Logistisk regression

Aalborg Intelligence

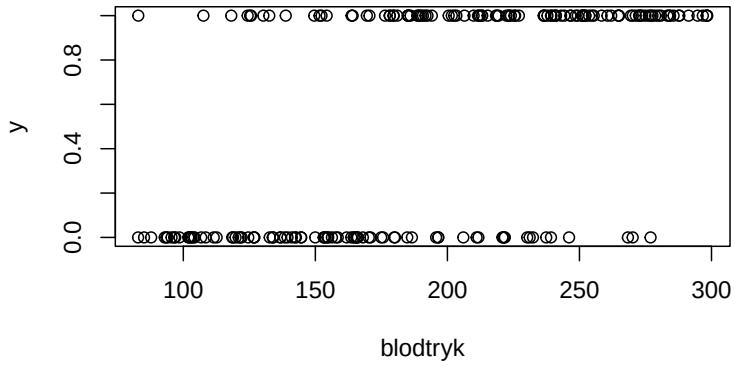
Logistisk regression

Denne note skal handle om 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. I slutningen af dette dokument forklarer

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[1]. 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 samhørende x- og y-værdier ind i et koordinatsystem for de første 200 personer i datasættet.

[1] 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 $p(x)$ 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

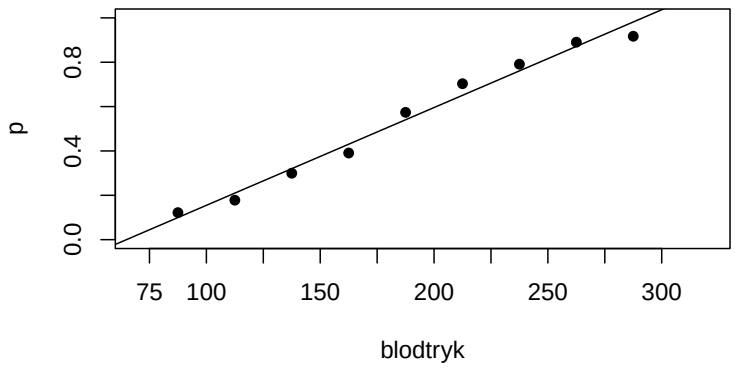
For hvert interval beregner andelen af patienterne, der lider af hjerte-kar-sygdom som estimat for sandsynligheden for hjerte-kar-sygdom i den gruppe. Fx er der 236 personer med et blodtryk i intervallet $]100, 125]$. Heraf er 42 syge. Andelen af syge i denne gruppe er derfor

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

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

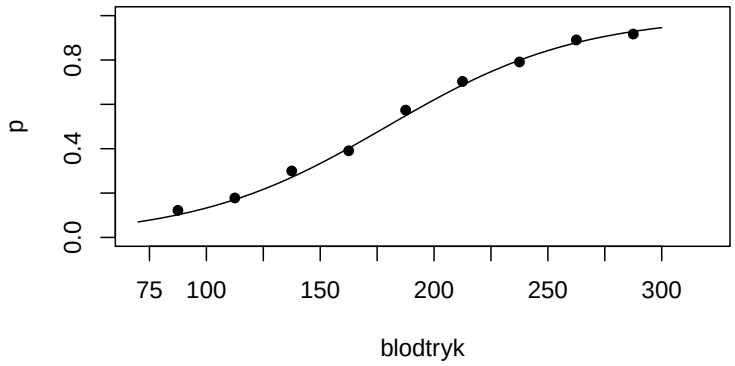
$$p(x) = ax + b.$$

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[2]. 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

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

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:

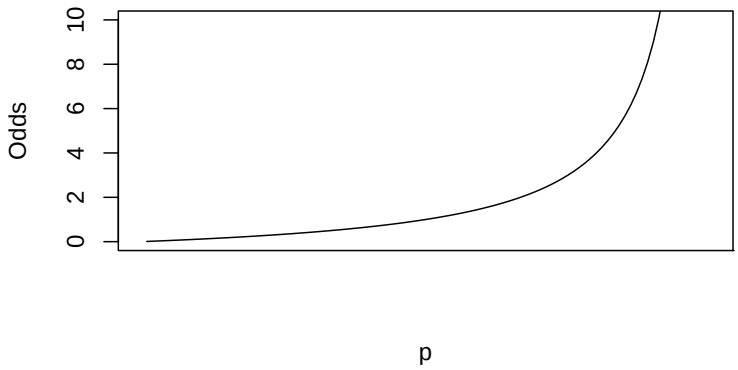
[2] Du kender måske begrebet odds fra sportsgambling. Det er dog en anden betydning af ordet end det vi bruger her. Inden for gambling angiver odds, hvor mange gange man får pengene igen, hvis en bestemt hændelse indtræffer(fx. et bestemt hold vinder). Gambling odds er naturligvis udregnet ud fra de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

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.

