Notater

Sverre Langaas

25. november 2020

Innhold

Kapittel 1

Sannsynlighet

Sannsynlighetsteori gir oss et rammeverk for å håndtere og kvantifisere usikkerhet. Det tar utgangspunkt i et stokastisk eksperiment. Dette består for det første av et utfallsrom Ω som er mengden av alle de ulike utfallene ω som kan bli realisert i eksperimentet. Utfallsrommet er fullstendig og gjensidig utelukkene slik at ett, og bare ett, utfall blir realisert. I eksperimentet vil det ikke være mulig å vite hva utfallet blir før det er realisert. Vi kan tenke oss at dette eksperimentet blir gjentatt N ganger. Den relative frekvensen av hvert utfall vil da konvergere mot sannsynligheten for det utfallet blir realisert når $N \to \infty$.

1.1 Aksiom og teknisk rammeverk

Dette er den enkle historien. Utfordringen er at vi kan ha utfallsrom som er ikke-tellbare. Det vil si at det ikke er mulig å liste opp de ulike utfallene som $\Omega = \{\omega_1, \omega_2, ...\}$. Et eksempel på et slikt utfallsrom er enhetsdisken

$$\Omega = \{(i, j) \in \mathbb{R}^2 : |i + j| \leq 1\} \tag{1.1}$$

Det er da ikke mulig å gi sannsynlighet til enkeltutfall (i, j) siden man kan vise at sannsynligheten nødvendigvis er null. Vi tallfester derfor kun sannsynlighet til delmengder og ikke enkelt-utfall. Merk at hvis utfallsrommet er tellbart kan vi kan vi også angi sannsynlighet til delmengder som kun inneholder enkeltutfall, eks: $A = \{\omega_k\}$. Vi kaller delmengder for hendelser. En hendelse A inntreffer hvis et utfall $\omega \in A$ blir realisert. Hvis sannsynligheten er uniform så kan vi i tellbare utfallsrom finne sannsynlighet for hendelser som antall gunstige utfall delt på antall mulige,

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|} \tag{1.2}$$

Den naturlige generalisering til ikke-tellbare mengder er det relative arealet til mengden A.

$$\mathbb{P}(A) = \frac{\lambda(A)}{\lambda(\Omega)} \tag{1.3}$$

der λ er en funksjon som gir arealet. Det er en utfordring at ikke alle delmengder har et veldefinert mål på areal. Vi avgrenser oss derfor til å se på delmengdene av Ω som oppfører seg bra. Såkalte σ-algebraer har egenskapene vi ønsker. $\mathscr{F}$ er en σ-algebra på Ω hvis

1. $A \in \mathscr{F} \implies A^C \in \mathscr{F}$
2. $A_1, A_2, ... \in \mathscr{F} \implies U_n A_n \in \mathscr{F}$
3. $\Omega \in \mathscr{F}$

Dette sikrer at delmengdene oppfører seg bra, men sikrer ikke at de inneholder alle hendelser vi er interessert i. Et trivielt eksempel er $\mathscr{F} = \{\varnothing, \Omega\}$. I praksis er utfallsmengden $\mathbb{R}^N$ og vi betrakter borel-mengdene $\mathscr{B}(\mathbb{R}^N = \mathscr{F}$. Uansett, vi kan nå definere en *probability measure* $\mathbb{P} : \mathscr{F} \to [0, 1]$ som ikke trenger å være uniform, men som må tilfredstille aksiomene

1. $\mathbb{P}(A) \geq 0, \quad \forall A \in \mathscr{F}$
2. $\mathbb{P}(\Omega) = 1$
3. $\mathbb{P}(A_1 U A_2, ...) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ for disjunkte delmengder

Disse egenskapene korrepsonderer med tolkningen av sannsynlighet som relativt frekvens av hendelser i uendelig gjenntatte forsøk, men det vil også være mulig å bruke en mer subjektiv oppfatning av sannsynlighet der sannsynlighetsfunksjon tilfredstiller aksiom. Tilsammen utgjør $(\Omega, \mathscr{F}, \mathbb{P})$ et *probability space*.

1.2 Sannsynlighetsregning

Fra aksiomene kan vi utlede diverse regneregler som kan brukes til å finne sannsynlighet for ulike hendelser. Et veldig enkelt og nyttig resultat er at

$$\mathbb{P}(\Omega) = \mathbb{P}(A \cup A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) = 1 \tag{1.4}$$

$$\implies \mathbb{P}(A) = 1 - \mathbb{P}(A^c). \tag{1.5}$$

I praksis kan det ofte være enklere å finne sannsynligheten for komplementet, spesielt hvis hendelsen har formen *minst én ....* Vi kan være interessert i om én eller flere hendelser

inntreffer. Dette er ikke helt trivielt siden hendelsene kan overlappe og vi vil ikke telle hvert utfall mer enn én gang.

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{1.6}$$

Vet ikke helt hvordan jeg kan utlede det resultatet formelt. Det er forholdsvis greit å utvide til 3 hendelser

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}((A \cup B) \cup C) \tag{1.7}$$
$$= \mathbb{P}(A \cup B) + \mathbb{P}(C) - \mathbb{P}((A \cup B) \cap C) \tag{1.8}$$
$$= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(C) - \mathbb{P}((A \cap C) \cup (B \cap C)) \tag{1.9}$$
$$= \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C) \tag{1.10}$$

Litt av intuisjonen her er at vi må legge til igjen elementene fra interseksjon der vi trakk fra for mange ganger. Ved induksjon er det mulig å utlede en generell formel for sannsynlighet for union av N vilkårlige hendelser, dette får eventuelt bli en annen gang.

1.2.1 Betinget sannsynlighet

Vår evne til å oppdatere oppfatning om sannsynlighet for ulike hendelser når vi får ny informasjon om utfallet er veldig sentralt. Formelt kan vi definere betinget sannsynlighet for en hendelse A gitt $\omega \in B$ som

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{B} \tag{1.11}$$

der vi merker at vi kan behandle B som det nye utfallsrommet og at $\mathbb{P}(\cdot|B)$ er et fullverdig probability measure siden det tilfredstiller aksiomene. Vi skal senere se at definisjonen over gir sammenhengen mellom simultanfordeling og betinget fordeling til til to tilfeldige variabler ettersom realiserte verdier av variablene implisitt avgrenser delmengder av utfallsrommet Ω. Vi kan også merke at det den betingede fordelingen er en skalering av simultanfordeling.

Vi sier at to hendelser er uavhengige dersom

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \tag{1.12}$$

som impliserer $\mathbb{P}(A|B) = \mathbb{P}(A)$ og $\mathbb{P}(B|A) = \mathbb{P}(B)$. Det medfører at informasjon om hvorvidt den ene hendelsen har inntruffet ikke gir oss noe informasjon om sannsynligheten for den andre hendelsen. Vi kan generalisere uavhengighet til en samling av N mengder, $A_1, A_2, ...A_N$, ved å kreve at $P(\cap_{j \in R} A_j) = \prod_{j \in R} P(A_j)$ for alle $R \subset I = \{1, 2, ..., N\}$. Det er altså ikke tilstrekkelig at de er parvis uavhengige.

Det kan også være praktisk å betinge av informasjon selv om vi er interessert i en ubetinget hendelse. Dette gjør at vi kan dele problem inn i enklere delproblem og finne løsningen som en vektet sum. Begynner med å observere at

$$\mathbb{P}(A) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A \cap (B \cup B^c)) = \mathbb{P}((A \cap B) \cup (A \cap B^c) \tag{1.13}$$

$$= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c) \tag{1.14}$$

$$= \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)(1 - \mathbb{P}(B)) \tag{1.15}$$

Mer generelt vil en mengde av delmengder $\{B_m\}_{m \geq 1}$ utgjøre en partisjonering av Ω hvis

1. $\cup_m B_m = \Omega$
2. $B_j \cap B_k = \emptyset$ hvis $j \neq k$

Vi kan da finne $\mathbb{P}(A) = \sum_m \mathbb{P}(A|B_m)\mathbb{P}(B_m)$. Dette resultatet er kjent som loven om total sannsynlighet. Vi kan representere fremgangsmåten grafisk med et tre.

Gitt at hendelsen A inntreffer kan vi også være interessert å finne sannsynlighet for utfallet er i de ulike mengdene av partisjoneringen, for eksempel

$$\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\mathbb{P}(B_j)}{\sum_m \mathbb{P}(A|B_m)\mathbb{P}(B_m)} \tag{1.16}$$

som er kjent som bayes regel.

1.2.2 Følge av hendelser

Vi sier at en følge av hendelser $A_1, A_2, ...$ er stigende dersom $A_n \subset A_{n+1}, n \geq 1$ og avtagende hvis $A_{n+1} \subset A_n, n \geq 1$. Grenseverdien til slike følger er definert ved

- $\lim_{n\to\infty} A_n = \cup_{i=1}^{\infty} A_i$, hvis stigende
- $\lim_{n\to\infty} A_n = \cap_{i=1}^{\infty} A_i$, hvis avtagende

Det er et poeng at

$$\mathbb{P}(\lim_{n\to\infty} A_n) = \lim_{n\to\infty} P(A_n) \tag{1.17}$$

uten at jeg helt forstår betydningen av dette. Kan bevises ved å definere $B_n := A_n \cap A_{n-1}^c$ når følgen er voksende og $B_n := A_n \cap A_{n-1}^c$ når den er avtagende. Det gir en følge $B_1, B_2, ...$ som er disjunkt og som dekker samme område, men må eventuelt ta dette senere. I praksis er det litt tungvindt å jobbe direkte med delmengder av utfallsrommet. Vi vil derfor finne en representasjon som gjør det enklere å få svar på de spørsmål vi er interessert i. I praksis er det enklere å jobbe med tall siden de har naturlig rangering, mål på avstand og vi kan blant annet bruke resultat fra kalkulus. Dette motiverer tilfeldige variabler.

1.3 Tilfeldige variabler

På tross av navnet er en tilfeldig variabel verken tilfeldig eller variabel. Det er en deterministisk funksjon x som mapper fra utfallsrommet til talllinjen, $x : \Omega \to \mathbb{R}$. Vi kan illustrere bruken av tilfeldige variabler med et eksmpel. Anta at vi ser på en uendelig coin-flips og la mynt være 1 og krone være 0. Da har vi uendelig antall utfall der hvert utfall er en uendelig følge. Vi er interessert i hvor mange kast det tar før det blir en mynt.

$$\Omega = \{(a_1, a_2, ...)|a_n \in \{0, 1\},\ \forall n \in \mathbb{N}\} \tag{1.18}$$

$$x(\omega) = min\{n \in \mathbb{N}|a_n = 1\} \tag{1.19}$$

Merk at vi da - for hver realisering i utfallsrommet - bare får det tallet som sier antall kast før første mynt i stedet for hele den uendelige følgen. Denne transformasjonen medfører et tap av informasjon, men vi får den informasjonen vi trenger. Det er et generelt poeng at for å løse problem må vi finne en egnet representasjon av informasjon.

Det faktum at tilfeldige variabler bare er en deterministisk funksjon og all action skjer i probability space er skjult av notasjonelle konvensjoner. Når vi skriver $\{x = 1\}$ så refererer vi implisitt til delmengden av utfallsrommet $\{\omega \in \Omega|x(\omega) = 1\}$. Det betyr at når vi snakker om sannsynligheten til en tilfeldig variabel P_x så er det egentlig $\mathbb{P}$ som jobber under the hood.

$$P_x(1) = \mathbb{P}(\{x = 1\}) = \mathbb{P}\{\omega \in \Omega|x(\omega) = 1\} = \mathbb{P}(A) \tag{1.20}$$

En annen konvensjon er at sammenligninger mellom tilfeldige variabler blir gjort punktvis i utfallsrommet

$$x = y \iff x(\omega) = y(\omega), \forall \omega \in \Omega \tag{1.21}$$

Det er altså ikke tilstrekkelig at de har samme fordeling. Eksempel: la $X \sim U(0, 1)$ slik at $F(x) = x$ når $x \in (0, 1)$, og la $Y = g(X) = 1 - X$. Har da at $F(y) = P(1 - X \leq y) = P(X \geq 1 - y) = 1 - P(X < 1 - y) = 1 - (1 - y) = y$. Dette medfører at $X \overset{d}{=} Y$, men $X \neq Y$.

Kan også nevne at bineære tilfeldig variabler er mye brukt siden vi ofte er interessert i om et eller annet inntreffer eller ikke

$$\mathbb{I}_A(\omega) = \mathbb{I}\{\omega \in A\} \tag{1.22}$$

1.4 Fordelinger

Vi har sett at vi kan definere en tilfeldig variabel x på et sannsynlighetsrom $(\Omega, \mathscr{F}, \mathbb{P})$ som gjør at vi for hver $B \in \mathscr{B}(\mathbb{R})$ kan tallfeste $\mathbb{P}(x \in B) = \mathbb{P}\{\omega \in \Omega | x(\omega) \in B\}$. Dette er litt omstendelig. Vi kan også bare omdefinere utfallsrommet slik at $\Omega = \mathbb{R}$. Den tilfeldig variabelen gjør denne transformasjonen eksplisitt. Alternativt kan vi abstrahere vekk fra transformasjonen og jobbe direkte med sannsynlighetsrommet $(\mathbb{R}, \mathscr{B}(\mathbb{R}), P)$. *Fordelingen* P er *probability measure* der $\Omega = \mathbb{R}$ og $\mathscr{F} = \mathscr{B}(\mathbb{R})$. Vi sier at P er *supported* av S hvis $P(S) = 1$.

Fordelingen P er i likhet med andre probability measures $\mathbb{P}$ en funksjon som mapper mengder til $[0, 1]$. Det er veldig fleksibelt og generelt, men det er litt vanskelig å karakterisere funksjonen P. Det er enklere å jobbe med funksjoner som som mapper tall. Dermed er det veldig greit at er en-til-en korrespondanse mellom P og en kumulativ fordelingsfunksjon F der

$$F(s) = P(x \leq s) = P((-\infty, s]), s \in \mathbb{R} \tag{1.23}$$

Denne funksjonen må oppfylle en del egenskaper

1. $\lim_{s \to \infty} F(s) = 1$
2. $\lim_{s \to -\infty} F(s) = 0$
3. $b > a \implies F(a) \leq F(b)$
4. Den er høyre-kontinuerlig, $\lim_{s \to s^+} F(s) := F(s^+) = F(s)$

Merk at definisjonsmengden er hele tallinjen uavhenig av om fordelingen er supported av mindre delmengde. De fleste fordelinger er enten diskret eller absolutt kontinuerlige. Fordelingen er diskret hvis den er støttet av en tellbar mengde, altså at det eksisterer en mengde $\{s_j\}_{j \geqslant 1}$ der $P(\{s_j\}_{j \geqslant 1}) = 1$. Sannsynlighetsmengden på et gitt element s_j i mengden er $p_j := P(\{s_j\})$ og følgen $\{p_j\}_{j \geqslant 1}$ utgjør en *pmf*. Hvis fordelingen derimot er absolutt kontinuerlig kan den representeres med en tetthet p som er en ikke-negativ funksjon på $\mathbb{R}$ som integrerer til 1, der

$$P(B) = \int_B p(s) ds, \quad \forall B \in \mathscr{B}(\mathbb{R}) \tag{1.24}$$

Det er poeng at med lebesgue integral trenger vi ikke alltid skille mellom diskret og absolutt kontinuerlig fordeling siden vi kan integrere begge. Dette er en fordel når vi utvikler teori. På en annen side er distinksjonen vesentlig når vi anvender teori.

$$f_x(s) = F_x'(s) \tag{1.25}$$

$$p_x(s) = F(s) - F(s^-) \tag{1.26}$$

Tilfeldige variabler er funksjoner som transformerer vilkårlige utfallsrom til tallinjen som er enklere å jobbe med. Vi har sett at vi kan velge hvorvidt vi vil være eksplisitt om denne transformasjonen. Hvis vi velger å være eksplisitte kan vi betegne fordelingen P som fordelingen til x, der

$$P(B) = \mathbb{P}(\{x \in B\}) \tag{1.27}$$

For et gitt probability space så vil hver tilfeldig variabel definere en fordeling. Tilsvarende vil det for hver fordeling være mulig å finne en tilfeldig variabel som har denne fordelingen. Vi kan bruke notasjonen $\mathcal{L}(x)$ for å betegne fordelingen til x. Merk at når vi snakker om fordelingen til en tilfeldig variabel så eksisterer det alltid et underliggende sannsynlighetsrom.

1.4.1 Momenter

Vi har lyst på sammendragsmål som beskriver egenskap til funksjon. Forventningsverdi er første moment

$$\mathbb{E}[X] = \int x d(f(x)) = \begin{cases} \int x f(x) dx, & \text{hvis kontinerlig} \\ \sum x f(x), & \text{hvis diskret} \end{cases} \tag{1.28}$$

Det første integralet er noe lebesgue integral som i prinsippet kan evalueres, men jeg bruker det bare for enhetlig notasjon. Forventningsverdi gir et vektet gjennomsnitt av utfallene til tilfeldig variabel og er et mål på sentraltendensen. Det er forholdsvis enkelt å finne forventnignsverdi til transformasjoner av X dersom vi kjenner fordelingen til denne, fordi

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_x g(x) p_X(x) \tag{1.29}$$

For å få et mål på spredningen kan vi definere en ny variabel som angir avvik fra forventningsverdi og se på hvor stor størrelsen på dette avviket er i gjennomsnitt.

$$\mathbb{E}[Y^2] = \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - \mu^2 \tag{1.30}$$

Merk generelt at forventningsverdien til en transformasjon av X ikke tilsvarer transformasjonen evaluert i forventningsverdien, altså

$$\mathbb{E}f(X) \neq f(\mathbb{E}X) \tag{1.31}$$

Forsikringsselskap tjener penger fordi $\mathbb{E}f(X) < f(\mathbb{E}X)$ er lavere når f er konkav. Folk er derfor villig til å betale for å redusere variasjon i X. Kan knytte dette til *Jensens ulikhet*.

1.4.2 Momentgenererende funksjoner

Den momentgenererende funksjonen til en tilfeldig variabel er gitt ved

$$M(t) = \mathbb{E}[e^{tX}], \quad \text{for } t \in (-h, h) \tag{1.32}$$

Funksjonen er definert dersom uttrykket over konvergerer for verdier av t på et åpent intervall om 0. Kan ta et raskt eksempel på utledning av mgf. La $p(x) = \frac{1}{6}\left(\frac{5}{6}\right)^{x-1}$ være sannsynlighet for antall kast det tar å få sekser på terningen.

$$\mathbb{E}[e^{tX}] = \sum_{x=1}^{\infty} \frac{1}{6}\left(\frac{5}{6}\right)^{x-1} e^{tx} \tag{1.33}$$

$$= \frac{1}{6}e^t \sum_{x=1}^{\infty} \left(\frac{5e^t}{6}\right)^{x-1} \tag{1.34}$$

$$= \frac{e^t}{6}\left(1 - \frac{5e^t}{6}\right)^{-1} \tag{1.35}$$

som konvergerer dersom

$$\frac{5e^t}{6} < 1 \implies \log(t) < \log\left(\frac{6}{5}\right) \tag{1.36}$$

For fordelingene med definert mgf er det én-til-én korrespondanse mellom fordeling og mgf.[1] Med andre ord så vil

$$M_x(t) = M_y(t), \quad \forall t \in (-h, h) \tag{1.37}$$

$$\implies F_x(s) = F_y(s), \quad \forall s \in Z \tag{1.38}$$

Vi kan bruke denne alternative representasjonen til å beskrive egenskaper til fordelingen. Merk at[2]

$$M'(t) = \frac{\partial}{\partial t}\mathbb{E}[e^{tX}] \tag{1.39}$$

$$\begin{cases} \int_Z \frac{\partial}{\partial t} e^{tx} f(x)dx = \int_Z x e^{tx} f(x)dx \\ \sum_{x \in Z} \frac{\partial}{\partial t} e^{tx} p(x) = \sum_{x \in Z} x e^{tx} p(x) \end{cases} \tag{1.40}$$

som medfører at $M'(t)|_{t=0} = \mathbb{E}[x]$ og mer generelt at $M^{(k)}(t)|_{t=0} = \mathbb{E}[x^k]$.

[1] Det har noe sammenheng med laplace-transformasjon av funksjon. Det finnes også en såkalt karakteristisk funksjon som er en generalisering som er definert for alle fordelinger som har sammenheng med fourier-trasnformasjon.

[2] Har brukt at vi kan flytte derivasjonstegn inn og ut av sum og intergral. Er noen tekniske betingelser som må være oppfylt for at dette skal være gyldig operasjon (i betydning at uttrykkene har samme løsningsmengde).

1.4.3 Kvantiler

Vi har sett at vi kan karakterisere fordelinger med den kumulative fordelingen

$$F_x(s) = \begin{cases} \int_{-\infty}^{s} f_x(t)dt, \text{hvis absolutt kontinuerlig} \\ \sum_{j:x_j \leq s} p_j, \text{hvis diskret} \end{cases} \tag{1.41}$$

Litt usikker på hvilken notasjon jeg vil bruke. Tenker at det er greit å spesifisere hvilken variabel vi betrakter fordelingen til slik at vi ikke må bruke så mange ulike bokstaver til å betegne funksjonene. Tror også jeg foretrekker å betegne tilfeldige variabler med stor bokstav. Må avklare dette senere. Uansett, vi kan ofte være interessert i å finne ζ der $F_X(\zeta) = \tau$. Det finner vi ved å evaluere den inverse av cdf i τ. For å håndtere tilfellet der cdf ikke er strengt voksende kan vi definere

$$F_X^{-1}(\tau) = \inf\{s : F_X(s) \leq \tau\} \tag{1.42}$$

Vi får blant annet bruk for kvantilfunksjonen når vi vil finne kritisk verdi i hypotesetester. Anta at vi har en standardnormalfordelt testobservator Z. Vi vil finne et (sentrert) intervall (l, u) der $P(l \leq Z \leq u) = 1 - \alpha$. Vi utnytter at fordeling er symmetrisk slik at $F_Z(-z) = 1 - F(z)$. Dette medfører at

$$F_{|Z|}(s) = P(-s \leq Z \leq s) = F_Z(s) - F_Z(-s) = F_Z(s) - (1 - F_Z(s)) = 2F_Z(s) - 1 \tag{1.43}$$

La $F_{|Z|} := F$. Vi vil finne kritisk verdi c der

$$c = F^{-1}(1 - \alpha/2) \tag{1.44}$$

$$P(|Z| \leq c) = 2F(c) - 1 = 2F[F^{-1}(1 - \alpha/2)] - 1 = 1 - \alpha \tag{1.45}$$

Vi betegner ofte $c := z_{\alpha/2} := \Psi^{-1}(1 - \alpha/2)$

1.5 Transformasjon av tilfeldig variabel

Anta at vi har en variabel X med kjent fordeling $\mathcal{L}(X)$. Vi definerer er ny variabel $Y = g(X)$. Kan vi utlede fordelingen til denne variabelen? Ja.

1.5.1 Diskret

$$p_Y(y) = P(Y = y) = P(g(X) = y) = P(X = g^{-1}(y)) = p_X(g^{-1}(y)) \tag{1.46}$$

La oss ta antall kast før første kron som eksempel. Ha da $p_X(x) = 0.5^x$. Anta nå at vi allerede vet at første kastet aldri gir treff. Vi vil derfor finne fordeling til $Y = g(X) = X+1$. Begynner med å finne $g^{-1}(c) = c - 1$. Det medfører da at

$$p_Y(y) = p_X(g^{-1}(y)) = 0.5^{g^{-1}(y)} = 0.5^{y-1} \tag{1.47}$$

Intuisjonen er at vi evaluerer pmf til X i det tallet som mapper til y-verdien vi er interessert i, altså i $g^{-1}(y)$.

1.5.2 Kontinuerlig

La igjen $Y = g(X)$. Anta at den er monotont voksende.

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)) \tag{1.48}$$

$$f_Y(y) = \frac{\partial}{\partial y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{dx}{dy} \tag{1.49}$$

der $\frac{dx}{dy} = \frac{d}{dx} g^{-1}(y)$. Hvis funksjonen derimot er monotont avtagende må vi snu ulikhetstegnet,

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X > g^{-1}(y)) = 1 - F_X(g^{-1}(y)) \tag{1.50}$$

$$f_Y(y) = \frac{\partial}{\partial y}[1 - F_X(g^{-1}(y))] = -f_X(g^{-1}(y)) \frac{dx}{dy} \tag{1.51}$$

Merk at $\frac{dx}{dy} < 0$ slik at uttrykket er positivt. Vi kan få felles uttrykk for begge tilfellene med

$$f_Y(y) = f_X(g^{-1}(y)) |\frac{dx}{dy}| \tag{1.52}$$

Dette gir oss en formel med størrelser vi må plugge inn. Vi trenger inverse og den deriverte av den inverse. Eksempel: $f(x) = x^2/9, x \in (0, 3), y = g(x) = x^3$

$$g^{-1}(y) = y^{\frac{1}{3}}, \quad \frac{dx}{dy} = \frac{1}{3} y^{-\frac{2}{3}} \tag{1.53}$$

$$f_Y(y) = \frac{\left(y^{\frac{1}{3}}\right)^2}{9} \frac{1}{3} y^{-\frac{2}{3}} = \frac{1}{27}, \quad y \in (0, 27) \tag{1.54}$$

$$\tag{1.55}$$

Husk at vi kan finne $f(x)$ ved å transformere tilbake, så ingen unnskyldning for å gjøre feil!

1.6 Flere variabler

Jeg begynner med å betrakte bivariat fordeling med to variabler. Deretter generaliserer jeg til N variabler.

1.6.1 Bivariat fordeling

Vi kan definere flere tilfeldige variabler på samme utfallsrom Ω. Anta for eksempel at vi kaster et kronestykket to ganger slik at $\Omega = \{HH, HT, TH, TT\}$ og la X være antall heads på første kast og Y være antall heads totalt. Vi kan ordne disse tilfeldige variablene i en tuple slik hver realisering utgjør et punkt i $\mathbb{R}^2$, f.eks: $(X(HT), Y(HT)) = (1, 1)$. På samme måte som med én variabel har denne vektoren en fordeling $\mathcal{L}(X, Y) = P$ som angir sannsynlighet til delmengder av $\mathbb{R}^2$. Denne kan beskrives med en kumulativ fordeling F som tar vektor som argument og der

$$F(x, y) = P(\{X \leq x\} \cap \{Y \leq y\}) \tag{1.56}$$

$$= \mathbb{P}\{\omega \in \Omega : X(\omega) < x \wedge Y(\omega) < y\} \tag{1.57}$$

Hvis fordelingen er kontinuerlig er det en simultanfordeling $f(x, y)$ der

$$F(x, y) = \int^x \int^y f(w_1, w_2) dw_1 dw_2 \tag{1.58}$$

Hvis fordelingen derimot er diskret er sammenheng mellom simultan og kumulativ gitt ved

$$F(x, y) = \sum_{w_1 < x} \sum_{w_2 < y} p(w_1, w_2) \tag{1.59}$$

Vi kan utlede de marginale fordelingene ved å observere at

$$P(X = x) = P(\{X = x\} \cap \{Y < \infty\} = \mathbb{P}(\{X = x\} \cap \Omega \tag{1.60}$$

$$= \sum_y p(x, y) \tag{1.61}$$

Betinget fordeling

Vi har nå definert simultanfordeling og sett hvordan vi kan få ut igjen marginal fordeling fra dette. Det er veldig interessant å betrakte fordelingen til én av variabelene gitt verdien

av de andre... Må bli neste gang.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \implies p_{X|Y}(x,y) = \frac{p_{X,Y}(x,y)}{p_Y(y)} \tag{1.62}$$

1.6.2 Multivariat fordeling

Vi kan utvide til flerdimensjonale fordelinger der $P : \mathscr{B}(\mathbb{R}^N) \to [0,1]$, $\mathbf{x} : \Omega \to \mathbb{R}^N$ og

$$F_{\mathbf{x}} : \mathbb{R}^N \to [0,1] \tag{1.63}$$

$$\mathbf{s} \mapsto P(\times^N(-\infty, s)) \tag{1.64}$$

som karakteriserer fordelingen med den simultane kumulative fordelingen. For absolutt kontinuerlig fordelte variabler er det også en simultan tetthetsfunksjon $p(\cdot)$ som oppfyller egenskapen

$$P(B) = \int_B p(\mathbf{s}) d\mathbf{s} \tag{1.65}$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbb{I}_B(s_1, ..., s_N) p(s_1, ..., s_N) ds_1 ... ds_N \tag{1.66}$$

Den marginale fordelingen til variabel n i vektoren og dens marginale cdf er gitt ved

$$P_n(B) = P(\mathbb{R} \times \cdot \times \mathbb{R} \times B \times \mathbb{R} \cdot \times \mathbb{R}) \tag{1.67}$$

$$F_n(s) = P_n((-\infty, s)) \tag{1.68}$$

Den simultane kumulative fordelingen karaktersierer hele fordelingen. Med utgangspunkt i denne kan vi finne marginale kumulative fordelinger. Vi kan også finne betingede fordelinger ved å skalere simultanfordelinger med marginale. Det er derimot som oftest

1.6.3 Momenter i flerdimensjonale fordelinger

Forventningsverdi til $\mathbf{x} = (x_1, ..., x_N)$ er bare en vektor der hver komponent er forventningsverdi til den tilhørende tilfeldige variabelen. Variansen er

$$var(\mathbf{x}) := \Sigma = \mathbb{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'] \tag{1.69}$$

$$= \mathbb{E}\mathbf{x}\mathbf{x}' - \mu\mu' \tag{1.70}$$

der elementene $\Sigma_{ij} = cov(x_i, x_j)$

$$var(A\mathbf{x}) = A\Sigma A' \tag{1.71}$$

Kovariansen til to variabler er forventningsverdien til produktet av avviket fra forventningsverdi til hver av variablene

$$cov(X,Y) := \sigma_{X,Y} = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \tag{1.72}$$

Kan få litt intuisjon av at kovarians er positiv hvis det er tendens til at positive avvik skjer samtidig i begge variablene. Fanger opp om det er lineær sammenheng. Kan skalere ved å dele på produktet av standardavvikene og få korrelasjonskoeffisient som er begrenset av (-1,1)

$$\rho :=:= \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{1.73}$$

Momentgenerende funksjon

Kan merke oss kort at mgf til tilfeldig variabel $\mathbf{x} = [x_1, ..., x_N]$ er

$$M(\mathbf{t}) = \mathbb{E}[exp\{\mathbf{t}'\mathbf{x}\}] = \int \cdots \int exp\{\mathbf{t}'\mathbf{x}\} f(\mathbf{x}) dx_1 \dots dx_N \tag{1.74}$$

og det følger at vi kan bruke dette til å finne marginal mgf til x_n ved evaluere den i $\mathbf{t} = (t_1, \dots, t_n, \dots, t_N) = (0, \dots, t_n, \dots, 0)$.

1.6.4 Betinget fordeling

Kan først merke oss at $\mathbb{P}(\cdot|A)$ er en fullverdig probability measure. Det avgrenser universet av mulige utfall, men gitt at det har skjedd kan vi jo bare tenke på det som vårt nye univers. Vi snakker om tilfeldige variabler her, så

$$\mathbb{E}[Y|X=x] = \int y f(y|x) dy \tag{1.75}$$

For gitt $X = x$ er dette et uttrykk som i prinsippet kan evalueres og gir oss et tall. Før X er realisert er den betingede forventningen en tilfeldig variabel der $\mathbb{E}[Y|X=x] = g(X)$. I tillegg til betinget forventning har vi også betinget varians

$$\mathbb{V}(Y|X=x) = \int (y - \mathbb{E}[Y|X=x])^2 f(y|x) dy \tag{1.76}$$

kan dekomponere den samlede variansen i ..

$$\mathbb{V}(Y) = \mathbb{E}[\mathbb{V}(Y|X)] + \mathbb{V}[\mathbb{E}(Y|X) \tag{1.77}$$

Skal nå se på alternativ utledning som gir mer intuisjon ...

1.6.5 Projeksjon av tilfeldige variabler

Det er mulig å konstruere vektorrom som består av andre objekter enn tradisjonelle vektorer (tupler av reelle tall). Skal nå konstruere vektorrom og utvikle analoge resultat for ortogonal projeksjon og minimering av avstand mellom objekter i rommet.

Vi betrakter en mengde av tilfeldige variabler. Det hadde vært veldig greit å ha et mål på avstand mellom objektene i mengden. Et naturlig mål er $RMSE(x,y) = \sqrt{E[(x-y)^2]}$. Hvis vi definerer indre produkt mellom objekter som $\langle x, y\rangle = E[xy]$ får vi det analoge resultat at $\|x\| = \sqrt{\langle x, x\rangle}$ og vi kan definiere $x \perp y = 0 \iff E[xy] = 0$. På tilsvarende måte utgjør delmengder underrom dersom de er lukket for skalering og addisjon. La $\mathbb{1}_\Omega := \mathbb{1}$ være tilfeldig variabel som tar verdi 1 med sannsynlighet 1. Et eksempel på underrom er da

$$span(X) = \{\alpha\mathbb{1} + \beta x : \alpha, \beta \in \mathbb{R}\} \tag{1.78}$$

På samme måte som i tradisjonelle vektorrom har ortonormale basiser gode egenskaper. En mengde U er ortonormal basis for S hvis

$$E[u_j u_k] = \mathbb{1}\{j = k\} \quad \text{og} \quad span\{u_1, ..., u_k\} = S \tag{1.79}$$

Jeg vil derfor lage en ortonormal basis for $S = span\{\mathbb{1}, x\}$. I tråd med gram schmidt algoritmen trekker jeg fra komponenten til x som går i retning til den konstante tilfeldige variabelen, nemlig $E[x] := \mu$. Da sitter jeg igjen med residualen $x - \mu$. Lengden til denne variabelen, i henhold til normen som ble definert over, er $\sigma_x := \sqrt{E[(x-\mu)^2]}$. Bruker dette til å skalere og har ortonormal basis

$$U = \{\mathbb{1}, \frac{x-\mu}{\sigma_x}\}, \quad span(U) = S \tag{1.80}$$

Ser da at ved å standardisere variablene i data så er det utvalgsanalog til å lage ortogonal basis for de tilfeldige variablene. Gitt definisjonene over er det ortogonale projeksjonstheoremet helt analogt så gidder ikke gjenta dette. Det er veldig kult at vi kan overføre teori om vektor til R.V siden vi får veldig mye gratis og det gir oss geometrisk intuisjon. Tenk for eksempel at jeg vil projektere y på S som jeg har gitt en ortonormal basis. La

den være u_1, u_2. Det følger da at

$$\mathbf{P}y = \langle y, u_1\rangle u_1 + \langle y, u_2\rangle u_2 \tag{1.81}$$

$$= E[y] + E\left[\frac{x-\mu}{\sigma_x}y\right]\frac{x-\mu}{\sigma_x} \tag{1.82}$$

$$= E[y] + \frac{cov(x,y)}{var(x)}(x-\mu) \tag{1.83}$$

$$= (E[y] - \beta\mu) + \beta x \tag{1.84}$$

merk at $E[(x-\mu)y] = cov(x,y)$ fordi $E[x-\mu] = 0$. Dette er populasjonsversjonen av den bivariate lineære regresjonslinjen der jeg brukte ortonormal basis. Skal nå generalisere til å finne $\mathbf{P}y = proj_S y = \mathbf{x}'\beta$. Merk at jeg alltid kan slenge inn en konstant tilfeldig variabel i mengden siden vi alltid "observererdenne uavhengig av hvilke data vi har.

$$E[(y - \mathbf{x}'\beta)\mathbf{x}] = 0 \tag{1.85}$$

$$\implies \beta = E[\mathbf{x}\mathbf{x}']E[\mathbf{x}y] \tag{1.86}$$

veldig nice. TODO: knytte til prediksjon, informasjonmengde og utvide til arbitrære funksjoner ved å definere underom $L_2(X)$.

1.6.6 Projektering i L_2

Begynner med å definere L_2 som mengden av tilfeldige variabler med definert andre moment, $L_2 = \{x|\mathbb{E}(x^2) < \infty\}$. Vi kan definere et indre produkt mellom elementer i mengden, $\langle x, y\rangle = \mathbb{E}(xy)$ der $\mathbb{E}(xy) = 0 \implies y \perp x$. Normen til elementer i mengden er $\|x\| = \sqrt{\langle x, x\rangle}$. Merk at $\|x-y\| = \sqrt{\mathbb{E}[(x-y)^2]}$ som tilsvarer *root mean square error*. Delmengder utgjør et lineært underrom hvis de er lukket under skalering og addisjon, $x, y \in S \implies \alpha x + \beta y \in S, \quad \forall \alpha, \beta \in \mathbb{R}$. For en mengde tilfeldige variabler $\mathbf{x} = (x_1, ..., x_K)$ vil $span(\mathbf{x}) = \{\mathbf{x}'\mathbf{b}|\mathbf{b} \in \mathbb{R}^{\mathbb{K}}\} \equiv S(\mathbf{x})$ være et underrom som består av alle lineære kombinasjoner av $(x_1, ..., x_K)$. Mer generelt kan vi betrakte arbitrære deterministiske funksjoner av $\mathbf{x}$. En variabel z er $\mathbf{x}$-*measurable* hvis det finnes en funksjon $h : \mathbb{R}^{\mathbb{K}} \to \mathbb{R}$, der $h(\mathbf{x}) = z \in L_2$. Mengden av disse variablene utgjør også et underrom som vi kan kalle $L_2(\mathbf{x})$.

