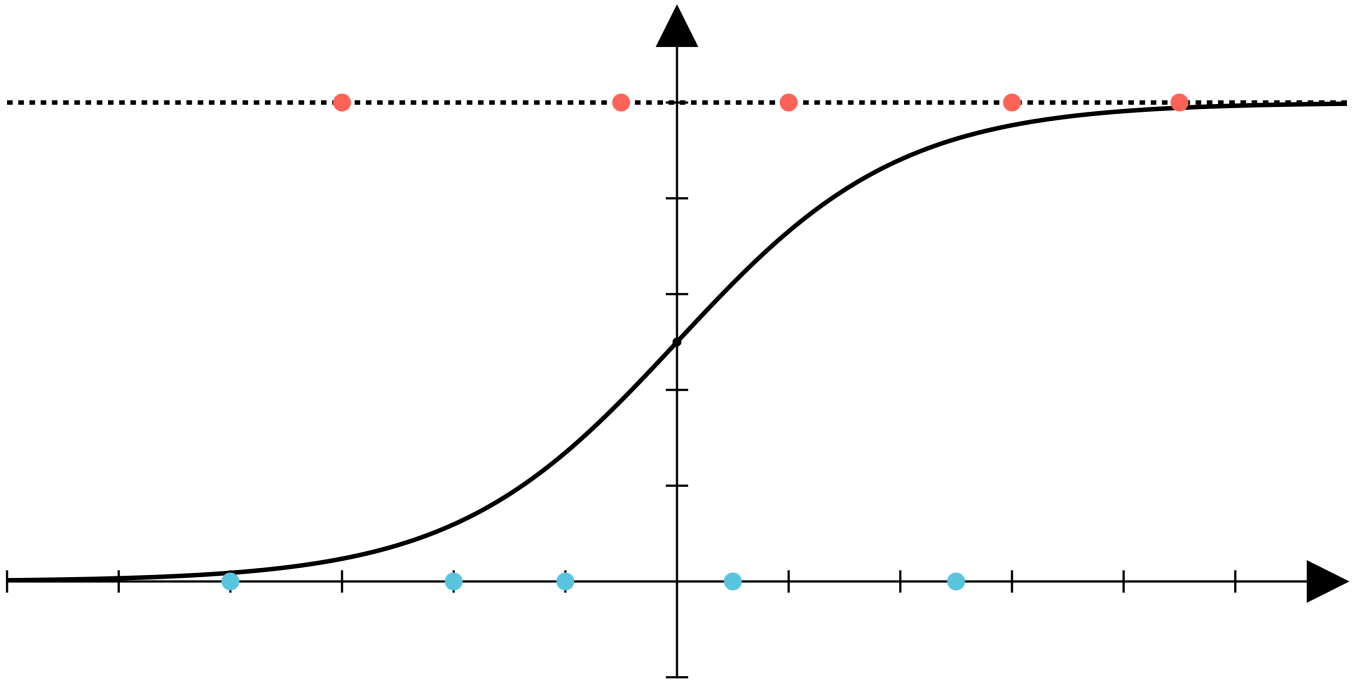
AARHUS BSS
SCHOOL OF BUSINESS AND SOCIAL SCIENCES
AARHUS UNIVERSITET



Logistisk Regression og Kreditrisiko

Et let tilgængeligt supplement til Matematik A-undervisningen
på gymnasiet og handelsskolen, med forklarende videoer og eksempler.

OPGAVEVERSION

Aarhus Universitet · Institut for Økonomi
2026, 1. udgave

Indhold

1 Introduktion til kreditrisiko

1.1 Hvad er kredit, og hvorfor er det vigtigt for samfundet?

Kredit handler om tid. I stedet for at spare op i mange år, før man kan dække en stor udgift, kan man låne pengene nu og afdrage dem over tid. Det gælder i privatøkonomien, når man for eksempel køber bolig, tager et SU-lån eller betaler en større udgift på afbetaling. Det gælder også for virksomheder, der bruger lån til at investere i udstyr, ansætte medarbejdere eller udvide produktionen.

Kredit bygger også på tillid: långiveren skal tro på, at låntageren kan og vil betale tilbage. Derfor skaber kredit både muligheder og risiko. Når en bank låner penge ud, har den brug for at vurdere kreditrisiko systematisk. I dette kompendium vil vi se, hvordan man kan omsætte oplysninger om en låntager til en sandsynlighed for misligholdelse.

Definition: Kredit og kreditrisiko

Kredit er en aftale, hvor en långiver (typisk en bank) stiller penge til rådighed for en låntager, mod at låntageren betaler pengene tilbage over tid, typisk med renter.
Kreditrisiko er risikoen for, at låntageren ikke kan eller vil betale lånet tilbage som aftalt.

Kredit fylder meget i samfundsøkonomien, og derfor kan selv små fejl i kreditvurdering blive dyre. Tabellen nedenfor viser størrelsesordener for udlån i Danmark.

Type	Beløb
Realkreditlån (bolig)	ca. 3.200 mia. kr
Banklån til private	ca. 500 mia. kr
Banklån til erhverv	ca. 1.200 mia. kr
I alt	**ca. 4.900 mia. kr**

Kilde: Danmarks Nationalbank og Finanstilsynet (2024).

Hvis banken i gennemsnit tabte 1% af udlånet, ville det svare til ca. 49 mia. kr. Det er baggrunden for, at banker bruger modeller til at vurdere sandsynligheden for misligholdelse, før de giver et lån.

1.2 Bankernes daglige udfordring: Hvem skal have lån?

Hver dag træffer banker beslutninger om lån, som har stor betydning for mennesker og virksomheder. Det kan være den unge familie, der vil købe bolig, iværksætteren, der vil investere i en idé, eller den studerende, der har brug for transport for at kunne passe et job. I alle tilfælde skal banken vurdere det samme grundspørgsmål: Hvor stor er risikoen for, at lånet ikke bliver betalt tilbage?

Det er ikke en beslutning, der kun handler om at sige ja eller nej. Banken forsøger at undgå to typer fejl, som trækker i hver sin retning.

Hvis banken siger ja til for mange kunder, som ender med at misligholde deres lån, får banken direkte tab. Det kan være tab på renter, tab på selve lånebeløbet og ekstra omkostninger til inddrivelse. Omvendt, hvis banken siger nej til for mange kunder, som faktisk godt kunne betale tilbage, går banken glip af en god forretning. Det kan også have konsekvenser for kunden, som mister muligheden for at købe bolig, investere i en virksomhed eller klare en nødvendig udgift.

Kreditvurdering handler derfor om at finde en rimelig balance: Banken vil gerne låne ud, men kun når risikoen er acceptabel. Det kræver, at banken kan regne en kundes oplysninger om til en sandsynlighed for misligholdelse. Det er præcis den type problem, som logistisk regression er designet til at håndtere.

Når man søger om lån, samler banken typisk en række oplysninger, som tilsammen siger noget om betalingsevne og risiko. En del handler om økonomi her og nu, for eksempel indkomst, faste udgifter, opsparing og eksisterende gæld. Andre oplysninger handler om stabilitet, for eksempel beskæftigelse, jobsikkerhed, hvor længe man har været i sit job, og om man har forsørgerpligt. Endelig indgår historik, for eksempel betalingshistorik, tidligere lån og eventuelle registreringer for manglende betaling.