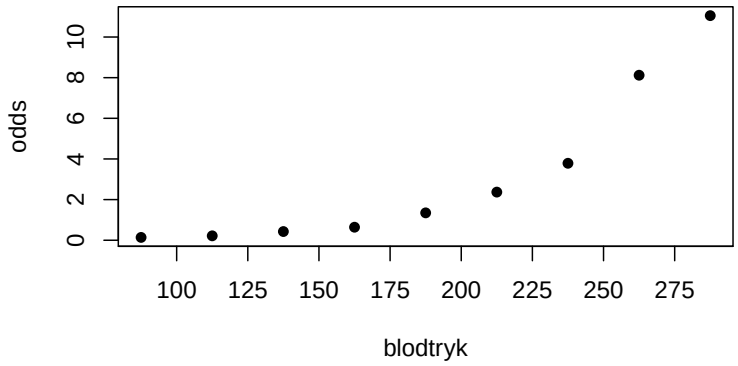


Vi ser at f altid er positiv, da både tæller og nævner er positive. Når p nærmer sig 0, nærmer tælleren sig 0 mens nævneren nærmer sig 1, så $f(p)$ går mod 0. Når p nærmer sig 1, nærmer tælleren sig 1 og nævneren nærmer sig 0, så hele brøken $f(p)$ går mod uendelig. Værdimængden for f er altså hele den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

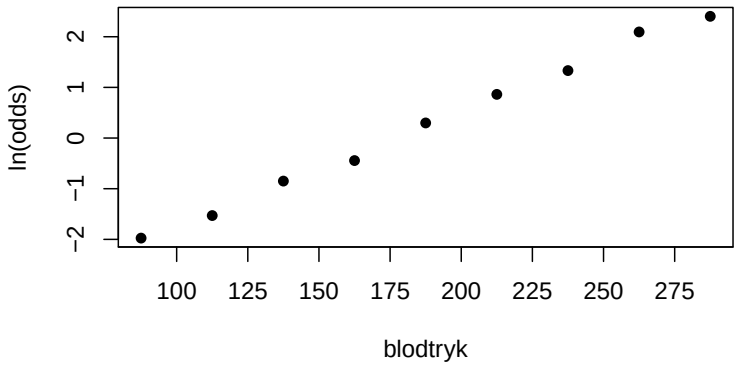
I vores dataeksempel, hvor sandsynligheden for hjerte-karsygdom er en funktion $p(x)$, bliver oddsene for hjerte-karsygdom også en funktion af x

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

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.

Logit-funktionen og den logistiske funktion

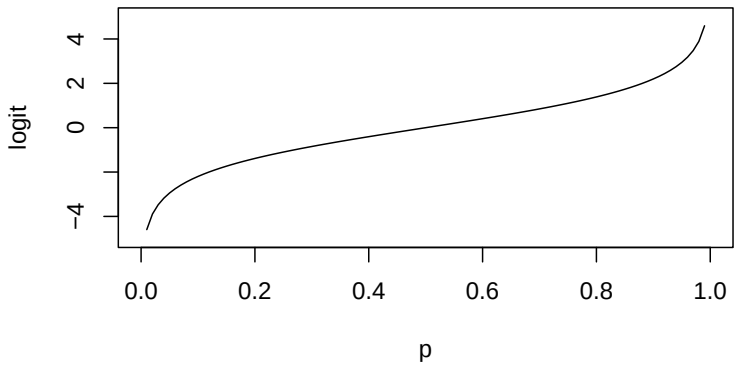
Når vi tager den naturlige logaritme til oddsene får vi

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

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.

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

Vi kan så lægge pe^y til på begge sider og sætte uden for parantes

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

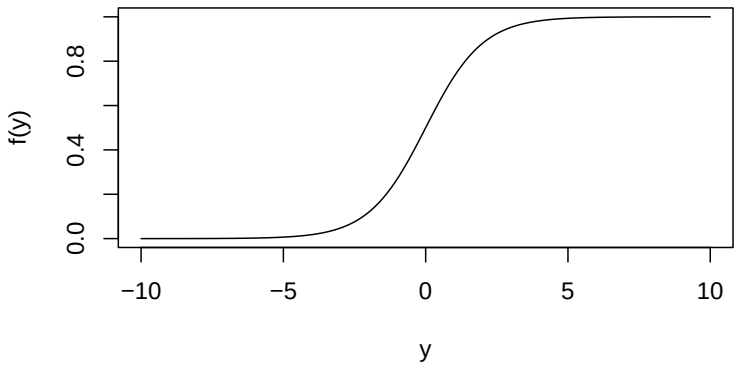
Endelig kan vi isolere p og få

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

I sidste udregning forkortede vi brøken med 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.