Anta nå at vi for $y \in L_2$ og et underrom $S \subset L_2$ vil vi finne elementet i S som minimerer avstanden til y. Analogt til tradisjonelle vektorrom kan L_2 dekomponeres i et underrom S og dets ortogonale komplement $S^\perp$, og der $S^\perp = \{z|\langle z, x\rangle = 0, \forall x \in S\}$. Alle element i L_2 kan da skrives som en sum av et element i hvert underrom, $y = \hat{y} + \hat{u}$, der $\hat{y} \in S$ og $\hat{u} \in S^\perp$. Denne $\hat{y}$ er løsningen på minimeringsproblemet $\arg\min_{z\in S}\|y - z\|$. Vi vet at løsningen eksisterer, er unik og har egenskapen $\langle (y - \hat{y}), x\rangle = 0, \forall x \in S$. Det eksisterer også en lineær transformasjon P slik at $P(y) = \hat{y}$. Denne transformasjonen

utfører den ortogonale projeksjonen av y på underrommet S.

Betrakt tilfellet der $S = S(\mathbf{x})$. Løsningen er da gitt ved $\hat{y} = \mathbf{x}'\mathbf{b}^*$, der $\langle x_k, y - \mathbf{x}'\mathbf{b}^* \rangle = 0, k = 1, ..., K \iff \mathbb{E}(\mathbf{x}(y - \mathbf{x}'\mathbf{b}^*) = \mathbf{0} \iff \mathbf{b}^* = \mathbb{E}(\mathbf{x}\mathbf{x}')^{-1}\mathbb{E}(\mathbf{x}y)$, hvis $\mathbb{E}(\mathbf{x}\mathbf{x}')$ er inverterbar.

Så langt er det helt analogt til tradisjonelle vektorrom, men vi kan nå knytte ortogonal projeksjon til forventning. Vi kan definere

$$\mathbb{E}(y|\mathbf{x}) \equiv \underset{z \in L_2(\mathbf{x})}{\arg\min} \|y - z\| \tag{1.87}$$

Den betingede forventningsverdien av y gitt $\mathbf{x}$ er den $\mathbf{x}$-*measurable* variabelen som minimerer avstanden til y. Hvis vi definerer en konstant tilfeldig variabel $\mathbb{1}_\Omega \equiv \mathbb{I}\{\omega \in \Omega\}$, så vil ubetinget forventning være gitt ved

$$\mathbb{E}(y) \equiv \underset{z \in L_2(\mathbb{1}_\Omega)}{\arg\min} \|y - z\| \tag{1.88}$$

Den konstante tilfeldige variabelen som minimerer avstanden til y.

Payoff

Okay, hva er gevinsten ved å tenke på forventning som ortogonale projeksjoner?

1. $y = \mathbb{E}(y|\mathbf{x}) + u \implies \mathbb{E}[g(\mathbf{x})u] = 0, \forall g$ følger direkte av at $\mathbb{E}(y|\mathbf{x})$ er ortogonal projeksjon av y på $L_2(\mathbf{x})$. Alle andre variabler $g(\mathbf{x})$ ligger i underromet $L_2(\mathbf{x})$ og u er derfor ortogonal med disse. Siden $\mathbb{E}(u) = 0$ er u ukorrelert med alle deterministiske funksjoner av $\mathbf{x}$.

2. $y = \mathbf{x}'\mathbf{b}^* + u \implies \mathbb{E}[\mathbf{x}'\gamma u] = 0, \forall \gamma \in \mathbb{R}^{\mathbb{K}}$. Av samme argument er feilledd fra projeksjon på $S(\mathbf{x})$ ortogonal på alle lineære kombinasjoner av $\mathbf{x}$.

3. Har generelt at hvis underrommene $V_2 \subset V_1$ så vil det være ekvivalent om man projekterer y direkte på V_2 eller først projekterer på V_1 og deretter på V_2. Ettersom $S(\mathbf{x}) \subset L_2(\mathbf{x})$ vil da $\mathbf{x}'\mathbf{b}^*$ være beste lineære tilnærming til $\mathbb{E}(y|\mathbf{x})$.

4. Kan bruke et tilsvarende argument for å utlede *law of iterated expectations*. Merk at $L_2(\mathbb{1}_\Omega) \subset L_2(\mathbf{x})$ slik at $\mathbb{E}[\mathbb{E}(y|\mathbf{x})] = \mathbb{E}(y)$.

5. Kan bruke ortogonal dekomponering av underrom,$S(\mathbf{x}) \subset L_2(\mathbf{x}) \subset L_2$, og pythagoras' setning til å dekompene den forventede feilen ved å predikere y med $\mathbf{x}'\mathbf{b}$. La $\mathbb{E}(y|\mathbf{x}) \equiv f^*(\mathbf{x})$

$$\|y - \mathbf{x}'\mathbf{b}\|^2 = \|y - f^*(\mathbf{x})\|^2 + \|f^*(\mathbf{x}) - \mathbf{x}'\mathbf{b}^*\|^2 + \|\mathbf{x}'\mathbf{b}^* - \mathbf{x}'\mathbf{b}\|^2 \tag{1.89}$$

$$\mathbb{E}[(y - \mathbf{x}'\mathbf{b})^2] = \mathbb{E}[(y - f^*(\mathbf{x})^2] + \mathbb{E}[(f^*(\mathbf{x}) - \mathbf{x}'\mathbf{b}^*)^2] + \mathbb{E}[(\mathbf{x}'\mathbf{b}^* - \mathbf{x}'\mathbf{b})^2] \tag{1.90}$$

6. Kan tilsvarende dekomponere *mean square error* mellom en estimator og parameter i varians og kvadrert bias ved å projektere på $L_2(\mathbb{1}_\Omega)$

$$\|\hat{\theta} - \theta\|^2 = \|\hat{\theta} - \mathbb{E}(\hat{\theta})\|^2 + \|\mathbb{E}(\hat{\theta}) - \theta\|^2 \tag{1.91}$$

7. Linearitet til forventning følger av linearitet til ortogonale projeksjoner.

1.7 Samling av regler

Lov om total sannsynlighet. Hvis $\{B_j\}$ er en partisjonering av Ω er

$$\mathbb{P}(A) = \sum_j \mathbb{P}(A|B_j)\mathbb{P}(B_j) \tag{1.92}$$

Kan omskrive simultanfordeling

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{1.93}$$

Kan utlede analoge resultat med tilfeldige variabler ved å merke at vi kan omskrive $A = \{X = x\} := \{\omega : X(\omega) = x\}$. Betingede fordelinger er fullverdige fordelinger, så har samme egenskaper bare at vi må dra med oss mengden vi betinger av

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|B \cap C)\mathbb{P}(B \cap C) \tag{1.94}$$

Lov om iterated expectations

$$E[X] = E[E[X|Z]] = \sum_z E[X|Z = z]P(Z = z) \tag{1.95}$$

$$E[Y|X] = E[E[Y|X, Z]] = \sum_z E[Y, X, Z = z]P(Z = z) \tag{1.96}$$

Kapittel 2

Stokastiske prosesser

En stokastiske prosesser på $\mathbb{R}^K$ er en samling tilfeldige vektorer $(\mathbf{x}_t)_{t\in T}$ definert på samme sannsynlighetsrom $(\Omega, \mathscr{F}, \mathbb{P})$. Vi betegner utfallsrommet T som indeksmengden og utfallsrommet til de tilfeldige vektorene er såkalt *state space*. Det at de lever i samme sannsynlighetsrom medfører at de har en simultanfordeling. Denne simultanfordelingen kan ofte være ganske komplisert fordi

$$f(x_1, ..., x_N) = f(x_1)f(x_2|x_1)f(x_3|x1, x2)\cdots f(x_N|x_1, ..., x_{N-1}) \tag{2.1}$$

$$= \prod f(x_n|\text{historie}_n) \tag{2.2}$$

der $\text{historie}_n = (x_1, ..., x_{n-1})$. Stort sett har vi unngått dette problemet ved å anta uavhengighet, men vi skal utvikle vektøy for å håndtere avhengighet mellom observasjoner. Først litt motivasjon

1. Det åpner for avhengighet i sampling. Dette er spesielt relevant når vi observerer samme enhet på flere tidspunkt (panel/tidsserie), men det kan også være avhengighet i krysseksjon. Tror det kan være avhengighet for observasjoner i samme gruppe (klasse,geografisk område, mm.), men veldig usikker på dette. Tror en tilnærming til dette er å skalere standardfeil med cluster.

2. Vi kan bruke det til å studere egenskap til estimator. For utvalg med gitt N kan vi utlede fordeling til estimator under gitt antagelser, men vi kan også være interessert i hvordan denne fordelingen endres når vi endrer N. Vi kan da betrakte det som en stokastisk følge der vi i hvert ledd observerer én ny realisering. Ved å undersøke egenskaper til følgen når N går mot uendelig kan vi utlede egenskaper under svakere antagelser og bruke dette som tilnærming for store utvalg.

3. Vi kan bruke det til å modellere systemer. Verden er dynamisk, ikke statisk. Mer om dette senere.

2.1 Asymptotisk teori

Jeg betrakter en følge av tilfeldige variabler $(Z_1, Z_2, ...) = (Z_N)_{N\geq 0}$. Vi kan tenke at egenskapene til variablene avhenger av plassering i følgen, altså av N. Vi kan være interessert i egenskap for en gitt N eller for alle N som er større enn en gitt verdi. I praksis er vi ofte interessert i grenseverdiene når N går mot uendelig... altså om følgen konvergerer. I motsetning til kalkulus har vi ulike former for konvergens.

1. Konvergens i sannsynlighet: $X_N \xrightarrow{p} X \iff \lim_{N\to\infty} P(\|Z_n - Z\| > \epsilon) = 0, \forall \epsilon > 0$
2. Konvergens i fordeling: $X_N \xrightarrow{d} X \iff \lim_{N\to\infty} F_N(t) = F(t)$ for alle t der $F(\cdot)$ er kontinuerlig.
3. Konvergens i *mean square* (L2): $X_N \xrightarrow{m.s} X \iff \lim_{N\to\infty} \mathbb{E}(X_n - X)^2 = 0$

Konvergens i sannsynlighet impliserer konvergens i fordeling. Tror jeg bruker konvergens i m.s. til å bevise konvergens i sannsynlighet. Merk at en konstant c bare er special case av R.V. X der $\mathbb{P}(X = c) = 1$. Det er et veldig fint resultat at konvergens er bevart av kontinuerlig transformasjoner $g(\cdot)$, slik at $Z_n \xrightarrow{p} Z \implies g(Z_n) \xrightarrow{p} g(Z)$. Noen regneregler:

1. $Z_N \xrightarrow{p} Z \implies aZ_N \xrightarrow{p} aZ$
2. $Z_N \xrightarrow{p} Z$ og $X_N \xrightarrow{p} X \implies Z_N X_N \xrightarrow{p} ZX$
3. $X_N \xrightarrow{p} X \wedge Y_N \xrightarrow{p} Y \implies X_N + Y_N \xrightarrow{p} X + Y$
4. $X_N \xrightarrow{p} X \wedge Y_N \xrightarrow{p} Y \implies X_N Y_N \xrightarrow{p} XY$
5. $X_N \xrightarrow{d} X \wedge Y_N \xrightarrow{p} c \implies X_N Y_N \xrightarrow{d} cX$
6. $A_N \xrightarrow{p} A$ og $\mathbf{x_N} \xrightarrow{d} \mathbf{x_N} \implies A_N \mathbf{x_N} \xrightarrow{d} A\mathbf{x}$. Medfører at dersom $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$ så er $A\mathbf{x} \sim N(\mathbf{0}, A\Sigma A')$

Tror det er noen regler til og de generaliserer til vektorer/matriser. Noen av disse er slutskys theorem.

2.1.1 Store talls lov

Store talls lov som sier at utvalgsmoment fra *iid* prosess konvergerer i sannsynlighet til populasjonsmoment. Så lenge første moment er definert så vil

$$\mathbb{E}_{P_N}[X_n] = \frac{1}{N}\sum X_n \xrightarrow{p} \mathbb{E}[X] \tag{2.3}$$

For å bevise dette vil jeg først utlede Markovs og Chebyshevs ulikheter. La X være ikke-negativ tilfeldig variabel og $g(\cdot)$ er transformasjon som flytter tyngde nedover

$$Y = g(X) = \begin{cases} a, \text{ hvis } Y \geq a \\ 0, \text{ ellers} \end{cases} \tag{2.4}$$

Det følger da at

$$\mathbb{E}[X] \geq aP(X \geq a) \tag{2.5}$$

$$\implies P(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{2.6}$$

som er Markovs ulikhet. Kan utlede Chebyshevs ulikhet som spesialtilfelle der $X = (\bar{X}_N - \mu)^2$ og $a = \epsilon^2$.

$$P[(\bar{X}_N - \mu)^2 \geq \epsilon^2] \leq \frac{\mathbb{E}[(\bar{X}_N - \mu)^2]}{\epsilon^2} = \frac{\mathbb{V}[\bar{X}_N]}{\epsilon^2} \tag{2.7}$$

$$\implies P(\mid \bar{X}_N - \mu \mid \ \geq \epsilon) \leq \frac{\sigma^2}{N\epsilon^2} \to 0 \tag{2.8}$$

når $N \to \infty$. Okay, nå var jo beviset ferdig uten at jeg definerte Chebyshevs ulikhet... Kunne også tatt beviset fra *mean square*. Uansett, LLN er et veldig fundamentalt resultat fordi det enkelt kan utvides. Det gjelder også for utvalgsmoment til kontinuerlige funksjoner av X.

$$\frac{1}{N}\sum g(X_n) \xrightarrow{p} \mathbb{E}[g(X)]. \tag{2.9}$$

Det kan også brukes til å bevise at relativ andel i utvalg konvergerer til sannsynlighet

$$\frac{1}{N}\sum I\{X_n \in B\} \xrightarrow{p} \mathbb{E}[I\{X \in B\}] = P(B) \tag{2.10}$$

2.1.2 Sentralgrenseteoremet

Store talls lov sier at i uendelig store utvalg er hele tyngden av fordelingen til utvalgsmomentet konsentrert på de populasjonsmomentet. Det har sammenheng med at empirisk fordeling konvergerer til teoretisk fordeling. Det er et viktig teoretisk resultat, men i praksis har vi aldri uendelig store utvalg så vi vil også vite noe om hvor raskt tyngden til utvalgsfordeling konvergerer: mer presist ønsker vi å angi sannsynlighet for avvik mellom utvalgsgjennomsnitt og forventningsverdi. Her kommer sentralgrenseteoremet oss til

unnsetning. Det sier at så lenge $\{X_n\}$ er *iid* og $\mathbb{E}[X^2] < \infty$

$$\sqrt{N}(\bar{X}_N - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{2.11}$$

$$\frac{\bar{X}_N - \mu}{\sqrt{\mathbb{V}[\bar{X}_N]}} \xrightarrow{d} N(0,1) \tag{2.12}$$

$$\tag{2.13}$$

Ser at dette er veldig nyttig siden differansen av utvalgsmoment og populasjonsmoment (oppskalert med rate of convergence) blir normalfordelt uavhengig av underliggende fordeling. Normalfordeling er jævlig nice fordi den blir bevart av transformasjoner. Så det for lineær trasnformasjon i sted og kan generalisere til alle differensierbare transformasjoner med delta-metoden. Dette kan brukes til å utlede fordeling til estimatorerer!

2.1.3 Delta-metoden

Husker at $A_N \xrightarrow{p} A$ og $\mathbf{x_N} \xrightarrow{d} \mathbf{x} \implies A_N\mathbf{x_N} \xrightarrow{d} A\mathbf{x}$. Dette er veldig nice siden CLT kan gi oss at at $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$. Jeg vet allerede hvordan jeg finner fordeling til lineær transformasjon. Nå skal jeg også kunne finne fordeling til transformasjoner som er lokalt lineære (ie. kontinuerlige..).

$$Y_N \xrightarrow{d} N(\mu, \frac{\sigma^2}{N}) \implies g(Y_N) \xrightarrow{d} N\left(g(\mu), (g'(\mu))^2\frac{\sigma^2}{N}\right) \tag{2.14}$$

2.1.4 LNN og CLT med flere variabler

2.1.5 LLN og CLT med avhengighet mellom observasjoner

Hvis prosessen er *iid* så er $\mathcal{L}(\mathbf{x}_t) = P, \forall t$ og simultanfordelingen er produkt av marginal. Slike følger er veldig greie å jobbe med siden vi kan bruke LLN og CLT, men det er litt restriktivt siden det kan være litt persistens i størrelse over tid. Hvis vi vil jobbe med tidsserier er det derfor relevant å få finne en større klasse av stokastiske prosesser som *nesten* er *iid* og har de samme gode egenskapene. Det viser seg at LLN holder for prosesser som er stasjonære og ergodiske. Stasjonæritet medfører at simultanfordeling til del-tupler av simultanfordelingen ikke endres av å forskyves. Altså:

$$\mathcal{L}(\mathbf{x}_{t1}, ...\mathbf{x}_{tk}) = \mathcal{L}(\mathbf{x}_{1t+m}\mathbf{x}_{tk+m}) \tag{2.15}$$

Det finnes litt ulike og kompliserte definisjoner av ergodisitet. Jeg velger å si at en stasjonær prosess er ergodisk dersom LLN holder. Dette flytter målposten til å si noe om tilstrekkelige betingelser for ergodisitet. En uformell definisjon på ergodisitet er at gjennomsnitt av observasjon over tid omtrent samsvarer med gjennomsnitt på et gitt tidspunkt.

Det er et poeng at vi kan få CLT som kun krever at prosess er martingale difference sequence. Jeg skal forsøke å utlede litt mer formelt hva dette er for noe. Jeg vet ikke hvor relevant det er, men jeg kjører på og håper at det blir litt payoff senere. Trenger først å introdusere noen konsepter.

2.2 Markov-kjeder

Markovkjeder har diskret tilstandsrom og sannsynlighet for ulike tilstander i en periode kun avhenger av tilstand i perioden før,

$$\mathbb{P}(X_n = x|X_1, ..., X_{n-1}) = (X_n = x|X_{n-1}) \tag{2.16}$$

dette forenkler uttrykket for simultanfordelingen

$$f(x_1, ..., x_N) = f(x_1)f(x_2|x_1)f(x_3|x1, x2)\cdots f(x_N|x_1, ..., x_{N-1}) \tag{2.17}$$

$$= f(x_1)\prod_{n=2}^{N} f(x_n|x_{n-1}) \tag{2.18}$$

Markovkjeder gir et rammeverk med nok struktur til at det er mulig å utlede teoretiske resultat og samtidig har det tilstrekkelig fleksibilitet til å beskrive systemer i virkeligheten.

2.2.1 Overgangssannsynlighet

De sentrale størrelsene i en markovkjede er tilstandsrommet og sannsynlighet for å bevege seg mellom tilstander. Hvis markovkjeden er overgangssanynlighetene uavhengig av n slik at vi kan definiere

$$p_{ij} := \mathbb{P}(x_{n+1} = j|x_n = i). \tag{2.19}$$

og organisere disse i en matrise P. Sentrale spørsmål: $p_{ij}(n)$ og konvergens.. mer om dette senere.

2.3 Annet

En filtrasjon er en økende følge av informasjonsmengder, altså $(\mathscr{F}_t) = (\mathscr{F}_1, \mathscr{F}_2, .., \mathscr{F}_t, ...)$ der $\mathscr{F}_t \subset \mathscr{F}_{t+1}$ $\forall t$. Merk at en informasjonsmengde bare er en mengde av tilfeldige variabler. Filtrasjonen som er generert av (x_t) er

$$\mathscr{F}_0 = \varnothing \text{ og } \mathscr{F}_t = \{x_1, ..., x_t\} \text{ for alle } t \geqslant 1 \tag{2.20}$$

En stokastisk prosess (z_t) er *adapted* til en filtrasjon $(\mathscr{F}_t)$ hvis z_t er $\mathscr{F}_t$-measurable for alle t. Altså at vi kan regne ut z_t fra størrelsene i informasjonsmengden som blir realisert på samme tidspunkt. Prosessen (z_t) er i tillegg en martingale wrt $(\mathscr{F}_t)$ hvis

$$\mathbb{E}[z_{t+1}|\mathscr{F}_t] = z_t \tag{2.21}$$

La differansen $d_t = z_t - z_{t-1}$. Dette definerer en stokastisk prosess (d_t) som er *martingale difference sequence* wrt $(\mathscr{F}_t)$ hvis

$$\mathbb{E}[d_{t+1}|\mathscr{F}_t] = 0 \tag{2.22}$$

okay... får håpe det blir noe payoff en dag LOL. Tror jeg kommer til å studere dette til høsten så forhåpentligvis blir det en søt syntese. Tidsserier og sånn.. vi får se. Men er glad for at jeg fikk sannsynlighet på litt tryggere grunn. Litt usikker på hva jeg skal gjøre fremover... får sjå.

Kapittel 3

Noen kjente fordelinger

3.1 Normalfordeling

3.1.1 Truncated normalfordeling

Generelt er en truncated fordeling betinget av at variabelen tar utfall i et gitt intervall. Vi vil for eksempel finne betinget tetthet til en variabel y gitt at vi vet at $y > c$. Det kan vises at dette bare er ubetinget tetthet til y skalert med sannsynligheten for å være i intervallet slik at det integrerer til 1,

$$f(y|y > c) = \frac{f(y)}{P(y > c)} = \frac{f(y)}{1 - F(c)} \tag{3.1}$$

som gir mening siden

$$\int_c^\infty f(y)dy = 1 - \int_{-\infty}^c f(y)dy = 1 - F(c) \tag{3.2}$$

Dersom y er standardnormalfordelt kan vi utlede enkle uttrykk for hvordan momentene til den avkuttede fordelingen til y avhenger av c,

$$\mathbb{E}[y|y > c] = \frac{\phi(c)}{1 - \Phi(c)} \tag{3.3}$$

og har også et uttrykk for varians. Disse vil jeg utlede og det vil også være enkelt å utvide andre normalfordelte variabler.

3.2 Fordelinger assosiert med normalfordeling

3.2.1 χ^2-fordeling.

3.2.2 t-fordeling

3.2.3 F-fordeling

3.3 Fordelinger fra bernoulli-prosess

3.3.1 Binomialfordeling

Sannsynlighet for k treff av N uavhengige $Y_n \sim Bernoulli(P)$ er

$$P_x(k) = \binom{N}{k} p^k (1-p)^{(N-k)} \tag{3.4}$$

Merk at vi kan visualisere utfallene med et tre (graf) siden det deler i to i hvert steg. Forventningsverdi er

$$\mathbb{E}X := \mu_X = \mathbb{E}g(Y_1, ..., Y_N) = \mathbb{E}\sum Y_n = \sum \mathbb{E}Y_n = np \tag{3.5}$$

Variansen er

$$\mathbb{V}X := \sigma_X^2 = \mathbb{V}g(Y_1, ..., Y_N) = \sum \mathbb{V}Y_n = np(1-p) \tag{3.6}$$

3.3.2 Geometrisk fordeling

Tar utgangspunkt i en *iid* bernoulli prosess med sannsynlighet for treff ρ. Kan definere X som antall forsøk som kreves for å oppnå et treff og $Y = X - 1$ som antall feil før første treff. Vi kan ganske enkelt utlede sannsynlighetsfordelingen til disse.

$$P(X = k) = (1-\rho)^{k-1}\rho, k = 1, 2, ... \tag{3.7}$$

$$P(Y = k) = (1-\rho)^{k}\rho, k = 0, 1, ... \tag{3.8}$$

3.3.3 Negativ binomialfordeling

Antall feil før vi oppnår r treff.

3.3.4 Multinomialfordeling

Generalisering av binomialfordeling der det er K kategorier i stedet for bare 2. Kan tenke på det som sannsynlighet for antall baller med ulike farger fra en urne etter N trekk med

andeler $p_1, p_2, ..., p_K$. Tenker at det er en tilfeldig vektor $(X_1, X_2, ..., X_K)$ som angir antall i hver kategori og at vi kan beskrive pmf til denne.

3.4 Fordelinger fra poisson-prosess

3.4.1 Poissonfordeling

3.4.2 Eksponetialfordeling

Lengde mellom treff i poissonfordeling.

3.5 Andre fordelinger

3.5.1 Uniformfordeling

Det er en kontinuerlig fordeling med like stor sannsynlighet for utfall i alle delmengder som er like store. I én dimensjon kan vi spesifisere $X \sim U(a, b)$, uniform på intervallet $[a, b]$. I flere dimensjoner kan vi være litt mer kreative med geometriske objekt, f.eks. disk eller kradrat. Tetthetsfunksjonen vil uansett være en konstant siden den ikke avhenger av hvor i mengden vi er. For å finne denne konstanten i én dimensjon kan vi bruke

$$\int_a^b k dx = k(b-a) = 1 \implies k = \frac{1}{b-a} \tag{3.9}$$

Dette følger også av at vi skal ha et rektangel der ene siden er bredde er k, lengde er $b-a$ og areal skal være 1. Forventningsverdi er

$$\mathbb{E}X = \int_a^b x f(x) dx = \frac{1}{b-a} \int_a^b x dx = \frac{a^2 - b^2}{2(b-a)} = \frac{a+b}{2} \tag{3.10}$$

Kan også merke oss at cdf til enhetsuniform er $F(x) = \int f(x) = \int 1 = x$. Hvis jeg har N uavhengige uniforme variabler og jeg vil finne forventningsverdi til $Y = \max\{X_1, ...X_N\}$ kan jeg bruke at

$$F(Y_n = y) = P(X_1 < y, ..., X_N < y) = \Pi F_X(y) = y^N \tag{3.11}$$

$$\mathbb{E}Y = \int_0^1 y f(y) dy = \int_0^1 N y^{N-1} y dy = \int_0^1 N y^N dy \tag{3.12}$$

$$= \frac{N}{N+1} \tag{3.13}$$

3.5.2 Gammafordeling

3.5.3 Betafordeling

Kapittel 4

Inferens

I sannsynlighetsteori tar vi utgangspunkt i et veldefinert probabilistisk eksperiment som er beskrevet av sannsynlighetsrommet $(\Omega, \mathscr{F}, \mathbb{P})$ og bruker dette til å beregne sannsynlighet for ulike hendelser. I praksis jobber vi ofte med $\Omega = \mathbb{R}$ slik at vi kan bruke kumulativ fordelingsfunksjon til beregne sannsynlighet for ulike utfall gitt fordelingen P. Dette forutsetter at fordelingen er kjent. I statistikk er fremgangsmåten omvendt; vi observerer utfall og med utgangspunkt i dette vil vi si noe om egenskapene til fordelingen som genererte utfallene. De realiserte utfallene gir ikke tilstrekkelig informasjon til å entydig bestemme egenskaper ved P; ulike fordelinger kan generere samme utfall og hvis vi observerer nye realiseringer fra samme fordeling vil det være tilfeldig variasjon. Det er derfor vesentlig å kvantifisere usikkerheten til våre mål på egenskapen til P.

4.1 Motivasjon

Vi tar alltid utgangspunkt i et datasett som egentlig bare er en mengde av tall organisert i en matrise. Målet er å beskrive egenskaper til fordelingen som genererte utvalget.[1] Gitt at utvalget er representativt vil de realiserte verdiene gir oss informasjon om prosessen, men med endelige mengde data vil informasjonen være ufullstendig. Hvis vi observerer et nytt datasett fra den samme fordelingen ville vi fått andre verdier. Dette medfører tilfeldig variasjon, og vi må utlede et rammeverk for å håndtere denne variasjonen for å kvantifisere hvor mye vi kan lære om egenskaper til den fordelingen fra det éne gitte datasettet vi har.

I utgangspunktet kan vi være ganske agnostisk om fordelingen til variablene i utvalget så lenge vi har tilstrekkelig antall observasjoner, og disse er uavhengige og identisk

[1] I økonometri forsøker vi å beskrive en egenskap til den sanne prosessen som genererte utfallet y. Dette kan avvike fra simultanfordeling mellom variablene vi observerer. Vi kan tenke at det er en sann, deterministisk prosess $y = f(x_1, ..., x_K)$ som bestemmer utfallet, men vi observer bare en delmengde av disse variablene. Utfordring blir da å si noe om den kausale $\frac{\partial f}{\partial x_k}$ gitt den doble utfordringen at vi kun har delmengde av variabler og begrenset antall observasjoner. Teori om inferens kan brukes til å håndtere det siste. For førstnevnte trenger vi forskningsdesign med tilfeldig variasjon x_k. Mer om dette senere.

fordelt. Store talls lov sier oss at gjennomsnitt i utvalget konvergerer mot gjennomsnitt i populasjonen og sentralgrenseteoremet lar oss kvantifisere sannsynlighet for størrelsen på avviket. Sannsynlighetsfordelingen til avviket er normalfordelt slik at små avvik er betydelig mer sannsynlig enn store avvik. Hvor stor spredningen til fordelingen er avhenger av antall observasjoner i utvalget og hvor mye vi lærer om egenskapen fra hver observasjon. Som vi skal se kan disse to teoremene også anvendes på andre estimatorer enn rene gjennomsnitt.

Fordelen med å være agnostisk er at konklusjonene vi gjør er gyldig nesten uavhengig av egenskapene til den ukjente P. I praksis vil det være hensiktsmessig å påføre litt mer struktur ved å legge avgrensinger på hvilken form fordelingen kan ta. Vi kan betegne dette som modellering.

4.1.1 Modellering

Generelt er modeller er forenklede representasjoner av virkeligheten som er enklere å tolke og manipulere, og dermed kan brukes til å analysere spesifikke mekanismer og svare på gitte spørsmål. Hvorvidt en modell er sanneller ikke er derfor litt *besides the point*; det avgjørende er hvorvidt det er egnet til sitt formål. Modeller har en struktur - altså, den består av størrelser og relasjoner mellom disse - som gjør at vi kan manipulere de på logiske konsekvente metoder og trekke entydige konklusjoner. Dette er mulig fordi vi er eksplisitt om premissene. Nedsiden med dette er at premissene nødvendigvis innebærer forenklinger og ikke er oppfylt i virkeligheten. Konklusjonene vi trekker fra modell avhenger av disse premissene. De er derfor sanne for modellen, men det er ikke gitt at konklusjonen kan overføres på virkeligheten. Her kreves det dømmekraft og tester for å vurdere om antagelsene gir en rimelig tilnærming av virkeligheten.

Modellen påfører en struktur på prosessen som genererte utvalget. Dette kan for eksempel være at regresjonslinjen $E[Y|X = x] = h(x)$ er glatt (eller enda sterkere: lineær). En fordel med denne forenklede strukturen er at den resulterende modellen blir både enkel å tolke og å manipulere. Det gjør også at vi kan bruke data fra hele utvalget til å beregne verdier av $h(x)$ for ulike x. Siden vi kan bruke informasjon fra flere observasjoner blir det mindre variasjon i verdiene av $h(x)$ fra ulike utvalg enn om vi kun bruker gjennomsnitt fra lokale observasjoner. Anta for eksempel at vi vil ser på sammenheng mellom gjennomittlig høyde og alder. For hver alder kan vi dekomponere høyde til gitt observasjon som gjennomsnitt for den alderen (signal) pluss avvik fra gjennomsnittet (støy). Hvis vi har få observasjoner for hver alder vil gjennomsnitt per alder i utvalget være sensitivt for mengden støy i det gitte utvalget vi observerer. Dersom vi bruker en parametrisk funksjonell form, for eksempel $E[høyde|alder = x] = h(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ vil estimatet for hver enkelt alder bruke informasjon fra hele utvalget slik at det blir mindre sensitivt for variasjon mellom utvalg. Mer formelt er det en bias-varianse tradeoff, der vi kan forbedre-

de estimatene ved å påføre (potensielt feil) struktur fordi det reduserer variasjon mellom utvalg. I valg av struktur vil det være en avveining som avhenger av hvor mye data vi har i utvalg og hvor mye kunnskap vi har om fordelingen fra før.[2]

4.1.2 Usikkerhet

Vi kan skille mellom to typer usikkerhet når vi bruke et gitt datasett til å besvare ulike spørsmål.

Stokastisk usikkerhet

Dette er usikkerhet i estimering av modellen gitt at den er riktig spesifisert. Denne usikkerheten skyldes at vi har begrenset antall observasjoner slik at vi ikke er helt sikre på egenskapene i prosessen; nye utvalg av samme størrelse ville gitt annen føyning og det er en kvantifiserbar usikkerhet som følge av dette. Vi kan i prinsippet håndtere denne usikkerheten på en konsekvent måte. Tenker det er viktig å klare å *propagere*/videreføre denne usikkerheten når vi bruker modellen til å beregne sannsynlighet for hendelser. Anta for eksempel at vi har en parametrisk fordeling P_θ og vi bruker $P_{\hat{\theta}}$ som estimert fordeling. Når vi beregner sannsynlighet $P_{\hat{\theta}}(A)$ så må vi ta hensyn til fordeling til $\hat{\theta}$ og sannsynlighet for hendelsen med fordelingsfunksjon med de ulike verdier av parameter. Tror dette er enklere å håndtere i bayesiansk rammeverk.

Induktiv usikkerhet

Det vil også være mer generell usikkerhet om i hvilken grad det gitte utvalget og modellen gjør at vi kan svare på spørsmål vi er interessert i. Her kreves det kontekst-spesifikk kunnskap.

1. Hvordan er data generert; er det skjevheter i utvalget? Er det missing values? Målefeil?

2. Er det riktig spesifikasjon av modell? Her er det jo mye fleksibilitet, slik at andre valg kan føre til andre konklusjoner. Viktig å være transparent og gjøre goodness of fit test på antagelser i modell som påvirker konklusjon. Konklusjoner er mer troverdig dersom de er robuste.[3]

3. Med prediksjon kan ekstern validitet i prinsippet testes direkte med kryssvalidering, men det er fortsatt vesentlig å forstå hvilke variabler/egenskaper ved observasjon som gjør at hypotesefunksjonen trekker konklusjoner om y for å bygge kredibilitet

[2] Jeg tror det er litt problemet med å forsøke å lære om struktur til fordeling fra gitt utvalg... kan finne struktur som passer til det gitte realiserte utvalget, men ikke nødvendigvis beskriver fordeling. Omtalt som p-hacking...

[3] Robust i betydningen at de ikke er sensitive for antagelser og andre (mer eller mindre) vilkårlig valg.

og sikre at det kan generaliseres til nye settinger.*En klassifier som skiller mellom ulv og husky ut i fra om det er snø på bildet kan ha gode resultater for et gitt datasett, men ikke nødvendigvis så nyttig for alle settinger.* Induktiv usikkerhet blir enda mer vesentlig når vi vil lære om datagenereringsprosses i stedet for simultanfordeling til de observerte variablene. Da må vi gjerne ha en troverdig historie som sannsynliggjør tilfeldig variasjon i behandling.

4.2 Formelt rammeverk

Vi begynner med å betrakte det enkleste tilfellet der vi observerer *iid* realiseringer $(\mathbf{z}_1, ..., \mathbf{z}_N)$ fra en ukjent fordeling $P \in \mathscr{P}$, der $\mathscr{P}$ er mengden av fordelinger vi betrakter. Denne mengden består i utgangspunktet av fordelinger, men dette korresponderer med mengde av tilhørende kumulative fordelinger som igjen korresponderer med pmf/pdf. Vi kan generelt indexe mengden med en parametervektor, slik at mengden kan beskrives som

$$\mathscr{P} = \{P_\theta : \theta \in \Theta\} \tag{4.1}$$

eller

$$\mathscr{P} = \{f(\cdot; \theta) : \theta \in \Theta\} \tag{4.2}$$

der f har kjent form.[4] Vi kan bruke dette rammeverket til å beskrive strukturen vi påfører inferensproblemet gjennom valg av f samt potensielt en avgrensing av Θ. Rammeverket er generelt og omfatter både parametriske og ikke-parametriske tilnærminger

- Parametrisk hvis fordeling kan blir indexet med parametervektor θ med endelig dimensjon.
- Ikke-parametrisk hvis θ trenger uendelig dimensjon for å karakterisere fordeling.
- Semi-parametrisk hvis vi kan dekomponere $\Theta = \Theta_0 \times \Theta_1$, der kun den første rommet har endelig dimensjon og inneholder parameter vi er interessert i.

4.2.1 Identifiserbarhet

En modell er identifiserbar hvis det eksisterer en injektiv (én-til-én) funksjon $g : \theta \mapsto \mathbb{P}_\theta$. Hvis den ikke er injectiv kan vi ikke vite parameteren selv om vi observerer selve fordelingen som har generert observasjonene. Eksempler:

- $X \sim bern(g(\theta))$, kan bare lære θ dersom $g : \theta \mapsto \rho$ er injektiv.

[4] Tror vi også kan bruke andre representasjoner dersom vi kun er interesert i spesifikk egenskap ved fordeling som genererte data..

- $Y = I\{X > a/2\}$ og $X \sim U(0, a)$. Vi observerer bare Y; kan vi lære a fra cdf til Y? Nei, fordi $P(Y = 1) = 1 - P(X < a/2) = 1 - \int_0^{a/2} \frac{1}{a} dx = 1 - (\frac{a/2}{a} = 1/2$ for alle a.

4.2.2 Fordeling til estimatorer

La oss innføre litt notasjon.