I praksis skal banken omsætte disse oplysninger til en beslutning. Hvis vurderingen alene bygger på mavefornemmelse, kan to rådgivere ende med at vurdere den samme kunde forskelligt. Det er en vigtig grund til, at banker bruger matematiske modeller: Modeller kan give samme vurdering uanset hvilken rådgiver, der bruger dem, og kan gøre processen systematisk.

Det er vigtigt at være opmærksom på, hvilke oplysninger man overhovedet bør bruge. Nogle faktorer som alder, køn, civilstand eller etnicitet kan godt korrelere med betalingsevne, men de kan også føre til diskrimination eller uretfærdig behandling. Derfor forsøger man ofte at balancere modellens præcision med krav om fairness, lovgivning og etiske hensyn.

1.3 Historisk perspektiv: Fra bankrådgiver til algoritme

Kreditvurdering har eksisteret lige så længe som lån. Det, der har ændret sig, er måden beslutningen bliver truffet på: Fra personligt kendskab og mavefornemmelse til data, scorekort og algoritmer.

I 1800-tallet kendte bankdirektøren ofte sine kunder personligt, og beslutningen om lån byggede i høj grad på ry, relationer og lokale fortællinger. Hvem var kendt som pålidelig, og hvem "passede ind"? Den tilgang kunne virke intuitiv, men den havde en klar bagside: Den var sårbar over for bias og forskelsbehandling. Kvinder, indvandrere og mennesker fra "forkerte" familier kunne blive vurderet negativt, uanset deres reelle betalingsevne. Det var også svært at gøre beslutninger konsistente, fordi de afhængte af, hvem man talte med.

I 1950'erne opstod en ny idé, som stadig præger kreditvurdering i dag: I stedet for at vurdere, om en person virker pålidelig, kan man bruge historiske data til at se, hvad der typisk kendetegner dem, der betaler tilbage, og dem, der ikke gør. I USA blev denne tilgang kendt gennem Fair Isaac Corporation, og den førte til den såkaldte **FICO-score**, som stadig er udbredt. I Danmark bruger man andre systemer og skalaer, men princippet er det samme: Man forsøger at omsætte oplysninger om en låntager til en samlet score, der kan bruges i en beslutning.

I dag har banker adgang til langt flere data end tidligere og kan bruge mere avancerede modeller. I nogle lande bruger man også signaler fra digital adfærd og forbrugsmønstre, ud over indkomst,

gæld og betalingshistorik. Det kan give mere præcise vurderinger, men rejser samtidig spørgsmål om privatliv, fairness og hvilke data, der overhovedet er rimelige at bruge.

Der er også en afvejning mellem menneskelig og algoritmisk vurdering. Mennesker kan tage højde for særlige omstændigheder, men kan være inkonsistente og påvirket af bias. Algoritmer er hurtige og ensartede, men kan være svære at forklare og kan også gentage den bias, der findes i de historiske data. Derfor bruger mange banker en kombination, hvor modellen giver en anbefaling, og et menneske vurderer grænsetilfælde.

1.4 Case: De tre kunder

Forestil dig, at du sidder som bankrådgiver i en dansk bank og skal tage stilling til tre kunder, der alle søger om et **billån på 200.000 kr** til køb af en brugt bil.

	Kunde A (Emma)	**Kunde B (Peter)**	**Kunde C (Sofie)**
Alder (år)	28	45	22
Beskæftigelse	Sygeplejerske (fast)	Selvstændig IT-konsulent	Studerende + deltidsjob
Indkomst (kr/md)	28.000	45.000 (gns.; 35–55.000)	12.000 (SU + job)
Boligtype	Leje	Ejer	Kollegium
Familie	Single, ingen børn	Gift, 2 børn	Kæreste, ingen børn
Gældstype (primær)	Ingen	Realkreditlån	SU-lån
Gæld (kr)	0	2.000.000	40.000
Tidligere misligholdelse	nej	ja (tidligere konkurs)	nej
Formål med lånet	Transport til vagter/familie	Nødvendig bil til familien	Bil til praktikophold

Video 1 – Kreditrisiko: Problemet og konsekvenserne

Varighed: 5 minutter
Indhold: Videoen introducerer kreditrisiko gennem virkelige eksempler fra danske banker. Den viser:

- Hvordan banker traditionelt har vurderet låneansøgere
- Konsekvenserne af fejlvurderinger, både for banker og samfund
- De tre eksempelkunder præsenteres
- Eleverne opfordres til at gætte, hvem der bør få lån
- Selv med samme information har folk forskellig opfattelse af kundernes kreditrisiko: derfor har vi brug for en matematisk model.

▶ Se videoen

https://www.youtube.com/watch?v=_ey1mw-RXQc&list=PLIvNrmh87wtA&index=2

Når de tre ansøgere sammenlignes, bliver det tydeligt, hvorfor kreditvurdering ikke kan bero på mavefornemmelse alene. Med de samme oplysninger kan forskellige rådgivere nå til forskellige

konklusioner, og så bliver bankens beslutninger inkonsistente. Det er en central grund til, at banker bruger matematiske modeller, som kan omsætte en kundes oplysninger til en systematisk vurdering.

Selv gode modeller kan dog fejle, og fejlene kan blive dyre, hvis de gentager sig mange gange. Det blev blandt andet tydeligt under finanskrisen, hvor kreditvurdering og risikomodeller flere steder viste sig at være for optimistiske. En vigtig forklaring var, at modellerne i høj grad var formet af en periode med stigende boligpriser og derfor havde svært ved at håndtere et boligmarked i kraftigt fald. Derudover undervurderede man risikoen for, at mange tab kan ramme samtidigt. Når økonomien vender, kan mange få problemer på samme tid, og mange misligholdelser kan opstå på én gang.

Danmark klarede sig relativt godt, men vi var ikke immune. Flere banker fik alvorlige problemer i perioden efter 2008, blandt andet Roskilde Bank (2008), Fionia Bank (2009) og Amagerbanken (2011), og i årene 2008–2012 kollapsede eller blev en række banker overtaget af andre. Et gennemgående træk var høj koncentration af udlån til ejendomme og for optimistiske kreditvurderinger.

For det første er kredit vigtig, og banker har brug for at træffe ensartede beslutninger i et problem, hvor der altid er en afvejning mellem risiko og indtjening. For det andet skal modeller bruges med omtanke, fordi de kan fejle, hvis antagelserne er for stærke, eller hvis de data, modellen bygger på, ikke dækker de situationer, der kan opstå. Næste skridt er at se, hvordan logistisk regression kan bruges til at lave mere konsistente vurderinger af kreditrisiko ved at knytte en kundeprofil til en sandsynlighed.

Video 2 – Fra lineær til logistisk: Hvorfor vi har brug for S-kurven

Varighed: 5 minutter
Indhold: Videoen viser problemet med lineær regression gennem animationer:

- Starter med et velkendt eksempel: forudsige huspriser (virker fint!)
- Prøver samme tilgang på kreditdata – viser hvordan modellen fejler
- Demonstrerer sandsynligheder over 100% og under 0%
- Introducerer S-kurven som løsning
- Intuition: Hvorfor S-formen giver mening

▶ Se videoen

https://www.youtube.com/watch?v=h9wV_RdwEL8&list=PLIvNrmh87wtA&index=3

2 Fra mavefornemmelse til matematik

I forrige afsnit så vi, at vurderinger baseret på mavefornemmelse fører til inkonsistente beslutninger. Nu skal vi se, hvordan matematik kan hjælpe. Det kræver først, at vi præciserer, hvad banken forsøger at forudsige.