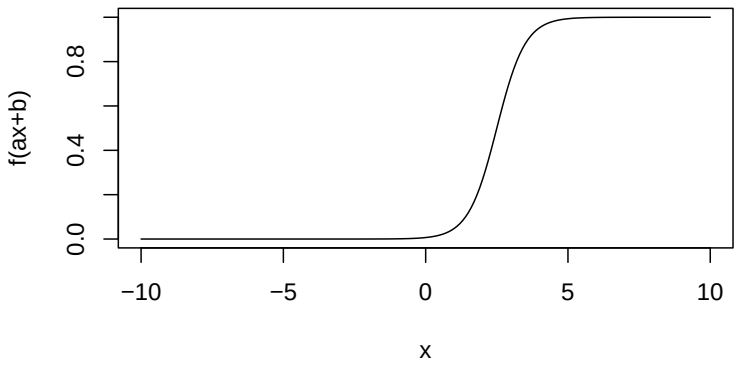
Fra Allan:

Hvis man har en funktion h kan vækshastigheden øges med en faktor a ved at se på $h(ax)$. Hvis desnæst ønsker at forskyde med grafen med $-\frac{b}{a}$ i x-aksens retning kan det gøres ved at se

på $h(ax+b)$. Hvis man gør dette for den standard logistiske funktion, får man den generelle logistiske funktion

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

Sammenlignet med den standard logistiske funktion får man altså en S-formet kurve, der vokser a gange så hurtigt og er forskudt med $-b$. 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 ?.

NOTE: Dette kunne også gøres interaktivt.

NOTE: Opgaver?

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

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

Vi ser altså, at oddsene vokser eksponentielt. Fremskrivningsfaktoren (kalder I den det?) er e^a. 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

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

Vi får altså et logistisk funktionsudtryk for p, deraf navnet "den logistiske regressionsmodel".

Maksimum likelihood estimation

NOTE: Skal flettes sammen med tekst fra Allan

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 $i = 1, \ldots, n$ 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 $i$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 $y_1, \ldots, y_n$, som vi faktisk har observeret, når vi ved at patienternes blodtryk er givet ved $x_1, \ldots, x_n$. 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:

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

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

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

(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 $y_1, \ldots, y_n$ som vi faktisk observerer på samme tid er fælleshændelsen $A_1 \cap A_2 \cap \cdots \cap A_n$. Da vi antog at de n personer er udvalgt uafhængigt af hinanden, kan vi bruge produktformlen:

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

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

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

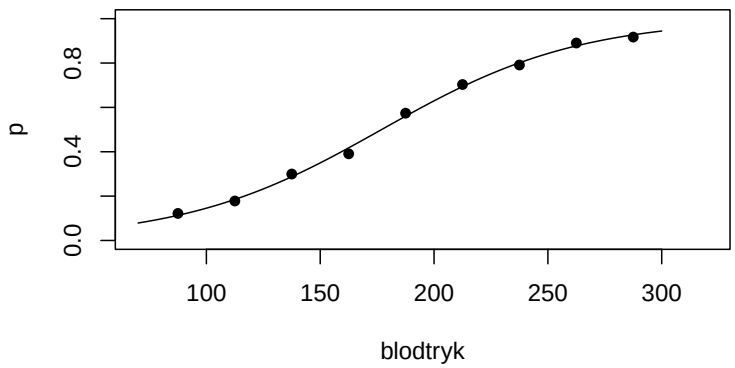
Denne funktion kaldes likelihoodfunktionen. Idéen med maksimum likelihood metoden er at vælge de værdier af a og b, der

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**

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

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

Grafen for p er vist nedenfor. Vi har en odds-ratio på $e^{0.03} \approx 1,03$. Odds for hjerte-kar-sygdom stiger derfor med en faktor 1,03 (altså 3%) for hver gang blodtrykket stiger med 1 mmHg.



NOTE: Der er to tilgange til maksimering af funktionen: enten gradient descent eller ved at differentiere. Skal vi give dem begge, eller er det nok med en af dem?

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?)[3]. Fordelen ved at bruge logistisk regression er, at man ikke bare får en prædiktion, men også en model for, hvordan sandsynligheden $p(x)$ afhænger af variablen 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 prædiktion, uden at brugeren af algoritmen ved hvor den kommer fra. Inden for medicin er det ofte vigtigt at kende baggrunden for en given prædiktion, 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 prædiktion.

[3] 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 formlen for log-odds ved

$$\ln O(x_1, x_2, x_3) = a_1 x_1 + a_2 x_2 + a_3 x_3 + b,$$

dvs.

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

og

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

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 a^{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

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

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.

Sammenhæng mellem logistisk regression og logistisk vækst

Allans tekst...