- $\mathbf{z}_n \in Z = \mathbb{R}^d$, der Z er utfallsrommet til observasjonene.
- Utvalget $\mathbf{z}_D = (\mathbf{z}_1, ..., \mathbf{z}_N)$ får realiserte verdier i et utvalgsrom $Z_D = \bigtimes_{n=1}^{N} Z = \mathbb{R}^{N \times d}$

Noen ganger vil vi dekomponere hver observasjon for å se på relasjon mellom $\mathbf{x}$ og y.[5]

- Har observasjoner $\mathbf{z}_n = (x_{1n}, ..., x_{Kn}, y_n)$ sin får realiserte verdier i et utfallsrom $\mathbf{z}_n \in Z = \mathbb{R}^{d+1}$
- Har utvalg $\mathbf{z}_D = (\mathbf{z}_1, ..., \mathbf{z}_N)$ som får realiserte verdier i et utvalgsrom $\mathbf{z}_D \in Z_D = \bigtimes_{n=1}^{N} Z = \mathbb{R}^{N \times (d+1)}$

Utvalget har en simultanfordeling $\mathcal{L}(\mathbf{z}_D) = P_D = \Pi_n \mathcal{L}(z_n) = \Pi_n P_n$ siden observasjonene er *iid*. Med utgangspunkt i de realiserte verdiene i utvalget kan vi forsøke å si noe om P. Generelt kan vi betrakte en egenskap til P som $\gamma(P)$, der $\gamma : \mathcal{P} \to S$ og S er en vilkårlig mengde for å holde det generelt. En statistikk er en funksjon T som som mapper fra utvalgsrommet til en mengde, $T : Z_D \to S$. Hvis funksjonen mapper til samme mengde som egenskapen γ så kan vi bruke den til å si noe om egenskapen. Vi kaller det da for en estimator for $\gamma(P)$ og betegner funksjonen som $\hat{\gamma}$.[6]

Estimatoren er en funksjon av utvalget. Mer presist er observasjonene $\mathbf{z}_1, .., \mathbf{z}_N$ tilfeldige variabler på et sannsynlighetsrom $(\Omega, \mathscr{F}, \mathbb{P})$ som tar verdier $\mathbf{z}_n(\omega) \in \mathbb{R}^d$. De har en fordeling P der $P\{\mathbf{z}_n \in A\} = \mathbb{P}\{\omega \in \Omega | x_n(\omega) \in A\}$. Så hver av observasjonene har en fordeling og utvalget har en simultanfordeling. Estimatoren er en funksjon av utvalget og har derfor også en fordeling. Fordelingen til estimatoren er $\mathcal{L}(\hat{\gamma})$. Evaluert på en mengde $B \subset S$ gir der

$$\mathcal{L}(\hat{\gamma})(B) = P_D\{\mathbf{z}_D \in Z_D | T(\mathbf{z}_D) \in B\} \tag{4.3}$$

Merk at fordelingen er entydig bestemt av P_D og T, men i praksis er simultanfordelingen til utvalget ukjent. Skal se på tre måter å beregne utvalgsfordelingen.

[5] Finnes veldig mye overlappende terminologi for å beskrive dette, vet ikke helt hvilket begrept jeg liker best..

[6] Merk at estimatorer per definisjon ikke kan avhenger av den ukjente egenskapen, siden det er den vi vil bruke funksjonen til å lære noe om! Statistikker mer generelt kan avhenge av ukjente egenskaper, f.eks. testobservasjoner.

4.2.3 Normalfordelt utvalg

Nå som vi har innført notasjon kan vi ta et eksempel for å gjøre dette mer konkret. La $x_1, ..., x_N$ være *iid* på $\mathbb{R}$ med $\mathcal{L}(x_n) = P = N(\mu, \sigma^2)$ for alle n. Da vet vi at $P_D = N(\mathbf{I}\mu, \sigma^2\mathbf{I})$ slik at den er kjent opp til ukjente parametre. La $\hat{\gamma}(\mathbf{x}_N) = \bar{x}_N$. Kan da finne

$$\mathcal{L}(\hat{\gamma}) = N\left(\mu, \frac{\sigma^2}{n}\right) \tag{4.4}$$

og estimere de ukjente parametrene for å finne den estimerte fordelingen.

4.2.4 Bootstrap

Metoden over krever ganske sterke antagelser. En alternativ fremgangsmåte er å anta at observasjone er *iid* og bruke den empiriske fordelingen fra de N realiseringene i utvalget som estimator på P. Vi kan sample fra dette for å observere empirisk simultanfordeling. Mer formelt er

$$\mathcal{L}_P(\hat{\gamma}) := \text{ fordelingen til } \hat{\gamma}(x_1, ..., x_N) \text{ når } x_1, ..., x_N \overset{iid}{\sim} P \tag{4.5}$$

Dette er ukjent fordi vi ikke kjenner P. Hvis vi observerer et utvalg $(x_1^0, ..., x_N^0)$ vil den empiriske fordelingen $\hat{P}_N$ være en estimator. Bootstrapfordelingen er

$$\mathcal{L}_{\hat{P}_N}(\hat{\gamma}) := \text{ fordelingen til } \hat{\gamma}(x_1, ..., x_N) \text{ når } x_1, ..., x_N \overset{iid}{\sim} \hat{P}_N \tag{4.6}$$

Denne fordelingen vil avvike litt fra den sanne fordelingen. Den er også litt vanskelig å jobbe med analytisk. I praksis kan vi simulere realiseringer fra den kjente fordelingen $\mathcal{L}_{\hat{P}_N}(\hat{\gamma})$. Dette er veldig enkelt fordi fordelingen $\hat{P}_N$ legger lik sannsynlighetstyngde på hver av realiseringene $(x_1^0, ..., x_N^0)$ slik at vi kan sample med tilbakelegging fra utvalget vårt. Algoritmen ser da slik ut

1. $\hat{P}_N$ = empirisk fordeling av observert data $(x_1^0, ..., x_N^0)$
2. for m in 1,...,M do:
3. trekk $(x_1^b, ..., x_N^b)$ fra $\hat{P}_N$
4. set $\hat{\gamma}_m^b = \hat{\gamma}(x_1^b, ..., x_N^b)$
5. end for, returner utvalget $\hat{\gamma}_1^b, ..., \hat{\gamma}_M^b$

Vi har da M realiseringer av estimatoren fra den empiriske fordelingen. Disse realiseringene er igjen en empirisk fordeling. Vi kan bruke standardavvik og forventningsverdi til denne empiriske fordelingen som estimat for de tilsvarende egenskapene til den sanne fordelingen til estimatoren.

4.2.5 Asymptotisk teori

Fordelingen til estimatorer avhenger av størrelsen på utvalget. Flere observasjoner gir mer presise estimat. Vi kan betrakte en følge av estimatorer, $(\hat{\theta}_N)_{N\in\mathbb{N}} = (\hat{\theta}_1, ..., \hat{\theta}_N, ...)$, for å se på hvordan fordelingen endrer seg når utvalget vokser. I asymptotisk teori ser vi på fordelingen i grensetilfellet der $N \to \infty$ der vi kan utlede eksakt fordeling under svakere antagelser enn det som er nødvendig for å utlede fordeling som gjelder for vilkårlig N. Denne asymptotiske fordelingen vil være en god tilnærming så lenge vi har tilstrekkelig stort utvalg...

4.3 Egenskaper til estimatorer

En estimator er en funksjon $\hat{\gamma} : \mathbf{z}_d \to S$ som vi vil bruke til å lære om $\gamma(P) \in S$. Ettersom estimatoren er en funksjon av utvalget vil den ha en utvalgsfordeling siden den tar ulike verdier for ulike realiserte utvalg fra P. Vi vil at tyngden av fordelingen til estimatoren er konsentrert rundt $\gamma(P)$. Sentrale begrep for å beskrive dette er bias og varians. Den forventningsrette estimatoren med lavest varians omtales ofte som den *effektive* estimatoren, men det er ikke nødvendigvis så godt begrunnet at forventningsrette estimatorer skal ha så priviligert status. I praksis er vi ofte bare interessert i å minimere forventet avvik (RMSE) og at det er en trade-off mellom varians og bias. Vi kan gjøre dette litt mer formelt ved å innføre notasjon

$$\|\hat{\gamma} - \gamma\|^2 = MSE(\hat{\gamma}, \gamma) = E[(\hat{\gamma} - \gamma)^2] \tag{4.7}$$

Kan vise at dette kan dekomponeres i kvadret bias og varians ved å skrive ut uttrykket og eliminere ledd. Men dette følger av at $\gamma \in L_2(\mathbb{I}_\Omega)$. Hvis jeg projekter $\hat{\gamma}$ ned på denne mengden finner jeg $E[\hat{\gamma}]$ som er konstanten som minimerer avstand til $\hat{\gamma}$. Denne kvadrerte avstanded er variansen. Denne konstanten avviker fra parameteren som ligger i samme mengde. Denne avstanden er biasen. Disse to differansene er ortogonale fordi den ene er i $L_2(\mathbb{I}_\Omega)$ og den andre i det ortogonale komplementet til denne mengden. Det følger da fra pythagoras at den samlede kvadrerte avstanden er summen av kvadratene til katenene. TODO: Ta bias variance trade-off her slik at jeg er ferdig med det.

4.4 Punktestimat

Punktestimat er den realiserte verdien av $\hat{\gamma}(\mathbf{z}_d)$ for vårt gitte utvalget. Dette utgjør gjerne vår beste gjetning på egenskap til fordeling P som genererte data. På en annen side har estimatoren en utvalgsfordeling og realiserte verdier vil avvike fra den sanne egenskapen til P som vi vil estimere. Vi kjenner ikke den sanne egenskapen og observerer bare én realisert verdi av estimatoren. Hvor mye konfidens har vi på at dette er et godt mål på

egenskapen? Det vil jo blant annet avhenge av størrelsen på utvalget og hvor mye vi lærer fra hver observasjon. Jeg vil kvantifisere dette slik være like konfident som data tillater; verken mer eller mindre. En god start er å rapportere standardfeil i tillegg, men vi kan også bruke kunnskap om hele fordelingen til estimatoren til å konstruere mengde som vi med gitt sannsynlighet kan påstå at γ ligger innenfor.[7]

4.5 Konfidensmengder

Hvis vi antar at sannsynlighetsfordeling er kjent opp til en ukjent parameter kan vi avgrense en parametrisk klasse $P_\theta \in \mathscr{P}_\theta = \{P_\theta | \theta \in \Theta \subset \mathbb{R}^K\}$. Anta videre at vi har et utvalg $(z_1, ..., z_N) = \mathbf{z}_D$ der z_n er *iid* på $\mathbb{R}$ med fordeling P_θ. En tilfeldig mengde $C(\mathbf{z}_D) \subset \Theta$ er en $1 - \alpha$ konfidensmengde for θ hvis

$$\mathbb{P}\{\theta \in C(\mathbf{z}_D)\} = P_D\{\mathbf{z}_D \in Z_D | \theta \in C(\mathbf{z}_D)\} \geqslant 1 - \alpha \tag{4.8}$$

for alle $\theta \in \Theta$ når $\mathcal{L}(\mathbf{z}_D) = P_\theta^D = \prod_n P_\theta$. Dette er litt generell notasjon, men kan ta et klassisk eksempel for å gjøre det litt mer konkret. Vi kan ofte konstruere en slik mengde analytisk gitt antagelse om at prossess er *iid* og vi kjenner parametrisk klasse. Tar ofte utgangspunkt i en statistikk som har kjent fordeling for ulike verdier av parametre. Anta at $P = N(\mu, \sigma)$, at σ er kjent og at vi vil finne konfidensmengde for μ.

$$\mathcal{L}(\bar{x}_N) = N\left(\mu, \frac{\sigma^2}{N}\right) \tag{4.9}$$

$$\mathcal{L}\left(\sqrt{N}\frac{\bar{x}_N - \mu}{\sigma}\right) = N(0, 1) \tag{4.10}$$

$$\tag{4.11}$$

dette medfører at:

$$\mathbb{P}\{|\ \sqrt{N}\frac{\bar{x}_N - \mu}{\sigma}\ | \leqslant z_{\alpha/2}\} \quad \text{når} \quad z_{\alpha/2} := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \tag{4.12}$$

som kan brukes til å vise at

$$C(\mathbf{x}) = (\bar{x}_N - e, \bar{x}_N + e) \tag{4.13}$$

[7] Noe om tolkning av sannsynlighet og konfidensintervall..

for $e = \frac{\sigma}{\sqrt{N}} z_{\alpha/2}$.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \tag{4.14}$$

$$\implies P\left(z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha \tag{4.15}$$

$$\implies P\left(z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X} < -\mu < z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X}\right) = 1 - \alpha \tag{4.16}$$

$$\implies P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \tag{4.17}$$

$$\tag{4.18}$$

Opplegget er ganske greit, men utnyttet at fordeling til punktestimator var kjent. Asymptotisk kan vi få kjent fordeling til punktestimator uten å måtte spesifisere P. Kunne vært en grei øvelse å gjøre det litt formelt, men det får bli en annen dag. Det er også et poeng at man kan finne ikke-parametriske mengder. Kan for eksempel ønske å estimere kumulativ fordeling. Vet da at empirisk kumulativ fordeling er en naturlig estimator, men hvordan kvantifisere usikkerhet til denne? Det er visst mulig.

4.6 Hypotesetester

Vi tenker at vi observerer N realiseringer fra $P \in \{P_\theta : \theta \in \Theta\}$. Vi vil undersøke om data gir sterkt nok bevis for at vi kan forkaste hypotesen om at den sanne parameteren er i $\Theta_0 \subset \Theta$. Hvis Θ_0 er ett enkelt element (singleton) er den en enkel hypotese. Hvis det ugjør en mengde er den en sammensatt hypotese. Det kan entent være fordi vi har en-sidet test eller fordi det er flere parametre og vi kun vil teste avgrensing på én av de.

En naiv fremgangsmåte vil være å bruke en punktestimator $\hat{\theta}$ og forkaste hypotesen dersom $\hat{\theta} \notin \Theta_0$. Problemet er at $\hat{\theta} \neq \theta$. Det er en tilfeldig variabel som varierer mellom ulike utvalg, slik at den kan ta andre verdier selv om hypotesen er sann. Vi trenger en buffer slik at realisert verdi av estimator må være veldig usannsynlig dersom data faktistisk var generert av en $P \in \{P_\theta : \theta \in \Theta_0\}$. Vi har et godt utviklet rammeverk som formaliserer denne intuisjonen. La oss innføre litt begreper og notasjon:

- En *test* er en funksjon $\psi : \mathbf{z}_D \mapsto \{0, 1\}$ der verdi 1 medfører at hypotesen forkastes.
- Sannsynligheten for at testen forkaster avhenger av den sanne parameteren i fordelingen. Dette er beskrevet med styrkefunksjonen $\beta_\psi(\theta) = \mathbb{P}_\theta\{\psi(\mathbf{z}_D) = 1\}$.
- Ideelt sett vil vi ha $\beta_\psi(\theta) = 0$ for $\theta \in \theta_0$ og at sannsynligheten er 1 ellers. Fordi vi med endelig data ikke kan estimere parameteren helt presist vil ikke dette være mulig. Det er da to typer feil vi kan gjøre.

1. Type I: Forkast hypotesen selv om $\theta \in \Theta_0$.
2. Type II: Ikke forkast hypotesen selv om $\theta \notin \Theta_0$.

- Vi er mest redd for å gjøre type I feil. Hypotesen representerer status quo og bevisbyrden ligger på de som vil forkaste den. Vi vil derfor utforme testen slik at det er lite sannsynlig å gjøre den feilen. Merk at det er en tradeoff her: ved å gjøre vanskeligere å forkaste vil vi øke sannsynlighet for å beholde selv om feil.
- Nivået *(size)* til testen er den største sannsynligheten for å gjøre type I feil, $\alpha_\psi(\theta) = \sup_{\theta \in \Theta_0} \beta_\psi(\theta)$
- I praksis konstruerer vi tester med tre steg:
  1. Velger nivå (α). I praksis 0.01 eller 0.05, avhenger litt av konsekvens hvis man tar feil.
  2. Finner en såkalt testobservator T med kjent fordeling gitt θ.
  3. Finner kritisk verdi c slik at $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\{T(\mathbf{z}_D) > c\} = \alpha$
- Dette gir en test som er beskrevet med (T, c) og der $\psi(\mathbf{z}_D) = \mathbb{I}\{T(\mathbf{z}_D) > c\}$

Fremgangsmåten som ble beskrevet over er litt *old school* og mer er relevant for klassiske eksperiment i naturvitenskap enn det jeg holder på med i praksis. I stedet for å formulere en test apriori er det mulig å ta utgangspunkt i den realiserte verdien av testobservatoren og se med hvilket nivå α det ville vært mulig å forkaste hypotesen. Formelt kan vi definere *p-verdi* til den (implisitte) testen som

$$p(\mathbf{z}_D) := \quad \text{nivået } \alpha \text{ slik at } c(\alpha) = T(\mathbf{z}_D) \tag{4.19}$$

Altså hva nivået på testen var dersom vi lot den krititiske verdien være den realiserte verdien slik at hypotesen akkurat blir forkastet. Dette tilsvarer sannsynlighet for å se en observasjon som er minst like ekstremgitt at hypotesen er sann.

4.6.1 Eksempler

Anta at vi har $X_1, ..., X_N$ som er *iid* fra $\mathcal{L}(X) = N(\mu, \sigma)$ med kjent σ. Jeg vil teste om vi kan forkaste $H_0 : \mu < 0$. For å konstruere testen må jeg ta utgangspunkt i en størrelse

som sier noe om μ gitt observerte data. Jeg bruker $\bar{X}_N$.

$$\psi(X_1, ..., X_N) = \mathbb{I}\{\bar{X}_N > c\} \tag{4.20}$$

$$\beta_\psi(\mu) = \mathbb{P}_\mu\left(\bar{X}_N > c\right) \tag{4.21}$$

$$= \mathbb{P}_\mu\left(\sqrt{N}\frac{\bar{X}_N - \mu}{\sigma} > \frac{\sqrt{N}(c-\mu)}{\sigma}\right) \tag{4.22}$$

$$= 1 - \Phi\left(\frac{\sqrt{N}(c-\mu)}{\sigma}\right) \tag{4.23}$$

Ser litt indirekte at styrken øker når μ øker og reduseres når c øker. Valg av c avhenger av nivået til testen

$$\alpha_\psi = \sup_{\theta\in\Theta_0} \beta_\psi(\mu) = \beta_\psi(0) = 1 - \Phi\left(\frac{\sqrt{N}c}{\sigma}\right) \tag{4.24}$$

$$\implies c = \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{N}} \tag{4.25}$$

Dette medfører at vi forkaster når

$$\bar{X}_N > \frac{\sigma\Phi^{-1}(1-\alpha)}{\sqrt{N}} \iff \frac{\sqrt{N}\bar{X}_N}{\sigma} > \sigma\Phi^{-1}(1-\alpha) := z_\alpha \tag{4.26}$$

I praksis bruker vi i stedet en standardnormalfordelt testobservator, men det er jo poeng at vi kan oversette til kritisk verdi av fordelingen til $\bar{X}$ for å få det i samme måleenhet. Merk koblingen til konfidensmengder: vi forkaster dersom μ_0 ikke er i $1-\alpha$ konfidensmengde rundt $\hat{\mu}$.

Målefeil

Anta at vi er interessert i x, men observerer $\tilde{x} = x + u$ der $u \sim N(0, \sigma^2)$. Det kan for eksempel være promilletest der vi vil sjekke om noen har $x > 0.8$. Vi vil ikke straffe noen som er uskyldig, så vil at maks 5% sannsynlig at straff dersom $x \leq 0.8$. Hvor må vi da sette grense for observert $\tilde{x}$? Vel, vi tar utgangspunkt i $x = 0.8$ og finner $P\left(\frac{\tilde{x}-0.8}{\sigma} < c\right) = 1-\alpha = 0.95$ og finner $c = z_\alpha$ og kritisk målt promille som $\tilde{x}' = z_\alpha\sigma + 0.8$.

Kapittel 5

Momentestimatorer

5.1 Utvalgsanalogprinsippet

Den empiriske sannsynlighetsfordelingen gir tyngde $\frac{1}{N}$ til hver av observasjonene i utvalget, $\hat{P}_N(B) \equiv \frac{1}{N}\sum \mathbb{I}\{\mathbf{z_n} \in B\}$. Det er et sentralt resultat at dersom observasjonene er *iid* vil $\hat{P}_N(B) \xrightarrow{p} P(B)$. Dette følger av store talls lov. La $h(\mathbf{z}_n) \equiv \mathbb{I}\{\mathbf{z_n} \in B\}$ og merk at forventningsverdi til en indikatorfunksjon tilsvarer sannsynligheten.

$$\frac{1}{N}\sum h(\mathbf{z}_n) \xrightarrow{p} \mathbb{E}[h(\mathbf{z})] \tag{5.1}$$

$$\implies \frac{1}{N}\sum \mathbb{I}\{\mathbf{z}_n \in B\} \xrightarrow{p} P(B) \tag{5.2}$$

For spesialtilfelle der $B = (-\infty, s]$ så følger det at $\hat{F}_N \xrightarrow{p} F$, der $\hat{F}_N(s) = \frac{1}{N}\sum \mathbb{I}\{X_n < s\}$. Dette motiverer utvalgsanalogprinsippet. I mange tilfeller kan vi finne gode estimatorer ved å evaluere γ på den empiriske sannsynlighetsfordelingen, $\hat{\gamma} = \gamma(\hat{P}_N)$.

$$\gamma(P) = \mathbb{E}_P(x) \tag{5.3}$$

$$\gamma(\hat{P}_N) = \mathbb{E}_{\hat{P}_N}(x) = \frac{1}{N}\sum x_n \equiv \bar{x} \tag{5.4}$$

For å gjøre dette litt mer operativt kan vi merke at egenskaper ofte er funksjon av kumulativ fordeling, $\gamma = \gamma(F)$. Vet ikke hvor vesentlig det poenget var. Uansett, dette er en fleksibel ikke-parametrisk fremgangsmåte der vi erstatter F med empirisk CDF $\hat{F}_N$ og egenskapene vi måtte være interessert i. Lurer på om vi har noen generelle fremgangsmåter til å kvantifisere usikkerhet til disse estimatene uten å pålegge mer struktur..

5.1.1 Motivere OLS som utvalgsanalog

Vi kan nå bruke dette rammeverket til å studere relasjonen mellom inputvektor $\mathbf{x}$ og avhengig variabel y. Den betingede forventningen $\mathbb{E}[y|\mathbf{x}]$ gir et godt sammendragsmål

på relasjonen, men den kan være vanskelig å estimere og å tolke. I praksis estimerer vi ofte den lineære populasjonsregresjonsfunksjonen, $\gamma(P) = \mathbf{b}^*$. Hvorvidt dette gir en god tilnærming avhenger om CEF er tilnærmet lineær. Det er likevel ganske fleksibelt siden vi kan transformere inputvektor til et *feature space*, $\Phi : \mathbb{R}^K \to \mathbb{R}^L$. Da kan vi ofte få tilnærmet lineær relasjon i forhold til den transformerte inputvektoren, $\mathbb{E}[y|\Phi(\mathbf{x})]$. Kan eventuelt også transformere y.

Dersom $\mathbb{E}(\mathbf{x}(y - \mathbf{x}'\mathbf{b}^*) = \mathbf{0}$, observasjonene er *iid* og $\mathbb{E}(\mathbf{x}\mathbf{x}')$ er inverterbar gir utvalgsanalogsprinsippet en konsistent estimator for $\mathbf{b}^* \equiv \beta$

$$\widehat{\beta} = \mathbb{E}_{\hat{P}_N}(\mathbf{x}\mathbf{x}')^{-1}\mathbb{E}_{\hat{P}_N}(\mathbf{x}y) = \left(\frac{1}{N}\sum \mathbf{x_n}\mathbf{x_n'}\right)^{-1}\frac{1}{N}\sum \mathbf{x_n}y_n \tag{5.5}$$

Utvalgsanalogprinsippet er intuitivt. Det er naturlig å bruke den relative andelen av observasjoner som havner i en mengde som estimat på sannsynlighet for den mengden i populasjonen. Asymptotisk kan vi da observere P og lære egenskaper ved prosessen uten å måtte gjøre antagelser. Vi kan la data snakke for seg selv. Hvorfor trenger vi andre måter å utlede estimatorer? For det første har vi aldri uendelig store utvalg. Hele poenget er å generalisere fra utvalg og da må vi håndtere det faktum at $\hat{P}_N \neq P$. For det første er $\hat{P}_N$ alltid diskret selv om fordelingen er absolutt kontinuerlig. For det andre kan vi få estimatorer med bedre egenskaper ved å påføre struktur a priori. Dette motiverer empirisk risikominimering som gir oss et rammeverk for å kombinere utvalgsanlog med struktur.

5.2 Momentestimator

Den enkleste momentestimatoren følger direkte fra utvalgsanalogprinsippet

$$\theta = \mathbb{E}[X] = \int xf(x)dx \tag{5.6}$$

$$\hat{\theta} = \mathbb{E}_{\hat{P}_N}[X] = \sum x_n f_N(x_n) = \frac{1}{N}\sum x_n \tag{5.7}$$

Mer generelt kan anta parametrisk form $f(x;\theta)$ der $\theta = g(\mathbb{E}X)$. Fremgangsmåten er da å finne moment og løse ligningen med hensyn på parameter for å finne estimator. Eksempel: $X \sim geo(p)$, der $p = 1/\mathbb{E}X$.

$$g(p) = \mathbb{E}[X] = \int xf(x)dx \tag{5.8}$$

$$g(\hat{p}) = \mathbb{E}_{\hat{P}_N}[X] = \sum x_n f_N(x_n) = \frac{1}{N}\sum x_n \tag{5.9}$$

$$\hat{p} = \frac{N}{\sum x_n} \tag{5.10}$$

Kan utvide til å estimere flere parametre som da gir oss et ligningssystem. Jeg er litt usikker på hvilken notasjon jeg ønsker å bruke.

5.2.1 Egenskaper

Gitt regularitetsbetingelser er estimatorene konsistente og asymptotisk normale med varians som det er mulig å beregne.

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} N(0, Avar(\hat{\theta})) \tag{5.11}$$

der

$$Avar(\hat{\theta}) = g\mathbb{E}[xx']g' \tag{5.12}$$

tror det generelt har en sandwich form og at g består av derivater som kan gis enkel form dersom momentbetingelse er lineær, men ser på dette senere når jeg får notasjon på plass.

5.3 GMM

TODO: motivere GMM. Tror jeg vil motivere momentestimatorer i samme slengen. GMM er utvidelse av momentestimator til overdeterminerte ligningsystem der vi har flere instrument enn endogene variabler. Fordelen med å inkludere flere instrument er at vi få mer effektiv estimator (lavere asymptotisk varians) og kan bruke overidentifikasjon for å teste gyldighet til instrument siden vi kan observere u med $L-1$ instrument. Det er mye notasjon, så det er viktig å være ryddig. Tenker vi har en prosess som genererer observasjoner til utvalget vårt, $\{y_n, \mathbf{x_n}, \mathbf{z_n}\} = \{\mathbf{w_n}\}$ som er ergodisk og stasjonær. Jeg har en momentbetingelse som definererer verdien på parameteren jeg vil finne. I praksis er det ofte ortogonalitetsbetingelse. Kan være greit å tenke på hvordan jeg kan relatere dette til L_2, men annen gang. Parameter sier noe om relasjon mellom variabler i prosessen

$$g_n(\delta) = g(\mathbf{z_n} u_n(\delta)) = g(\mathbf{z_n}(y - \mathbf{x_n}'\delta) \tag{5.13}$$

$$\mathbb{E}[g_n(\delta)] = \mathbf{0} \tag{5.14}$$

der vi kan tenke på $g_n(\delta)$ som en tilfeldig variabel. Det korresponderene empiriske momentet er

$$\mathbb{E}_{P_N}[\mathbf{g}_n(\tilde{\delta})] = \frac{1}{N}\sum \mathbf{g}_n(\tilde{\delta}) \equiv \mathbf{g}_N(\tilde{\delta}) \tag{5.15}$$

$$= \frac{1}{N}\sum \mathbf{z_n}(\mathbf{y_n} - \mathbf{x_n}'\tilde{\delta}) \equiv \mathbf{S}_{zy} - \mathbf{S}_{zx}\tilde{\delta} \tag{5.16}$$

Hvis eksakt identifisert kan jeg løse dette for $\tilde{\delta}$ som da blir min estimator $\hat{\delta}$. Hvis overidentifisert er ikke $\mathbf{S}_{zx}$ inverterbar. Det er ikke mulig å få alle utvalgsmomentene lik 0. En naturlig løsning er å minimere det samlede avviket fra 0, altså minimerere lengden av vektoren $g_N(\tilde{\delta})$. Merk at $\|\mathbf{x}\|^2$ er $\mathbf{x}'\mathbf{x}$. Men vi få lavere asymptotisk varians ved å vekte momentene, slik at moment med lavere varians får høyere vekt. Litt analogt til vektet minste kvadrat. Uansett, i stedet for å minimere lengden av vektoren direkte setter vi opp en kvadratisk form

$$J(\tilde{\delta}, \hat{\mathbf{W}}) = \mathbf{g}_N(\tilde{\delta})'\hat{\mathbf{W}}\mathbf{g}_N(\tilde{\delta}) \tag{5.17}$$

der vi kan finne closed form løsning på minimeringsproblemet som gir et eksplisitt uttrykk for GMM-estimatoren.

$$\hat{\delta}_{GMM} = \arg\min_{\tilde{\delta}} J(\tilde{\delta}, \hat{\mathbf{W}}) \tag{5.18}$$

$$= \left(\mathbf{S}_{zx}\hat{\mathbf{W}}\mathbf{S}_{zx}\right)^{-1}\mathbf{S}'_{zx}\hat{\mathbf{W}}\mathbf{S}_{zy} \tag{5.19}$$

Har nå funnet estimatorene. Neste steg blir å utlede den asymptotiske fordelingen. Fremgangsmåten er å substituere inn for y og bruke dette til å få utrykk for utvalgsfeilen $\hat{\delta} - \delta$.

$$\mathbf{S}_{zy} = \frac{1}{N}\sum \mathbf{z_n}\mathbf{y_n} = \frac{1}{N}\sum \mathbf{z_n}(\mathbf{x_n}'\delta + u_n) \tag{5.20}$$

$$= \mathbf{S}_{zx}\delta + \bar{\mathbf{g}} \tag{5.21}$$

der $\bar{\mathbf{g}} \equiv \frac{1}{N}\sum g_n(\delta) = \frac{1}{N}\sum \mathbf{z_n}u_n$. Det følger da at

$$\hat{\delta} = \delta + \left(\mathbf{S}'_{zx}\hat{\mathbf{W}}\mathbf{S}_{zx}\right)^{-1}\mathbf{S}'_{zx}\hat{\mathbf{W}}\bar{\mathbf{g}} \tag{5.22}$$

slik at $\hat{\delta} - \delta = \mathbf{A}_N\bar{\mathbf{g}}$. Den er da konsistent hvis $\mathbf{A}_N \xrightarrow{p} \mathbf{A}$ og $\bar{\mathbf{g}} \xrightarrow{p} \mathbb{E}[\mathbf{g}_n(\delta)] = \mathbf{0}$. Den asymptotiske fordelingen er da

$$\sqrt{N}(\hat{\delta} - \delta) = \mathbf{A}_N\sqrt{N}\bar{\mathbf{g}} \xrightarrow{d} N(\mathbf{0}, \mathbf{ASA}') \tag{5.23}$$

hvis $\{\mathbf{g_n}\}$ er *mds* slik at at $\sqrt{N}\bar{\mathbf{g}} \xrightarrow{d} N(\mathbf{0}, \mathbf{S})$ og der $\mathbf{S} = var(\mathbf{g_n}) = \mathbb{E}[\mathbf{g_n}\mathbf{g_n}']$. Dette ser ganske ryddig ut, men $\mathbf{A}$ skjuler masse dritt.

$$\mathbf{A} = \left(\mathbf{\Sigma}'_{ZX}\mathbf{W}\mathbf{\Sigma}_{ZX}\right)^{-1}\mathbf{\Sigma}'_{ZX}\mathbf{W} \tag{5.24}$$

Kan få ryddet opp hvis $\hat{\mathbf{W}} \xrightarrow{p} \mathbf{S}^{-1}$ som er den asymptisk effektive estimatoren. Følger da at

$$Avar(\hat{\delta}(\hat{\mathbf{S}}^{-1})) = \left(\boldsymbol{\Sigma}_{ZX}'\mathbf{S}^{-1}\boldsymbol{\Sigma}_{ZX}\right)^{-1}\boldsymbol{\Sigma}_{ZX}'\mathbf{S}^{-1}\mathbf{S}\mathbf{S}^{-1}\boldsymbol{\Sigma}_{ZX}\left(\boldsymbol{\Sigma}_{ZX}'\mathbf{S}^{-1}\boldsymbol{\Sigma}_{ZX}\right)^{-1} \tag{5.25}$$

$$= \left(\boldsymbol{\Sigma}_{ZX}'\mathbf{S}^{-1}\boldsymbol{\Sigma}_{ZX}\right)^{-1} \tag{5.26}$$

Dette kan vi finne dersom vi finne dersom vi har konsistent estimator for $\mathbf{S}$

$$\mathbf{S} = \mathbb{E}[\mathbf{g}_n\mathbf{g}_n'] = \mathbb{E}[\mathbf{z}_n u_n(\mathbf{z}_n u_n)'] \tag{5.27}$$

$$\hat{\mathbf{S}} = \frac{1}{N}\sum \hat{u}_n^2\mathbf{z}_n\mathbf{z}_n' \tag{5.28}$$

Det er ikke helt opplagt hvorfor denne estimatoren fungerer, men tror dette er white sin hetero-robuste greie som jeg ser på senere. Problemet nå er at vi trenger $\hat{\delta}$ for å finne $\hat{u}_n$ fordi

$$\hat{u}_n = y_n - \mathbf{x}_n'\hat{\delta} \tag{5.29}$$

men vi trenger vektematrise for å finne $\hat{\delta}$! Kan vår *catch 22* med to-stegs estimator.

1. Velger en default vektematrise, som oftest $\hat{\mathbf{W}} = \mathbf{I}$ eller $\hat{\mathbf{W}} = \mathbf{S}_{zz}^{-1}$. Bruker dette til å finne $\delta(\hat{\mathbf{W}})$. Bruker dette til å finne $\hat{\mathbf{S}}$.
2. Bruker dette til å finne $\delta(\hat{\mathbf{S}}^{-1})$

Tror jeg alternativt jeg kunne brukt en algoritme som gjentar prosess til konvergens. Uansett, jeg har nå et veldig fleksibelt rammeverk som lar meg utlede asymptotisk effektive estimatorer for en stor mengde av DGPer. Hvorfor sitter ikke alle og kjører GMM? I praksis har vi ofte upresis estimering av $\mathbf{S}^{-1}$ slik at det blir lite gevinst i forhold til 2sls. Kan påføre litt ekstra antagelser og utlede de vanlig estimatorerene som special case av asymptotisk effektive GMM og slipper da 2-stegs opplegget over.

5.3.1 2SLS

Innfører antagelse om betinget homoskedastisitet

$$\mathbb{E}[u_n^2|\mathbf{z}_n] = \sigma^2 \tag{5.30}$$

Det følger da at

$$\mathbf{S} = \mathbb{E}\left[\mathbb{E}(\mathbf{z}_n u^2\mathbf{z}_n'|\mathbf{z}_n)\right] = \sigma^2\mathbb{E}[\mathbf{z}_n\mathbf{z}_n'] \tag{5.31}$$

$$\hat{\mathbf{S}} = \hat{\sigma}^2\frac{1}{N}\sum \mathbf{z}_n\mathbf{z}_n' = \hat{\sigma}^2\mathbf{S}_{zz} \tag{5.32}$$

Slenger dette inn i GMM-estimatoren og får 2SLS

$$\hat{\delta}(\hat{\mathbf{S}}^{-1}) = \left(\mathbf{S}_{zx}(\hat{\sigma}^2\mathbf{S}_{zz})^{-1}\mathbf{S}_{zx}\right)^{-1}\mathbf{S}'_{zx}(\hat{\sigma}^2\mathbf{S}_{zz})^{-1}\mathbf{S}_{zy} \tag{5.33}$$

$$= \left(\mathbf{S}_{zx}(\mathbf{S}_{zz}^{-1}\mathbf{S}_{zx}\right)^{-1}\mathbf{S}'_{zx}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zy} \tag{5.34}$$

$$= \hat{\delta}(\hat{\mathbf{S}}_{zz}^{-1}) = \hat{\delta}_{2SLS} \tag{5.35}$$

Det at estimatoren er konsistent og asymptotisk normalfordelt følger av at det er special case av GMM. Kan finne asymptotisk varians med homoskedastisitet

$$Avar(\hat{\delta}_{2SLS}) = \left(\mathbf{\Sigma}'_{zx}(\sigma^2\mathbf{S}_{zz})^{-1}\mathbf{\Sigma}_{zx}\right)^{-1} \tag{5.36}$$

$$\widehat{Avar(\hat{\delta}_{2SLS})} = \hat{S}^2\left(\mathbf{\Sigma}'_{zx}\mathbf{S}_{zz}^{-1}\mathbf{S}_{zx}\right)^{-1} \tag{5.37}$$

Kan jo også utvide til robuste standardfeil her. Kan bruke 2SLS (og OLS) selv om antagelse om homoskedastisitet ikke er oppfylt, men da må vi leve med at estimatorene ikke er asymptotisk effektive. For spesialtilfelle med eksakt identifisering er det tilstrekkelig å bruke at $\mathbf{\Sigma}_{zx}$ er inverterbar til å utlede IV og OLS fra GMM.