Når en bank giver et lån, er bankens største bekymring ikke, om kunden bliver "lidt dyr" at have,

men om kunden slet ikke betaler tilbage. Det er misligholdelse. Hvis en kunde misligholder, kan banken miste renter, bruge tid og penge på at inddrive lånet og i værste fald tabe en del af selve lånebeløbet. Derfor er det naturligt at formulere problemet som et spørgsmål om en sandsynlighed:

$$p = P(\text{misligholdelse})$$

Hvis banken kan knytte en ansøger til en sandsynlighed p, kan banken derefter træffe en beslutning ved at sammenligne p med en grænse, for eksempel "vi godkender, hvis p er lav nok". Det passer også godt til den måde, man taler om risiko på i praksis: som noget, der ligger mellem 0% og 100%.

I virkeligheden kombinerer banken mange oplysninger, når den vurderer en ansøger. Men for at gøre idéen så enkel som muligt starter vi med én oplagt forklarende variabel: **indkomst**. Intuitionen er, at en højere indkomst ofte giver bedre betalingsevne og dermed lavere risiko.

En naturlig måde at gøre vurderingen mere systematisk på er at give hver ansøger en **score**, som afhænger af indkomsten. Den simpleste version er en lineær sammenhæng, hvor man starter med en grundværdi og derefter justerer op eller ned alt efter indkomstniveauet. Det er den type model, man kender fra lineær regression, og den er et naturligt første forsøg, når man vil lave en matematisk model for risiko.

2.1 En første idé: lineær regression på sandsynlighed

En enkel første model er at gøre det, I allerede kender fra Matematik A, nemlig at bruge en lineær sammenhæng. Hvis vi lader x være indkomst (målt i 1.000 kr pr. måned), kan vi skrive en univariat regressionsmodel som:

$$p = \beta_0 + \beta_1 x$$

Her er β_0 en startværdi (det, modellen giver, når $x = 0$), og β_1 beskriver, hvor meget risikoen ændrer sig, når indkomsten stiger med 1.000 kr.

Hvis vi forventer, at højere indkomst giver lavere risiko, vil β_1 typisk være negativ.

Lad os se et konkret eksempel. Antag, at en bank (naivt) bruger følgende lineære model:

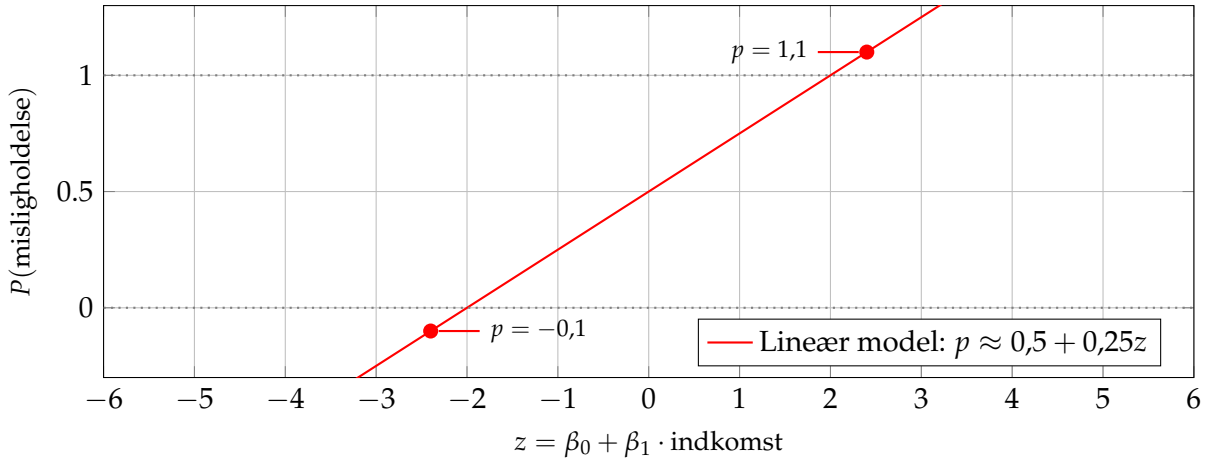
$$p = 0{,}80 - 0{,}02 \cdot x$$

hvor x er indkomst målt i 1.000 kr pr. måned. Modellen siger altså, at risikoen falder med 2 procentpoint for hver 1.000 kr mere i indkomst.

Opgaver

Opgaver om beregning af risiko med en lineær model er i Kapitel 6.1.

Lineære modeller har dog et alvorligt problem: Modellen kan give negative værdier for p. Men en sandsynlighed kan ikke være mindre end 0. Tilsvarende kan en lineær model også ende med at give værdier over 1 (over 100%), hvis man bevæger sig i den anden retning.



Figur 1: En lineær model $p = 0{,}5 + 0{,}25z$ giver umulige sandsynligheder uden for $[0, 1]$ for store eller små værdier af z.

Der er også et mere subtilt problem. I en lineær model har en ændring i indkomst altid den samme effekt på p. Hvis $\beta_1 = -0{,}02$, så reducerer en ekstra 1.000 kr indkomst risikoen med 2 procentpoint, uanset om udgangspunktet var 10% eller 90%. Det er ofte urealistisk, fordi en sandsynlighed har en nedre grænse på 0 og en øvre grænse på 1, og fordi forbedringer typisk "flader ud" nær 0% og nær 100%. Man vil ofte forvente, at effekten er størst midt i intervallet og mindre i yderpunkterne.

Vi har derfor brug for en model, som stadig kan kombinere information på en systematisk måde, men som altid giver et svar i intervallet $0 \leq p \leq 1$. I næste afsnit introducerer vi en funktion, der gør dette: den logistiske funktion.

2.2 S-kurven som løsning

Den lineære model fejler af én årsag: Vi prøver at bruge en model, der kan give alle tal på tallinjen, til noget, der skal være en sandsynlighed. Derfor har vi brug for en anden type funktion, som kan oversætte en score til en sandsynlighed.

Idéen er at gøre det i to trin. Først beregner vi en **lineær score**, der opsummerer informationen om kunden. Den kaldes ofte z og kan for eksempel være

$$z = \beta_0 + \beta_1 x,$$

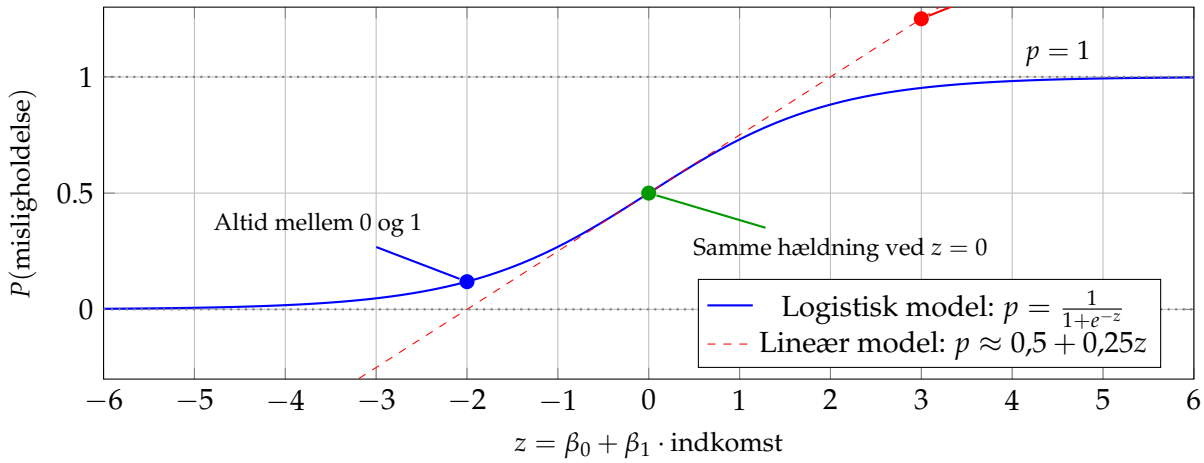
hvor x er indkomst (målt i 1.000 kr pr. måned). Fordi z er lineær, kan den i princippet antage alle værdier fra $-\infty$ til ∞. Negative værdier kan vi tænke som lavere risiko, og positive værdier som højere risiko, men z kan ikke i sig selv være en sandsynlighed. Derfor skal z oversættes til en sandsynlighed p med en funktion, der altid giver værdier mellem 0 og 1.

