, b)$  kan vi optimere ud fra funktionen  $\ln(L(a, b))$  som nu betegnes med  $l(a, b)$ . Dette kan gøres da  $\ln(x)$  er en voksende funktion.

Ved at benytte at  $\ln(a \cdot b) = \ln(a) + \ln(b)$  adskillige gange får vi at

$$l(a, b) = \ln(L(a, b)) = \sum_{i=1}^n (\ln(p(x_i)^{y_i}) + \ln((1 - p(x_i))^{1-y_i})).$$

Ved at benytte at  $\ln(a^k) = k \cdot \ln(a)$  adskillige gange får vi at

$$l(a, b) = \sum_{i=1}^n (y_i \cdot \ln(p(x_i)) + (1 - y_i) \cdot \ln(1 - p(x_i))).$$

Ved at ophæve parentesen  $(1 - y_i)$  fås

$$l(a, b) = \sum_{i=1}^n (y_i \cdot \ln(p(x_i)) + \ln(1 - p(x_i)) - y_i \cdot \ln(1 - p(x_i))).$$

I to af ledene indenfor sumtegnet har vi  $y_i$  som en faktor, og  $y_i$  sættes udenfor parentes

$$l(a, b) = \sum_i (\ln(1 - p(x_i)) + y_i \cdot (\ln(p(x_i)) - \ln(1 - p(x_i)))).$$

Ved hjælp af logaritme-regnereglen  $\ln(a/b) = \ln(a) - \ln(b)$  får vi

$$l(a, b) = \sum_{i=1}^n \left( \ln(1 - p(x_i)) + y_i \cdot \ln \left( \frac{p(x_i)}{1 - p(x_i)} \right) \right).$$

Her opsplitter vi til to summer hvor den ene ikke afhænger af  $y_i$ .

$$l(a, b) = \sum_{i=1}^n \ln(1 - p(x_i)) + \sum_{i=1}^n y_i \cdot \ln\left(\frac{p(x_i)}{1 - p(x_i)}\right).$$

Nu har vi fået styr over udtrykket for  $l(a, b)$  og vi har et udtryk for  $p(x_i)$  som dog afhænger af  $a$  og  $b$ .

Ved at kombinere ligningerne ??? får vi at

$$l(a, b) = \sum_{i=1}^n \ln\left(1 - \frac{e^{a \cdot x_i + b}}{1 + e^{a \cdot x_i + b}}\right) + \sum_{i=1}^n y_i \cdot (ax_i + b).$$

Udtrykket i logaritmen laves om til en samlet brøk

$$l(a, b) = \sum_{i=1}^n \ln\left(\frac{1}{1 + e^{a \cdot x_i + b}}\right) + \sum_{i=1}^n y_i \cdot (ax_i + b).$$

Her benytter vi igen regnereglen  $\ln(a - b) = \ln(a) - \ln(b)$ .

$$l(a, b) = \sum_{i=1}^n (\ln(1) - \ln(1 + e^{a \cdot x_i + b})) + \sum_{i=1}^n y_i \cdot (ax_i + b).$$

Da  $\ln(1) = 0$  har vi endelig

$$l(a, b) = \sum_{i=1}^n -\ln(1 + e^{a \cdot x_i + b}) + \sum_{i=1}^n y_i \cdot (ax_i + b).$$

Hvis man ser på bidraget fra  $p_i(a, b)$  bliver det til

$$\ln(p_i(a, b)) = y_i \cdot \ln(p(x_i)) + (1 - y_i) \cdot \ln(1 - p(x_i)).$$

## Partiel afledte for $l(a, b)$

Lad os først se på  $\frac{\partial l(a, b)}{\partial b}$ . Ved den første sum skal vi se udtrykket som en sammensat funktion hvor den indre funktion har et led som også er en sammensat funktion.

$$\frac{\partial l(a, b)}{\partial b} = \sum_{i=1}^n -\frac{1}{1 + e^{a \cdot x_i + b}} \cdot (0 + e^{a \cdot x_i + b}) \cdot (0 + 1) + \sum_{i=1}^n y_i \cdot (0 + 1).$$

Ved at reducere fås

$$\frac{\partial l(a, b)}{\partial b} = \sum_{i=1}^n -\frac{e^{a \cdot x_i + b}}{1 + e^{a \cdot x_i + b}} + \sum_{i=1}^n y_i$$

Ved at bruge ligning ?? i forbindelse med den første sum og efterfølgende omskrive til en sum fås

$$\frac{\partial l(a, b)}{\partial b} = \sum_{i=1}^n -p(x_i) + \sum_{i=1}^n y_i = \sum_{i=1}^n (y_i - p(x_i)).$$

Nu ser vi på  $\frac{\partial l(a, b)}{\partial a}$ . på tilsvarende måde.

$$\frac{\partial l(a, b)}{\partial a} = \sum_{i=1}^n -\frac{1}{1 + e^{a \cdot x_i + b}} \cdot (0 + e^{a \cdot x_i + b}) \cdot (1 \cdot x_i + 0) + \sum_{i=1}^n y_i \cdot (1 \cdot x_i + 0)$$

Der reduceres

$$\frac{\partial l(a, b)}{\partial a} = \sum_{i=1}^n -\frac{e^{a \cdot x_i + b}}{1 + e^{a \cdot x_i + b}} \cdot x_i + \sum_{i=1}^n y_i \cdot x_i.$$

Igen bruges ligning ?? til at få

$$\frac{\partial l(a, b)}{\partial a} = \sum_{i=1}^n -p(x_i) \cdot x_i + \sum_{i=1}^n y_i \cdot x_i = \sum_{i=1}^n (y_i \cdot x_i - p(x_i) \cdot x_i).$$

Endelig kan  $x_i$  sættes udenfor parentes hvorved vi har

$$\frac{\partial l(a, b)}{\partial b} = \sum_{i=1}^n (y_i - p(x_i)) \cdot x_i.$$

For at lave optimering og finde minimum skal vi derved løse følgende to ligninger med to ubekendte

$$0 = \sum_{i=1}^n (y_i - p(x_i)) \cdot x_i \quad \text{og} \quad 0 = \sum_{i=1}^n y_i - p(x_i).$$

Dette ligningssystem er dog ikke bare lige til at løse, så enten skal man igang med at benytte numeriske metoder til løsninger af dette eller benytte sig af gradient nedstigning (reference) for at få det optimale bud på parametrene  $a$  og  $b$  som fastlægger funktionen  $p(x)$ .