5.3.2 IV

5.3.3 OLS

5.4 Utvidelser

Vi trenger bare momentbetingelsene for å konsistent estimere helningskoeffisientene og de er asymptotisk normalfordelte fra CLT. I utgangspunktet trenger vi ikke bry oss så mye om fordelingen til $\mathbf{u} := [u_1, u_2, ..., u_N]'$ dersom vi kun vil ha mål sentraltendens i betingede fordelinger. Når vi estimerer med MLE antar vi gjerne at $\mathbb{V}[u|x] = \sigma^2$ og at observasjon er iid slik at $\mathbb{V}[\mathbf{u}|X] = \sigma^2 I$. Hvis det er heteroskedastisitet så blir dette målet på standardfeilen ikke riktig. I praksis har det ikke så mye å si og jeg skal vise at vi kan finne en mer generell formel for standardfeilen som ikke avhenger av den antagelsen. Videre kan vi også finne alternativ estimator som utnytter at noen observasjoner er mer informative om verdien til β for de feilleddet til observasjonen har mindre varians. Mer generelt så kan det være ønskelig å transformere variablene slik at de oppfyller $G - M$ antagelser og vi får mer effektiv estimering... tror ikke dette er så veldig relevant i praksis, men jeg tar det litt raskt.

5.4.1 Generalisert minste kvadrat

Vi kan skrive $\mathbb{V}[\mathbf{u}|X] = \sigma^2\psi$. Vi har altså en parameter som er skalert med en matrise som kan avhenge av X. Jeg vil transformere variabelen $\mathbf{u}$ slik at skaleringsfaktoren reduserer

til I. Merk først hvordan vi kan gå fram for å standardisere i én dimensjon,

$$\mathbb{V}[u|x] = a\sigma^2 \tag{5.38}$$

$$\implies \mathbb{V}[\frac{u}{\sqrt{a}}|x] = \frac{1}{a}\mathbb{V}[u|x] = \sigma^2 \tag{5.39}$$

Jeg bare skalerer variabelen med den inverse av kvadratoren av skaleringsfaktoren i uttrykket for variansen. Kan greit generalisere dette ved å finne A slik at $\psi^{-1} = A'A$.[1] Kan da transformere modellen

$$A\mathbf{y} = A[X\beta + \mathbf{u}] \tag{5.40}$$

$$\mathbf{y}^* = X^*\beta + \mathbf{u}^* \tag{5.41}$$

og observere at

$$\mathbb{V}[Au|X] = A\mathbb{V}[u|X]A' \tag{5.42}$$

$$= A\sigma^2\psi A' \tag{5.43}$$

$$= \sigma^2 A(A'A)^{-1}A' \qquad = \sigma^2 AA^{-1}(A')^{-1}A' = \sigma^2 I \tag{5.44}$$

Vektet minste kvadrat

Hvis vi utelukker seriekorrelasjon er ψ en diagonalmatrise. I mange tilfeller er det rimelig at varians til feilledd[2] avhenger av de observerte variablene. For eksempel er spredningen til mange variabler større for menn enn for kvinner. Det kan også være slik at størrelse avhenger av kontinuerlig variabel (mer variasjon i forbruk for rikinger). Vi kan generelt modellere dette som

$$\mathbb{V}[u|X] = \sigma^2\psi = \sigma^2 diag(h_n)^2 \tag{5.45}$$

der $h_n^2 := h(x_n)$. Tror det vil være en eller annen deterministisk funksjon av variablene.. Uansett, bare skalrerer alle observasjoner med kvadratet av den inverse for å redusere ψ til I,

$$\frac{y_n}{h_n} = \frac{x_n}{h_n}\beta + \frac{u_n}{h_n} \tag{5.46}$$

I praksis så er det stor utfordring at h_n må estimerers. Hvis vi ikke påfører mer struktur så blir det like mange parametre som observasjoner. Litt usikker på hvordan jeg går frem i praksis. Et enkelt eksempel er grupperte observasjoner det jeg kun ser gjennomsnitt i

[1] Dette kan f.eks. gjøres med cholesky dekomponering. Eksisterer alltid fordi ψ er positiv semi definit (analog til positiv skalar), men er ikke unik.

[2] Siden feilledd kan tolkes som avvik fra sentraltendens i betinget fordeling så tilsvarer det varians i betinget fordeling.

gruppen, men antar at $u \sim IID(0, \sigma^2)$. Da vil $u_i := \frac{1}{N(i)} \sum u_i$ og $h_i^2 = \frac{\sigma}{N}$. Legger med vekt på grupper med flere observasjoner.

Feasible generalisert minste kvadrat

5.4.2 Robust estimering

Kapittel 6

Maximum likelihood

Vi ønsker å lære om en fordeling P med utgangspunkt i realiserte verdier fra fordelingen. Dette er det generelle utgangspunktet i statistisk inferens. I likelihood tilnærmingen gjør vi sterke antagelser ved å anta at fordelingen tilhører en parametrisk klasse $P := P_{\theta_0} \in \mathscr{P} = \{f(\cdot;\theta) : \theta \in \Theta \subset \mathbb{R}^k\}$ der $f(\cdot;\cdot)$ har kjent form.[1] Med andre ord antar vi at fordelingen er kjent opp til en ukjent parameter og at denne parameteren fullt ut beskriver hele fordelingen som genererer data vi observerer. Funksjonen $f(\cdot;\theta)$ evaluert i en gitt verdi $\theta = \theta'$ er en sannsynlighetstetthetsfunksjon.[2] Hvis vi derimot holder den realiserte verdien konstant og betrakter det som en funksjon av θ kan vi betegne det som likelihoodfunksjonen, $f(\cdot;z) := L(\cdot;z) := L(\cdot)$. Vi kan betraktre z enten som en gitt realisert verdi eller som en tilfeldig variabel $z := z(\omega)$ som tar en konstant verdi når tilstanden ω blir avslørt. I den første tolkningen har likelihoodfunksjonen en ganske konkret tolkning som relativ sannsynlighet for at en fordeling med de ulike parameterverdiene kan ha generert den observerte verdien z.[3] Hvis vi derimot betrakter situasjonen situasjonen før verdi er realisert er likelihoodfunksjonen utfallet av $g : \Omega \to S = \{L(\cdot;z) : z^{-1}(\omega) \in \Omega\}$. Utfallet er en funksjon! Men hvis vi bare evalurer dette i gitte verdier av θ blir det en tilfeldig variabel med fordeling som vi kan beskrive og analysere på vanlig måte.

Denne tilnærmingen krever sterke antagelser, men har til gjengeld en veldig rik teori. Estimatoren vil også kunne ha gode egenskaper selv om modellen er feilspesifisert.

[1] Denne mengden burde kanskje bestått av fordelinger, ikke tetthetsfunksjoner. Men så lenge modell er identifiserbar eksisterer det injektive funksjoner $\theta \mapsto P_\theta$ og $P_\theta \mapsto f_\theta$ slik at de ulike representasjonene av fordelingen er kjent når θ er kjent.

[2] Eller en pmf. Distinksjonen er ikke viktig og hvis jeg kunne litt measure theory tror jeg fremstillingen kunne blitt gjort mer ryddig.

[3] Vet ikke om tolkningen er helt presis. Merk at det ikke er en sannsynlighetsfordeling siden det ikke oppfyller aksiom. Siden den kun betegner relativ sannsynlighet er den bare unik opp til en multiplikativ konstant siden slike skaleringer inneholder samme informasjon. Det er praktisk siden det medfører at likelihoodfunksjonen er invariant til valg av måleenhet på observasjonen.

6.1 Begreper

Likelihoodfunksjonen for en gitt observasjon har en konkret tolkning. Formen på funksjonen beskriver i hvilken grad ulike parameterverdier korresponderer med den realiserte verdien vi observerer. Det er uansett begrenset hvor vi kan lære fra én enkelt realisering. Heldigvis er det enkelt å kombinere informasjon fra ulike realiseringer fra samme fordeling så lenge disse er uavhengige. Anta at vi obseverer $(z_1, ..., z_n)$ der $\mathcal{L}(z_n) = P_{\theta_0}$ for $n = 1, ..., N$. Vi kan da kalle likelihoodfunksjonen for hver av observasjonene, $L_n(\cdot; z_n)$ for *likelihood contribution* til observasjon n. Vi kan kombinere informasjonen ved å betrakte hele utvalget som én realisering fra simultanfordelingen $\mathcal{L}(z_1, ..., Z_n) = \pi_n \mathcal{L}(z_n)$ slik at $L(\cdot; z_1, ...z_n) = \pi_n L_n(\cdot; z_n)$. Den samlede likelihoodfunksjonen er da

$$L : \Theta \to [0, \infty) \tag{6.1}$$

$$: \theta \mapsto \Pi f(z_n; \theta) = \Pi f(z_n; \theta) \tag{6.2}$$

Likelihoodfunksjonen er riktignok ikke unik; alle skaleringer med positiv konstant inneholder akkurat like mye informasjon om θ siden vi kun kan vurdere relativ sannsynlighet for ulike parameterverdier gitt observert utvalg. Dette har litt sammenheng med at vi ønsker at likelihood skal være invariant for én-til-én transformasjoner av data, for eksempel valg av måleenhet. La $y = g(x)$ og $x = g^{-1}(y) := x(y)$. Da er

$$F_Y(y) = P(Y < y) = P(g(X) < y) = P(X < x(y)) = F_x(x(y)) \tag{6.3}$$

$$f_Y(y) = \frac{\partial}{\partial y} F_x(x(y)) = f_x(x(y)) |\frac{dx}{dy}| \tag{6.4}$$

Det følger derfor at $L(\theta|x) = f_X(\theta; x)$ og $L(\theta|y) = f_Y(\theta; y) = L(\theta|x)|\frac{dx}{dy}|$. Disse er ulik med positiv skalar, men relativ likelihood evaluert i to ulike verdier θ_1 og θ_2 vil være den samme for begge funksjonene og er dermed begge like gode likelihoodfunksjon for $\mathcal{L}(X)$.

I praksis er det enklere å jobbe med logaritmen av likelihood når vi kombinerer informasjon fra ulike kilder siden

$$\log L(\cdot; z_1, ...z_n) = \log[\pi_n L_n(\cdot; z_n)] \tag{6.5}$$

$$= \Sigma_n \log[L_n(\cdot; z_n)] \tag{6.6}$$

$$= \Sigma_n log L_n(\cdot; z_n) \tag{6.7}$$

$$\tag{6.8}$$

Logaritmen er en positive monoton transformasjon som bevarer arg max, gjør det enklere å kombinere informasjon fra avhengige observasjoner og optimere numerisk. Funksjonsverdiene har ikke like umiddelbar tolkning som i likelihoodfunksjonen, men vi skal se at de sentrale teoretiske størrelsene er knyttet til denne såkalte *loglikelihood-funksjonen*. Merk

også at verdien for hver θ er summen av N uavhengige realiseringer fra P_{θ_0} slik at hvis vi skalerer det med $1/N$ så vil det konvergere mot[4]

$$\mathbb{E}_{\theta_0}[logL(\theta, z)] := \int_Z logL(\theta, z) f(z; \theta_0) dz \tag{6.9}$$

6.1.1 Score

Helningen[5] til loglikelihood-funksjonen betegnes som dens *score*,

$$S_n(\theta) = \frac{\partial}{\partial\theta} logL_n(\theta) \tag{6.10}$$

$$\implies S(\theta) = \frac{\partial}{\partial\theta} logL(\theta) = \Sigma_n S_n(\theta) \tag{6.11}$$

Hvis vi betrakter denne størrelsen som en funksjon av θ for gitt realisert verdi av z betegnes det som *score-funksjonen* og hvis vi derimot betrakter hvordan verdien for en gitt θ avhenger av tilfeldig z betegnes det som *score-statistic*.

Det gir mening at P_{θ_0} asymptotisk generer et utvalg som korresponderer med θ_0 i betydningen at av alle kanditat-verdier av θ så er det mest sannsynlig at det ble generert fra fordeling med θ_0. Det medfører at forventningsverdien til log-likelihoodfunksjonen er størst når den er evaluert i θ_0 og tilsvarende at forventningsverdi til score-statistikken i θ_0 er 0,

$$\mathbb{E}_{\theta_0} S_n(\theta_0) : = \int S_n(\theta_0) f_{\theta_0}(x) dz \tag{6.12}$$

$$= \int \frac{\partial}{\partial\theta} logL_n(\theta_0) f_{\theta_0}(x) dz \tag{6.13}$$

$$= \int \frac{\frac{\partial}{\partial\theta} L_n(\theta_0)}{L_n(\theta_0} f_{\theta_0}(x) dz \tag{6.14}$$

$$= \int \frac{\partial}{\partial\theta} L_n(\theta_0) dz \tag{6.15}$$

$$= \frac{\partial}{\partial\theta} \int L_n(\theta_0) dz = 0 \tag{6.16}$$

der vi har brukt at $L(\theta_0) := L(\theta_0; z) := f_{\theta_0}(z) := f(z; \theta_0)$. Utvalgsanalogen til dette er å velge

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta} \mathbb{E}_{\hat{P}_N}[S_n(\theta)] \tag{6.17}$$

$$= \arg\max_{\theta\in\Theta} \frac{1}{N} \sum S_n(\theta, z_n) = 0 \tag{6.18}$$

som vi i enkle eksempler kan finne en *closed form* løsning på slik at vi får et eksplisitt ut-

[4]Det er direkte resultat av store talls lov: $\mathbb{E}_{\hat{P}_N}[g(z)] \xrightarrow{p} \mathbb{E}[g(z)]$.

[5]Skal generalisere til parametervektor med flere dimensjoner senere. Da vil det være gradienten.

trykk $\hat{\theta}_{MLE} = g(z_1, ..., z_N)$ og som ofte vil tilsvare momentestimatoren.[6] Vi vil maksimere forventen log-likelihood, men vi observerer det ikke så vi maksimerer i stedet gjennomsnitt loglikelihood fra observasjonen generert av sann fordeling og lener oss på at størrelsene konvergerer asymptotisk.

6.1.2 Informasjon

Hvor mye lærer vi om θ_0 fra å observere én realisering fra P_{θ_0}? Som nevnt lener vi oss på at $\mathbb{E}_{\theta_0} S_n(\theta_0) = 0$ og at gjennomsnittet i mitt utvalg bestående av N *iid* observasjoner konverger mot denne sentraltendensen. Men jeg har et begrenset antall observasjoner og vil derfor være interessert i mål på spredningen til den tilfeldige variabelen $S_n(\theta_0)$. Dette angir den teoretiske *fisher-informasjonen* til den ukjente fordelingen P_{θ_0},

$$
\begin{aligned}
I(\theta) &:= \mathbb{V}_{\theta_0}[S(\theta_0)] && (6.19)\\
&= \mathbb{E}_{\theta_0}[S(\theta_0)^2] && (6.20)\\
&:= \int_Z S(\theta_0)^2 f_{\theta_0}(z) dz && (6.21)\\
&:= \int_Z S(\theta_0, z)^2 f(z; \theta_0) dz && (6.22)\\
& && (6.23)
\end{aligned}
$$

Denne spredningen avhenger av hvor spiss toppen til $\mathbb{E}_{\theta_0}[logL(\theta)]$ er i $\theta = \theta_0$. For at det skal være mye informasjon om θ i hver observasjon av z vil vi at den skal ha en spiss topp i θ_0. Det viser seg at vi kan bruke denne intuisjonen til å finne en alternativ utledning av fisher-informasjonen. Generelt vil hesse-matrisen beskrive krumming til score-funksjon,

$$
H(\theta) = \frac{\partial}{\partial \theta} S(\theta) = \frac{\partial^2}{\partial \theta \partial \theta'} logL(\theta). \tag{6.24}
$$

Den teoretiske fisher-info tilsvarer den negative verdien av den forventede hessematrisen evaluert i θ_0,

$$
\begin{aligned}
I(\theta_0) &= -\mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial \theta} S(\theta_0)\right] && (6.25)\\
&= -\mathbb{E}_{\theta_0}\left[H(\theta_0)\right] && (6.26)\\
&= -\int_Z H(\theta_0, z) f(z; \theta_0) dz && (6.27)
\end{aligned}
$$

Mer informasjon gir bedre presisjon av $\hat{\theta}_{MLE}$. Noe kontra-intuitivt er det derfor bedre med høyere varians til $S(\theta_0)$. Et veldig sentralt resultat, som jeg forhåpentligvis kommer

[6] Kanskje finne noe måte å koble de sammen.

tilbake til senere, er at for alle MLE-estimatorer vil

$$\hat{\theta}_{MLE} - \theta_0 \xrightarrow{d} N\left(0, I(\theta)^{-1}\right) \tag{6.28}$$

Så langt har jeg betraktet fisher-informasjon som en ren teoretisk størrelse ved P_{θ_0}, men informasjon i utvalget avhenger i tillegg av antall observerte verdier fra fordelingen. For et utvalg tar jeg utgangspunkt i den samlede log-likelihoodfunksjonen som er sum av log-likelihood-contribution fra $n = 1, ..., N$. Det medfører at utvalgsstørrelse blir bakt inn i fisher-informasjonen slik at det ikke er en eksplisitt N med i uttrykket.

6.1.3 Alternativ utledning

Log-likelihood contribution for gitt parameterverdi, $logL_n(\theta; z_n)$, er en tilfeldig variabel med forventningsverdi $\mathbb{E}_{\theta_0}[logL_n(\theta)] := \int_Z \log[f(z;\theta)]f(z;\theta_0)dz$ som er en skalar. For å understreke at $\mathbb{E}_{\theta_0}[logL_n(\cdot)]$ er en helt vanlig funksjon som mapper $\mathbb{R}^d \to \mathbb{R}$ vil jeg betegne den som g.[7]. Denne funksjonen g kan utledes fra fordelingen P_{θ_0} og vi kan finne egenskaper ved denne helt streite funksjonen som dermed er egenskaper ved den sanne fordelingen.

- Score: $\frac{\partial}{\partial\theta}g(\theta)|_{\theta_0}$
- Fisher-informasjon: $\frac{\partial^2}{\partial\theta\partial\theta}g(\theta)|_{\theta_0}$

Vi kan ikke observere P_{θ_0} og kjenner dermed ikke g. Men vi kan bruke utvalgsanaloget til å tilnærme oss funksjonen siden $\mathcal{L}(z_n) = P_{\theta_0}$ medfører at

$$\mathbb{E}_{\hat{P}_N}[logL_n(\theta)] := \frac{1}{N}\Sigma_n logL_n(\theta; z_n) \tag{6.29}$$

$$\xrightarrow{p} \mathbb{E}_{\theta_0}[logL_n(\theta)] = g(\theta) \tag{6.30}$$

Vi kan konsistent estimere $g(\cdot)$ med gjennomsnitt i utvalget og bruke det til å beregne egenskaper til funksjonen. Den teoretiske fisher-informasjonen gir oss for eksempel et mål på hvor mye vi lærer om θ_0 fra én observasjon. For å finne den estimerte fisher-informasjonen i utvalget trenger vi bare å skalere gjennomsnittet med N observasjoner. Dette gir oss tilbake hessematrisen fra $logL(\cdot)$ i utvalget.

[7] Den mapper fra parametermengden som generelt er delmengde av $\mathbb{R}^d$.

6.2 Eksempler

6.2.1 Bernoulli

Likelihoodfunksjonen kan skrives kompakt på én linje,

$$L_n(\rho) = \rho_n^x(1-\rho)^{1-x_n} \tag{6.31}$$

$$logL_n(\rho) = x_n \log(\rho) + (1-x_n)\log(1-\rho) \tag{6.32}$$

Vi deriverer med hensyn på parameter og finner score contribution,

$$\frac{\partial}{\partial\rho} logL_n(\rho) := S_n(\rho) = \frac{x_n}{\rho} - \frac{1-x_n}{1-\rho} \tag{6.33}$$

$$= \frac{x_n - \rho}{\rho(1-\rho)} \tag{6.34}$$

Vi kan anta at $\mathcal{L}(x_n) = bernoulli(\rho_0)$. Forventningsverdi til score contribution er da

$$\mathbb{E}_{\rho_0}\left[S_n(\rho)\right] = \mathbb{E}_{\rho_0}\left[\frac{x_n - \rho}{\rho(1-\rho)}\right] \tag{6.35}$$

$$= \frac{\rho_0 - \rho}{\rho(1-\rho)} \tag{6.36}$$

som medfører at $\mathbb{E}_{\rho_0}[S_n(\rho)|_{\rho_0}] = 0$. Vi kan også finne variansen til score contribution evaluert i sann parameter

$$\mathbb{V}_{\rho_0}\left[S_n(\rho_0)\right] = \mathbb{E}_{\rho_0}\left[S_n(\rho_0)^2\right] \tag{6.37}$$

$$= \frac{1}{(\rho_0(1-\rho_0))^2}\mathbb{E}_{\rho_0}\left[(x_n-\rho_0)^2\right] \tag{6.38}$$

$$= \frac{1}{(\rho_0(1-\rho_0))^2}\mathbb{V}_{\rho_0}[x_n] \tag{6.39}$$

$$= \frac{1}{\rho_0(1-\rho_0)} \tag{6.40}$$

Kan også vises at dette tilsvarer den negative forventningsverdien til hessen evaluert i sann parameter, men hessen er ganske stygg siden vi må bruke kvotientregel. I stedet utvider jeg til å se på utvalg som består av summen av N contributions.

$$logL(\rho) = \sum_{n=1}^{N} logL_n(\rho) = \sum_{n=1}^{N}\left[x_n\log(\rho) + (1-x_n)\log(1-\rho)\right] \tag{6.41}$$

$$S(\rho) = \sum_{n=1}^{N} S_n(\rho) = \frac{\sum_{n=1}^{N}\left[x_n - \rho\right]}{\rho(1-\rho)} \tag{6.42}$$

Dette medfører at $\mathbb{E}[logL_n(\rho)] = \mathbb{E}[\frac{1}{N}logL(\rho)]$. Forventningsverdi til størrelsene er den samme, men med flere observasjoner så gir evaluering av forventningsverdi med hensyn på empirisk fordeling en bedre tilnærming.*Dette er reflektert i høyere fisher-informasjon.* Vi finner $\hat{\rho}_{MLE}$ ved å løse

$$\mathbb{E}_{\hat{P}_N}[S(\hat{\rho})] = 0 \tag{6.43}$$

$$\implies \hat{\rho} = \bar{x}_N \tag{6.44}$$

For å finne variansen til denne punktestimatoren må vi beregne fisher-informasjonen i utvalget. Med uavhengige variabler kan vi enkelt summere variansene slik at

$$\mathbb{V}_{\rho_0}[S(\rho_0)] = N \cdot \mathbb{V}_{\rho_0}[S_n(\rho_0)] \tag{6.45}$$

$$= \frac{N}{\rho_0(1-\rho_0)} := I(\rho_0) \tag{6.46}$$

Kunne kanskje evaluert variansen med hensyn på empirisk fordeling. Tror vi får samme resultat av å plugge inn punktestimator. Ble litt usikker. Uansett har vi nå at

$$(\hat{\rho} - \rho_0) \xrightarrow{d} N\left(0, I(\rho_0)^{-1}\right) = N\left(0, \frac{\rho_0(1-\rho_0)}{N}\right) \tag{6.47}$$

og vår estimasjon av denne fordelingen fra vårt éne realiserte utvalg er

$$N\left(0, \frac{\hat{\rho}(1-\hat{\rho})}{N}\right) \tag{6.48}$$

Dette er vårt beste forsøk på å tilnærme oss den ukjente asymptotiske fordelingen som igjen uansett bare vil være en tilnærming for vårt endelige utvalg. Med endelig antall observasjoner kan $\hat{\rho}$ bare ta et endelig antall verdier, så den eksakte fordelingen kan ikke være kontinuerlig. Vi kan vise at fordelingen til $N \times \hat{\rho}$ er $binom(N, \rho)$ og brukt dette resultatet i stedet, men er jo kjekt at MLE gir et generelt rammeverk til å finne asymptotisk fordelingen til stor klasse av estimatorer med gode asymptotiske egenskaper!

6.2.2 Normalfordeling med kjent varians

$$f(x;\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \tag{6.49}$$

$$\log L_n(\mu) = -\frac{(x-\mu)^2}{2\sigma^2} + ... \tag{6.50}$$

$$S_n(\mu) = \frac{\partial}{\partial\mu} \log L_n(\mu) = \frac{x-\mu}{\sigma^2} \tag{6.51}$$

$$\mathbb{E}[S_n(\mu)] = 0 \implies \mu = \mathbb{E}[x] \tag{6.52}$$

$$I(\mu) := -\mathbb{E}\left[\frac{\partial}{\partial\mu} S_n(\mu)\right] = \mathbb{E}[\sigma^{-2}] \tag{6.53}$$

$$Avar(\hat{\mu}) = I(\mu)^{-1} = \sigma^2 \tag{6.54}$$

$$\implies \sqrt{N}(\hat{\mu} - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{6.55}$$

6.2.3 Andre hendelser

Vi observerer utfall fra P_θ og på bakgrunn av dette vil vi kvantifisere relativ sannsynlighet for ulike verdier av den ukjente θ. Likelihoodfunksjonen fikser dette og kan håndtere ulike typer utfall. La oss ta normalfordeling med kjent $\sigma = 1$ som eksempel

1. Observere utfall direkte, f. eks. $X = 1.3$, $L(\theta) = \phi(1.3 - \theta)$
2. Observere at utfall er i et intervall, f.eks. $X \in (1, 3)$, $L(\theta) = P_\theta(X \in (1, 3)) = \Phi(3 - \theta) - \Phi(1 - \theta)$
3. Observere en funksjon av utfall, f.eks. $Y = g(X) = \max(X_1, ..., X_N) := X_N = 4$, $L(\theta) = P(g(X) = 4) = N\Phi(4 - \theta)^{N-1}\phi(4 - \theta)$

der siste følger av at $P(X_N < s) = P(X_1 < s, ..., X_N < s) = \Phi(s-\theta)^N$ og $L(\theta) = f_\theta(s) = \frac{\partial}{\partial s}F(s)$. Ser generelt at jeg vil evaluere tettheter og eventuelt integral over tettheter dersom jeg ikke har eksakt verdi. Kan også enkelt kombinere informasjon fra ulike kilder så lenge observasjonen er uavhengige. Med log-likelihood er det bare å summere opp funksjonene. Det kan enten være enkeltobservasjoner eller fra ulike utvalg. Trenger ikke justere for antall observasjoner som ingikk for å konstruere funksjonen, siden all informasjon er oppsumert i selve funksjonen.

6.3 Oppsummere informasjon fra likelihoodfunksjonen

For et gitt utvalg vil likelihoodfunksjonen angi relativ sannsynlighet for ulike parameterverdier. Dette gir både informasjon om hvilke verdier som er mest sannsynlig og samt hvor sikre vi er på at den sanne, ukjente parameteren er i ulike intervall. Det er en utfordring at det kan være vanskelig å kommunisere denne informasjonen, spesielt hvis parameteren

er en vektor slik at likelihood blir funksjon av flere variabler. Det kan dessuten være litt vanskelig å jobbe med funksjoner. Vi vil derfor ønske å finne alternative måter å oppsummere informasjon i likelihoodkurven. Ser generelt at det er både enklere å jobbe med log-likelihood numerisk og at analytiske resultat bruker denne formen.

Vi vet for det første at maksimumsverdien gir det best punktestimatet. Videre vil spissheten til funksjonen rundt $\hat{\theta} = \arg\max L(\theta)$ gi et mål på hvor sikre vi er på at den gitte $p_{\hat{\theta}}$ har generert utvalget. Hvis det er flatt på toppen er det mange ulike kandidater som er omtrent like sannsynlige, eg. kan korrespondere med observasjonene vi har observert, slik at det er mye usikkerhet knyttet til $\hat{\theta}$.

6.3.1 Kvadratisk tilnærming

Vi kan bruke størrelsene over til å beregne en andre ordens taylor ekspansjon av $logL(\theta)$ i $\theta = \hat{\theta}$.

$$logL(\theta) \approx logL(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{I(\hat{\theta})^{-1}}{2}(\theta - \hat{\theta})^2 \tag{6.56}$$

$$= logL(\hat{\theta}) - \frac{I(\hat{\theta})^{-1}}{2}(\theta - \hat{\theta})^2 \tag{6.57}$$

Hele formen på funksjonen er da beskrevet av punktet $(\hat{\theta}, logL(\hat{\theta}))$ samt den estimerte fisher-informasjonen. Det er mulig å vise at dette gir en eksakt beskrivelse av loglikelihood-funksjonen til forventningsverdien av normalfordeling. For andre likelihoodfunksjoner vil det være en god tilnærming dersom de er såkalt *regulære*. Det kan vises at de fleste log-likelihoodfunksjoner konvergerer mot denne kvadratiske formen når antall observasjoner øker og dette har litt sammenheng med CLT. Jeg skal nå utvikle teoretiske resultat som har utgangspunkt i denne forenklede representasjonen. Disse resultatene vil holde eksakt for gjennomsnitt av normalfordeling og være en asymptotisk tilnærming for andre fordelinger. Først skal jeg bare utlede to enkle sammenhenger til. Det første er en forenklet representasjon av normalisert likelihood.

$$\log\left(\frac{L(\theta)}{L(\hat{\theta}}\right) = logL(\theta) - logL(\hat{\theta}) \tag{6.58}$$

$$\approx \frac{I(\hat{\theta})^{-1}}{2}(\theta - \hat{\theta})^2 \tag{6.59}$$

Den andre sammenhengen er bare første ordens taylor ekspansjon av score funksjonen som kan brukes til å visualisere om den kvadratiske tilnærmingen er god ved å se om

sammenhengen under faktisk er lineær.

$$S(\theta) \approx S(\hat{\theta}) - I(\hat{\theta})^{-1}(\theta - \hat{\theta}) \tag{6.60}$$
$$= -I(\hat{\theta})^{-1}(\theta - \hat{\theta}) \tag{6.61}$$

6.3.2 Konfidensintervall

Jeg kan ha lyst til å konstruere et intervall $\Theta_c \subset \Theta$ der det virker rimelig at $P \in \{P_\theta : \theta \in \Theta_c\}$ kan ha generert det utvalget jeg observerer. Et greit utgangspunkt kan være å betrakte mengden

$$\Theta_c = \left\{ \theta : \frac{L(\theta)}{L(\hat{\theta})} > c \right\} \tag{6.62}$$

Spørsmålet nå er hvordan vi skal velge c. Vil forsøke å knytte det til noe som i prinsippet er en observerbar sannsynlighet. Vil manipulere uttrykket slik at jeg får en størrelse med kjent fordeling. Begynner med å observere at

$$\log\left(\frac{L(\theta)}{L(\hat{\theta}}\right) = \frac{\sigma^2}{2N}(\bar{x} - \theta)^2 \tag{6.63}$$

for gjennomsnitt av normalfordeling og at dette kan være god tilnærming for andre regulære likelihoods. Jeg vet at

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{N}\right) \implies 2 \cdot \frac{L(\hat{\theta})}{L(\theta)} \sim \chi^2(1) \tag{6.64}$$

Manipulerer begge sider av ulikheten over og får

$$P\left(2 \cdot \frac{L(\hat{\theta})}{L(\theta)} < -2\log(c) = \chi^2_{1-\alpha}\right) = 1 - \alpha \tag{6.65}$$

6.4 Likelihood i flere dimensjoner

Har nå sett at jeg kan anta at observasjonene i utvalget mitt er *iid* realiseringer fra $P_\theta \in \mathscr{P}_\theta$ og bruke parameterverdien som gjør det mest sannsynlig å observere verdiene i utvalget mitt som estimat. Dette kan vi generalisere til observasjoner $\mathbf{z} \in \mathbb{R}^d$ der P_θ nå blir en simultanfordeling. Det er en utfordring at simultanfordelinger er veldig komplekse objekter. Da skal angi skal angi sannsynlighet for utfall i ganske vilkårlige delmengder av $\mathbb{R}^d$. Det er både vanskelig å estimere og beskrive. I praksis vil vi ofte heller si noe om betinget sannsynlighet.

6.4.1 Betinget likelihood

I praksis vil vi ofte dekomponere $\mathbf{z} = (\mathbf{x}, y)$ og se på hvordan $\mathbf{x}$ påvirker fordeling av y. Da får vi bruk for at

$$f(\mathbf{x}, y) = f(y|\mathbf{x})f(\mathbf{x}) \tag{6.66}$$

der vi er interessert i $f(y|\mathbf{x})$. Vi kan parametrisere tetthetene over slik at

$$f(\mathbf{x}, y; \theta, \gamma) = f(y|\mathbf{x}; \theta)f(\mathbf{x}; \gamma) \tag{6.67}$$

Hvis vi tar log-likelihood får vi

$$logL(\theta, \gamma) = \log(f(y|\mathbf{x}; \theta)) + \log(f(\mathbf{x}; \gamma)) \tag{6.68}$$

Hvis vi bare er interessert i θ så er andre leddet en uvesentlig konstant. Får samme estimat ved å kun betrakte første del som om vi betraktet likelihood til hele simultanfordelingen.[8]

6.4.2 Generell fremgangsmåte til å finne likelihood til betinget fordeling

Vi har en regresjonsmodell

$$y = \mathbf{x}'\beta_0 + u, \quad u|\mathbf{x} \sim N(0, \sigma_0^2) \tag{6.69}$$

Vi begynner med å finne betinget kumulativ sannsynlighet

$$P(Y \leq y|\mathbf{x}) = F(\mathbf{x}'\beta_0 + u < y|\mathbf{x}) \tag{6.70}$$

$$= F(u < y - \mathbf{x}'\beta_0|\mathbf{x}) \tag{6.71}$$

$$= F\left(\frac{u}{\sigma_0} < \frac{y - \mathbf{x}'\beta_0}{\sigma_0}|\mathbf{x}\right) \tag{6.72}$$

$$= \Phi\left(\frac{y - \mathbf{x}'\beta_0}{\sigma_0}\right) \tag{6.73}$$

[8] Gitt at det ikke er funksjonell relasjon mellom parameterene (θ, γ)... i praksis vil vi neppe ønske å modellere dette.

der $F(c|\mathbf{x}) := \int_{-\infty}^{c} yf(y|\mathbf{x})dy$. Merk at selv om $\mathbf{x}'\beta$ er tilfeldig så kan vi behandle det som en konstant når vi betinger av $\mathbf{x}$. Vi kan deretter enkelt finne tetthet ved å derivere

$$f(y|x) = \frac{\partial}{\partial y}\Phi\left(\frac{y - \mathbf{x}'\beta_0}{\sigma_0}\right) \tag{6.74}$$

$$= \frac{1}{\sigma_0}\phi\left(\frac{y - \mathbf{x}'\beta_0}{\sigma_0}\right) \tag{6.75}$$

Denne fremgangsmåten bruker eksplisitt antagelse om betinget fordeling til feilledd i stedet for å modellere betinget fordeling til y direkte.. Tror det er litt ulike måter man kan gjøre dette på.

6.4.3 Betinget normal

Et konkret eksempel er $y|\mathbf{x} \sim N(\mu_x, \sigma_x^2) = N(g(\mathbf{x}), h(\mathbf{x})^2)$ og spesifisere hvordan parametrene avhenger av $\mathbf{x}$. Vanlig valg er $g(\mathbf{x}) = \mathbf{x}'\beta$ og $h(\mathbf{x}) = \sigma$, altså at vi varians ikke avhenger av $\mathbf{x}$. Det er ingenting i veien for at vi modeller hvordan varians avhenger av $\mathbf{x}$, men ofte er vi bare interessert i betinget forventningsverdi.[9] Vi kan da skrive opp likelihood-funksjonen.

$$L(\beta, \sigma) = \prod_n f(\mathbf{x}_n, y_n) = \prod_n \frac{1}{\sqrt{2\pi}\sigma} exp\{-\frac{(y_n - \mathbf{x}_n\beta)^2}{2\sigma^2} \cdot f(\mathbf{x}_n) \tag{6.76}$$

Kan merke da at parameter i betinget fordeling ikke avhenger av $f(\mathbf{x}_n)$ slik at vi kan se bort i fra dette.. og vise at OLS maksimerer likelihood. hm.

6.4.4 Betinget bernoulli

Vi vil se på hvordan ulike egenskaper x til en person påvirker sannsynlighet for at hen deltar i arbeidsmarkedet.

$$q(s) = \mathbb{P}\{y = 1|x = s]\} \tag{6.77}$$

For at funksjonen skal tilfredstiller aksiom til sannsynlighetsfunksjoner må $q(s) \in [0, 1] \forall s \in \mathbb{R}^k$. En type funksjoner som tilfredstiller det kraver er kumulative sannsynlighetsfunksjoner. I praksis bruker vi derfor cdf med lineær parametrisering, $q(s) = F(s'\beta)$. Kan ikke estimere det med OLS siden det er en ikke-lineær funksjon mhp parametrene. For å gjøre MLE operativt må vi ha en spesifisert log likelihood funksjon som vi kan optimere. Første steg er betinget pmf. I utgangspunktet er det en piecewise funksjon, men jeg kan bruke

[9] Tror dette er eksempel på semi-parametrisk estimering der σ er såkalt nuisance-parameter.

triks for å få det på én linje:

$$P(y = i|x = s) = F(s'\beta)^i(1 - F(s'\beta)^{1-i} \tag{6.78}$$

$$\implies \log L(\beta) = \sum y_n \log(F(s'\beta)) + \sum(1 - y_n)\log(1 - F(s'\beta)) \tag{6.79}$$

Dette kan jeg løse og få logit eller probit avhengig av valg av F. Ble litt ukomfortabel notasjon fordi jeg ikke vil bruke store bokstaver, men skal helst sikkert se nærmere på dette senere.

6.5 Prinsipp for å utlede tester

Vi vil ofte teste om teste om data i utvalg gir tilstrekkelig bevis til at vi kan forkaste påstand om at $\theta \in \Theta_0$. Vi forkaster dersom det er lite sannsynlig at de faktiske observasjonene i utvalget har blitt generert fra en fordeling med parameter fra nullhypotesen. Videre gjør vi ofte avgrensinger av hvilke fordelinger $\mathscr{P}$ vi vil betrakte. Det er da nyttig å gjøre spesifikasjonstester for se om vi kan forkaste at $\mathbb{P} \in \mathscr{P}$ slik at modell er feilspesifisert, for eksempel ved at feilledd er heteroskedastisk. Har tre ulike prinsipp for å utlede testobservator fra likelihoodfunksjon som er asymptotisk ekvivalente og alle gir χ^2-fordelte testobservator. Tror at at $t - ogF - fordeling$ bare er justering som tar hensyn til at utvalg er begrenset, men litt usikker på dette.