Denne oversættelse skal også give en rimelig sammenhæng: Hvis z bliver større, skal risikoen stige, men ikke uendeligt. Når risikoen allerede er tæt på 0% eller 100%, bør ekstra ændringer i z kun ændre sandsynligheden lidt. Det er nøjagtigt det, S-kurven gør.

Den logistiske funktion (S-kurven) er

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

Den gør to ting på én gang: Den sikrer, at p altid ligger mellem 0 og 1, og den giver en glidende overgang fra lav risiko til høj risiko.



Figur 2: S-kurven (blå) holder sig altid mellem 0 og 1, mens en lineær model (rød stiplet) kun approksimerer lokalt og giver umulige sandsynligheder for ekstreme værdier af z

I figuren viser x-aksen z, altså scoren: store negative værdier svarer til lav risiko, og store positive værdier svarer til høj risiko. y-aksen viser p, altså sandsynligheden for misligholdelse af lånet. Den blå S-kurve ligger altid mellem 0 og 1. Den går gennem $p = 0{,}5$, når $z = 0$, fordi en neutral score giver en "midt i mellem"-risiko. Når z bliver meget negativ, nærmer kurven sig 0, og når z bliver meget positiv, nærmer den sig 1. Den rammer aldrig præcis 0 eller 1, hvilket passer godt til, at man sjældent kan være helt sikker i praksis.

I figuren er den røde stiplede linje en lineær approksimation omkring $z = 0$. Den kan ligne S-kurven lokalt, men den kan ikke holde sig inden for intervallet $[0, 1]$ for store positive eller negative z. Det er den grundlæggende forskel: S-kurven er lavet til at håndtere sandsynligheder.

Opgaver

Opgaver om sammenligning af lineær model og S-kurven er i Kapitel 6.2.

Man kan tænke på S-kurven som en slags "oversætter" mellem to verdener. I input-verdenen har vi scoren z, som kan være et hvilket som helst tal på tallinjen, altså fra $-\infty$ til ∞. I output-verdenen vil vi derimod have en sandsynlighed p, og den skal altid ligge mellem 0 og 1. S-kurvens job er at tage et vilkårligt z og oversætte det til en gyldig sandsynlighed.

S-kurven forklarer desuden, hvorfor ændringer i z ikke bør have samme effekt overalt. Når risikoen allerede er tæt på 0 eller tæt på 1, skal der meget information til, før vi ændrer mening. Hvis risikoen ligger omkring 5%, betyder lidt ekstra information typisk ikke, at vi pludselig springer til 30%. Og hvis risikoen ligger omkring 95%, ændrer lidt ekstra information heller ikke meget, fordi vi allerede er ret sikre på misligholdelse. Omkring 50% er vi mest usikre, og derfor er det netop her, at ny information har størst betydning. Det er denne intuition, S-kurven udtrykker i sammenhængen mellem z og p.

Vi kan nu samle idéen i én simpel model. Først beregner vi en lineær score ud fra indkomst,

$$z = \beta_0 + \beta_1 x,$$

hvor x er indkomst (målt i 1.000 kr pr. måned). Derefter oversætter vi scoren til en sandsynlighed med S-kurven,

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

Tilsammen giver det den univariate logistiske model

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}.$$

Forskellen til lineær regression er, at p nu altid ligger mellem 0 og 1. Omkring $z = 0$ (hvor $p = 0{,}5$) ændrer sandsynligheden sig relativt hurtigt, mens kurven flader ud i enderne, når p er tæt på 0 eller 1.

Video 3 – S-kurvens egenskaber: hvorfor den passer til sandsynligheder

Varighed: 6 minutter
Indhold: Videoen forklarer S-kurvens vigtigste egenskaber med animationer:

- Hvorfor $0 < p < 1$ altid, og hvorfor kurven nærmer sig 0 og 1 i enderne
- Hvad der sker når $z \to \infty$ og $z \to -\infty$
- Hvorfor større z altid giver større p (monotoniforhold)

▶ Se videoen

https://www.youtube.com/watch?v=j4Fykk21-hs&list=PLIvNrmh87wtA&index=4

Lad os undersøge den logistiske funktion systematisk, præcis som I har lært i differentialregning. Vi går tre klassiske spørgsmål igennem: definitionsmængde og værdimængde, monotoniforhold og symmetri.

Egenskab 1: Definitionsmængde og værdimængde

Definitionsmængde: $D_f = \mathbb{R}$ (alle reelle tal kan sættes ind i funktionen)
Værdimængde: Vi skal finde alle mulige outputværdier fra funktionen.
Sæt forskellige værdier af z ind i $f(z) = \frac{1}{1+e^{-z}}$:

- Når vi sætter et tal z ind i funktionen, får vi outputtet $f(z)$
- Uanset hvilken værdi af z vi vælger, er e^{-z} altid positiv
- Dette betyder, at nævneren $1 + e^{-z}$ altid er større end 1
- Derfor er brøken $\frac{1}{1+e^{-z}}$ altid mellem 0 og 1

Grænseværdier:

- Når z bliver meget stor ($z \to \infty$):
  - e^{-z} nærmer sig 0
  - $f(z) = \frac{1}{1+e^{-z}}$ nærmer sig $\frac{1}{1+0} = 1$
- Når z bliver meget negativ ($z \to -\infty$):
  - e^{-z} bliver meget stor
  - $f(z) = \frac{1}{1+e^{-z}}$ nærmer sig $\frac{1}{\infty} = 0$

(Bemærk: $\frac{1}{\infty}$ er teknisk set forkert, da ∞ ikke er et tal. Med korrekt grænseværdinotation skriver vi $\lim_{z\to-\infty} f(z) = 0$ og $\lim_{z\to\infty} f(z) = 1$.)
Konklusion: Funktionen kan give alle værdier mellem 0 og 1 (men aldrig præcis 0 eller 1). Derfor er værdimængden $V_f =]0;1[$ – perfekt til at repræsentere sandsynligheder!

Vi ved nu, at $f(z)$ tager værdier i $]0;1[$. Næste skridt er at undersøge, hvordan funktionen ændrer sig, når z vokser, altså dens monotoniforhold.