6.5.1 Wald-test

Denne fremgangsmåten tar utgangspunkt i en (asymptotisk) normalfordelt MLE estimator $\hat{\theta}$ og bruker at

$$\mathbf{R}\hat{\theta} - \mathbf{q} \xrightarrow{d} N(\mathbf{0}, \mathbf{R}\Sigma\mathbf{R}') \tag{6.80}$$

hypotesen $\mathbf{R}\theta = \mathbf{q}$ er sann.. Kan da forkaste nullhypotese hvis testestimator gir stort avvik fra 0, der vi bruker kvantiler av normalfordeling til å konkretisere hva som utgjør tilstrekkelig stort avvik for våre formål. Med enkelt parameter har testen form

$$W = \frac{\hat{\delta} - \delta_0}{\hat{se}} \xrightarrow{d} N(0, 1) \tag{6.81}$$

Med flere parametere tror jeg den fremgangsmåten gir flervariabel normalfordeling, men det er mye greiere å få tilbake et enkelt tall slik at vi kan forkaste hvis langt fra null. Alle tre prinsippene for å utlede asymptotiske tester gir oss derfor generelt testobservatorer som er χ^2-fordelt. For den enkle testen over gir det

$$\xi_w = Z^2 = (\hat{\delta} - \delta_0)[\hat{var}]^{-1}(\hat{\delta} - \delta_0) \xrightarrow{d} \chi^2(1) \tag{6.82}$$

I praksis er vi ofte interessert i forskjell mellom parametre siden vi ikke har et kjent benchmark vi tester mot. Det er relativt greit dersom utvalgene er uavhengige av hverandre siden varians til differansen av estimatorer er sum av variansen til hver av de. Et eksempel er forskjell i parameter i bernoulli-fordelt variabel. Vi har $\bar{X_1} \sim binom(p_1, n_1)$ og $\bar{X_2} \sim binom(p_2, n_2)$. Da er $\hat{\delta} = \hat{p_2} - \hat{p_1}$. Gjenstår bare å finne $\hat{se} := \widehat{se(\hat{\delta})}$. Vet at $\hat{p_j}$ gjennomsnitt. Vet at varians til gjennomsnitt er $\frac{\sigma^2}{N}$. Vet at $\sigma = p(1-p)$ i bernoulli, som er fordeling til X. Dette er tilstrekkelig til å finne $\hat{se}$, men gidder ikke skrive. Ta det som oppgave når du leser dette. Tilsvarende kan vi teste differanse mellom normalfordelte. En utvidelse et t-test i stedet for z-test.

6.5.2 Likelihood ratio

Likelihood ratio tar utgangspunkt i forskjellen i maksiumum av log-likelihood fra ubetinget og betinget optimering.

$$\lambda = 2 \log \left(\frac{\sup_{\theta \in \Theta} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} \right) = 2 \left(\frac{L(\hat{\theta})}{L(\hat{\theta}_0)} \right) \tag{6.83}$$

der $\hat{\theta}$ er MLE-estimatoren og $\hat{\theta}_0$ er MLE-estimator avgrenset til $\Theta = \Theta_0$. Dette har visst en χ^2 fordeling. Intuisjon for dette er at hvis forskjellen er stor så er det lite sannsynlig at avgrensingen ikke er bindene, altså at lite sannsynlig at $\theta \in \Theta_0$

6.5.3 Lagrange multiplier

Nullhypotesen medfører en restriksjon av parameterromet. Undersøker i hvilken grad restriksjon er bindene ved å se på lagrangemultiplier assosiert med restriksjonen av de ulike parameterne, skyggepris. Hvis stor skyggepris er lite sannsynlig at sann parameter i Θ_0 som vi har avgrenset til å velge løsning innenfor..

6.6 Egenskaper ved feilspesifikasjon

Det er ganske sterk antagelse at $P_0 \in \mathscr{P}$, så kjekt at ikke alt rakner dersom denne antagelsen er feil.

6.6.1 Total variation distance og KL-divergence

Jeg vil ha et mål på avstand mellom to probability measures $D(\mathbb{P}_{\theta_0}, \mathbb{P}_\theta)$. Et ganske naturlig mål er *Total variation distance*

$$TV(\mathbb{P}_{\theta_0}, \mathbb{P}_\theta) = \max_{A \subset \Omega} |\mathbb{P}_{\theta_0}(A) - \mathbb{P}_\theta(A)| \tag{6.84}$$

Dette gir et tall i intervallet $[0, 1]$ og tilfredstiller egenskapene til en avstand (symmetrisk, trekantulikhet..). Hvis vi ser på fordelinger på $\mathbb{R}$ kan vi beregne størrelsen med

$$TV(P_{\theta_0}, P_\theta) = \begin{cases} \frac{1}{2} \sum |p_{\theta_0}(x) - p_\theta(x)| \\ \frac{1}{2} \int |f_{\theta_0}(x) - f_\theta(x)| dx \end{cases} \tag{6.85}$$

Dette tilsvarer areal av mellom kurvene i området der den ene er større enn den andre. Det er symmetri siden areal under begge kurvene summerer til 1. Dette er et naturlig mål med gode egenskaper, men det litt vanskelig å gjøre operativt. Dette motiverer Kullback-Leibler (KL) divergence som har noe av de samme gode egenskapene, men som vi kan estimere fra utvalg med observasjoner fra P_{θ_0}.

$$KL(P_{\theta_0}, P_\theta) = \begin{cases} \sum p_{\theta_0}(x) \log\left(\frac{p_{\theta_0}(x)}{p_\theta(x)}\right) \\ \int f_{\theta_0}(x) \log\left(\frac{f_{\theta_0}(x)}{f_\theta(x)}\right) dx \end{cases} \tag{6.86}$$

Dette målet er ikke symmetrisk og tilfredstiller ikke triangelulikhet, men er i likhet med TVD 0 når funksjonene er like og vokser når avstanden øker. Selve tallet har ikke naturlig tolkning. Merk at dette er forventningsverdi av en funksjon med hensyn på P_{θ_0}.

$$KL(P_{\theta_0}, P_\theta) = E_{\theta_0}[log(f_{\theta_0})] - E_{\theta_0}[log(f_\theta)] \tag{6.87}$$

Et naturlig valg av θ er den verdien som minimerer den empiriske analogen til KL-divergence,$\widehat{KL}(P_{\theta_0}, P_\theta)$. Merk at første ledd er konstant som ikke påvirker arg min.

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \widehat{KL}(P_{\theta_0}, P_\theta) \tag{6.88}$$

$$= \arg\min_{\theta \in \Theta} \{E_{\theta_0}[log(f_{\theta_0}(x_n))] - E_{\hat{P}_N}[\log(f_\theta(x_n))]\} \tag{6.89}$$

$$= \arg\min_{\theta \in \Theta} \text{konstant} - \frac{1}{N} \sum \log(f_\theta(x_n)) \tag{6.90}$$

$$= \arg\max_{\theta \in \Theta} \log(\Pi_n f_\theta(x_n)) \tag{6.91}$$

$$= \arg\max_{\theta \in \Theta} \log(\Pi_n L_n(\theta; x_n) \tag{6.92}$$

Løsningen på dette optimeringsproblemet tilsvarer MLE-estimatoren. Så langt har vi antydet at den sanne fordelingen tilhører den parametriske klassen som vi søker over, men siden $E_{\theta_0}[log(f_{\theta_0}(x_n))]$ uansett bare er en konstant som vi kan se bort i fra i optimeringen, så sier dette resultatet oss at MLE asymptotisk gir oss θ som korresponderer med den fordelingen innenfor $\mathscr{P}$ som minimerer avstand til den sanne P, både i betydningen av TVD og KL. Kunne jo selvsagt kommet nærmere ved å søke over en riktig spesifisiert $\mathscr{P}$, men det er jo uansett et greit resultat.

6.6.2 MLE fra empirisk risikominimering

Kan vise at MLE-estimatoren kan utledes som spesialtilfelle av empirisk risikominimering.[10] Anta at vi vil estimerne en ukjent tetthetsfunksjon $q(\cdot)$ med utgangspunkt i observerte realiseringer fra fordeling med den tettheten. Definererer tapsfunksjon til kandidat $p(\dot{c})$ ved $L(p, x) := -\log(p(x))$. Hvis vi observerer realisering $x = s$ så vil det realiserte tapet være større desto lavere verdi av $p(s)$, altså jo lavere tyngde vår kandidat plasserer på den realiserte verdien. Risikofunksjonen er dermed gitt ved

$$R(p) = \mathbb{E}_q[L(p, x)] = -\int \log(p(s))ds \tag{6.93}$$

For å knytte dette til MLE avgrenser vi til å betrakte et parametrisert hypoteserom $\mathcal{P}_\theta = \{P_\theta : \theta \in \Theta\}$. Antar at modellen er identifisert slik at $\theta \mapsto P_\theta$ er bijektiv (én-til-én) slik at vi ekvivalent kan løse minimeringsproblemet med hensyn på θ.

$$P_{\hat{\theta}} = \arg\min_{p \in \mathcal{P}_\theta} R_{emp}(p) \tag{6.94}$$

$$\implies \hat{\theta} = \arg\min_{\theta \in \Theta} -\sum \log(p_\theta(s_n)) \tag{6.95}$$

$$\implies \hat{\theta} = \arg\max_{\theta \in \Theta} \sum \log(p(\theta; s_n)) \tag{6.96}$$

Ser at løsningen tilsvarer $\hat{\theta}_{MLE}$ når vi bruke $-\log(p(s))$ som tapsfunksjonen. Vi kan også dekomponere risikoen assosiert med denne tapsfunksjonen for å knytte det til KL-divergence,

$$R(p) = -\int \log(p(s))q(s)ds \tag{6.97}$$

$$= \mathbb{E}_q[(-\log(p(s)) + \log(q(s)) - \log(q(s)))] \tag{6.98}$$

$$= \mathbb{E}_q\left[\log\left(\frac{q(s)}{p(s)}\right)\right] - \mathbb{E}_q[\log(q(s))] \tag{6.99}$$

der første ledd er KL-divergence og andre ledd er entropy. Ser at entropy tilsvarer risiko når fordelingen er kjent. Får tilbake igjen MLE-estimatoren ved å minimere utvalgsanalogen til KL-divergence med hensyn på parametrisert p som beskrevet i seksjonen over.

Liten digresjon, må endres eller slettes

Kan utlede estimering av logistisk regresjon m.m. ved å bruke såkalt logistisk tap,

$$-\{y \log h_\theta(x)) + (1 - y) \log(h_\theta(x))\} \tag{6.100}$$

[10] I hvert fall for estimering av tetthetsfunksjon... skal se om jeg kan utvide til regresjon.

men denne tapsfunksjonen kan jeg jo uansett utlede fra loglikelihood til bernoulli-fordelingen. Det er jo litt poeng at jeg kan motivere dette uten MLE, men er uansett bedre å gjøre det innenfor.

6.6.3 Kvasi-MLE

Forventningsverdien til score-funksjonen evaluert i sann parameter er 0.

$$\mathbb{E}_{\theta_0} S(\theta_0) := \int S(\theta_0) f(x;\theta_0) dx = 0 \tag{6.101}$$

$$\tag{6.102}$$

Utvalgsanalogen er å finne $\hat{\theta}$ som gjør at $\mathbb{E}_{\hat{P}_N}[S(\theta) = 0$. Dette gir oss et ligningssystem av momentbetingelser som vi kan løse og estimatoren kan betraktes som en momentestimator. Hvis modellen er riktig spesifisert er estimatorene ekvivalente, men egenskapene til momentestimatoren vil være gyldig for en større klasse av fordelinger som har de samme første-ordens betingelsene. Hvis vi bruker F.O.B fra MLE til å betrakte fordelingen til estimatoren fra denne utvidede mengden kan vi betegne det som kvasi-likelihood. Fra store talls lov vet vi at estimatoren er konsistent og fra CLT får vi normalfordelingen, men asymptotisk varians er ikke lenger $I(\theta)^{-1}$. Vi må bruke såkalt sandwich estimator,

$$(\hat{\theta}_{QMLE} - \theta_0) \xrightarrow{d} N(0, V) \tag{6.103}$$

der

$$V = ... \tag{6.104}$$

Vil knytte dette til robust standardfeil i regresjon..S

6.6.4 Extremum estimators

Kan betrakte en klasse av estimator som er løsning på et optimeringsproblem av en objektfunksjon $Q_N(\cdot)$ som har verdimengde i $\mathbb{R}$,

$$\hat{\theta} = \arg\max_{\theta \in \Theta} Q_N(\theta; \mathbf{z}_d) \tag{6.105}$$

En viktig underklasse er M-estimatorer der $Q_N(\cdot)$ har formen

$$Q_N(\theta; \mathbf{z}_d) = \frac{1}{N} \sum_n m(\theta; \mathbf{z}_n). \tag{6.106}$$

Et annet eksempel er GMM der

$$Q_N(\theta; \mathbf{z}_d) = -\frac{1}{2} g_N(\theta; \mathbf{z}_d)' \hat{W} g_N(\theta; \mathbf{z}_d) \tag{6.107}$$

der $g_N(\theta; \mathbf{z}_d) := \frac{1}{N} g(\theta; z_n)$ og $g(\cdot)$ er momentbetingelse. Tror vi kan bruke denne mer generelle klassen av estimatorer til å dra koblinger mellom MLE og GMM, blant annet type *Limited information maxmimum likelihood* og finne asymptotiske tester for GMM. Jeg mistenker at de greiene der må bli i neste liv.

Kapittel 7

Lineær regresjon

Noen av fordelene med å avgrense til linære funksjoner er at vi kan få parametre med enkle tolkninger og det empiriske risikominimeringsproblemet med kvadratisk tap har en analytisk løsning (MKM). Kjenner teoretisk egenskap, stabil, prediksjonsrisiko. Stor klasse av additative modeller... tenker det finnes parametrisk representasjon av splines og lignende... litt usikker på hva jeg sier om dette. Nært knyttet til MKM, men dette er bare én av flere måter å estimere koeffisientene. (Vekte observasjonene i tapsfunksjonen ut fra varians... legge til regularisering... si noe om IV?).

Lineær regresjon bruker i ulike fagfelt, for ulike formål og kan motiveres på ulike måter. Beste tilnærmede løsning på et overdeterminert ligningssystem. I økonometri motiveres det gjerne av såkalt *conditional independence assumption* der behandling er tilfeldig fordelt innad i delgrupper og kan betraktes som analog til *matching estimator*. I statistisk modellering (MLE/Bayes) gjør vi eksplisitte antagelser om parametrisert struktur til fordeling som generer data. I ren prediksjonssetting kan vi være mer agnostisk om fordeling og bare løser risikominimeringsproblem. Jeg begynner med siste tilnærming.

7.1 Populasjon

beskrive egenskap ved simultanfordeling. ting vi kan forsøke å lære fra realiserte observasjoner

Vi kan betegne den betingede forventningsfunksjonen $E[Y|X] := \int yf(y|X)dy$ som (populasjons) regresjonsfunksjonen.[1] Dette er en tilfeldig variabel som for hver $X = x$ angir forventningsverdi til den betingede fordelingen av y. Det er den ortogonale projeksjonen av y ned på underrommet som består av alle tilfeldige variabler som kan skrives som en deterministisk funksjon av X. Dette medfører at det minimerer forventet avvik

[1] Merk at populasjonsregresjonsfunkjsjonen (PRF) også brukes som betegnelse på den beste lineære tilnærmingen. Jeg tenker det er bedre å betegne dette som den lineære populasjonsregresjonsfunksjonen.

og at vi kan dekomponere

$$Y = E[Y|X] + Y - E[Y|X] = E[Y|X] + U \tag{7.1}$$

der feilleddet U per konstruksjon er ortogonal på alle funksjoner av X, altså $E[g(X)U] = 0$. I praksis bruker vi små bokstaver av notasjonell konvensjon.

Den lineære populasjonsregresjonsfunksjonen tilsvarer den ortgonale projeksjonen av y ned på mengden av tilfeldige variabler som kan skrives som en lineær funksjon av x. Det medfører at vi kan dekomponere

$$y = \beta' x + y - \beta' x = \beta' x + u \tag{7.2}$$

der feilleddet u per konstruksjon er ortogonal på alle lineære funksjoner av x, altså $E[(b'x)u] = b'E[xu] = 0$. Dersom det inkluderer et konstantledd så medfører det også at $E[1u] = E[u] = 0$ og $cov(x_k, u) = 0$. Vi kan også motivere dette som beste lineære tilnærming til $E[y|x]$ som er det vi egentlig er interessert i. Husk at dette er en tilfeldig variabel, så det er ingenting i verien for å betrakte det som den avhengige variabelen i regresjonen.

7.1.1 Projeksjon

7.1.2 Dekomponering av varians

knytte noe til avstand, geometri.. mest mulig analog

7.1.3 Tolkning av feilledd

avvik, E[u|x], strukturelt eller ikke. tolkning av parameter.

7.2 Algebra i utvalg

Utvalget er egentlig ikke det vi er interessert i, men det er alt vi har. Beskrive algebraisk egenskap. Egenskaper ved MKM løsning.

7.2.1 Ortogonal projeksjon

Minimeringsproblemet med kvadratisk tapsfunksjon er

$$h_{\hat{b}} = \underset{h_b \in \mathcal{H}_l}{\arg\min}\, \mathbb{E}_{P_{\hat{N}}}[(y_n - h_b(\mathbf{x}_n)^2] \tag{7.3}$$

$$\hat{\mathbf{b}} = \underset{b \in \mathbb{R}^K}{\arg\min} \frac{1}{N} \sum_n (y_n - \mathbf{x}_n' \mathbf{b})^2 \tag{7.4}$$

Generelt må vi minimerer tapsfunksjonen numerisk ved å bruke algoritme som søker over parameterromet. Her kan vi løse det analytisk.[2] Vi begynner med å sette det opp på matriseform ved å stacke input-vektorene,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_N' \end{bmatrix} \tag{7.5}$$

slik at $col_k(\mathbf{X})$ gir verdi av feature k til hver av de N observasjonene i utvalget. Omskriver tapsfunksjon[3],

$$\frac{1}{N}\sum_n (y_n - \mathbf{x}'\mathbf{b})^2 = \frac{1}{N}(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \tag{7.6}$$

Kan knytte dette til avstand og ortogonal projeksjon. Uansett, finner

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7.7}$$

7.2.2 Frisch-Waugh-Lovell

Vi projekterer $\mathbf{y}$ på $S(\mathbf{X})$.

$$\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{M}\mathbf{y} \tag{7.8}$$

$$= \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \mathbf{M}\mathbf{y} \tag{7.9}$$

FWL-theoremet sier at vi får samme $\hat{\beta}_2$ ved å projektere $\mathbf{y}$ på residualen av projeksjonen av $\mathbf{X}_2$ på $S(\mathbf{X}_1)$. Merk at matrise kan transformere flere vektorer om gangen, men når jeg snakker om residualer så begrenser jeg implisitt til å se på én vektorer.. Kan få noe intuisjon ved å tenke på bivariat regresjon. Hvis jeg prosjekterer $\mathbf{x}$ på $S(\mathbf{1})$ så får jeg $\bar{x}\mathbf{1}$. Residualen er da den sentrerte vektoren der hver komponent er avvik fra gjennomsnitt. Hvis jeg regger $\mathbf{y}$ på den sentrerte variabelen uten konstantledd får jeg samme helning som i den bivariate regresjonen. Hm.

FLW-theoremet gjør at vi kan isolere enkeltkomponentner i helningskoeffisienten og betrakte det som analog bivariat regresjon.

$$\beta_1 = \frac{cov(y, \tilde{\mathbf{x}_k})}{var(\tilde{\mathbf{x}}_k)} \tag{7.10}$$

der $\tilde{\mathbf{x}}_k$ er residualen fra regresjonen av $\mathbf{x}_k$ på $S(\mathbf{x}_{-k})$.

[2]Det betyr at vi kan skrive $\hat{b} = g(\{(y_n, \mathbf{x_n}) : n = 1, ..., N\}$ der vi kjenner g. I praksis er det bedre å organisere utvalget i matrise som er helt analogt til representasjon i tabulær form som vi er vant til å se.

[3]Lurer på om jeg vil ha eget begrep for tap på hele utvalget i stedet for enkeltobservasjon... kostnad?

Tror det er enklest å tenke på dette som en to stegs prosess

$$\mathbf{M}_1\mathbf{y} = \mathbf{M}_1[\mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \mathbf{M}\mathbf{y}] \tag{7.11}$$

$$= \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 + \mathbf{M}\mathbf{y} \tag{7.12}$$

Merk at $S(\mathbf{M}) \subset S(\mathbf{M}_1)$. Hjelper dette..?

7.2.3 In-sample fit

Vi dekomponerer $\mathbf{y}$ i komponent i $span(\mathbf{X})$ og dets ortogonale komplement. Vi vil si noe om hvor god vår tilnærmede løsning er. Det avhenger den relative størrelsen på komponentene; hvor stor avviket er i forhold til *størrelsen* på $\mathbf{y}$. Alt dette er jo vektorer så bedre å snakke om lengde enn størrelse...

- $TSS = \|\mathbf{y}\|^2$
- $RSS = \|\mathbf{M}\mathbf{y}\|^2$
- $ESS = \|\mathbf{P}\mathbf{y}\|^2$

fra pythagoras følger det at TSS = RSS + ESS. Vi kan definere enkel R^2 som andelen av den totale lengden som går i retning av kolonnerommet til $\mathbf{X}$...

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - N\frac{R_{emp}(\hat{b}}{TSS} \tag{7.13}$$

Det følger at $R^2 \in [0, 1]$ og at det gir mål på in-sample fit.[4] Men målet vårt er å generalisere til nye data; ikke lære mest mulig om det gitte utvalget vi besitter. Maksimering av R^2 er dårlig kriterie i modellselskjon siden vi alltid kan få den høyere ved å legge til nye variabler enten de er relevante eller ikke,

$$span(\mathbf{X}_a) \subset span(\mathbf{X}_b) \implies R_a^2 \leq R_b^2 \tag{7.14}$$

Det er heller ikke problem i seg selv at R^2 er lav... hvis vi estimerer kausal sammenheng så kan det være at eksponering for behandling forklarer liten andel av variasjon i utfall. Kan medføre problem med presisjon til koeffisientestimatene, men ser på dette under statistisk egenskaper.

7.3 Inferens

Betrakte ting som estimator for populasjonsstørrelse, se på sammenheng, utvalgsfordeling. Har allerede sett litt på GMM, men hmm. vet ikke helt hvordan jeg skal organisere

[4] Hvis vi bruker annen estimering enn OLS og velger helt arbritære R^2 så kan vi få negative verdier.

7.3.1 Små utvalg

7.3.2 Store utvalg

7.3.3 Presisjon til koeffisient

7.3.4 Presisjon til prediksjon

Antar at jeg har en respons som er betinget normalfordelt og at CEF er lineær.

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\beta, \sigma^2\mathbf{I}) \tag{7.15}$$

Hvis jeg observerer én realisering av denne simultanfordelingen får jeg én realisering av $\hat{\beta}$. Jeg kan beregne utvalgsfordelingen til $\hat{\beta}$ gitt antagelsen jeg har gjort over,

$$\hat{\beta} \sim N(\beta, \sigma^2\mathbf{I}) \tag{7.16}$$

Hvis jeg får en ny observasjon $\mathbf{x}_{new}$ så vil min prediksjon fra den estimerte modellen være $\hat{y}_{new} := \mathbf{x}'_{new}\beta$ siden dette er sentraltendensen i betinget fordeling av $y|\mathbf{x}_{new}$. På en annen side er det jo slik at dersom jeg hadde observerte en annen realisering av simultanfordelingen ville jeg hatt en annen $\hat{\beta}$ og gitt annen prediksjon. Jeg vil forsøke å kvantifisere variasjonen i prediksjonen.[5]

$$\begin{aligned}
V(\hat{y}_{new}|\mathbf{X}) &= V(\mathbf{x}'_{new}\hat{\beta}|\mathbf{X}) && (7.17)\\
&= E[(\mathbf{x}'_{new}\hat{\beta})^2|\mathbf{X}] - E[\mathbf{x}'_{new}\hat{\beta}|\mathbf{X}]^2 && (7.18)\\
&= E[\mathbf{x}'_{new}\hat{\beta}\hat{\beta}'\mathbf{x}_{new}|\mathbf{X}] - (\mathbf{x}'_{new}\beta)^2 && (7.19)\\
&= \mathbf{x}'_{new}E[\hat{\beta}\hat{\beta}'|\mathbf{X}]\mathbf{x}_{new} - \mathbf{x}'_{new}\beta\beta'\mathbf{x}_{new} && (7.20)\\
& && (7.21)
\end{aligned}$$

Bruker nå at

$$\begin{aligned}
E[\hat{\beta}\hat{\beta}'|\mathbf{X}] &= V(\hat{\beta}|\mathbf{X}) + E[\hat{\beta}|\mathbf{X}]E[\hat{\beta}|\mathbf{X}]' && (7.22)\\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \beta\beta' && (7.23)
\end{aligned}$$

[5] Jeg tror målet mitt er å kunne angi et intervall der jeg med gitt sannsynlighet kan påstå at y_{new} vil ligge i. Det avhenger både av variasjon i $\hat{y}_{new}$ og fordeling til avvik fra sentraltendens... Begynner i hvertfall med å se på variasjon til prediksjonen.

slik at

$$V(\hat{y}_{new}|\mathbf{X}) = \mathbf{x}'_{new}(\sigma^2(\mathbf{X}'\mathbf{X})^{-1} + \beta\beta')\mathbf{x}_{new} - \mathbf{x}'_{new}\beta\beta'\mathbf{x}_{new} \tag{7.24}$$

$$= \sigma^2\mathbf{x}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{new} + \mathbf{x}'_{new}\beta\beta'\mathbf{x}_{new} - \mathbf{x}'_{new}\beta\beta'\mathbf{x}_{new} \tag{7.25}$$

$$= \sigma^2\mathbf{x}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_{new} \tag{7.26}$$

Tror det bør samsvare med variasjon jeg observerer når jeg sampler verdier av $\hat{\beta}$ og plotter linjer.. det vet jeg ikke om det gjør ..

7.3.5 Residual

Bruker residual til å estimere varians til feilledd,

$$\hat{\sigma}^2 = \frac{1}{N}\sum \hat{u}_n^2 \tag{7.27}$$

$$= \frac{1}{N}\sum (y_n - \mathbf{x}'_n\hat{\beta})^2 \tag{7.28}$$

$$= \frac{1}{N}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) \tag{7.29}$$

$$= \frac{1}{N}(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\hat{\beta} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \tag{7.30}$$

der

$$\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{7.31}$$

$$= \mathbf{y}'\mathbf{X}\hat{\beta} \tag{7.32}$$

slik at

$$\hat{\sigma}^2 = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\hat{\beta} \tag{7.33}$$

7.4 Hypotesetester

Knytte til test-prinsipp fra MLE...

7.4.1 Lineære restriksjoner av koeffisient

tror jeg vil ha t-test som special case av F-test

7.4.2 Diagnose

funksjonell form. vet ikke helt hva annet jeg vil teste..

7.5 Utvidelser

7.5.1 Funksjonell form

Vi forsøker å tilnærme oss funksjonen $\mathbb{E}[y|\mathbf{x}]$. Vi har sett litt på i hvilke tilfeller forskjellene i observert utfall for ulik nivå av *behandling* kan ha kausal tolkning. Uansett er jeg interessert i å beskrive hvordan funksjonen endrer seg når variabler endres. I praksis bruker vi lineær regresjon. Det kan håndtere ikke-linearitet ved å først transformere variablene. Tolke den estimerte funksjonen; endring i forhold til opprinnelige størrelser.

Logaritmer

Vi kan transformere den avhengige variabelen. Mange utfall er strengt positive, så da kan ikke feilledd være normalfordelt siden det er begrenset nedenfra. OLS er også veldig sensitiv for ekstremverdier, så greit å få skalert ned høye numerisk verdier av utfallet. Dessuten vil det ofte være slik at gjennomsnittlig utfall vokser omtrent proposjonalt med med variablene i stedet for lineært. Skal nå se hvordan vi tolker koeffisient der vi har tatt logaritmisk transformasjon av forklaringsvariabel og/eller utfall.

Vi bruker at logaritmer har omtrent prosentvis tolkning for små endringer. Dette følger av første-orders taylor ekspansjon av logaritmen evaluert i $x = 1$,

$$log(x) \approx log(1) + \frac{d}{dx} log(x)|_{x=1}(x-1) = x - 1, \tag{7.34}$$

slik at $\Delta log(x) := log(x_1) - log(x_0) = log(x_1/x_0) \approx x_1/x_0 - 1 := \Delta x/x_0$.

Log-level

Har modell $\log(y) = \beta_0 + \beta_1 x + u, E(u|x) = 0$. Det følger at $\beta_1 = \frac{d}{dx} E[\log(y)]$. Okay, men vi er interessert i hvordan det endrer forventningsverdi til y. Observerer at

$$\Delta E \log(y) = \beta_1 \Delta x \tag{7.35}$$

$$\implies 100 \cdot \beta \approx 100 \cdot E\frac{\Delta y}{y_0} := \%E\Delta y \tag{7.36}$$

Det kan tolkes som prosentvis endring i forventet utfall når x endres med én enhet. Hvis jeg vil finne eksakt prosentvis endring så kan jeg bruke

$$\Delta E \log(y) = \beta_1 \Delta x \tag{7.37}$$

$$\exp(\Delta E \log(y)) = \exp(\beta_1 \Delta x) \tag{7.38}$$

$$\Delta E \log(y) = \exp(\beta_1 \Delta x) \tag{7.39}$$

$$\%\Delta E \log(y) = 100 \cdot (\exp(\beta_1 \Delta x) - 1) \tag{7.40}$$

Det er nødvendig å være litt forsiktig når vi bruker denne spesifikasjonen til å predikere verdi av y. Det kan være fristende å bruke

$$\widehat{logy} = \mathbf{x}'\hat{\beta} \tag{7.41}$$

$$\hat{y} = \exp(\mathbf{x}'\hat{\beta}) \tag{7.42}$$

men dette er dårlig estimat på $E[y|x]$. For å se dette, observer at

$$E[y|x] = E[\exp(\mathbf{x}'\beta + u)|\mathbf{x}] = \exp(\mathbf{x}'\beta)E[\exp(u)] \tag{7.43}$$

der $E[e^u] = \alpha_0 \neq 0$. For å bruke spesifikasjonen til å predikere verdi av y må vi skalere opp $\hat{y} = \mathbf{x}'\hat{\beta}$ med $\hat{\alpha}_0 = E_{\hat{P}_N}[e^u] = \frac{1}{N}\sum_n e^{\hat{u}_n}$

Log-log

Kan toles som elastistet. Følger av at

$$\Delta E \log(y) = \beta_1 \Delta \log(x) \tag{7.44}$$

$$\implies \beta_1 \approx \%\Delta Ey/\%\Delta x \tag{7.45}$$

Regresjonsanatomi

$$cov(y_i, \tilde{x}_{ki}) = cov(\mathbf{x}'\beta + u, \tilde{x}_{ki}) = \beta_k var(\tilde{x}_{ki}) \tag{7.46}$$

$$\implies \beta_k = \frac{cov(y_i, \tilde{x}_{ki})}{var(\tilde{x}_{ki})} \tag{7.47}$$

hmmm. frischhh

Anova (flytt til betinget fordeling/L2)

$$V(y) = V(E[y|x] + u) = V(E[y|x]) + V(u) \tag{7.48}$$

der $V(u) = E[u^2] = E[E[u^2|x] = E[V[u^2|x]$.

Generalisert lineær modell, flytte hvor?

Utvide... fordeling til u (eg betinget fordeling av y); ikke bare normal, andre medlem av eksponentialfordeling. Dessuten utvide til linkfunksjon... mer utvidet måte å modellere E[y|x]. Vet ikke hvor relevant og spennende dette egentlig er.. vet faen ingen ting.

Saturated model

Benchmark, start punkt. Så forenkle. Utvide til eksponering av behandling som kan ta mange verdier. Se på heterogen behandlingseffekt..

Kapittel 8

Statistisk læring

Noen tanker fra Geron

Maskinlæring går ut på å lære masking å gjøre oppgaver uten å spesifisere eksplisitte regler for hva den skal gjøre. Tradisjonelle algoritmer er litt som en matoppskrift som vi i prinsippet kan få en maskin til å følge. Dette er greit for enkle og veldefinerte oppgaver, men i mange situasjoner er det bedre å utnytte at *computing power* gjør det mulig for maskinen å lære fremgangsmåten gjennom prøving og feiling i henhold til vurderingskriterie. Kan lære optimal steketid og potensielt hvordan det avhenger av andre variabler. Kan både oppnå bedre resultat og program som er enklere å utvikle og vedlikeholder, og som ikke trenger like mye *domain knowledge*.

Maskinlæringsalgoritmer bruker erfaring E til å bli bedre til å utføre oppgave T som målt ved kriterie P. Et viktig eksempel er prediksjon. Vi vil finne predikert output gitt input, $\hat{y} = h(\mathbf{x})$. Erfaring er $(\mathbf{x}_n, y_n)$, kriterie er tapsfunksjon $\frac{1}{N}\sum L(h(\mathbf{x}_n), y_n)$. Har labelled data og finne h slik at minimere tap. I praksis så kan vi indekse h_θ og søke over $\theta \in \Theta$. Ikke trivielt å optimere, gradient descent.

Kategorisere algoritmer: Supervisesed (reg og klassifikasjon) vs unsupervised (clustering, dimensjonsreduksjon, anamolie). Batch vs incremental learning. Instance vs model based.

Utfordringer: For lite data. Signal og støy... støy jevner seg ut (per konstruksjon/definisjon), avvik fra sentraltendens. Avhenger av hvor sterkt signal er i forhold til støy. Problem at komplekse/fleksible algoritmer er veldig flinke til å finne mønster. Finner selv om det bare skyldes tilfeldigheter ved utvalget (alle med navn som begynner på 's' og slutter på 'e' er veldig smarte.. i utvalget. Generaliserer ikke). Problem som kan løses", men legger begrensninger på løsning.

Annen utfordring er verre: Ikke representative data. Kan skyldes tilfeldigheter ved utvalg (denne risikoen kan i prinsippet kvantifiseres... men kan gi dårlig performance... mer data er bra). Skjevhet som ikke løses med mer data (seleksjonsproblem, respons bias). Generaliserer ikke til populasjon. Dårlig data (målefeil). Viktig med data cleaning og god

feature engineering... hvor mye modell klarer å lære avhenge av representasjon av data... teori + kryssvalidering.

Noen tanker fra Roger

Målet er å modellere assosiasjon[1] mellom input $\mathbf{x}$ og output y. Hvis relasjon så kan vi finne h slik at $y = h(\mathbf{x})$ men i praksis ikke mulig pga tre årsaker: manglende info, målefeil og ikke-determinisme.[2] All informasjon er i simultanfordeling; kan i praksis svare på alle spørsmål. I praksis kan vi ofte anta at $\mathbf{x}$ er kjent eller at vi driter litt i fordeling, så trenger bare betinget fordeling $Y|X$ til å svare på spørsmål. Dette er beskrevet av fordelingsfunksjon $g : \mathbb{R} \to [0, 1]$, men kan i utgangspunktet være vilkårlig komplisert og trenger funksjon for hver verdi av $\mathbf{x}$. Vanskelig å jobbe med selv om vi kjente fordeling og enda vanskeligere å lære fra begrenset data! Vil ha sammendragsmål som egenskap. I praksis forsøker vi å finne $E[y|\mathbf{x}]$. Har estimat $\widehat{E[y|\mathbf{x}]} = h(\mathbf{x})$. Kan da bruke h til å svare på spørsmål om populasjonen. Dette er en deterministisk funksjon så her er det bare å bruke kalkuluskunnskapene. På omtrent samme måte kan vi forsøke å estimere $V(y|\mathbf{x})$, men det er vanlig å anta at denne er konstant.[3]

Jeg synes det er en fin distinksjon mellom mekanisk og generativ perspektiv. Trenger generativ for kvantifisere usikkerhet. Bruker antagelser. Må sannsynliggjøre at de er rimelig, men det er forenklinger. Hjelper oss å svare på spørsmål. I økonometri er vi ute etter å svare på spesifikke spørsmål og vil helst unngå å gjøre antagelser som ikke er nødvendig for å svare på det gitte spørsmålet. Det gir litt ulike perspektiv og metodevalg.

Tenker at jeg har lyst til å knytte statistisk læring i større grad opp mot MLE siden det er vesentlig å kvantifisere usikkerhet. MLE er min tilnærming for å håndtere generativ tilnærming. Tror det hadde vært greit å få litt kobling mellom empirisk risikominimering og MLE hvis jeg kjører på dette.

Kan gå mekanisk til verks med tapsfunksjon, men vil knytte til generativ modell for å sikre generalisering og kvantifisere usikkerhet. Gir motivasjon for kryssvalidering. Litt usikker på om jeg kan koble kryssvaliderign til modellseleksjon i MLE, eller om det er nødvendig å bruke de andre informasjonskriteriene (AIC/..). Tror jeg tar anvendelse av MLE modellering for betinget fordeling under statistisk læring.

8.1 Hva er statistisk læring

Utgangspunktet vårt er at vi har et datasett med tall. Vi kan tenke på dette som et utvalg bestående N realiseringer fra en såkalt datagenereringsprosses (DGP). Målet vårt

[1] Bruker det bekrepet siden det ikke er en sammenheng eller tilknytning, men den er ikke eksakt

[2] Eksistens av sistnevnte er til dels et filosofisk spørsmål.

[3] Eller å behandle det som nuissance parameter og bruke heteroskedastisk robust estimator for standardfeil til helningskoeffisient som i økonometri.

er å prøve å lære noe om egenskapene til denne prosessen. Vi avgrenser til å betrakte en mengde av prosesser som oppfyller gitte egenskaper og kaller denne mengden $\mathbb{M}$ for modellen. For eksempel antar vi ofte at realiseringene er uavhengige.

Vi bruker estimatorer som er funksjoner av utvalget. Det gitte estimatet er bare én realisering og med nye utvalg ville vi fått andre tall. Vi vil at tyngden av denne utvalgsfordelingen skal være konsentrert rundt den sanne parameteren i DGP. Men hvordan kan man beregne utvalgsfordelingen med bare én realisering??

Ved å gjøre sterke antagelser og avgrense modellen vi betrakter kan vi utlede fordelingen analytisk for vilkårlig n. Ved å bruke asymptotisk teori kan vi være mer agonistisk og utlede resultater som holder eksakt når $n \to \infty$ og som gir god tilnærming i store utvalg. En tredje fremgangsmåte er å bruke bootstrap til å sample fra empirisk fordeling og på den måten observere empirisk fordeling til estimator.

Det er flere fordeler ved å modellere. Ved å påføre struktur reduserer vi hvor mye vi må lære fra data og kan dermed finne størrelsen mer presist. Anta for eksempel at vi vil finne $\mathbb{P}$. Hvis vi ikke kjør antagelser så må vi forsøke å estimere cdf på hele utfallsrommet. Hvis vi derimot antar at $\mathbb{P}_\theta$ så reduseres problemet til å finne θ og vi vil da kjenne cdf.

Når vi påfører struktur ved modellering kan resultatene være sensitive for om $DGP_0 \in \mathbb{M}$. Det kan derfor være lurt å teste i hvilken grad prosessen tilfredstiller antagelsen. Husk at modellen er en viljeløs golem som tar alt veldig bokstavelig og spytter ut tall stort sett uansett om de vi gjør er meningsfult eller ikke. Det krever derfor et kritisk blikk for å tolke modeller. Dessuten er det et poeng at modeller alltid er usanne, men dette er litt beside the point. Det er et verktøy for å lære noe om virkeligheten gjennom å forenkle det til noe vi kan forstå og håndtere.

Virkeligheten er komplisert og modellene er ofte veldig enkle. Det kan derfor være fristende å bruke mer fleksibel struktur som lar dataene snakke". Det er flere avveininger knyttet til dette. For det første kan det bli vanskeligere å beskrive og tolke den estimerte modellen. Dette er mindre problematisk dersom modellen skal brukes til prediksjon, men det kan uansett være interessant å bruke modellen til å lære om hvordan input er relatert til output. Selv om vi kun er interessert i best mulig prediksjon er ikke alltid mer fleksibilitet bedre. Problemet er at vi lærer *for mye* om utvalget, mens vi egentlig er interessert i egenskaper ved prosessen som genererte det.

Det signalet vi er interessert i å lære er ofte $E[Y|X] = f(X)$. For hver y_n i utvalget er $y_n = f(x_n) + u_n$. Vi vil lære $f(\cdot)$, men med men for mye fleksibilitet vil det i for stor grad også fange opp støyen u_n i det gitte utvalget. Dette er et eksempel på bias-variance-tradeoff. Modellen må være fleksibel nok til å fange signalet, men ikke så fleksibel at den fanger for mye støy. Det er et problem at vi ofte ikke vet hvilken struktur vi skal bruke apriori, men da er det veldig nice at kryssvalidering reduserer hele problemet til å tune/justere hyperparametre. Med prediksjon er *proof in the pudding*, så må ikke tenke så mye. Verre med kausalitet!

I statistikk er vi ofte interessert i $\mathbb{P}$. Vi avgrenser til en paramterisk klasse $\mathbb{P} \in \{\mathbb{P}_\theta : \theta \in \Theta\}$ og estimerer med MLE. Estimat av paramter gir da estimat på hele fordelingen. I økonometri er vi ofte kun interessert i parameter. Dessuten har vi ikke grunnlag for å gjøre antagelser om formen og vil ikke påføre unødvendig struktur som ikke er implisert av økonomisk teori. Derfor brukes i større grad (generaliserte) momentestimatorer.

Vi vil finne en funksjon f slik at $f(x) = y$. Dette gjør det mulig å vite y gitt at vi observerer x og vi kan beskrive relasjonen mellom variablene. I praksis er det som oftest ikke mulig å finne en slik eksakt, deterministisk funksjon. Dette skyldes for det første av vi kun observerer en delmengde av variablene som påvirker utfallet. Observasjoner som er like langs de observerte x kan være ulike langs andre dimensjoner og dermed ha ulik utfall. Hvis vi observerer flere variabler kan vi redusere usikkerheten, men det kan også være variabler som er fundamentalt uobserverbare, spesielt når vi analyserer menneskelig atferd. En annen kilde til usikkerhet er målefeil i variablene, slik at observerasjoner med samme observerte x kan ha ulik verdi av de reelle størrelsene som påvirker utfallet.

8.2 Empirisk risikominimering

Anta at vi observerer $(\mathbf{z_1}, ..., \mathbf{z_N})$ der $\mathbf{z_n} = (y_n, \mathbf{x_n})$ er realiseringer fra $\mathcal{L}(\mathbf{z}) = P$. Målet vårt er å predikere *output* y gitt at vi observerer *input* $\mathbf{x}$. Husk at vi kan dekomponere

$$y = f(\mathbf{x}) + u \tag{8.1}$$

slik at vi kan betrakte målet som å finne f som minimerer en norm av u. Vi begynner med å definere en tapsfunksjon som avhenger av størrelsen på avviket mellom vår prediksjon og faktisk verdi, $L(y, f(\mathbf{x}))$

- Kvadratisk tap: $L(y, f(\mathbf{x})) = (y - f(\mathbf{x})^2$
- Diskret tap: $L(y, f(\mathbf{x})) = I\{y \neq f(\mathbf{x})\}$

Dette er tilfeldige variabler som tar verdi for hver realisering av $\mathbf{z}$. Målet vårt er å minimere forventet tap som kan betegnes som prediksjonsrisikoen $R(f) = \mathbb{E}L(y, f(\mathbf{x}))$. Utfordringen er at vi ikke kjenner P slik at vi ikke kan evaluere prediksjonsrisikoen direkte. Vi må lære fra utvalget og bruker utvalgsanalogprinsippet til å motivere den empiriske risikoen

$$R_{emp}(f) \equiv \mathbb{E}_{\hat{P}_N} L(y, f(\mathbf{x})) = \frac{1}{N} \sum L(y_n, f(\mathbf{x_n})) \tag{8.2}$$

som proxy for prediksjonsrisikoen som vi egentlig vil minimere. En naiv tilnærming vil nå være å finne f som minimerer $R_{emp}(f)$, men dette vil ofte være en dårlig løsning. Dersom alle inputvektorene $\mathbf{x_n}$ tar ulike verdier vil det alltid være mulig å finne funksjoner f slik at $f(\mathbf{x_n}) = y_n, n = 1, ..., N$ og $R_{emp}(f) = 0$. Hvis vi minimerer empirisk risiko vil vi få en

funksjon som har lært *for mye* om det gitte utvalget vårt og generaliserer dårlig til nye observasjoner. Dette problemet kalles *overfitting*.

Det er mulig å vise at $f(\cdot)$ som minimerer forventet kvadratisk tap (MSE) er $\mathbb{E}(y|\cdot)$. Det gir en nedre grense på prediksjonsrisiko og vi kan betrakte målet vårt som å tilnærme oss denne funksjonen. Hver observasjon i utvalg kan derfor dekomponeres i signal og støy, $y_n = \mathbb{E}[y|x_n] + u_n$. Vi vil at funksjonen skal lære signalet og ignorere støyen. For å oppnå dette må vi påføre struktur som begrenser fleksibiliteten til f gjennom å modifisere tapsfunksjonen slik at det straffer wigglynesseller å avgrense mengden av funksjoner vi søker over. Eksempler på første strategi er LASSO og ridge som er alternative måter å estimere PRF. Jeg kommer tilbake til dette senere, men velger nå å se på valg av hypoteserom $\mathcal{H}$. Løsningen på det empiriske risikominimeringsproblemet kan nå uttrykkes som

$$\hat{f} = \arg\min_{f \in \mathcal{H}} R_{emp}(f) \tag{8.3}$$

der vi kan finne løsning $\hat{f}_j$ for hver $\mathcal{H}_j$. Hvordan velger vi hvilken $\mathcal{H}_j$ som er best? Vi kan ikke bruke empirisk risiko som mål fordi

$$\mathcal{H}_1 \subset \mathcal{H}_2 \implies R_{emp}(\hat{f}_1) \geq R_{emp}(\hat{f}_2) \tag{8.4}$$

Mer fleksibilitet vil alltid medføre at funksjonen kan lære mer fra data og få bedre *in-sample fit* og dermed lavere empirisk risiko. Målet vårt er derimot å generalisere til nye data og oppnå best mulig *out-of-sample fit*, altså prediksjonsrisiko. For å velge optimal struktur trenger vi en proxy for dette. Og det får vi med kryssvalidering som jeg kommer tilbake til senere. Kunne kanskje sagt litt om bias variance trade-off i valg av struktur siden dette er veldig generelt poeng, men det får bli en annen dag.

8.2.1 Dekomponering av risiko

Bias varians tradeoff. Projektering på underrom av L_2. Flytte greier fra fra første kapittel? Er greit skrevet, men litt out of place med min nye organisering. Vi får se.

8.3 hmm

Det beste vi kan gjøre er å finne en funksjon h slik at

$$y = h(x) + \epsilon \tag{8.5}$$

der ϵ er uavhengig av x. Det medfører at funksjonen h fanger opp all informasjon om verdi av y slik at det resterende feilleddet er uavhengig av x. Vi bruker da $\hat{y} = h(x)$ som

predikert verdi av y. Vi bruker forventet kvadrert avvik som mål på prediksjonsfeil, og siden dette er det beste vi kan oppnå er den såkalte *irreducible error*

$$E[(y - h(x))^2] = Var[\epsilon] \tag{8.6}$$

Jamført med diskusjon om projeksjon i L_2 er $h(x)$ projeksjonen av y på underromet som består av alle tilfeldige variabler som kan skrives som en deterministisk funksjon av x. Dette tilsvarer den betingede forventningsfunksjonen. I praksis så må vi estimere h fra realiserte verdier i utvalg. Vi finner da en annen $\hat{h}$ i underrommet. Den kvadrerte avstanden fra $\hat{h}$ til y er

$$E[(h(x) + \epsilon - \hat{h}(x))^2] = E[(h(x) - \hat{h}(x))^2] + Var[\epsilon] \tag{8.7}$$

der første ledd er prediksjonsfeilen vi kan ha håp om å redusere gitt x. Det er flere måter å vise denne dekomponeringen, men det følger av ortogonal projeksjon og pythagoras. Mye av statistisk læring handler om å finne best mulig $\hat{h}$. I praksis avgrenser vi oss ofte til å se på et underrom av mengden av tilfeldige variabler som kan skrives som funksjon av x; for eksempel alle lineære funksjoner. Biasen vil være avstand mellom $h(x)$ og $h^*(x)$ som er beste variabel i den delmengden. Det er i tillegg varians i estimeringen.

Utvalgsanalogen til MSE er

$$E_{P_N}[(y - \hat{h}(x))^2] \tag{8.8}$$

Den kan virke rimelig å minimere dette for å finne $\hat{h}$. Problemet er at dette er en forventningsskjev estimator av MSE og alltid vil foretrekke mer fleksible funksjoner som kan lære mønster i utvalget. Men målet vårt er ikke å memorisere utvalget! Målet er å predikere fremtidige data. Vi bruker derfor kryssvalidering til å estimere MSE: se hvor god jobb hypotesefunksjonen gjør på usette data. Vi vil da se at forholdet mellom MSE og fleksibilitet har en U-form. Bias reduseres og varians øker.

8.4 Modellseleksjon

Vi har mange ulike modeller/algoritmer for å gjøre prediksjoner. Hvordan skal vi velge? Det avhenger av hva målet vårt er og hva modellen skal brukes til. For det første har vi et mål på treffsikkerheten til prediksjonene. Hvordan vi definerer dette målet avhenger igjen av kontekst, men dette kan enkelt kvantifiserers og kan estimereres med kryssvalidering. I tillegg er det andre hensyn. Vi kan ha lyst til å beskrive sammenhengen mellom input og output. Hva er det som gjør at modellen gjør en gitt prediksjon? Hvor stor er usikkerheten til prediksjonen? Dette gjør at vi kan stole mer på modellen enn hvis jeg kommer med en black-box som spytter ut tall, der eneste begrunnelse for at det fungerer er høy accuracy

på mine test-data (selv om dette er bra utgangspunkt...). I mange tilfeller er det bra nok, men vi kan gjøre bedre.

Et eksempel på at det ikke alltid er tilstrekkelig med høy test accuracy på data vi har tilgjengelig er klassifikasjon mellom ulv og hund (eg. husky). Hvis bilde av ulv er i område med snø, så vil snø være input som gjør at modell kan oppnå gode prediksjon. Men da bygger vi en modell som er god til å oppdage snø og det er kanskje ikke så nyttig for oss.

Ved å se på sammenheng mellom input og output kan vi avdekke om det er såkalt irrelevante features (partikulære egenskaper ved de gitte dataene) som har stor forklaringskraft. Dette vil indikere at modellen ikke kommer til å generalisere så bra. Ved å analysere de uriktige prediksjonene kan vi også lære mer om hvilke nye variabler / features som kan være nødvendig for å oppnå bedre prediksjon.

8.4.1 Kryssvalidering

Vi vil finne funksjon $\hat{h}$ som minimerer prediksjonsrisiko

$$R(\hat{h}) = \mathbb{E}_P[L(y, \hat{h}(\mathbf{x}))] \tag{8.9}$$

som vi ikke kan evaluere siden vi ikke observerer P. Det er heller ingen god idé å evaluere det på den empiriske fordelingen $\hat{P}_N$ siden vi brukte den til å finne $\hat{h}$ og dermed bruker samme data to ganger. Hvis vi får nye data kan vi evaluere med hensyn på disse

$$\widehat{R(\hat{h})} = \frac{1}{j} \sum L(y_j, \hat{h}(\mathbf{x}_j)). \tag{8.10}$$

Dette vil ikke systematisk favoriesere med fleksible funksjoner som i større grad lærer mønsteret i det spesisifikke utvalget, men samtidig så sløser vi litt med data siden de nye observasjonen kunne vært brukt til å finne en bedre $\hat{h}$. En bedre tilnærming er K-fold-kryssvalidering som partisjonerer data $\mathcal{D}$ i K like store deler som vi angir med $\mathcal{D}_k$ $(k = 1, ..K)$. Algoritmen blir da:

1. for k in 1,...,K:
2. fit $\hat{h}$ på $\mathcal{D}_{-k}$
3. finn $\hat{R}_k = \frac{1}{|\mathcal{D}_k|} \sum_{n:n \in \mathcal{D}_k} L(y_n, \hat{h}(\mathbf{x}_n)$
4. end for, finn $\widehat{R(\hat{h})} = \frac{1}{K} \sum \hat{R}_k$

Målet vårt er å minimerer prediksjonsrisiko. I praksis har vi en mengde med kandidatfunksjoner $\hat{h}_m \in \mathcal{M}$. De kan for eksempel være estimert med ulike metoder (regresjon, knn, beslutningstrær mm), med ulike hypoteserom $\mathcal{H}$ eller fittet med ulike hyperparametre. Merk at selve funksjonene er tilfeldige fordi de avhenger av data $\mathcal{D}$. Vi er interessert i å

minimere prediksjonsrisiko gitt det konkrete utvalget vi har, så vi vil estimere $R(\hat{h}_m|\mathcal{D})$ for $m = 1, ...M$ og velge funksjonen med lavest estimert prediksjonsrisiko. Algoritmen blir da:

1. for m in 1,...,M:
2. Gjør K-fold på $\hat{h}_m$ og lagre $\widehat{R(\hat{h}_m)}$
3. velg m med lavest $\widehat{R(\hat{h}_m)}$

8.5 Lineær regresjon

I lineær regresjon avgrenser vi oss til å betrakte et hypoteserom $\mathcal{H}_l := \{h : h(\mathbf{x}) = \mathbf{x}'\mathbf{b}$ for noen $\mathbf{x} \in \mathbb{R}^K\}$. Vi skal se at dette faktisk er ganske fleksibelt siden vi kan gjøre vilkårlige transformasjoner av $\mathbf{x}$ før vi estimerer den linære funksjonen. Selv om regresjonsfunksjonen i praksis sjeldent er eksakt lineær kan vi uansett finne den beste lineære tilnærmingen.

8.5.1 Feature space

Vi kan gjøre en transformasjon

$$\Phi : \mathbf{x} \mapsto \Phi(\mathbf{x}) = \begin{bmatrix} \Phi_1(\mathbf{x}) \\ \vdots \\ \Phi_J(\mathbf{x}) \end{bmatrix} \tag{8.11}$$

og behandle $\Phi(\mathbf{x})$ som om det var inputvektoren.[4] De individuelle transformasjonene $\Phi_j(\cdot)$ betegnes som basistransformasjoner og verdimengden til $\Phi(\cdot)$ betegnes som feature space.[5] Dette gir oss et nytt hypoteserom

$$\mathcal{H}_\Phi = \{l \circ \Phi : \Phi : \mathbb{R}^K \to \mathbb{R}^J \text{ og } l\text{er lineær funksjon} l : \mathbb{R}^J \to \mathbb{R}\} \tag{8.12}$$

Kan empiriske risikominimeringsproblemet kan da skrives som

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^j}{\arg\min} \frac{1}{N} \sum (y_n - \gamma' \Psi(\mathbf{x}))^2 \tag{8.13}$$

[4] Hvis vi tolker de estimerte koeffisientene så vil vi ofte se på endring i forhold til opprinnelig input. Skal se på tolkning senere.

[5] I praksis er det ofte slik at basistrasnfomasjonen bare avhenger av verdi til én av komponenetene i inputvektoren, men kan avhenge av flere verdier eller ingen. Et grunnlegende eksempel er konstantledd, $\Phi(\mathbf{x}) = \begin{bmatrix} 1 \\ 1\mathbf{x} \end{bmatrix}$.

som vi løser for gitt transformasjon $\Phi(\cdot)$. I praksis er valg av $\Phi(\cdot)$ og dermed form på feature space vi søker over en viktig del av risikominimeringsproblemet. Det er bias-varians tradeoff og vi finner beste kandidat med kryssvalidering.

En mye brukt transformasjon er polynom for å modellere ikke-lineær sammenheng. Fra Weierstrass' theorem vet vi at kan oppnå vilkårlig god tilnærming av kontinuerlig funksjon med polynom av tilstrekkelig høy orden. Det kan derfor være tilforlatelig å gjøre en transformasjon

$$\Phi(x_n) = \begin{bmatrix} x_n^0 \\ x_n^1 \\ \vdots \\ x_n^J \end{bmatrix}, \quad \gamma'\Phi(x_n) = \sum_{j=0}^{J} \gamma_j x_n^j \tag{8.14}$$

men skal se at det finnes bedre måter å modellere ikke-lineær sammenheng dersom andre orden ikke er tilstrekkelig til å fange mønster.

8.5.2 Valg av featurespace

Målet vårt er å nærme oss y ved å finne en funksjon $h(\mathbf{x})$ som minimerer $\|u\|_{L_2}$. Minste kvadrats metode er konsistent[6] estimator av den lineære regresjonsfunksjonen $h^* := \arg\min_{h\in\mathcal{H}_l}\|u(h)\|_{L_2}$, der $\mathcal{H}_l := \{h(\cdot) : h(\mathbf{x} = \mathbf{x}'\beta\}$. Vet å forlenge $\mathbf{x}$ ved å legge til nye variabler kan vi utvide hypoteserommet $\mathcal{H}_l$ slik at vi i teorien kan komme nærmere y. Det kan dermed være fristende å tenke at flere input-variabler alltid er et gode og kaste hele kjøkkenskapet inn i algoritmen. Problemet er at variansen øker og at det generalisere dårlig til nye data.

Det er problem med høy varians dersom antallet features er stort i forhold til antallet observasjoner. Hvordan skal vi gå frem for å velge featurespace? I økonometri bruker vi teori/argument til å velge kontrollvariabler som gjør at *conditional independence assumption* er oppfylt.[7] I statistisk læring er målet vårt enklere å måle fra data og vi kan bruke algoritme til å finne hvilken kombinasjon som generaliserer best. [8].

1. Kan teste alle mulige kombinasjoner, men det blir fort ganske mange... 2^K kombinasjoner

2. Greedy algoritme (forward/backward selection). I hvert steg legger til variabel som fører til størtst reduksjon i RSS, deretter ta kryssvalidering.

[6] og forventningsrett? Fortsatt ikke skjønt hvorfor ikke forventningsrett..

[7] Det betyr i praksis at for gitt $\mathbf{x}$ så vil ikke observert eksponert behandling si oss noe om potensielle utfall ved andre eksponeringer. Dette er ganske vilkårlig. Tror det er gjort arbeid for å automatisere valg av kontrollvariabler i setting med høydimensjonal data (greiene til Athey).

[8] Som målt ved kryssvalidering i henhold til metric eller andre kriterier som pålegger straff for fleksibilitet (justert R^2, AIC, BIC,...). Vet ikke hvordan jeg utleder disse kriteriene eller hvorfor jeg skulle ønske å bruke det over kryssvalidering.

Jeg synes fremgangsmåten over virker ganske slitsom. Bedre å kaste alt inn og la regulariseringsparameter ta seg av problemet.

8.5.3 Regularisering

Jeg tenker at regularisering går ut på å glatte/flate ut funksjonen. Vi vil ta hensyn til at det er sannsynlighet for utfall som som vi ikke observerer i det partikulære utvalget som vi trener modellen på, og at utfallet til verdi vi ikke observerer sannsynligvis er ganske like de nærmeste naboene vi observerer. Vi oppnår dette ved å straffe høye koeffisientverdier.

Vi kan se litt intuisjon for dette ved å betrakte to features som er høyt korrelert. Ettersom prediksjon er vektet gjennomsnitt av features så kan vi oppnå samme $\hat{y}$ ved å skalere opp den éne koeffisienten og ned den andre. Dette kan gi bedre føyning i utvalget, men gjørt at både $\hat{y}$ og koeffisient blir ustabile. Vi kan gjøre det mer stabil ved å presse begge mot null. Dette oppnår vi ved å legge til et straffeledd i tapsfunksjonen. Har tre ulike måter i lineær regresjon.

Ridge

Kostnadsfunksjonen er

$$C(\theta) = MSE(\theta) + \alpha \sum_{k=1}^{K} \theta_k^2 \tag{8.15}$$

$$= MSE(\theta) + \alpha \mathbf{w}'\mathbf{w} \tag{8.16}$$

Merk at vi ikke straffer konstantleddet. Må standardisere features før vi regularisere slik at det ikke avhenger av måleenhet. Finnes en closed form løsning, men vet ikke hvor interessant det er.[9]

Lasso

Kostnadsfunksjoen er

$$C(\theta) = MSE(\theta) + \alpha \sum_{k=1}^{K} |\theta_k| \tag{8.17}$$

Bruker L_1-norm i stedet. Vet ikke hvordan jeg kan skrive det på matriseform. Fordelen med denne regulariseringen er at det setter koeffisienter lik 0 slik at den velger ut de relevante featurene. Her er marginalgevinsten ved å redusere koeffisient konstant. Kan også illustrere forskjell mellom ridge og lasso ved å sette det opp som betinget optimeringsproblem.

[9] Kan motivere at det har gode numeriske egenskaper som gjør inverse mer stabil eller noe sånt.

Elastic net

Kostnadsfunksjon har vektet gjennomsnitt av de to ulike regulariseringene,

$$C(\theta) = MSE(\theta) + \alpha \sum_{k=1}^{K} \theta_k^2 + (1-r)\alpha \sum_{k=1}^{K} |theta_k| \tag{8.18}$$

hmhm.

8.6 Andre regresjonsmetoder

8.6.1 Splines

8.6.2 Ikke-paramerisk regresjon

I lineær regresjon antar vi at $E[y|\cdot] := f(\cdot)$ er kjent opp til ukjent parameter. Vet ikke hvordan den ser ut på forhånd, men kan tilnærme arbritrære kontinuerlige funksjoner med basistransformasjoner og bruke kryssvalidering til å vurdere ekstern validitet. Dessuten har jeg sett av vi kan bruke splines til å partisjonere inputrommet slik at vi får mer fleksibilitet til å fange opp lokale sammenhenger. Sånn sett er parametriske metoder rimelig fleksible samtidig som de i prinsippet er mulige å tolke gjennom den parametriske representasjonen.

Har noen ikke-parametriske metoder som også kan være relevant å bruke i økonometri, men tror relevansen er avgrenset til regresjonsdiskontinuitet der det viktig å fange eksakte funksjonelle relasjonen. Metodene generaliserer veldig dårlig til input i flere dimensjoner; både fordi det er vanskelig å kommunisere den funksjonelle formen dersom den ikke kan visualiseres grafisk og fordi dimensjonalitetens forbannelse gjør at lokale metoder fungerer dårligere.

K nærmeste naboer

Kernelmetoder

Sieve?

8.6.3 Kvantilregresjon

Jeg tenker at den τ'te kvantilen til en variabel y med cdf F er gitt ved $Q(\tau)$ der

$$\tau = F(Q(\tau)) \implies Q(\tau) = F^{-1}(\tau) \tag{8.19}$$

Vi må utvide definisjonen til å håndtere at ikke alle F er monotont voksende slik at den inverse ikke er definert på hele verdimengden til F.

$$Q(\tau) = \inf\{t : F(t) \geq \tau\} \tag{8.20}$$

Dette er den vanlige definisjonen, men vi kan også definere det som løsningen på et minimerings problem som involverer forventningsverdi av parametrisert funksjon av y. Skal da se at vi kan få det inn i ERM-rammeverket og kan estimere kvantiler fra utvalgsanalog.

$$L_\tau(y, \xi) = |(y - \xi)(\tau - I\{y < \xi\})| \tag{8.21}$$

$$= piecewise \tag{8.22}$$

Kan vise at

$$Q(\tau) = \min_\xi \mathbb{E}[L_\tau(y, \xi)] \tag{8.23}$$

8.7 Klassifikasjon

I klassifikasjon er y kategorier som vi kan kode med tall, $y \in G = \{1, ..., K\}$. Siden tallene bare er kode for kategori kan vi ikke bruke de numeriske verdiene til å si noe om avstanden eller rangering av kategorier. Vi kan likevel behandle klassifikasjon innenfor rammeverket med empirisk risikominimering og vi skal se at metodene er ganske analoge til regresjon siden vi ofte vil modellere den betingede sannsynligheten som en kontinuerlig funksjon av $\mathbf{x}$. Begynner med å anta at det er to kategorier $(0, 1)$ og spesifiserer tapsfunksjon

$$L(h) = I\{h(\mathbf{x}) \neq y\} \tag{8.24}$$

$$R(h) = \mathbb{E}L(h) = \mathbb{P}\{h(\mathbf{x}) \neq y\} \tag{8.25}$$

$$R_{emp} = \mathbb{E}_{\hat{P}_N} L(h) = \frac{1}{N} \sum I\{h(\mathbf{x}) \neq y\} \tag{8.26}$$

Dette er eksempel på tapsfunksjon som gir feilrate. Vi kan velge andre tapsfunksjoner avhengig av målet vårt, som gjerne avhenger av kostnad ved ulike typer feil. For å operasjonalisere dette må vi finne en funksjon $h(\cdot)$, der $h(x) \in G$. En naturlig fremgangsmåte er å gjøre dette i to steg: estimere betinget sannsynlighet for de ulike kategoriene som funksjon av x, $p(x)$, og deretter bruke dette til å klassifisere:

$$h(x) = I\{p(x) > k\} \tag{8.27}$$

der parameteren k er treshold. Trenger tapsfunksjon for å lære $p(\cdot)$ og bruker såkalt logistisk tap som vi kan utlede fra loglikelihood-funksjonen til betinget bernoulli.[10] Hvis målet er å minimere tapsfunksjonen over (feilrate), så plasserer vi observasjonen i kategori med høyest betinget sannsynlighet og velger da $k = 0.5$. Dette impliserer en partisjonering av inputmengden i delmengder som predikerer ulike kategorier og grensene mellom delmengdene er decision boundary

$$D(h) = D(p, k) = \{x : p(x) = k\} \tag{8.28}$$

som både avhenger av den estimerte betingede sannsynligheten og valg av treshold. Mer generelt så vil $p(\cdot; \theta) : \mathbb{R}^K \to \mathbb{R}$ angi sannsynlighet for positiv kategori for hver inputvektor. Kan tegne nivåkurver i inputrommet.[11]

8.7.1 Flere kategorier

Kan utvide til flere kategorier. Med mange av modellene er det veldig greit å utvide dersom vi gjør ren prediksjon og bare vil minimerer error rate

$$h(\mathbf{x}) = \arg\min_k \mathbb{P}(y = k|x) \tag{8.29}$$

$$\arg\min_k f_k(\arg\min_k)\pi_k \tag{8.30}$$

Det er litt vanskeligere dersom vi skal gjøre inferens og se på sammenheng mellom input og betinget sannsynlighet for output. Tror jeg ser mer på dette i økonometri...

8.7.2 Logit

Med to kategorier kan variabelen y ta to verdier. Den er da nødvendigvis bernoulli-fordelt. Fordelingen kan være betinget av $\mathbf{x}$ slik at parameteren p er en funksjon av $\mathbf{x}$ og det kan være ulike bernoulli-fordelinger for de ulike $\mathbf{x}$-verdiene, $y|\mathbf{x} \sim bernoulli(g(\mathbf{x}))$. Denne funksjonen g må tilfredstille $g(\mathbf{x}) \in [0, 1], \forall \mathbf{x}$. I praksis vil vi paramterisere funksjonen med $g(\mathbf{x}'\beta)$ slik at vi kan si noe om hvordan betinget sannsynlighet endrer seg når vi endrer input. Vi trenger deretter en funksjon g som transformerer tallinjen til $[0, 1]$. Kumulative fordelingsfunksjoner har denne egenskapen, og de to vanlige valgene er

$$g(\mathbf{x}) = \begin{cases} \Phi(z) = \int_{-\infty}^{x} (2\pi)^{0.5} exp\{-s^2/2\} ds \\ \Lambda(z) = \frac{e^z}{1+e^z} \end{cases} \tag{8.31}$$

[10] Det er poeng at tapsfunksjon ikke alltid samsvarer med denne risikoen (metric) vi vil optimalisere. Dette er fordi det er kan være vanskelig å operasjonalisere risiko direkte (finne gradient mm.). Det blir litt to-stegs prosess som ikke er like ryddig som i enkel regresjon.

[11] Litt usikker på hvordan jeg håndterer det med flere kategorier. Tenker at hver kategori får sin egen funksjon $p^{(k)}$...

der første kalles probit og andre logit. Kan utvide til flere kategorier ved å lage egen parametervektor for hver kategori, $\theta^{(k)}, k = 1, .., K$. Den predikerte sannsynligheten for at input $\mathbf{x}_n$ tilhører kategori j er da

$$\hat{p}_j = \frac{\exp\left(\theta^{(j)}\mathbf{x}_n\right)}{\sum_{k=1}^{K}\exp\left(\theta^{(k)}\mathbf{x}_n\right)} \tag{8.32}$$

Dette har visstnok noe med *softmax* og *cross entropy* å gjøre, men det må bli annen dag.[12]

8.7.3 LDA og QDA

Bayes-regel gjør at vi kan estimere betinget sannsynlighet på en annen måte:

$$f(y = k|\mathbf{x}) = \frac{f(\mathbf{x}|y = k)\mathbb{P}\{y = k\}}{f(\mathbf{x})} \tag{8.33}$$

$$= \frac{f_k(\mathbf{x})\pi_k}{\sum_j f_j(\mathbf{x})\pi_j} \tag{8.34}$$

$$\propto f_k(\mathbf{x})\pi_k \tag{8.35}$$

denne metoden er enklere å bruke på flere kategorier og dette må jeg si litt om.. Uansett, må nå i stedet velge parametrisk klasse til $\mathbf{x}|y$ og får kvadratisk discriminant analysis hvis jeg antar at $\mathbf{x}|y = k \sim N(\mu_k, \Sigma_k)$ og lineær hvis jeg i tillegg antar at $\Sigma_k = \Sigma, \forall k$. Med utgangspunkt i dette kan jeg finne decision boundary som funksjon av enkle størrelser som jeg kan estimere. Skal si mer om dette senere.

8.7.4 Sammenheng mellom Logit og LDA/QDA

Jeg tror jeg kan utlede begge metodene med MLE. Vil si noe om hvilken simultanfordeling f som gjør det mest sannsynlig å observere de realiserte verdiene $\{(\mathbf{x}_n, y_n) : n = 1, ..., N\}$, som jeg har i utvalget mitt. Vil unngå å parametrisere hele simultanfordelingen direkte siden jeg er interessert i betinget sammenheng og dette er et enklere problem. Bruker at

$$f(\mathbf{x}, y) = f(y|\mathbf{x})f(\mathbf{x}) = f(\mathbf{x}|y)f(y) \tag{8.36}$$

Oppgave: hvordan er de parametrisert?

8.7.5 KNN

En ikke-parametrisk metode for å estimere betinget sannsynlighet. Bruker relativ av kategorier i nabolag til $\mathbf{x}$ som estimat på betinget sannsynlig, der nabolaget $N_K(\mathbf{x})$ består

[12] Også et poeng at vi kan kjøre log-reg som one-versus-all, men det skal i teorien være mulig å tolke $\theta^{(k)}$ fra multinomial... må prøve å få dette operativt i økonometri-delen, tror Cameron og Trivedi er best på dette.

av K observasjoner med minst $\|\mathbf{x}_n - \mathbf{x}\|$.

$$\hat{P}(y = j|\mathbf{x}) = \frac{1}{k} \sum_{n \in N_K(\mathbf{x})} I\{y_n = j\} \tag{8.37}$$

Hvis målet er å minimere feilrate blir klassifiseringsregelen h å predikere kategori som er mode i nabolag.

8.7.6 Naiv bayes

For å implementere bayes-regel kan vi estimere

$$f_k(\mathbf{x})\pi_k. \tag{8.38}$$

Har sett at lda/qda gir en måte å gjøre dette på. Den metoden har ganske sterke parametriske antagelser. Vil bruke svakere antagelser, men problem å estimere betinget simultanfordeling $f_k(\mathbf{x})$. Blir mye enklere dersom vi antar at de er uavhengige slik at

$$f_k(\mathbf{x}) = \prod_j f_{kj}(x_j) \tag{8.39}$$

mer om dette senere.

8.7.7 Support vector machines

8.7.8 Beslutningstrær

Beslutningstrær er en fleksibel og transparent metode som kan brukes til både regresjon og klassifikasjon. Metoden går ut på å partisjonere inputmengden og bruke mode eller gjennomsnitt i hver delmengde som predikert kategori for observasjon med input der. Formelt finner vi R_j der

$$\cup_{j=1}^{J} R_j = \mathcal{X}, \quad R_j \cap R_k = \emptyset, j \neq k \tag{8.40}$$

og predikert verdi kan representeres parametrisk i lineær modell som

$$\hat{y}_n = \sum_j I\{\mathbf{x}_n \in R_j\}\hat{\beta}_j \tag{8.41}$$

der $\hat{\beta}_j = avg(\{y_n : \mathbf{x}_n \in R_j\})$. Partisjonering kan representeres med et tre som er en special case av en graf.[13] Treet består av nodes og koblinger mellom nodes som vi kan

[13] Må ha litt formell definisjon av graf som jeg får fra algoritme/datastrukturer. Poeng at det har nodes og verticies.

betegne som greiner (eller *branches*). Det begynner i root-node og splitter i to eller flere child nodes ut i fra verdi av en variabel. Nodes som ikke har childs betegnes som blader (eller *leaves*). Bladene på treet utgjør den endelige partisjoneringen og det er dette vi bruker til å gjøre prediksjoner.

For nye inputs kan vi bevege oss gjennom treet. Dette gjør estimatoren transparent siden vi ser hvilke inputs som fører til hvilke inputs. Mer spesifikt kan vi både se hvilke inputs som er viktige for å forklare forskjeller i observert kategori (tidlig split) og hvilken retning det påvirker predikert sannsynlighet for kategori. Vi kan observere predikert sannsynlighet i hver node og se hvordan den endres mens vi beveger oss i treet gjennom å gi gradvis mer informasjon om input til observasjon.

En nedside med beslutningstrær er at de er veldig sensitive for treningsdata. Små endringer i data den blir opplært på kan få store konvekvenser for partisjoneringen (som bestemmer decision boundary og prediksjoner). Dette har til dels sammenheng med at den lager rektangulære partisjoner som er ortogonalt på aksene. Vi skal senere se at vi kan glatte ut boundaries ved å kombinere mange trær i en såkalt tilfeldig skog. Først skal vi se litt kort på algoritme for å konstruere hvert enkelt tre.