Egenskab 2: Monotoniforhold

Vi skal vise, at funktionen er strengt voksende.
Bevis via den afledte:

$$f'(z) = \frac{d}{dz}\left[(1+e^{-z})^{-1}\right] \tag{1}$$
$$= -1 \cdot (1+e^{-z})^{-2} \cdot (-e^{-z}) \quad \text{(kædereglen)} \tag{2}$$
$$= \frac{e^{-z}}{(1+e^{-z})^2} \tag{3}$$

Da både tæller og nævner er positive for alle z, er $f'(z) > 0$ overalt. Derfor er funktionen strengt voksende.

Elegant form af den afledte: Vi kan skrive $f'(z)$ kun ved hjælp af funktionsværdien selv. Vi har $f(z) = \frac{1}{1+e^{-z}}$, og ved at finde fælles nævner får vi:

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

Derfor

$$f'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{1+e^{-z}} \cdot \frac{e^{-z}}{1+e^{-z}} = f(z)\big(1-f(z)\big).$$

Det er en elegant identitet: den afledte kan udtrykkes alene ved funktionsværdien selv.

Praktisk betydning: Højere z giver højere p (større misligholdelsesrisiko). Bedre kreditværdighed svarer til lavere z i vores opsætning (fx højere indkomst giver negativt bidrag til z).

Funktionen er altså strengt voksende fra 0 til 1. Tilbage står spørgsmålet om, hvad der sker midt på kurven, der hvor sandsynligheden er 0,5.

Egenskab 3: Symmetri og midtpunkt

S-kurven har et naturligt midtpunkt: når $z = 0$, gælder

$$f(0) = \frac{1}{1+e^0} = \frac{1}{1+1} = \frac{1}{2}.$$

Altså svarer $z = 0$ præcis til en sandsynlighed på 50%. Det er også her, kurven er stejlest. Den afledte $f'(z)$ har sit maksimum ved $z = 0$, hvor

$$f'(0) = f(0)\big(1 - f(0)\big) = \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{4}.$$

Funktionen er desuden symmetrisk om dette midtpunkt:

$$f(-z) = 1 - f(z).$$

Bevis: Vi har

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

Ved at gange tæller og nævner med e^{-z} får vi

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

Men netop denne brøk er $1 - f(z)$ (jf. Egenskab 2). Altså $f(-z) = 1 - f(z)$.

To z-værdier med samme størrelse, men modsat fortegn, giver sandsynligheder, der summerer til 1.

De tre egenskaber forklarer den vigtigste intuition: S-kurven oversætter enhver score $z \in \mathbb{R}$ til en gyldig sandsynlighed $p \in]0;1[$, højere score giver altid højere risiko, og kurven har sit naturlige midtpunkt ved $z = 0$, hvor sandsynligheden er 0,5. Figuren viser også, at kurven flader ud tæt på 0 og 1. Derfor har ændringer i z typisk størst betydning midt i intervallet (omkring $p \approx 0{,}5$) og mindre betydning, når risikoen allerede er meget lav eller meget høj.

En vigtig bemærkning om koefficienter: Bemærk, at koefficienterne i den lineære model (fx $\beta_1 = -0{,}02$ tidligere i kapitlet) og koefficienterne i den logistiske model (som vi snart indfører) ikke er direkte sammenlignelige. Den lineære β arbejder på *sandsynlighedsskalaen*, mens den logistiske β arbejder på *log-odds-skalaen*. Samme variabel (fx indkomst) får derfor forskellige talværdier afhængigt af, hvilken model den indgår i.

3 Sandsynlighed, odds og log-odds

Indtil nu har vi gået én vej: fra en lineær score z til en sandsynlighed p via S-kurven. Men når vi senere skal *fitte* sådan en model til data, viser det sig at være lettere at tænke baglæns (fra p til z). Det kræver, at vi kan udtrykke en sandsynlighed på en skala, der spænder over hele tallinjen. Det er motivationen bag de næste tre repræsentationer: sandsynlighed, odds og log-odds.

Video 4 – Odds og log-odds: Hvorfor skifter vi skala?

Varighed: 10 minutter
Indhold: Videoen viser med animationer, hvordan man kan beskrive den samme usikkerhed på tre skalaer:

- Fra sandsynlighed p til odds og log-odds
- Hvorfor odds-skalaen er asymmetrisk, og hvorfor log-odds bliver symmetrisk omkring 0
- Sammenhængen $p = 0{,}5 \Leftrightarrow \text{odds} = 1 \Leftrightarrow \text{log-odds} = 0$
- Idéen: Log-odds kan ligge på hele tallinjen, og derfor kan man modellere log-odds lineært

▶ Se videoen

https://www.youtube.com/watch?v=njJ7W6fMhpM&list=PLIvNrmh87wtA&index=5

I daglig tale bruger vi sandsynligheder, for eksempel "der er 20% chance for regn". I logistisk regression møder man ofte **odds** og **log-odds** i stedet. Grunden er, at sandsynligheder er bundet til intervallet $[0, 1]$, mens log-odds kan antage alle værdier på tallinjen. Det gør det muligt at lave en lineær model for log-odds, og det er netop dér, logistisk regression bliver enkel.

Repræsentation	Formel	Værdiområde
Sandsynlighed	p	$[0, 1]$
Odds	$\dfrac{p}{1-p}$	$[0, \infty[$
Log-odds (logit)	$\ln\left(\dfrac{p}{1-p}\right)$	$\mathbb{R}$

De tre størrelser beskriver den samme usikkerhed, men på forskellige skalaer. Når p stiger, stiger odds, og log-odds stiger også.

3.1 Hvad er odds?

Odds er forholdet mellem sandsynligheden for, at noget sker, og sandsynligheden for, at det ikke sker:

$$\text{odds} = \frac{p}{1-p} = \frac{P(\text{sker})}{P(\text{sker ikke})}.$$

Hvis $p = 0{,}5$, får vi odds $= 1$, hvilket svarer til "lige store chancer". Hvis p er lille, bliver odds mindre end 1, og hvis p er stor, bliver odds større end 1.

Et simpelt eksempel er en terning. Sandsynligheden for en sekser er

$$p = \frac{1}{6} \approx 0{,}167, \qquad 1 - p = \frac{5}{6} \approx 0{,}833.$$

Oddset for en sekser er derfor

$$\frac{p}{1-p} = \frac{1/6}{5/6} = \frac{1}{5} = 0{,}2.$$

Det kan læses som "1 til 5": for hver gang du får en sekser, vil du i gennemsnit få omkring fem ikke-seksere.

Opgaver

Opgaver om konvertering fra sandsynlighed p til odds er i Kapitel 6.3.

3.2 Hvad er log-odds?

Efter at have regnet odds ud fra en sandsynlighed kan man spørge, hvorfor vi overhovedet behøver endnu en størrelse. Grunden er, at odds ligger på en skala fra 0 til ∞, og den skala er ikke særlig symmetrisk. For eksempel ligger odds 0,5 og odds 2 lige langt fra 1 (som svarer til $p = 0{,}5$), men tallene ser meget forskellige ud.

Derfor tager man ofte logaritmen af odds. Det giver log-odds (også kaldet logit):

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

Når vi tager logaritmen, flytter vi fra en skala med værdier i $[0, \infty[$ til en skala, der kan antage alle reelle tal $(-\infty, \infty)$. Skalaen bliver desuden symmetrisk omkring 0, fordi $\ln(1) = 0$, og odds $= 1$ svarer præcis til $p = 0{,}5$.