Algoritme for å konstruere trær

I hver node så søker vi over alle mulige cut-points for å finne partisjonering som fører til størst mulig reduksjon i såkalt *impurity*. Vi vil at labels i hver delmengde skal være mest mulig homogene. De to vanligste målene på impurity er *gini* og *entropy*. I praksis bruker vi en greedy algoritme som søker over grid og kalkulerer reduksjon i impurity for hver punkt i grid og bruker dette til å partisjonere. Deretter gjøres dette rekursivt helt til det ikke lenger er mulig å redusere impurity (alle inputs har enten samme features eller samme labels) eller til treet har nådd en spesifisert grense for dybde.

Jeg tror det er best å vokse ut hele treet og deretter trimme (*prune*) det ex-post for å fjerne oppdelinger som ikke fører til bedre fit out-of-sample. I praksis tror jeg vi bruker maks-dybde som regularisering selv om dette ikke er optimalt...

8.7.9 Neurale nettverk

8.8 Ensemble

Vi kan oppnå bedre prediksjoner ved å kombinere flere estimatorer. Litt av intuisjonen bak dette er at idiosynkratiske feil jevner seg ut når vi tar gjennomsnitt.[14] Dette argumentet bygger på at det er variasjon i prediksjonene til de ulike estimatorene. Den enkleste måten

[14] Kan koble til forsikring... Selv om hver enkelt estimator kun predikere riktig kategori i 51% av tilfellene vil andel som predikere riktig konvergere i sannsynlighet mot 51% slik at majoriteten tar riktig i 100% av tilfellene dersom de er uavhengige.

å oppnå dette er å bruke algoritmer til å trene opp estimatorer på treningsdata og deretter bruke en avstemming til å predikere nye inputs. Vi kan da enten bruke simpel majoritet (såkalt hard voting) eller vi kan vekte ut i fra den predikerte sannsynligheten til de ulike estimatorene (såkalt soft voting). I praksis bruker vi to andre fremgangsmåter for å skape variasjon: vi sampler fra treningsdata for å skape variasjon i data eller vi trener estimatorer sekvensielt der det blir lagt større vekt på observasjonene som ble feilpredikert av forrige estimator.

8.8.1 Bagging og tilfeldig skog

Bagging er kort for bootstrap aggregering. Ved å sample med replacement fra treningsdata så sampler vi fra empirisk fordeling. Dette innfører litt mer bias, men ved å ta gjennomsnitt av estimatorene så kan vi redusere varians. Fremgangsmåten er spesielt egnet for beslutningstrær siden de er sensitive for treningsdata. I praksis bruker vi derfor såkalte tilfeldig skog som er baggete beslutningstrær med litt ekstra triks for å oppnå mer varians. Det er for eksempel vanlig å avgrense mengden av variabler den kan bruke til å partisjonere til en tilfeldig delmengde.

Den tilfeldige skogen er litt mindre transparent enn et enkelt beslutningstre siden vi ikke kan følge hvordan input beveger seg langs greinene, men vi kan få et mål på feature importance ut fra gjennomsnittlig bidrag til reduksjon i impurity fra tra trærene i skogen. Det er også en fordel at trærene kan blir trent opp parallelt.

8.8.2 Boosting

Dette er en alternativ fremgangsmåte der estimatorene blir trent opp sekvensielt og som dermed ikke kan paralleliseres. I stedet for å bruke mange unbiased estimatorer med høy varians så forsøker vi å sekvensielt redusere bias ved å legge mer vekt på observasjonene som ble feilpredikert av forrige estimator. Det finnes i hovedsak to fremgangsmåter for å booste: Ada(ptive)Boost og gradient boosting.

Adaptive Boosting

Eksplisitt endringer av vekter i kostnadsfunksjon, må gjør algoritmen formell en annen gang.

Gradient Boosting

Fitter residual av forrige, vet ikke om det fungerer på klassifikasjon.

8.8.3 Stacking

Kan også forsøke å lære den beste mulige måten å kombinere de ulike estimatorene... Hard vs soft voting osv; alt kan læres og valideres...

8.9 Vurderingskriterier

8.9.1 Confusion matrix

I klassifikasjon er ikke vurderingskriteriet like entydig som i regresjon. Det mest intuitive kriteriet er *accuracy* som angir andelen av observasjoner som blir plassert i riktig kategori, men dette er ofte ikke et godt mål på hvor egnet modellen er. For det første kan vi med ubalanserte kategorier oppnå høy treffsikkerhet ved å alltid predikere majoritetskatogrien som ikke er så nyttig. Dessuten er det ofte ulike kostnader assoiert med ulike *typer* feil. Vi kan kategoriesere ulike typer feil gjennom en en såkalt *confusion matrix* som deler inn observasjoner ut fra faktisk kategori og predikert kategory. Med binær klassifikasjon gir dette fire muligheter og vi bruker dette til å lage vurderingskriterier

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{TN}+\text{FN}} \tag{8.42}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8.43}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{8.44}$$

Anta nå at vi bruker algoritmen til å diagnostisere om personer har en gitt sykdom (f.eks. covid). Presisjon til algoritmen angir andelen av de som tester positiv som faktisk er smittet. Denne kan vi få arbritært høy gjennom å kun diagnostisere de som helt klart er syke. Recall (sensitivitet) angir derimot andelen av de som faktisk er syke som tester positivt. Hvis testen er lite sensitiv så er det mange syke som vil gå under radaren. Vi kan igjen få denne arbritært høy ved å si at alle som tester seg er syke.

Isolert sett gir disse kriteriene ikke noe godt mål siden vi kan lage rimelig trivielle algoritmer som maksimerer kriterium uten å være nyttig. Et alternativ kan være å ta et gjennomsnitt av presisjon og sensitivitet. F1-score tar harmonisk gjennomsnitt. [15] Dette kan gi et greit sammendragsmål for å sammenligne ulike algoritmer, men i praksis er det bedre å undersøke tradeoff mellom presisisjon og sensitivitet for å finne den beste balansen til vårt formål.

[15] litt usikker på hvorfor vi ikke tar artimetisk snitt. Harmonisk vekter slik at det blir større straff dersom én av de er lav..

8.9.2 Presisjon vs Recall trade-off

Algoritme gir gjerne et mål på hvor sikker den er at en observasjon tilhører en kategori.

$$\widehat{P(y_n = 1}|\mathbf{x}_n) = \hat{p}_n \tag{8.45}$$

$$\hat{y} = I\{\hat{p}_n \geq k\} \tag{8.46}$$

ved å øke *treshold* k kan vi øke presisjon og redusere recall. Vi kan visualisere dette gjennom å tegne $(k, Pres(k))$ og $(k, Recall(k))$ i et diagram. Det er mer nyttig å tegne output fra $f : k \mapsto (Pres(k), Recall(k))$ som angier en såkalt *mulighetskurve* med recall vi kan oppnå for gitt presisjon. Hvis kurven er bratt så må vi gi opp masse recall for å oppnå litt mer presisjon. Vi kan bruke denne kurven til å finne k som korresponderer med balansen av presisjon og recall som vi foretrekker. Kan også tegne kurver fra ulike algoritmer. Ulike algoritmer kan ha ulik performance på ulike deler av kurven. Så dersom vi er veldig opptatt av å ha f.eks. over 90% recall, så kan vi se hvilken algo som oppnår høyest presisjon i det intervallet.

Et annet mye brukt mål er den såkalte *ROC-kurven* som plotter True Positive Rate (?) vs False Positive Rate (?). Igjen kan vi undersøke kurven eller bruke areal under kurven (AUC) som et sammendragsmål for valg av algoritme og treshold k.

8.10 Annet

Algoritme består av tre deler

1. Representasjon av egenskaper den lærer. Hypoteserom: kandidater av funksjoner h den søker over
2. Evaulering: mapper hver kandidat til et mål på fit. F.eks: $g : \beta \mapsto RSS(\beta)$ i lineær regresjon
3. Optimisering: må ha måte å effektivt søke over hypoteserommet for å finne gode kandidatfunksjoner

Må si noe om representasjon av data. Feature selection, dimensjonalitet,.. Her eier neurale nettverk på data som ikke har opplagt struktur.. lager egne features (informative representasjoner). Curse of dimensionality: vanskelig å finne nabo i høy dimensjon, påvirker algoritme som bruker avstand.

Kapittel 9

Læring uten tilsyn

Data uten labels.

9.1 Dimensjonalitetsreduksjon

Vi kan ønske å finne en lavere dimensjonal representasjon av data som bevarer mest mulig informasjon.[1] Hva som utgjør informasjon har ikke en eksakt definisjon og tenker at det avhenger litt av kontekst.[2] En måte å redusere dimensjonene er å droppe variabler som vi anser som irrelevante. Vi skal nå se på nye fremgangsmåte som i stedet konstruerer nye variabler som fanger opp informasjon fra de eksisterende variablene, slik at vi kan finne lavere dimensjonal representasjon som bevarer mest mulig info.

9.1.1 Dimensjonalitetens forbannelse

Det kan være en utfordring å jobbe med høydimensjonal data. Problemet er størrelsen på rommet vokser veldig raskt når vi øker antall dimensjoner slik den gjennomsnittlige avstanden mellom observasjone også øker. Dette medfører at vi må ekstrapolere kurver til områder av rommet der vi har lite informasjon slik at det blir fort gjort å overfitte. Det medfører også at nabolaget ikke er så veldig lokalt og metoder som bygger på avstand mellom observasjon ikke fungerer så bra.[3]

9.1.2 Principal component analysis

Metode for å finne et d-dimensjonalt underrom for datasett med k variabler der $d \leq k$. Vi vil bevare mest mulig informasjon. For hver d finner vi derfor underrommet som bevarer

[1] Når vi klassifiserer siffer så er mange av pixlene hvite for alle bildene slik at de ikke er så informative. Vi kunne droppet disse og få færre variabler (dimensjoner) per observasjon uten at vi taper info.

[2] Ren unsupervised eller preproccesing i supervised... hvilke variabler er informative i en regresjon liksom.

[3] Tror litt av grunnen til at neurale nettverk fungerer så bra på høydimensjonal data er at det klarer å lære meningsfull representasjon i lavere dimensjon...

mest mulig av variansen. Dette er ekvivalent med å at det minimerer MSE av projeksjon av data ned på underrommet.[4] Det som er veldig fint er at greedy algoritme også gir optmal løsning: kan finne k principale vektorer $[\mathbf{v}_1, ..., \mathbf{v}_k]$ og for $d < k$ så velger vi bare delmengde som består av d første.

For å finne disse principale komponentene kan vi bruke singulærverdidekomposisjon som er en måte å faktorisere en matrise,

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}' \tag{9.1}$$

der $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_k]$. Vi finner projektering ned på underrommet som er utspent av komponentene med

$$\hat{\mathbf{X}}_d = \mathbf{X}\mathbf{V}_d \tag{9.2}$$

der $\mathbf{V}_d = [\mathbf{v}_1 \dots \mathbf{v}_d]$. Vi kan også forsøke å gjøre invers transformasjon for å forsøke å gjenskape det opprinnelige datasettet med

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_d \mathbf{V}_d' \tag{9.3}$$

Veldig kjekt at vi kan finne andel av forklart variasjon til hver komponent slik at vi kan plotte dette og bruke til å bestemme hvor mange dimensjoner vi trenger til å representere informasjonen i datasettet.

9.1.3 Andre metoder

Kan bruke noe kernel eller manifold ... for noen representasjoner er det vesentlige greier som går tapt når vi projekterer på underrom.

9.2 Clustering

Metoder som forsøker å konstruerer clusters og plassere observasjoner i kategori. Vil at de skal være homogene innad og ha distinksjon mellom andre cluster. For gitte cluster blir det litt sånn som klassifikering bare at vi ikke kjenner tolkning til cluster-label.. Kan skille mellom hard cluster og soft cluster, der sistnevne beregner sannsynlighet for at observasjon tilhører de ulike clusterene.

[4] Tror jeg kan vise dette med dekomponering av varians... vil knytte til ting jeg kan om prosjektering fra før, men blir litt anderledes siden jeg nå projekterer en matrise i stedet for vektor...

9.2.1 K-means

Finner K *centroids* og angir label til observasjon ut fra nærmeste centroid. Algoritmen er enkel: bruk en tilfeldig initialisering, label data og oppdater plassering av centroids slik at det er i midten av observasjonene som ble angitt til den centroiden. Deretter angi labels ut fra oppdatert posisjon til centroids og fortsett slik til det konvergerer. Det er en utfordring at det kan konvergere mot lokal minimum som ikke er globalt optimalt, men dette kan vi håndtere ved å kjøre flere ganger og velger det som minimerer avstand innad i clusterene.

Den større utfordringen er valg av K. Vi har noen vektøy for å gjøre informert valg om dette: såkalt intertia og silhouette, men må se på dette en annen gang.

9.3 Tetthetsestimering

9.3.1 Histogram

Vi observerer realiseringer $X_1, ..., X_N$ på utfallsrom $Z = [0, 1]$ fra en fordeling med tetthet f. En veldig naiv tilnærming er å bruke relativ andel av hver observasjon som estimat på fordelingen, $\hat{f}(x) = \frac{1}{n} \sum I\{X_n = x\}$. Det vil jo vanligvis være sannsynlighet for verdier vi ikke observerer i det gitte utvalget vårt, og dette gjelder spesielt siden vi antar at den sanne fordelingen er kontinuerlig. Histogram gir en litt bedre tilnærming. Vi finner en partisjonering av Z som er en mendre $\{B_1, ..., B_K\}$ der $\cup B_k = Z$ og $B_k \cap B_j =$ for $k \neq j$. Dette er bins med lengde $1/K$. Sannsynligheten for å få utfall i hver bin er $p_k = \int_{B_k} f(x)dx$ og estimator er $\hat{p}_k = \frac{1}{N} I\{x \in B_j\}$. Dette gir en diskontinuerlig funksjon på Z som vi kan skrive som $\hat{f}(x) = \sum_k \hat{p}_k I\{x \in B_k\}$.

Hyperparameteren i histogrammet er antall bins som også bestemmer lengden på intervallene. Hvis det er for få bins klarer det ikke fange mønsteret i den sanne tettheten (høy bias), mens for mange bins gjør at estimat fra ulike utvalg blir veldig forskjellige (høy varians). Dette kan man til en viss grad ta på øyemål, men analogt til estimasjon av regresjonsfunksjon/betinget sannsynlighet kan vi definere en tapsfunksjon og forsøke å velge antall bins som minimerer risiko. Tar det senere.

En nedside med histogram er at det gir en diskontinuerlig funksjon. Skal nå se en måte å utlede kontinuerlig tetthetsfunksjon.

9.3.2 Kernel density estimation

Intuisjonen er at vi vil spre litt tetthet rundt de observerte realiseringene, der tyngden avtar mer avstand. For å gjøre bruker vi kernel funksjon $K(\cdot)$ med egenskaper $\int K(x)dx = 1$, $\int xK(x)dx = 0$, $K(x) = k(-x)$. Det er en tetthetsfunksjon som jeg vil at skal være symmetrisk. Sentrerer hver av kernelfunksjone på de realiserte verdiene, tar summen og

skalerer med $1/N$ slik at vi får ut en tetthet

$$\hat{f}(x) = \frac{1}{N} \sum K(x - X_N) \tag{9.4}$$

Kan også skalere med bandwith h for å justere spredningen til de individuelle tetthetene som dermed påvirker i hvilken grad tyngden til den estimerte tetthetsfunksjonen er konsentrert på de observerte verdiene i utvalget.

$$\hat{f}(x) = \frac{1}{Nh} \sum K\left(\frac{x - X_N}{h}\right) \tag{9.5}$$

Det er ikke opplagt for meg hvorfor vi deler på h og hvordan vi generaliserer til flere dimensjoner, så må se på det senere. Kan også ta diskusjon om dimensjonalitetens forbannelse. I høyere dimensjon trenger vi veldig mange observasjoner for å få god estimat siden "areal"av Z vokser raskt..

Vi har sett på måter å finne $\gamma(P)$, for eksempel en paramterer θ til fordelingen. Noen ganger er vi interessert i å estimere selve fordelingen P. En mulighet er å avgrense til en parametrisk klasse $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Hvis modellen er riktig spesifisert og vi konsistent kan estimere θ så kan vi også konsistent estimere P. Men hvordan definere konvergens av tetthetsfunksjoner og hva skjer hvis modell er feilspesifisert? For å gjøre denne diskusjonen mer presis må vi definere en norm på tetthetsfunksjoner. Kan definere en L_p norm på funksjoner

$$\|f\|_p := \left(|f|^p)^{1/p} := (\int |f(\mathbf{s})|^p d\mathbf{s}\right)^{1/p} \tag{9.6}$$

Dette gir et mål på avvik mellom funksjoner

$$d(f, g) = \|f - g\|_p. \tag{9.7}$$

En følge med tilfeldige tetthetsfunksjoner $(\hat{f}_N)_{N \in \mathbb{N}}$ er L_p konsistent for f hvis

$$\|\hat{f}_N - f\|_p \xrightarrow{p} 0 \tag{9.8}$$

når $n \to \infty$. Dersom modellen er feilspesifisert slik at $P_0 \notin \{P_\theta : \theta \in \Theta\}$ så har avviket en nedre begrensning

$$\delta(P_0, \hat{P}_\theta) = \inf_{\theta \in \Theta} P_0 - \hat{P}_\theta \|\|_p \tag{9.9}$$

Kernel density estimators gir en alternativ fremgangsmåte for å estimere tetthetsfunksjonen under svakere antagelsen (ie. ikke avgrenset til spesifikk parametrisk klasse). Estimatoren er en den skalerte summen av N tetthetsfunksjoner som hver er sentrert på de

observerte $\mathbf{x}_n$ $(n = 1, ..., N)$. Det er en såkalt *bandwidth* som justerer spredningen på de individuelle tetthetsfunksjone og dermed glattheten (*smoothness*) til summen. Med lavere bandwith blir større del av tyngden konsentrert på rundt de observerte verdiene. Formelt kan vi skrive estimatoren som

$$\hat{f}_N(\mathbf{s}) = \frac{1}{Nh^p} \sum K\left(\frac{\mathbf{s} - \mathbf{x}_n}{h}\right) \tag{9.10}$$

Kapittel 10

Økonometri

For meg er økonometri synonymt med programevaluering.[1] Vi bruker data til å estimere effekt av behandling på utfall til individer.[2] For å kvantifisere dette vil vi ideelt sett observere utfallene til hvert av individene i en verden der de blir eksponert for behandling og i en verden uten behandling. Dessverre er dette umulig siden kun ett av tilfellene kan intreffe, og vi kan dermed aldri kvantifisere individuelle behandlingseffekter. I praksis beregner vi gjennomsnittlige behandlingseffekter ved å sammeligne forskjell i gjennomsnittlig utfall til en behandlingsgruppe og en kontrollgruppe. For at denne observerte forskjellen skal gi et godt mål på behandlingseffekten må kontrollgruppen være en god *proxy* for det kontrafaktiske utfallet til behandlingsgruppen dersom de ikke ble eksponert for behandling.

10.1 Programevaluering

Det er rimelig å betrakte kontrollgruppen som en proxy dersom de to gruppene er omtrent like bortsett fra at den éne ble eksponert for behandling. I så fall kan observerte forskjeller i utfall tilskrives denne ene dimensjonen der gruppene er forskjellig.[3] Det sentrale spørsmålet i programevaluering er hvorvidt dette er rimelig antagelse med de dataene som foreligger i analysen. Ettersom mange variabler som påvirker utfallet er uobserverte (og uobserverbare) kan det ikke testes med de gitte dataene, og må i stedet sannsynliggjø-

[1] Økonometri omfatter også greier med (makroøkonomiske) tidsserier og estimering/kalibrering(?) av parametre i økonomiske modeller, men disse greiene vet jeg lite om. Det er forøvrig andre fagfelt som holder på med programevaluering og det er overlapp i problemstillinger og faglige tilnærminger. Det som kjennetegner økonometri er bruk av såkalte naturlige eksperiment og metoder for å analysere disse. I biostatistikk bruker de mer kontrollere eksperiment. Andre har mer naiv tilnærming for å isolere kausal effekt gjennom matching/justering for andre observerte egenskaper.

[2] Terminologi stammer fra medisinske eskperiment. Det som betegnes som behandling kan være andre former for tiltak og reformer. Enhetene som blir eksponert for disse kan være aggregerte størrelser som foretak.

[3] Noe som også impliserer at det ikke ville vært observerte forskjeller dersom de ikke fikk ulik eksponering for behandling. Med begrensede utvalg vil det alltid være litt tilfeldig variasjon, men dette abstraherer vi stort sett vekk fra når vi diskuterer kausalitet.

res gjennom en beskrivelse av hvordan data er generert. Vi skal nå se på tre ulike typer beskrivelser av hvordan eksponering for behandling er bestemt.

Den første kategorien er tilfeldige eksperimenter der eksponering for behandling blir bestemt av forskere i henhold til en randomiseringsregel.[4] Ettersom eksponeringen er tilfeldig vil det ikke være systematiske forskjeller mellom gruppene langs noen egenskaper, hverken observerte eller uobserverte. Tilfeldige eksperiment regnes som gullstandard i programevaluering siden antagelsen om at kontrollgruppen er god proxy er veldig troverdi. Det er likevel vesentlige begrensinger ved slike eksperiment. Det er mange interessante kausale spørsmål som ikke kan besvares på denne måten, enten fordi det er praktisk umulig, for dyrt eller uetisk. Det er også mange mange praktiske utfordringer, og det kan argumenteres for at de kan ha begrenset ekstern validitet. På tross av dette er det gjennomført en del eksperiment i stor skala og det har blitt gradvis mer fremtredende også i økonometri.[5]

En alternativ fremgangsmåte er å utnytte at ytre omstendinger kan skape variasjon i behandling selv om det ikke er planlagt som et tilfeldig eksperiment.[6] Det faktum at individene ikke selv velger egen eksponering for behandling gjør det ofte mer kredibelt at det ikke er systematiske forskjeller i uobserverte egenskaper. Et eksempel på en slik situasjon er at egenskaper ved institusjoner at eksponering for en behandling er bestemt ved om individ havner over eller under en noe abritrær *cut-off*.[7] I den grad det er vanskelig for individ å strategisk velge side så blir behandling som om tilfeldig fordelt i populasjonen i nærheten av cut-off. Dersom hele grupper blir eksponert for ulike behandlinger avhengig av geografi (fordi ulike policy på ulik sted) eller fødselsalder (fordi endring i policy som rammer personer født etter gitt dato) har vi også verktøy for å sammenligne forskjeller og vurdere i hvilken grad det skyldes effekt av ulik eksponering for behandling. Vi kan også håndtere omstendigheter som skaper noe variasjon i behandling uten å bestemme det eksakt. Ved hjelp av såkalte instrumentelle variabler kan vi i store utvalg isolere variasjonen i behandling som skyldes den ytre omstendigheten.[8]

I fravær av slik eksogen variasjon kan det være fristende å stratifisere observasjonsdata og aggregere forskjell mellom behandling og kontroll innad i hvert strata. Selv om behandlings- og kontrollgruppen samlet sett er systematisk forskjellig kan vi konstruere delutvalg der individer med ulik eksponering for behandling er omtrent like langs andre observerte egenskaper. Vi kan da estimere behandlingseffekter innad i hvert delutval og

[4]Kan være tilfeldig på hele utvalget eller tilfeldig innad i strata definert av observert egenskap, for eksempel kjønn.

[5]Eksempler på store eksperiment i er STAR som undersøkte effekt av klassestørrelse på barns utfall og noe greier med effekt av helseforsikring på pasientenes utgifter i USA. Eksperimenter er viktig i atferdsøkonomi og har blitt viktig del av utviklingsøkonomi.

[6]Tror vi betegner det som eksogen variasjon.

[7]Noen eksempel er karakterkrav for å komme inn på skole og helsetiltak som avhenger av nyfødt barns vekt.

[8]Liker dårlig denne formuleringen siden IV i praksis bare er skalering av redusert form..

forsøke å aggregere dette til en gjennomsnittlig behandlingseffekt i populasjonen. Denne fremgangsmåten kan motiveres med at individer som er like langs observerte egenskaper forhåpentligvis også er ganske like langs uobserverte egenskaper som kan påvirke utfallet. Problemet er at det alltid er en grunn til at individene velger ulik eksponering for behandling innad i hvert strata og det er lite kredibelt at denne grunnen ikke også påvirker utfallet.[9] For å publisere i gode tidskrift er det nødvendig å ha et forskningsdesign som isolerer eksogen variasjon i behandling. Ellers regnes det som lite troverdig at kontrollgruppen er proxy for det kontrafaktiske utfallet til behandlingsgruppen i fravær av behandling, slik at de observerte forskjellene i utfall ikke samsvarer med kausal effekt av behandling.[10]

Jeg skal nå utlede et rammeverk som formaliserer idéen om kontrollgruppe som proxy.

10.1.1 Potensielle utfall

Vi har nå et rammeverk som lar oss beskrive relasjon mellom variabler og estimere dette fra data. Denne relasjonen består både av en eventuell kausal relasjon mellom variablene og spuriøs korrelasjon som følge av andre variabler som er korrelert med både utfall og forklaringsvariabler. Den kausale effekten kan defineres som differansen i de potensielle utfallene med og uten behandling. Det grunnleggende problemet er at kun én av tilstandene blir realisert for hver observasjon. Vi innfører notasjonen

$$y_i = y_i^0 + D_i(y_i^1 - y_i^0) \tag{10.1}$$

Det er ikke mulig å estimere individuell kausal effekt, men vi kan forsøke å estimere gjennomsnittlig effekt for en avgrenset populasjon ved å se på differansen i utfall til de som blir eksponert for behandlingen og kontrollgruppen som ikke blir eksponert. Intuisjonen bak denne sammenligningen er at utfallet til kontrollgruppen gir en proxy for det kontrafaktiske utfallet til behandlingsgruppen slik at den observerte differansen tilsvarer differanse i potensielle utfall. Uten randomisering vil observert differanse bestå av både kausal effekt og seleksjonsskjevhet.

$$E[y_i|D_i = 1] - E[y_i|D_i = 0] = E[y_i^1|D_i = 1] - E[y_i^0|D_i = 1] \tag{10.2}$$

$$+E[y_i^0|D_i = 1] - E[y_i^0|D_i = 0] \tag{10.3}$$

[9] Individer er sånn omtrent rasjonelle og vi kan betrakte eksponering for behandling som løsning på et optimeringsproblem. Det er lite rimelig at forskjellene bare er tilfeldig. Det kan skyldes ulike preferanser: de som spiser vitaminer større preferanse for 'sunnhet' og vil gjerne dermed være sunnere uavhengig av eventuell behandlingseffekt. Eller kanskje de kompenserer for usunt kosthold. Uansett: vanskelig å isolere behandlingseffekt fra andre systematiske forskjeller.

[10] Dette er til dels en konsekvens av den såkalte kredibilitetsrevolusjonen.

For at denne naive sammenligningen mellom behandling og kontroll skal isolere kausal effekt trenger vi randomisering av behandling. Dette sikrer at potensielle utfall er uavhengig av behandling. Sagt på en annen måte; observert behandling gir oss ikke noe informasjon om kontrafaktisk utfall.

$$(y_i^1, y_i^0) \perp\!\!\!\perp D_i \implies E[y_i^j | D_i] = E[y_i^j], j = 0, 1 \tag{10.4}$$

Intuisjonen er at randomisering gir oss en eple-til-eple sammenling siden gruppene i forventning er like langs alle dimensjoner, inkludert uobserverbare, slik at differanse kan isoleres til å skyldes ulik eksponering for behandling. I praksis er det slik at folk velger eksponering for behandling som del av et optimeringsproblem. Som resultat er gruppene med ulik eksponering ulike slik at vi ikke får eple-til-eple sammenligninger og kan ikke isolere kausal effekt. En mulig strategi er å *kontrollere* for observerte egenskaper. Vi kan tenke på dette som en stratifisering av utvalget der vi tar vektet gjennomsnitt av test-kontroll sammenligninger for observasjonene med samme observerte egenskaper. Dette vil avdekke den kausale effekten dersom observasjoner som er like langs observerte dimensjoner også er like langs uobserverte dimensjoner, slik den obseverte eksponeringen for behandling ikke gir oss noe informasjon om potensielle utfall. Formelt er antagelsen

$$(y_i^1, y_i^0) \perp\!\!\!\perp D_i | X \implies E[y_i^j | D_i, X] = E[y_i^j | X], j = 0, 1 \tag{10.5}$$

Vi kan da bruke matching eller regresjon til å estimere denne effekten, noe jeg skal se på senere. I praksis er dette som oftest lite troverdig siden det er en grunn til at observasjonene i hver kategori valgte ulik behandling og det er lite troverdig at dette ikke også påvirker potensielle utfall. Vi trenger derfor et forskningsdesign som skaper tilfeldig (eksogen) variasjon i behandling. Vi kan analysere variasjon i behandling D direkte dersom vi har randomisert eksperiment eller vi kan se på variasjonen som skyldes et instrument Z. En viktig kilde til eksogen variasjon er såkalte *naturlige eksperiment*. En viktig kilde til slike eksperiment er kunnskap om institusjonelle regler... som vi skal analysere med regression discontinuity. Senere skal jeg også se på paneldata som lar oss kontrollere for uobservert heterogenitet som er konstant over tid.

10.1.2 Knytte regresjon til potensielle utfall

For å knytte dette til kausalitet kan vi bruke potensielle utfall der parameter kan korrespondere med (gjennomsnittlig) kausal effekt. Anta først at $y_i^1 - y_i^0 = \delta$.

$$y_i = y_i^0 + \delta D_i \tag{10.6}$$
$$= E[y_i^0] + \delta D_i + y_i^0 - E[y_i^0] \tag{10.7}$$
$$= \alpha + \delta D_i + u_i \tag{10.8}$$
$$\tag{10.9}$$

Ser nå at feilleddet har konkret innhold og hvis vi nå sier at $cov(D_i, u_i) = 0$ så er dette en veldig sterk antagelse som impliserer at $cov(y_i^0, D_i) = 0$.

10.1.3 Dårlig kontroll

Av og til kan det være fristende å kontrollere for variabler som er bestemt etter behandlingen. Dette er som oftest en dårlig idé siden behandling endrer sammensetning av undergruppene vi ser på slik at forskjellene i utfall ikke kan tilskrives en kausal effekt av behandlingen. Anta at myndighetene innfører et tiltak der et tilfeldig utvalg får gratis personlig trener ($D_i = 1$) og samtidig observerer utfallene til en kontrollgruppe ($D_i = 0$). Vi kan se på gjennomsnittlig effekt av tiltaket på ulike utfall y ved å beregne $E[y_i|D_i = 1] - E[y_i|D_i = 0] = E[y_i^1 - y_i^0]$. Det er ikke nødvendig å betinge for andre variabler, men det kan øke presisjon til estimat og vi kan også bruke stratifisering til å undersøke heterogentitet i behandlingseffekt. Anta nå at vi vil undersøke effekt av tiltaket på undergruppen av observasjoner som trener etter at tiltaket blir iverksatt ($T_i = 1$). Denne beslutningen kan avhenge av D_i så vi kan skrive det opp i potensielt utfall rammeverk,

$$y_i = y_i^0 + D_i(y_i^1 - y_i^0) \tag{10.10}$$
$$T_i = T_i^0 + D_i(T_i^1 - T_i^0) \tag{10.11}$$

Finner differanse i gjennomsnitt i undergruppe,

$$E[y_i|D_i = 1, T_i = 1] - E[y_i|D_i = 0, T_i = 1] \tag{10.12}$$
$$= E[y_i^1|T_i^1 = 1] - E[y_i^0|T_i^0 = 1] \tag{10.13}$$
$$= E[y_i^1|T_i^1 = 1] - E[y_i^0|T_i^1 = 1] \tag{10.14}$$
$$+ E[y_i^0|T_i^1 = 1] - E[y_i^0|T_i^0 = 1] \tag{10.15}$$

der siste linje er seleksjonseffekt. I dette tilfelle vil seleksjonseffekten sannsynligvis være negativ siden gruppen som trener uavhengig av eksponering av tiltak gjerne har bedre utfall en gruppen som trener på tiltak. Dette skaper seleksjonseffekt selv om behandling

i utgangspunktet var tilfeldig fordelt. Analogt så bør man være forsiktig med å legge til utfallsvariabler som kontroll i regresjoner også på observasjonsdata.

10.1.4 Målefeil

10.1.5 Utelatte variabler

Jeg har lyst til å undersøke sammenhengen mellom antall år med skolegang s og inntekt y. Det finnes ingen deterministisk funksjon som forklarer relasjonen siden personer med lik skolegang kan ha ulik lønn av andre grunner. Vi setter derfor opp en modell

$$y = \alpha + \beta s + \epsilon \tag{10.16}$$

der det stokastiske feilleddet ϵ blir et mål på vår uvitenhet. Koeffisientene i ligningen over er ikke entydig bestemt siden vi får enhver $f(\cdot)$ kan definere $\epsilon = y - f(\cdot)$. Generelt så vil vi finne $f(\cdot)$ som minimerer $\|\epsilon\|$ og jeg har vist over at dette er CEF. På en annen side vil vi ha en enkel funksjon med parametre som vi kan tolke. Jeg har derfor avgrenset til å se på lineære funksjoner som er parametrisert med β. Ved å påføre restiksjonen $cov(s, \epsilon) = 0$ korresponderer parameteren i ligningen over med PRF som er beste lineære tilnærming til CEF. Dette kan vi konsistent estimere med OLS og gir oss et greit sammendragsmål på *assosiasjonen* mellom skolegang og inntekt. Gitt at CEF er tilnærmet lineær vil β kunne tolkes som differansen i forventet inntekt mellom to individer med ett års differanse i skolegang. Denne differansen fanger både opp en eventuell kausal effekt av skolegang på inntekt og at individer med ulik skolegang er systematisk forskjellig på måter som påvirker inntekt. Det kan for eksempel tenkes at individ som velger lengre utdanning har høyere evner og motivasjon som vi kan benevne som a. Jeg skal nå undersøke mulighet til å isolere kausal effekt av skolegang og undersøke tolkningen av parametre i regresjonsmodeller.

Vi har formelt definert kausal effekt som differanse i potensielle utfall. For å muliggjøre estimering av kausale effekter fra observasjonsdata der behandling ikke er tilfeldig kan det være lurt å beskrive den kausale prosessen som genererer utfallet. I et laboratorium kan vi i noen sammenhenger beskrive eksakt hvordan et utfall avhenger av egenskaper ved eksperimentet, $y = f(\mathbf{x})$. På grunn av målefeil kan det være små avvik fra sammenhengen. Hvis dette er tilfeldig støy så er $y = f(\mathbf{x}) + \epsilon$ og $E[y|x] = f(\mathbf{x})$ slik at forventningsverdien gjenfanger den deterministiske, kausale sammenhengen. Det reduserer problemet til å estimere CEF. Virkeligheten er langt mer komplisert, men vi kan tenke på det som Guds laboratorium. I utgangspunktet kan vi tenke at fremtidige utfall er deterministisk bestemt av en fullstendig beskrivelse av verdens tilstand på et tidspunkt. La oss tenke på hva som bestemmer personers lønn.

$$y = f(\mathbf{z}) = \delta s + \gamma a + \beta' \mathbf{x} + \epsilon \tag{10.17}$$

Vi antar at den deterministiske $f(\cdot)$ eksisterer, men den kan være vilkårlig komplisert. Det er bare fantasien som setter grenser. Jeg har forenklet den ved å anta at inntekt avhenger lineært av skole, evne og noen andre variabler $\mathbf{x}$. På et eller annet tidspunkt må vi nesten slutte å liste opp variabler. Samler sammen bidraget til resten av variablene i et tilfeldig støyledd ϵ, der $E[\epsilon|s, a, \mathbf{x}] = \alpha$. Anta nå at vi observerer $(s, a, \mathbf{x})$. Ved å kjøre regresjon

$$y = \alpha + \delta s + \gamma a + \beta'\mathbf{x} + \epsilon \tag{10.18}$$

kan vi gjenfinne hele den kausale sammenhengen. Regresjon er verktøy for å estimere assosiasjon mellom variabler, men i dette tilfellet samsvarer det med den kausale sammenhengen. Anta nå at vi ikke kan observere $\mathbf{x}$. Vi legger det kumulative bidraget fra disse variablene inn i feilleddet, slik at den kausale sammenhengen nå er

$$y = \delta s + \gamma a + u \tag{10.19}$$

der $u := \beta'\mathbf{x} + \epsilon$. Hvis vi kjører regresjon

$$y = \alpha + \delta s + \gamma a + u \tag{10.20}$$

så vil OLS bestemme parametrene ved å konstruere et feilledd $\hat{u} := y - \hat{\alpha} + \hat{\delta}s + \hat{\gamma}a$ som er ukorrelert med s og a. Dette vil kun samsvare (asymptotisk) med de kausale parametrene dersom feilleddet u fra den kausale prosessen faktisk er ukorrelert med disse variablene. [11] Anta nå at $cov(s, u) = 0$ og $cov(a, u) \neq 0$. Kan vi fortsatt finne den kausale effekten av skolegang (δ)? Dette skjer dersom feilleddet konstruert ved OLS er ukorrelert med s. OLS lager et nytt feilledd $\hat{u}$.. hm... må se litt mer på dette.

$$\hat{u} = \phi a + v, cov(a, v) = 0 \tag{10.21}$$

slik at

$$y = \alpha + \delta s + \phi a + v, \quad cov(s, v) = cov(a, v) = 0 \tag{10.22}$$

Ser at vi fortsatt finner δ ,men ikke γ. Generelt kan vi bare håpe å finne en kausal parameter og er ikke noe problem om de andre variablene er korrelert med feilledd.