Det er denne egenskab, der gør log-odds praktisk i logistisk regression: Vi kan lade den lineære score være

$$z = \beta_0 + \beta_1 x$$

og fortolke z som log-odds. Derefter kan vi bruge S-kurven til at oversætte tilbage til en sandsynlighed:

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

Tabellen herunder viser symmetrien: sandsynligheder, der ligger lige langt fra 0,5, får log-odds med samme størrelse men modsat fortegn.

Sandsynlighed	Odds	Log-odds
0,10	0,11	$-2{,}20$
0,50	1,00	0,00
0,90	9,00	$+2{,}20$

Det er svært at se direkte på odds, men det bliver tydeligt på log-odds-skalaen.

Formler: Skift mellem p, odds og log-odds

Det er vigtigt at kunne skifte mellem de tre repræsentationer, fordi man ofte regner i odds eller log-odds, men gerne vil fortolke resultatet som en sandsynlighed.

Fra sandsynlighed til odds og log-odds:

$$\text{odds} = \frac{p}{1-p}, \qquad \text{log-odds} = \ln\left(\frac{p}{1-p}\right).$$

Fra odds til sandsynlighed:

$$p = \frac{\text{odds}}{1+\text{odds}}.$$

Fra log-odds til sandsynlighed (S-kurven):

$$p = \frac{1}{1+e^{-\text{log-odds}}}.$$

Opgaver

Opgaver om konvertering mellem sandsynlighed p, odds og log-odds er i Kapitel 6.4.

3.3 Den centrale model

Nu kan vi samle idéen i en egentlig model. I logistisk regression antager vi, at log-odds er lineære i de forklarende variable. Vi starter univariat med indkomst:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x,$$

hvor p er sandsynligheden for misligholdelse, og x er indkomst målt i 1.000 kr pr. måned. Venstre side er log-odds, og den kan antage alle reelle tal. Højre side er en lineær funktion, præcis som i lineær regression.

Koefficienterne β_0 og β_1 er nemmest at fortolke på log-odds-skalaen. Hvis β_1 er negativ, betyder det, at højere indkomst hænger sammen med lavere risiko, og hvis β_1 er positiv, hænger højere indkomst sammen med højere risiko.

Mere præcist betyder β_1, hvor meget log-odds ændrer sig, når x stiger med 1. Når x måles i 1.000 kr, svarer en stigning på 1 til 1.000 kr mere i månedlig indkomst. Hvis for eksempel

$$\beta_1 = -0{,}10,$$

så falder log-odds med 0,10 for hver ekstra 1.000 kr i indkomst. Det er fordelen ved log-odds-skalaen: effekten er lineær og den samme uanset udgangspunktet.

Man kan også oversætte den samme effekt til odds-skalaen. Hvis log-odds ændrer sig med β_1, så bliver odds ganget med faktoren e^{β_1} (ofte kaldet en odds-ratio):

$$\text{nye odds} = \text{gamle odds} \cdot e^{\beta_1}.$$

I eksemplet med $\beta_1 = -0{,}10$ får vi

$$e^{-0{,}10} \approx 0{,}905,$$

så odds bliver ganget med cirka 0,905 for hver ekstra 1.000 kr, altså et fald på cirka 9,5% i odds. Odds kan være mindre intuitive at aflæse som tal, men faktoren e^{β_1} giver en enkel måde at beskrive ændringen på.

Generelt gælder:

$$\begin{aligned}
\beta > 0 &\Rightarrow e^{\beta} > 1 \quad \text{(odds stiger)},\\
\beta < 0 &\Rightarrow e^{\beta} < 1 \quad \text{(odds falder)},\\
\beta = 0 &\Rightarrow e^{\beta} = 1 \quad \text{(ingen ændring i odds)}.
\end{aligned}$$

Log-odds-skalaen er praktisk, fordi den spænder over hele tallinjen. Det gør det muligt at bruge en lineær model på højre side uden at bekymre sig om, at output skal ligge mellem 0 og 1.

Det gør også koefficienterne stabile at arbejde med. Hvis indkomsten stiger med 10.000 kr, så stiger x med 10, og log-odds ændrer sig med $10\beta_1$. På odds-skalaen betyder det, at odds ganges med faktoren $e^{10\beta_1}$. På sandsynlighedsskalaen afhænger ændringen derimod af udgangspunktet: Den samme ændring i log-odds kan flytte p meget omkring 0,5, men kun lidt, hvis p allerede ligger tæt på 0 eller tæt på 1. Det er grunden til, at man typisk fortolker koefficienter på log-odds- eller odds-skalaen og derefter oversætter til sandsynligheder, når man skal forklare resultatet.

Opgaver

Opgaver om tolkning af koefficienter på odds-skalaen er i Kapitel 6.5.

Modellen kan udvides, så den inkluderer flere forklarende variable:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

Det er stadig log-odds, der modelleres lineært. Når modellen har flere forklarende variable, skal koefficienterne fortolkes *ceteris paribus*: β_1 beskriver, hvordan log-odds ændrer sig, når x_1 stiger med 1, *mens de andre variable holdes faste*. I modellen herover betyder det for eksempel, at β_1 for indkomst viser effekten af 1.000 kr mere i indkomst givet samme alder, samme betalingshistorik osv. Tilsvarende viser β_2 effekten af ét år mere i alder ved samme indkomst.

Interceptet β_0 er log-odds, når alle forklarende variable er 0. Det er ofte ikke en realistisk situation, men β_0 er et nødvendigt startpunkt, så modellen kan passe til data.

Eksempel med flere variable

En banks simple model for kreditrisiko:

$$\ln\left(\frac{p}{1-p}\right) = 1{,}5 - 0{,}08 \cdot \text{indkomst} - 0{,}02 \cdot \text{alder}$$

hvor indkomst måles i 1.000 kr/md og alder i år.
Kunde: Emma, 28 år, indkomst 28.000 kr/md
Beregning trin for trin:

1. **Beregn log-odds:**

$$z = 1{,}5 - 0{,}08 \times 28 - 0{,}02 \times 28 = 1{,}5 - 2{,}24 - 0{,}56 = -1{,}30$$

2. **Beregn odds:**

$$\text{odds} = e^{-1{,}30} \approx 0{,}27$$

3. **Beregn sandsynlighed:**

$$p = \frac{0{,}27}{1 + 0{,}27} = \frac{0{,}27}{1{,}27} \approx 0{,}21 = 21\%$$

4. **Beslutning:** 21% er acceptabelt. Godkend lånet.

Opgaver

Opgaver om beregning af log-odds z og sandsynlighed p for de to andre kunder er i Kapitel 6.6.

4 Evaluering af en kreditmodel

Indtil nu har vi brugt modellen til at beregne en sandsynlighed for misligholdelse. I praksis skal banken dog ofte ende med en beslutning: *godkend* eller *afvis*. Når en sandsynlighed bliver til en beslutning, opstår et nyt spørgsmål: Hvordan vurderer man, om modellen træffer gode beslutninger?

Video 5 – Klassifikationsmatrix: Når en sandsynlighed bliver til en beslutning

Varighed: 9 minutter
Indhold: Videoen viser med en simpel animation, hvad der sker, når vi bruger grænsen $p = 0{,}50$:

- Fra sandsynlighed p til beslutning (godkend/afvis) ved grænsen 0,50
- De fire felter i en klassifikationsmatrix (TN, FP, FN, TP), og hvordan de opstår
- Accuracy som andelen af korrekte beslutninger
- Et kort tankeeksperiment: Hvorfor FN og FP typisk har forskellige omkostninger i kredit

▶ Se videoen

https://www.youtube.com/watch?v=K99HOoTrKeg&list=PLIvNrmh87wtA&index=6

En logistisk model giver et tal p mellem 0 og 1. Som et første, simpelt udgangspunkt kan vi bruge reglen:

$$\text{Godkend hvis } p < 0{,}50, \qquad \text{Afvis hvis } p \geq 0{,}50.$$

Det svarer til, at banken kun afviser, når modellen vurderer, at misligholdelse er mere sandsynlig end tilbagebetaling. Grænsen 0,50 er valgt for at gøre eksemplet enkelt; i praksis vil en bank ofte vælge en lavere grænse, fordi en kunde med næsten 50% risiko sjældent er en acceptabel forretning.

Når modellen træffer en beslutning, kan vi sammenligne med, hvad der faktisk sker senere: Betalte kunden tilbage, eller misligholdt kunden? Det giver fire mulige udfald, som kan samles i en klassifikationsmatrix (confusion matrix). Her bruger vi "positiv" om *misligholdelse*. Selv om misligholdelse er den "dårlige" begivenhed for banken, er det en standardkonvention i statistik at lade det sjældne og dyre udfald være den positive klasse.

	Faktisk betaler tilbage	**Faktisk misligholder**
Model godkender ($p < 0{,}50$)	True negative (TN)	False negative (FN)
Model afviser ($p \geq 0{,}50$)	False positive (FP)	True positive (TP)

I kreditkonteksten kan man tænke på de to fejl sådan her:

- **FN (falsk negativ):** Banken godkender, men kunden misligholder. Det kan give et direkte tab.
- **FP (falsk positiv):** Banken afviser, men kunden ville have betalt tilbage. Det er tabt forretning.

En simpel måde at opsummere modellens præstation på er **accuracy**, som er andelen af korrekte beslutninger:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Eksempel: Klassifikationsmatrix på 100 låntagere

Antag, at vi har anvendt en logistisk model på 100 låntagere og brugt grænsen $p \geq 0{,}50$ til at klassificere dem. Resultatet bliver:

	Faktisk betalte tilbage	**Faktisk misligholdt**
Model godkender ($p < 0{,}50$)	TN $= 35$	FN $= 14$
Model afviser ($p \geq 0{,}50$)	FP $= 13$	TP $= 38$

Her bliver

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{38 + 35}{100} = 0{,}73 \ (73\%).$$

Det er det datasæt, I selv arbejder med i Excel-øvelsen i kapitel 5.

Accuracy kan være et godt udgangspunkt, men i kreditvurdering kan den også være misvisende, fordi de to typer fejl ikke nødvendigvis har samme konsekvens. En FN kan koste banken mange penge, fordi tabet kan være stort på et enkelt lån. En FP koster typisk mindre pr. kunde, men kan stadig være dyr, fordi banken mister renter og et fremtidigt kundeforhold.

Som tankeeksperiment kan vi antage, at en FN i gennemsnit koster banken 40.000 kr, mens en FP i gennemsnit koster banken 5.000 kr. Hvilken type fejl bør banken være mest villig til at acceptere?

Hvis banken især vil undgå FN, vil banken typisk blive mere konservativ og afvise flere kunder. Det kan ikke opnås ved at ændre modellen, men ved at ændre beslutningsreglen. I stedet for altid at bruge grænsen 0,50 kan banken vælge en anden grænse, så den afviser allerede ved lavere risiko. Valg af grænseværdi hænger altså sammen med de relative omkostninger ved FP og FN, og derfor kan man ikke evaluere en model ud fra accuracy alene.

5 Excel

I de foregående kapitler har vi bygget modellen op trin for trin: fra sandsynlighed og odds til log-odds, og videre til den centrale logistiske model. Indtil nu har vi anvendt modellen på enkelte kunder ad gangen og regnet med små tal. Når en bank arbejder i praksis, anvendes modellen derimod på hundreder eller tusinder af ansøgere på én gang.

I dette kapitel skifter vi derfor til Excel og anvender modellen på et helt datasæt. Filen `logit_excel.xlsx` indeholder 100 simulerede låntagere, hver med oplysninger om indkomst, alder og betalingshistorik, sammen med deres faktiske udfald (om de misligholdt deres lån eller ej). β-værdierne er estimeret på data og givet i arket, så I kan koncentrere jer om processen: tag inputtet, beregn sandsynligheden, klassificér og evaluér resultatet i en klassifikationsmatrix.

Målet er at tage hele rejsen fra rådata til beslutning og evaluering på én side i Excel.

Logistisk regression i praksis: beregn risiko i Excel

I denne øvelse skal I beregne sandsynligheden for misligholdelse i Excel og opsummere resultatet i en klassifikationsmatrix. Målet er at træne processen fra *input* til *sandsynlighed* og videre til en simpel beslutning ved grænsen 0,50.

↓ Hent Excel-arket

https://econ.au.dk/fileadmin/ECON/Collaboration/Tilbud_til_gymnasier/Logistisk_regression/Logit_excel.xlsx

Data:

- Kolonne B: **indkomst** (kr pr. måned)
- Kolonne C: **alder** (år)
- Kolonne D: **tidl_misligholdelse** (dummyvariabel, 0 eller 1)
- Kolonne E: **observeret misligholdelse** (0 eller 1)

Opgave:

1. Brug de givne β-værdier i arket til at beregne

$$z = \beta_0 + \beta_1 \cdot \text{indkomst} + \beta_2 \cdot \text{alder} + \beta_3 \cdot \text{tidl\_misligholdelse}.$$

2. Beregn sandsynligheden i **kolonne F** ved

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

3. Brug grænsen 0,50 til at klassificere:

$$p < 0{,}50 \Rightarrow \text{forvent "ingen misligholdelse"},$$
$$p \geq 0{,}50 \Rightarrow \text{forvent "misligholdelse"}.$$

4. Byg en 2×2 klassifikationsmatrix ved at tælle, hvor ofte hver kombination forekommer:

$$(E = 0 \text{ og } p < 0{,}50), \quad (E = 1 \text{ og } p < 0{,}50),$$
$$(E = 0 \text{ og } p \geq 0{,}50), \quad (E = 1 \text{ og } p \geq 0{,}50).$$

5. Beregn **accuracy** som andelen af korrekte beslutninger, dvs. summen af de to celler på diagonalen divideret med det samlede antal observationer.

Video 6 – Logistisk regression i Excel

Varighed: 14 minutter

Indhold:

- Beregning af p i Excel ud fra kundedata og givne β-værdier
- Klassifikation ved grænsen 0,50
- Klassifikationsmatrix og accuracy i Excel

▶ Se videoen

https://www.youtube.com/watch?v=avUgyj7ucbM&list=PLIvNrmh87wtA&index=7

6 Opgaver

6.1 Opgave 1: En lineær model for kreditrisiko

Opgave 1

Antag, at en bank (naivt) bruger følgende lineære model for sandsynligheden for misligholdelse:

$$p = 0{,}80 - 0{,}02 \cdot x$$

hvor x er indkomst målt i **1.000 kr pr. måned**, og p er en sandsynlighed.

1. Beregn p for følgende indkomster:
   - 0 kr
   - 10.000 kr
   - 30.000 kr
   - 50.000 kr
2. Hvilke af dine resultater kan tolkes som en sandsynlighed, og hvilke kan ikke? Forklar kort hvorfor.
3. Find den indkomst (i kr pr. måned), hvor modellen giver $p = 0$. Hvad fortæller det om modellen?
4. Hvad sker der med p, hvis indkomsten bliver endnu højere end 50.000 kr pr. måned? Hvad fortæller det om, hvorfor en lineær model er problematisk til at beskrive sandsynligheder?

6.2 Opgave 2: Test en lineær og en logistisk model

Opgave 2

Vi vil sammenligne to modeller, der begge bruger en score z.

- **Lineær model:**
$$p = 0{,}5 + 0{,}25z$$
- **Logistisk model (S-kurven):**
$$p = \frac{1}{1 + e^{-z}}$$

1. Beregn p for begge modeller for følgende z-værdier *Hint:* $e^{-4} \approx 0{,}018$ og $e^{4} \approx 54{,}6$:
   - $z = -4$
   - $z = 0$
   - $z = 4$
2. Sammenlign dine resultater. Hvilken model giver altid gyldige sandsynligheder? Forklar kort.

6.3 Opgave 3: Fra sandsynlighed til odds

Opgave 3

For en sandsynlighed p defineres odds som

$$\text{odds} = \frac{p}{1-p}.$$

1. Beregn odds for følgende sandsynligheder:
   - Møntkast (plat): $p = 0{,}50$
   - Lav kreditrisiko: $p = 0{,}10$
   - Høj kreditrisiko: $p = 0{,}80$
2. Forklar kort med egne ord, hvad et odds på f.eks. 4 betyder.

6.4 Opgave 4: Konvertering mellem p, odds og log-odds

Opgave 4

I denne opgave skal du øve at skifte mellem sandsynlighed, odds og log-odds.

Del A: Udfyld tabellen
Udfyld de manglende felter i tabellen (*Hint:* $\ln(4) \approx 1{,}39$ og $e^{1{,}39} \approx 4$):

Kunde	**Sandsynlighed** p	**Odds** $\frac{p}{1-p}$	**Log-odds** $\ln\left(\frac{p}{1-p}\right)$
Meget sikker	0,20		
På grænsen		1	
Risikabel			1,39

Del B: Fra log-odds til sandsynlighed
En banks model giver log-odds $= -1{,}5$ for kunde A og log-odds $= 0{,}8$ for kunde B.

1. Beregn sandsynligheden for misligholdelse for hver kunde.
2. Hvem har højest risiko?

6.5 Opgave 5: Tolkning af koefficient på odds-skalaen

Opgave 5

Antag, at en logistisk model for misligholdelse har koefficienten

$$\beta_1 = -0{,}08$$

for indkomst, hvor indkomst måles i 1.000 kr pr. måned.

1. Hvad sker der med log-odds for misligholdelse, når indkomsten stiger med 1.000 kr

pr. måned?

2. Beregn hvor meget odds bliver multipliceret med, når indkomsten stiger 1.000 kr pr. måned. Forklar i ord, hvad det betyder.
3. Hvad sker der med odds, når indkomsten stiger 5.000 kr pr. måned?
4. Hvad sker der med odds, når indkomsten stiger 10.000 kr pr. måned? *Hint:* $e^{-0{,}80} \approx 0{,}449$.

6.6 Opgave 6: Risiko for de to andre kunder

Opgave 6

I denne opgave bruger du samme logistiske model som i eksemplet med Emma:

$$z = \ln\left(\frac{p}{1-p}\right) = 1{,}5 - 0{,}08 \cdot \text{indkomst} - 0{,}02 \cdot \text{alder},$$

hvor indkomst måles i 1.000 kr pr. måned, og alder måles i år. Sandsynligheden findes ved

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

Del A: Beregn z og p for Kunde B og Kunde C

1. **Kunde B (Peter):** indkomst 45.000 kr/md og alder 45 år.
2. **Kunde C (Sofie):** indkomst 12.000 kr/md og alder 22 år.

For hver kunde skal du:

1. beregne log-odds z,
2. beregne sandsynligheden p.

Del B: Udvid modellen med tidligere misligholdelse (dummyvariabel) Nogle oplysninger er binære: enten har man tidligere misligholdt et lån, eller også har man ikke. Vi koder derfor

$$\text{tidl\_misligholdelse} = \begin{cases} 1 & \text{hvis kunden har tidligere misligholdelse,} \\ 0 & \text{ellers.} \end{cases}$$

Antag, at banken udvider modellen til:

$$z = 1{,}5 - 0{,}08 \cdot \text{indkomst} - 0{,}02 \cdot \text{alder} + 1{,}2 \cdot \text{tidl\_misligholdelse}.$$

1. Genberegn z og p for Kunde B. Som det fremgår af kundetabellen i Kapitel 1.4, har Peter tidligere konkurs, så vi sætter `tidl_misligholdelse`$= 1$.
2. Genberegn z og p for Kunde C. Sofie har, jf. samme tabel, ingen tidligere misligholdelse, så `tidl_misligholdelse`$= 0$.
3. Forklar med én til to sætninger, hvad koefficienten 1,2 betyder for risikoen, når `tidl_misligholdelse` går fra 0 til 1.

Tilbage til de tre kunder

I §1.4 mødte vi Emma, Peter og Sofie og blev bedt om at gætte, hvem banken bør låne til. Modellen giver os nu et systematisk svar:

- Emma (Kunde A): $p \approx 21\%$
- Peter (Kunde B): $p \approx 4{,}7\%$ (eller 14,2% når vi tager den tidligere konkurs med)
- Sofie (Kunde C): $p \approx 52{,}5\%$

Med grænsen $p < 0{,}50$ vil banken godkende Emma og Peter og afvise Sofie.

6.7 Opgave 7: Klassifikationsmatrix og fejltyper

Opgave 7

En bank har testet sin logistiske model på 100 låntagere og fået følgende klassifikationsmatrix:

	Faktisk betalte tilbage	**Faktisk misligholdt**
Model godkender ($p < 0{,}50$)	60 (TN)	7 (FN)
Model afviser ($p \geq 0{,}50$)	8 (FP)	25 (TP)

1. Beregn modellens **accuracy**.
2. Hvor mange af de 32 faktiske misligholdere fanger modellen?
3. Hvor mange af de 68 faktiske gode låntagere klassificerer modellen korrekt?
4. Banken antager, at en FN i gennemsnit koster 40.000 kr, mens en FP koster 5.000 kr. Beregn de samlede forventede fejl-omkostninger på de 100 låntagere.

6.8 Opgave 8: Valg af grænseværdi

Opgave 8

En bank kan vælge forskellige grænseværdier (cutoffs) til at klassificere låntagere. Tabellen nedenfor viser, hvor mange fejl af hver type modellen begår på 100 låntagere ved tre forskellige cutoffs:

Cutoff c	**Falsk negativ (FN)**	**Falsk positiv (FP)**
0,30	3	28
0,50	7	8
0,70	18	2

Antag, at en FN i gennemsnit koster banken 40.000 kr, og en FP koster 5.000 kr.

1. Beregn de samlede forventede fejl-omkostninger ved hver af de tre cutoffs.

2. Hvilken cutoff giver de laveste samlede omkostninger?
3. Forklar i 1–2 sætninger, hvorfor en lavere cutoff (fx 0,30) får banken til at afvise flere kunder.