Hva skjer dersom vi ikke observerer a, men likevel prøver å estimere kausal effekt fra observerte (y, s)? Vi kan vise at PRF i kort regresjon ikke samsvarer med kausale parameteren.

$$\frac{cov(s, y)}{var(s)} = \frac{cov(s, \alpha + \delta s + \phi a + v)}{var(s)} = \delta + \phi\frac{cov(s, a)}{var(s)} \tag{10.23}$$

[11] Merk at feilleddet u har en konkret tolkning utover å bare være det mekaniske avviket $y - f(\cdot)$.

Merk at OLS alltid konsistent estimerer PRF, men PRF i den korte regresjonen ikke samsvarer med den kausale parameteren og at OLS derfor er forventningsskjev estimator for den parameteren vi egentlig er interessert i. Når det er selvseleksjon til behandling er det lite troverdig at vi kan obsevere alle *confounding variables*. Vi trenger derfor en kilde til eksogen variasjon i behandling som gjør den ukorrelert med alle disse andre variablene. Dette bringer oss til instrumentelle variabler.

10.1.6 Matching

Matching er strategi for å estimere behandlingseffekt ved å konstruere undergruppe med samme covariates, finne forskjell i gjennomsnittlig utfall til behandling og kontroll innad i hver undergruppe og aggregere forskjellene ved å finne et vektet gjennomsnitt av forskjellene. La x være en diskret variabel og anta at CIA er oppfylt slik at $E[Y_i^j|D_i, X_i] = E[Y_i^j|X_i]$, $j = 0, 1$. Matching er fint siden vi kan knytte det direkte til CIA og finne estimator som er enkel utvalgsanalog.

$$\begin{aligned}
\delta_{ate} &= E(y_i^1 - y_i^0) && (10.24)\\
&= E[E(y_i^1 - y_i^0|x)] && (10.25)\\
&= E[E(y_i^1|, D_i = 1, x) - E(y_i^0|D_i = 0, x)] && (10.26)\\
&= E[E(y_i|, D_i = 1, x) - E(y_i|D_i = 0, x)] && (10.27)\\
&= E[\delta_x] && (10.28)\\
&= \sum_x \delta_x P(X = x) && (10.29)
\end{aligned}$$

Åpner for heterogenitet i behandlingseffekt avhengig av covariates (hvilken undergruppe). Vi er interessert i gjennomsnittlig behandlingseffekt for hele populasjon så gir større vekt til større undergrupper. Kan tilsvarende finne gjennomsnittlig behandlingseffekt for de som blir behandlet ved å i stedet vekte på betinget fordeling i stedet for marginal,

$$\delta_{att} = \sum_x \delta_x P(X = x|D = 1) \tag{10.30}$$

Kan finne utvalgsanalog til forventningene for å evaluere. Matching er greit å gjøre operativt dersom vi har et fåtall veldefinerte undergrupper, men hva hvis covariate er kontinuerlig? Hva hvis inndeling blir for fin slik at mange grupper der vi ikke både oppserverer behandling og kontroll? Kan delvis håndteres ved å minimere avstand. For hvert inidivid kan vi finne kontroll individ(er) som er mest mulig lik bortsett fra behandlingsstatus og

aggregere opp individuelle behandlingseffekter,

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{i:d_i=1} \left(y_i - \sum_{j \in N(i)} w_{ij} y_i \right) \tag{10.31}$$

der $N(i)$ er indeksene til observasjon i et nabolag til observasjon i og det vektene w_{ij} avhenger av avstand til observasjon. Summerer til én slik at det blir et vektet gjennomsnitt.

Propensity score matching

I stedet for å matche på $\mathbf{x} \in \mathbb{R}^k$ kan vi modellere sannsynlighet for at observasjon mottar behandling, $E[D_i = 1|\mathbf{x}] := p(\mathbf{x}) \in \mathbb{R}$, og matche på dette. Tror det kan forenkle problemet litt og det er i mange tilfeller enklere å modellere hvordan $\mathbf{x}$ påvirker sannsynlighet for behandling enn utfallet. På en annen side er fremgangsmåten litt mer *non-standard*. Det er ulike valg av vekting, konstruering av feilledd mm. slik at konklusjon kan avhenge av valg til forsker. Noe av fordelen med regresjon er at alt er standardisert!

Regresjon som matchinng

Hvis vi har modell er saturert i diskret x og antar homogen behandlingseffekt,

$$y = \sum_x d_{xi} \alpha_x + \delta_R D + u_i, \tag{10.32}$$

så kan det vises at

$$\delta_R = \frac{E[\sigma_D^2(X_i)\delta_X}{E[\sigma_D^2(X_i)} \tag{10.33}$$

der $\sigma_D^2(X_i) := E[(D_i - E[D_i|X_i)^2|X_i]$, betinget varians av D_i gitt undergruppe X_i. Dette betyr at regresjon - siden det er type effektiv estimator som utnytter informasjon - legger mer vekt på grupper der det er større variasjon i behandling. Med dummy behandling vil den legge mer vekt på grupper der $P(D = 1, X = x) \approx 0.5$.. i stedet for å bare se på størrelsen av gruppen. Har ikke så mye å si dersom behandlingseffekt er omtrent homogen, men hvis regresjon og matching gir veldig ulike resultat kan det være grunn til å tenke litt på hvorfor.

10.2 Instrumentelle variabler

Instrumentelle variabler ble først benyttet til å estimere parametre i simultane ligningssystem. Jeg begynner med å beskrive dette fordi mye av terminologien stammer derfra. I praksis brukes det som oftest til å håndtere problemet med utelatte variabler. Det er

enklest å motivere det i tilfeldige eksperimenter med delvis *coompliace*. Vi kan deretter bruke det til å analysere naturlige eksperimenter.

10.2.1 Simultane ligningssystem

Det klassiske eksempelet på simultant ligningssystem er tilbuds- og etterspørselskurven. Pris og kvantum i marked blir bestemt i samspill av tilbudskurve og etterspørselskurve,

$$Q^s = \alpha_0 + \beta_0 P + \gamma_0 z + u_0 \tag{10.34}$$

$$Q^d = \alpha_1 + \beta_1 P + u_1 \tag{10.35}$$

$$Q = Q^s = Q^d \tag{10.36}$$

der z er observerbar variabel som skifter kurven og vi antar at helning er konstant over tid.*Eventuelt kan vi betrakte det som en gjennomsnittlig helning og det blir litt analogt til heterogen behandlingseffekt...* Vi sier at tilbuds- og etterspørselskurven er strukturelle ligninger der parameterne at en kausal tolkning.[12] Konkret så sier det oss endring i henholdsvis tilbudt og etterspurt kvantum dersom vi endrer pris med én enhet og holder alt annet likt. Slike strukturelle sammenhenger er ofte ikke mulig å observere fra tilgjengelig data, så da må de nødvendigvis komme fra teori. Jeg skal nå se på muligheten til å lære (β_0, β_1) fra data. Utfordringen vår er at P er *endogen*.[13].

$$Q = \alpha_0 + \beta_0 P + \gamma_0 z + u_0 \tag{10.37}$$

$$Q = \alpha_0 + \beta_0[(Q - \alpha_1 - u_1)/\beta_1] + \gamma_0 z + u_0 \tag{10.38}$$

$$Q = \frac{1}{1 - \frac{\beta_0}{\beta_1}} \left[\alpha_0 - \frac{\beta_0}{\beta_1}\alpha_1 + \gamma_0 z + u_0 - \frac{\beta_0}{\beta_1} u_1 \right] \tag{10.39}$$

$$Q = \pi_0 + \pi_1 z + v \tag{10.40}$$

... det er intuitivt at vi ikke kan ha eksogen variasjon i P siden det også påvirker etterspørsel, men ser ikke med én gang korrelasjon i feilledd. Merk at siste ligning er såkalt *redusert form* fordi vi skriver en *endogen* variabel som funksjon av *eksogene* variabler og parametre. De ulike parametrene i redusert form er ikke-lineære funksjoner av underliggende strukturelle parametre og har ikke interessant tolkning i seg selv. Utfordringen er å lære atferdsparameterne fra enkeltligninger og det er her IV kommer inn.

Grafisk så gir det mening av vi kan bruke z til å lære β_1 fordi den skifter tilbudskurven opp og ned. Vil knytte dette til IV estimator.

[12] Vi kan også si at de inneholder såkalte *atferdsparametre* som sier noe om endring i atferd dersom vi endrer egenskap ved systemet

[13] Litt usikker på om endogen har en presis definisjon. Kan tenke at det er korrelert med feilledd, men endogenitet kan jo også henspille på at variabelen blir bestemt innenfor systemet...

10.2.2 Estimering

Har generelt to kriterier for at en variabel Z skal være et gyldig instrument for D

1. Relevans: $cov(D, Z) \neq 0$
2. Eksogenitet: $cov(D, \eta) = 0$

Skal vise at det lar oss identifisere strukturell parameter selv med utelatte variabler.

$$cov(y, Z) = cov(\alpha + \delta D + \eta, Z) = \delta cov(D, Z) \implies \delta = \frac{cov(y, Z)}{cov(D, Z)} \tag{10.41}$$

Kan også motivere det som en momentestimatorer på matriseform.

$$\mathbb{E}[\mathbf{z}\eta(\beta)] = \mathbb{E}[\mathbf{z}(y - \mathbf{x}'\beta] = \mathbf{0} \implies \beta = \mathbb{E}[\mathbf{x}\mathbf{z}']^{-1}\mathbb{E}[\mathbf{z}y] \tag{10.42}$$

Til slutt kan vi utlede estimatoren med utgangspunkt i simultant ligningssystem som beskrevet over. Vi har en strukturell ligning med en kausal parameter som vi vil estimere

$$y = \alpha + \delta D + \eta \tag{10.43}$$

I første omgang tar jeg ikke med kontrollvariabler, men jeg kan utvide senere. Jeg tenker at motivasjonen for å inkludere disse er litt analogt til kontrollvariabler i vanlig regresjon. For det første kan det øke presisjonen til estimatoren ved å redusere feilleddet. Videre kan det gjør antagelsen om eksogenitet mer kredibel ved at instrumentet er så godt som tilfeldig fordelt innenfor hvert strata (delgruppe). Betingelsen om relevans blir da et spørsmål om *partiell kovarians*; det vil si at det korrelasjon mellom instrument og behandling innenfor strata. Dette kan vi undersøke med helningskoeffisient i multivariat regresjon. Uansett, vi kan ikke lære parameter ved å konstruere residual som er ukorrelert med D i utvalget fordi D er *endogen*. Vi modellerer eksponering for behandling i såkalt *first stage*,

$$D = \pi_{00} + \pi_{01} z + v_0 \tag{10.44}$$

som ikke er en strukturligning. Dette beskriver bare deskriptiv sammenheng i data og det er derfor slik at $cov(z, v_0 = 0$ per konstruksjon. Det er et verktøy for å finne kausal parameter. Vi finner redusert form ved å plugge first stage inn i strukturligningen,

$$y = \alpha + \delta[pi_{00} + \pi_{01} z + v_0] + \eta \tag{10.45}$$

$$y = \alpha + \delta pi_{00} + \delta\pi_{01} z + \eta + \delta v_0 \tag{10.46}$$

$$y = \pi_{10} + \pi_{11} z + v_1 \tag{10.47}$$

der $\pi_{11} := \delta\pi_{01}$. Merk at $cov(z, v_0) = 0$ per konstruksjon, men trenger også at $cov(z, \eta) = 0$

for at $cov(z, v_1) = 0$ slik at vi kan lære π_{11}.[14] Kan nå se to nye strategier for å lære δ,

1. Kjøre first stage og redusert form separat, $\delta = \pi_{11}/\pi_{01}$
2. Plugge $\hat{D} = pi_{00} + \pi_{01} z$ inn for D i strukturligning og kjøre den.

Utlede wald som special case

$$y_i = \alpha + \delta D_i + \eta_i \tag{10.48}$$

betinger på instrumet

$$E[y_i|Z_i = 1] - E[y_i|Z_i = 0] = \tag{10.49}$$

$$\delta(E[D_i|Z_i = 1] - E[D_i|Z_i = 0]) + E[\eta_i|Z_i = 1] - E[\eta_i|Z_i = 0] \tag{10.50}$$

$$\implies \delta = \frac{E[y_i|Z_i = 1] - E[y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} \tag{10.51}$$

gitt at $E[\eta_i|Z_i = 1] - E[\eta_i|Z_i = 0] = 0$ og $E[D_i|Z_i = 1] - E[D_i|Z_i = 0] \neq 0$.

10.2.3 Heterogen behandlingseffekt

Vi kan tenke at instrumentet setter i gang en kausal kjedereaksjon. Instrumentet påvirker eksponering for en behandling som igjen påvirker utfallet. I utgangspunktet kan utfallet avhenge av både verdi til instrument og behandling slik at det må skrives $y_i(d, z)$. Den første antagelsen vi gjør er at instrumentet kun påvirker utfallet gjennom behandlingen, slik at

$$y_i(d, 1) = y(d, 0) \tag{10.52}$$

$$\implies y_i(1, 1) = y_i(1, 0) := y_i^1 \text{ og } y_i(0, 1) = y_i(0, 0) := y_i^0 \tag{10.53}$$

Denne antagelsen er ikke nødvendig for å finne kausal redusert form, altså kausal effekt av Z på y, men er nødvendig for å isolere effekt av behandling D. Vi kan også skrive opp en såkalt *first stage* som er analog til vanlig kausal effekt, men der *behandling* er instrumentet og utfallet er behandling,

$$D_i = D_i^0 + Z_i(D_i^1 - D_i^0) = \pi_0 + \pi_{1i} Z_i + v_i \tag{10.54}$$

[14] vet også at jeg trenger $\pi_{01} \neq 0$ men klarer ikke se hvordan det kommer inn her.

Den første antagelsen er at instrumentet er så godt som tilfeldig fordelt. Det betyr at den observerte eksponeringen ikke sier oss noen ting om de potensielle utfallene.

$$[Y_i^1, Y_i^0 D_i^1, D_i^0] \perp\!\!\!\perp Z_i \tag{10.55}$$

Dette er analog til tilfeldig fordeling av behandling og impliserer at $E[D_i|Z = j] = E[D_i^j]$ og $E[Y_i|Z = j] = E[Y_i^j]$ slik at vi kan lære kausal first stage og redusert form. For at dette skal være oppfylt i praksis må instrumentet ikke være informativt om uobserverte variabler som påvirker utfallet. Med andre ord så må observasjon med ulik eksponering for instrument ikke være systematisk forskjellige langs andre egenskaper som påvirker utfall. Vi kan utvide denne antagelsen til CIA ved å legge til covariates noen som kan gjøre det mer kredibelt hvis eksponering ikke kan betraktes som et rent eksperiment.

Videre må instrumentet være relevant for behandling, altså at eksponering for instrument endrer behandling for noen av observasjonene

$$E[D_i|Z_i = 1] - E[D_i|Z_i = 0] = E[D_i^1 - D_i^0] = E[\pi_{1i}] \neq 0 \tag{10.56}$$

Til slutt vil vi anta at first stage er monoton slik at $\pi_{1i} \geq 0$ eller $\pi_{1i} \leq 0$ for alle observasjonene. Det medfører at instrument enten øker eksponering for behandling eller reduserer det, men ikke begge deler. Vi kan dele populasjonen inn i fire kategorier ut fra hvordan instrument påvirker deres behandlingsstatus,

1. *Always takers* der $D_i^1 = D_i^1$ og $\pi_{1i} = 0$.
2. *Never takers* der $D_i^1 = D_i^0$ og $\pi_{1i} = 0$.
3. *Compliers* der $D_i^1 = 1$ og $D_i^0 = 0$ og $\pi_{1i} =$ .... hm.
4. *Defiers* der $D_i^1 = 1$ og $D_i^0 = 0$.

Når vi bruker instrument isolerer vi den variasjonen i eksponering i behandling som er skapt av instrumentet. Siden vi ekskluderer *defiers* per antagelse, så vil all all variasjon skyldes compliers og vi estimerer $E[y_i^1 - y_i^0 | complier]$. Med antagelse om konstant behandlingseffekt kan dette generaliseres som gjennomsnittlig behandlingseffekt for populasjonen, men i praksis kan compliers være systematisk forskjellig fra de andre gruppene. Med heterogen behandlingseffekt isolerer vi bare den lokale gjennomsnittlige behandlingseffekt for compliers. Gitt de fire antagelsene:

1. Eksklusjonskriteriet, eksklusiv kausal kanal gjennom behandling D_i
2. (Betinget) uavhengighetskriteriet, instrument er så godt som tilfeldig fordelt
3. Relevans, kausal first stage

4. Monotoniet

kan jeg vise at wald er LATE. Med andre ord:

$$\frac{E[y_i|Z_i=1]-E[y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]} = E[y_i^1 - y_i^1|D_i^1 > D_i^0] \tag{10.57}$$

For å ta beviset begynne vi med nevneren,

$$\begin{aligned}
&E[y_i|Z_i=1]-E[y_i|Z_i=0] && (10.58)\\
&= E[y_i^0 + D_i^1(y_i^1 - y_i^0)|Z_i=1] - E[y_i^0 + D_i^0(y_i^1 - y_i^0)|Z_i=0] && (10.59)\\
&= E[y_i^0 + D_i^1(y_i^1 - y_i^0)] - E[y_i^0 + D_i^0(y_i^1 - y_i^0)] && (10.60)\\
&= E[(D_i^1 - D_i^0)(y_i^1 - y_i^0)] && (10.61)\\
&= E[y_i^1 - y_i^0)|(D_i^1 > D_i^0)P(D_i^1 > D_i^0) && (10.62)
\end{aligned}$$

Ser at den reduserte formen fanger opp effekt på *compliers* skalert opp med andelen *compliers*. Skal deretter se at *first stage* i nevneren gir andel compliers,

$$\begin{aligned}
&E[D_i|Z_i=1]-E[D_i|Z_i=0] && (10.63)\\
&= E[D_i^0 + Z_i(D_i^1 - D_i^0)|Z_i=1] - E[D_i^0 + Z_i(D_i^1 - D_i^0)|Z_i=0] && (10.64)\\
&= E[D_i^1 - D_i^0] && (10.65)\\
&= E[D_i^1 - D_i^0|D_i^1 > D_i^0]P(D_i^1 > D_i^0) && (10.66)\\
&= P(D_i^1 > D_i^0) && (10.67)
\end{aligned}$$

Vi trenger antagelse om monotonitet fordi ...

Vi trenger ikke antagelsen dersom vi antar konstant behandlingseffekt fordi ...

Karakterisere compliant subpopulasjon

Vi finner en lokal behandlingseffekt for delpopulasjonen som faktisk responderer på instrumentet. Heldigvis for oss er dette gjerne effekten som er mest interessant fordi dette er personer som er på marginen i valg om å ta behandling og dermed er mest sensitiv for policy som gjør det enklere tilgjengelig. Vi vil uansett ønske å si noe om andelen *compliers* og hvordan de skiller seg fra resten av populasjonen langs andre observerte egenskaper.

Ettersom vi ikke kan observere både D_i^1 og D_i^0 kan vi aldri observere om gitt enhet er *compliant* eller *always-taker*. Vi kan aldri vite om en person uansett ville tatt behandlingen selv uten eksponering for instrumentet. Likevel kan vi forsøke å beskrive egenskaper til *compliers*. Vi kan finne andel fra first stage og vi kan visstnok også beskrive den betingede fordelingen av andre covariates.

10.2.4 Eksperiment med delvis compliance

I mange eksperiment er det tilfeldig utvalg som blir plassert i gruppen som får tilbud om behandling, men det er ikke alltid mulig å tvinge observasjonene til å eksponere seg for behandling. Dette medfører selv-seleksjon og det er dermed ikke mulig å finne den kausale behandlingseffekten ved å sammenligne utfall til de som blir behandlet og de som ikke. Vi kan finne kausal effekt av å bli *tilbudt* behandlings som betegnes som *Intention to treat*, men dette er ofte ikke like interessant. For å finne kausal effekt kan vi bruke instrumentvariabel der tildeling til behandlingsgruppe er instrument. Dette vil da være helt analogt med den mer generelle diskusjonen over, bortsett fra at vi nå kan ekskludere *always-takers* dersom kun behandlingsgruppen har tilgang på eksponering for behandling. Dette forenkler formelen til

$$\frac{E[y_i|Z_i=1]-E[y_i|Z_i=0]}{P[D_i|Z_i=1]} = E[y_i^1 - y_i^1|D_i=1] \tag{10.68}$$

10.2.5 Generalisering av wald

Jeg kan utvide dette til å se på flere instrument der vi finner den lineære kombinasjonen som er mest mulig korrelert med D. Dette finner vi uansett i first stage, så ikke noe problem å legge til flere. Det er ikke nødvendigvis så interessant siden hvert instrument estimerer en instrument-spesifikk lokal behandlingseffekt, det vil si behandlingseffekt fra delpopulasjon som complier til det gitte instrument. Når vi kombinerer får vi en saus; et vektet gjennomsnitt. Jeg starter heller enkelt med å se på wald-estimatorer som er spesialtilfelle av IV der både instrument og behandling er dummyvariabler.

$$\delta = \frac{E[Y|Z=1]-E[Y|Z=0]}{E[Y|D=1]-E[Y|D=0]} \tag{10.69}$$

10.3 Regresjonsdiskontinuitet

For at sammenligning av utfall til behandling- og kontrollgruppe skal avdekke kausal effekt må kontrollgruppen være en god proxy for det kontrafaktiske utfallet til behandlingsgruppen. I praksis krever det er eksogen variasjon i eksponering for behandling. Hvis individene selv bestemmer sin egen eksponering vil dette valget nesten alltid korrelere med andre uobserverte variabler som også påvirker utfallet, slik at at differansen ikke utelukkende skyldes kausal effekt av behandling. Regresjonsdiskontinuitet utnytter at regler som bestemmer eksponering for behandling har vilkårlige cutoffs og at det er begrenset mulighet for observasjonene til tilpasse seg eksakt hvilken side den havner på.[15] Mer formelt er sannsynlighet for å bli eksponert for behandling. er diskontinuerlig i en verdi s_0

[15] Det finnes vist to tilnærminger. Continuity based og basert på lokal randomisering. Vet ikke helt hva som er forskjell i praksis

av en såkalt *running variable* s. Dette kan for eksempel være en reform som ble innført på en gitt dato s, institusjonelle regler som gjør at enheter som ikke oppnår en gitt verdi av s må innføre tiltak eller at man en trenger en viss score s for å få en gevinst. I skarp rrd er behandlingen da for eksempel $D_i = I\{s_i \geqslant s_0)$. Hvis det er tilfeldig hvilken side av cutoff s_0 observasjone havner vil ikke dette si oss noe om potensielle utfall i fravær av behandling slik at differanse i utfall identifiserer kausal effekt. Dersom det ikke er mulig for enhetene å eksakt bestemme sin verdi av s er det kredibelt at det er tilfeldig for et intervall rundt s_0.

Det er en utfordring at forventet utfall uten behandling avhenger av s. Det kan være både fordi det eksisterer en direkte sammenheng mellom s og utfallet y, men også fordi andre variabler som påvirker utfall er korrelert med s. Vi modellerer derfor sammenhengen $\mathbb{E}[y|s]$ og bruker eventuell diskontinuitet i s_0 som estimat på kausal effekt. Det er også mulig å knytte til potensielle utfall. Anta at $\mathbb{E}[y_i^0|s_i] = f(s_i)$ og at det er konstant effekt slik at $y_i^1 = y_i^0 + \rho$. Da er

$$\mathbb{E}[y_i|s_i] = \begin{cases} \mathbb{E}[y_i^0|s_i] = f(s_i), & s_i < s_0 \\ \mathbb{E}[y_i^1|s_i] = f(s_i) + \rho, & s_i \geqslant s_0 \end{cases} \tag{10.70}$$

$$= f(s_i) + \rho D_i \tag{10.71}$$

der $D_i := I\{s_i \geqslant s_0\}$. Dette impliserer at

$$y_i = \mathbb{E}[y_i|s_i] + (y_i - \mathbb{E}[y_i|s_i]) \tag{10.72}$$

$$= f(s_i) + \rho D_i + u_i \tag{10.73}$$

Vi kan spesifisere en parametrisk form på $f(\cdot)$ som kan estimereres med OLS, for eksempel en polynom av orden p. Da blir ligningen

$$y_i = \alpha + \beta_1 s_{1i} + ... + \beta_p s_i^p + \rho D_i + u_i \tag{10.74}$$

Det er også mulig å generalisere til ulik $f(s)$ på hver side av s_0. Dette kan gjøres parameterisk ved å ha ulike parametre i polynom og interkasjon med indikator for om det er over s_0. Blir litt sånn som splines. Ellers kan vi også bruke ikke-parametrisk lokal regresjon.

Det er nødvendig å spesifisere et intervall rundt cutoff, $[s_{min}, s_{max}]$, som avgrenser hvilke observasjoner vi betrakter.[16] Lengden på intervallet omtales som *bandwidth*. Et lengre bandwith medfører flere observasjoner som kan mer presise estimat, men medfører også at observasjonene blir mer ulike. Jo lenger vekk fra s_0, jo mer sannsynlig at observasjoner er ulike langs andre dimensjoner. Dette gjelder spesielt hvis observasjonene kan tilpasse seg for oppnå payoff hvis $s > s_0$.

[16] Det finnes data-driven måter å gjøre dette på slik at vi slipper å velge det ad-hoc.

Det er viktig å treffe med funksjonell form på $f(\cdot)$ for å få riktig estimat på diskontinuiten. Hvis den modelleres som med en lineær likning vil ikke-linearitet kunne bli fanget opp som en diskontinuitet. Det er vanlig å bruke polynom, splines eller ikke-parametrisk lokal regresjon som er vektet med en kernel. Som alltid er det ikke én metode som dominerer; valg av struktur avhenger av antall observasjoner og hvor fleksibel funksjonen må være for å fange funksjonell form. Det er en klassisk bias-varians tradeoff.

For at strategien skal identifisere kausal effekt kan det ikke være slik at andre ting endres brått i s_0. For eksempel kan det være flere reformer som innføres samtidig, flere ting som endres når en person blir pensjonist. Da vil vi ikke diskontinuiteten identifisere akkurat den endringen vi er interessert i. Det må også være tilfeldig hvilken side observasjonen havner på. Hvis det er strategisk selv-selektering ved at observasjon bevisst velger $s \lesseqgtr s_0$ kan de være systematisk forskjellig langs uobserverte variabler.

Strategi for å underbygge kredibilitet til identifiserende antagelse:

1. Density test: Vil at sannsynlighetstetthet til s skal være kontinuerlig rundt s_0. Hvis de klumper seg sammen på éne siden gir det indikasjon på strategisk selvseleksjon.

2. Covariate balance: Vil at andre variabler ikke skal endres brått i s_0. Kan ikke observere alt, men kan kjøre opplegget med andre covariates vi observerer som utfall og se om det er diskontinuitet. Alternativt kan vi bare se om de har like gjennomsnitt.. men føler at vi kan håndtere at det varierer med s..

3. Placebo-tester: Vil ikke se hopp der vi ikke forventer hopp. Kan kjøre opplegget med andre verdier av s som cut-off.

tror jeg leser om dette i MHE i stedet :)

10.4 Tidsserier

Vi skal nå se på data der vi har gjentatt observasjon av samme enhet. Så langt har vi modellert sammenheng mellom variabler som blir realisert samtidig, $y = f(x) + u$. Hvis vi ønsker å predikere fremtidig verdi y_{t+k} når vi er i t så hjelper det oss ikke så mye å ha god f siden vi uansett må predikere x_{t+k} for å bruke den. En alternativ fremgangsmåte er å modellere stokastisk prosess $(y_t)_{t\in\mathbb{N}}$. I de fleste tilfeller vil fordeling til realisering t avhenge av realisering i tidligere perioder. Jeg skal begynne med å se på enkle måter å modellere denne avhengigheten. Deretter skal jeg inkludere andre forklaringsvariabler..

10.4.1 Lineær trend

Kanskje den enkleste måten å beskrive trenden er med en enkel lineær sammenheng der vi lar tidsperiode t være uavhengige variabel,

$$y_t = \alpha_0 + \alpha_1 t + \epsilon_t \tag{10.75}$$

der parametrene er definert slik at de konstruerer feilledd med egenskap $E[t\epsilon_t] = 0$ og $E[\epsilon_t] = 0$. Kan jo også påstå at $E[\epsilon_t|t] = 0$ men det er en ganske sterk påstand siden gjennomsnittlig utfall i hver periode sjeldent endrer seg helt linært. Dersom vi har få perioder så kunne vi gitt en indikator for hvert tidspunkt som gir fullstendig fleksibel beskrivelse av $E[y|t] = \alpha_t d(t)$,

$$y_t = \sum_t \alpha_t d_t + \epsilon_t \tag{10.76}$$

der $E[\epsilon_t|t] = 0$, men har jo bare én observasjon per tidspunkt så blir like mange parametre som observasjoner. Denne fremgangsmåten blir mer hensiktsmessig med paneldata der jeg har mange observasjoner i hver t. Vil da modellere trend for å *de-trende*, fjerne trend fra sammenheng mellom behandling og utfall jeg interessert i. Kan enkelt utvide til eksponentiell trend ved å ta logaritmisk transformasjon av sammenheng slik at den blir lineær,

$$y_t = e^{\beta_0 + \beta_1 t + \epsilon_t} \tag{10.77}$$

$$\log y_t = \beta_0 + \beta_1 t + \epsilon_t \tag{10.78}$$

10.4.2 Autoregressiv, AR(k)

I autoregressiv modell blir noe av utfall i forrige periode dratt med over til neste. Hvis det er spesielt høy verdi i én periode (høy ϵ_t) så blir det propagert videre i kommende perioder. Vi modellerer det med lagged utfall,

$$Y_t = \alpha + \rho Y_{t-1} + \epsilon_t \tag{10.79}$$

Hvor stor del av verdi som blir propagert videre avhenger av parameter som vi estimerer fra observert data. I praksis er det ofte enklere å jobbe med avvik fra gjennomsnitt, $y_t := Y_t - E[Y_t]$, og kan vises at

$$y_t = \rho y_{t-1} + \epsilon_t \tag{10.80}$$

Vil beskrive egenskap ved simultanfordeling til prosessen. Den er karakterisert ved såkalt autokovarians; korrelasjon mellom utfall på ulike tidspunkt.. Kan finne $var(y_t), cov(y_t, y_{t-1})$

og $cov(y_t, y_{t-k})$..

10.4.3 Moving average, MA(k)

Dette er en alternativ måte å beskrive sammenhengen mellom realisering på ulike tidspunkt. I stedet for at hele utfallet blir propagert videre er det nå bare selve sjokket ϵ_{t-1} som blir med å bestemme verdi i neste periode. Dette medfører at sjokket ikke påvirker verdi inn i evigheten,

$$Y_t = \mu + \alpha\epsilon_{t-1} + \epsilon_t \tag{10.81}$$

10.4.4 Lagged independent variable

Så langt har jeg sett på univariate tidsserier. Dette er ofte greiest dersom jeg vil bruke modellen til forecasting. Men jeg kan jo også være interessert i relasjon mellom variabler. Tenk for eksempel at jeg vil modellere hvor sulten jeg er og at jeg for hver time t rapporterer sult-nivå (y_t) samt egenskaper ved ting jeg har gjort i den aktutuelle perioden $\mathbf{x}_t$. La for eksempel x_t være antall kalorier jeg har spist. Dette vil ikke bare påvirke sultnivå i t, men også i fremtidige perioder $t+1, t+2, ..$ med gradvis mindre effekt. Modellen kan beskrives med

$$y_t = \alpha_0 + \beta_0 x_t + \beta_1 + x_{t-1} + \beta_2 x_{t-2} + \epsilon_t \tag{10.82}$$

Anta at jeg spiser 100 kalorier i $t = 1$ og ingen i de andre. Funksjonene blir da

$$y_1 = \alpha_0 + \beta_0 \cdot 100 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \epsilon 1 \tag{10.83}$$
$$y_2 = \alpha_0 + \beta_0 \cdot 0 + \beta_1 \cdot 100 + \beta_2 \cdot 0 + \epsilon 2 \tag{10.84}$$
$$y_3 = \alpha_0 + \beta_0 \cdot 0 + \beta_1 \cdot 0 + \beta_2 \cdot 100 + \epsilon 3 \tag{10.85}$$
$$\tag{10.86}$$

Hvis effekten er avtagende så vil $\beta_0 > \beta_1 > \beta_2$ slik at $E[y_1|(x_1, x_2, x_3) = (100, 0, 0)] > E[y_2|(x_1, x_2, x_3) = (100, 0, 0)]$ osv. Gir sånn passe mening dette her.

10.5 Paneldata

Så langt har vi sett på tilfeldig utvalg i en krysseksjon. Vi kan nå se på tilfelle der vi utnytter at observasjoner tilhører gruppe. Det kan for eksempel være at personene bor i samme by, går i samme klasse eller er i samme familie. Et annet eksempel er data der samme enhet blir observert på ulike tidspunkt. Med stor N og lav T kan dette analyseres på lignende måte, og de ulike observasjonene til hver enhet utgjør da gruppen. Senere

skal vi modellere dynamisk (tidsserie) aspekt også. Kan bruke to indexer it for å beskrive observasjon.. dette er litt notasjon og jeg kan redusere til én dimensjon, må se litt på det.

Fordelen med denne datastrukturen er at vi kan eliminere effekt av tidskonstant uobservert heterogenitet mellom ulike grupper ved å isolere sammenhengen mellom variasjon i behandling og variasjon i utfall innad i hver gruppe.

10.5.1 Dekomponering av feilledd

Anta at marginal kausal effekt av behandling x på utfall y er konstant $\partial y/\partial x = \beta$, som verken avhenger av mengden behandling eller andre egenskaper til enheten. Dette medfører at

$$y = \beta x + (y - \beta x) = \beta x + u \tag{10.87}$$

der $u := (y - \beta x)$ representerer det kumulative bidraget til utfallet av alle andre variabler. Jeg vil at feilleddet skal ha forventning lik null, så jeg sentrerer variabelen og omdefinerer,

$$y = \beta x + \mathbb{E}[u] + (u - \mathbb{E}[u]) = \beta x + \alpha + u_c \tag{10.88}$$

der jeg heretter lar $u_c := u$. La $d(i)$ være dummy som indikerer om observasjonen tilhører gruppe i. Vi kan da dekomponere feilleddet,

$$u = \mathbb{E}[u|d(i)] + (u - \mathbb{E}[u|d(i)]) = \alpha_i + \epsilon \tag{10.89}$$

slik at ligningen kan skrives som

$$y_{it} = \alpha + \beta x_{it} + \alpha_i + \epsilon_{it} \tag{10.90}$$

der α_i er en parameter som fanger opp uobservert heterogenitet. Det er bidraget av uobserverte variabler som er felles innad i gruppe. Tror jeg kan tenke på det som konkrete variabler eller bare dekomponering og projektering av tilfeldig variabel... Betegnes som fast-effekt fordi den er konstant innad i gruppe, f.eks. alle egenskaper innadd i familie som ikke varierer (foreldres utdanningsnivå?). Jeg kan estimere β konsistent med OLS dersom $cov(x_{it}, \epsilon_{it}) = 0$. Denne antagelsen er mer kredibel nå som jeg fått den fasteffekten ut av feilleddet, men det er fortsatt en ganske sterk antagelse på observasjonsdata. Hvis det er uobserverte ting som påvirker utfallet og samvarierer med behandlingen så kan vi ikke isolere effekten av denne. Tenke for eksempel om vi kjører to behandlingen samtidig og kun observerer eksponering for ene. Tror vi kan bruke instrument på samme måte som i krysseksjon ved å sette det opp på first-difference form, men vet lite om dette.

10.5.2 Identifikasjon

Det er en utfordring at grupper er systematisk forskjellig på måter som påvirker både utfall og eksponering for behandling. Dersom vi ikke kan observere de relevante egenskapene ved gruppen kan dette føre til skjevhet i estimat som følge av utelatte variabler. Ved å se på sammenheng mellom variasjon i behandling og i utfall *innad* i gruppene så eliminere vi effekt av egenskapene ved gruppen som er felles for medlemmene. Dette er blant annet intuisjonen bak tvillingstudier som gjør det mulig å kontrollere for genetiske egenskaper. Skal forsøke å formalisere dette med rammeverket for potensielle utfall. Anta heretter at i er individ og t er tid.

Begynner med å anta at at eksponering for behandling er så godt som tilfeldig fordelt betinget av egenskaper ved individet som er konstant over tid A_i, tidspunkt t og noen andre observerte covariates $\mathbf{x}_{it}$ som kan variere over tid for hvert individ,

$$y_{it}^0 \perp\!\!\!\perp D_{it} | \mathbf{x}_{it}, A_i, t \tag{10.91}$$

$$\implies E[y_{it}^0 | \mathbf{x}_{it}, A_i, t, D_{it}] = E[y_{it}^0 | \mathbf{x}_{it}, A_i, t] \tag{10.92}$$

Antar en lineær en lineær modell,

$$E[y_{it}^0 | \mathbf{x}_{it}, A_i, t] = \alpha + \delta_t + A_i'\gamma + \mathbf{x}_{it}\beta \tag{10.93}$$

Tror vi kun kan estimere en additativ kausal effekt,

$$E[y_{it}^1 | \mathbf{x}_{it}, A_i, t] = E[y_{it}^0 | \mathbf{x}_{it}, A_i, t] + \delta \tag{10.94}$$

$$\implies E[y_{it} | \mathbf{x}_{it}, A_i, t, D_{it}] = \alpha + \delta_t + A_i'\gamma + \mathbf{x}_{it}\beta + \delta D_{it} \tag{10.95}$$

$$\implies y_{it} = \alpha_i + \delta_t + \mathbf{x}_{it}\beta + \delta D_{it} + \epsilon_{it} \tag{10.96}$$

der $\alpha_i := \alpha + A_i'\gamma$ og $\epsilon_{it} := y_{it}^0 - E[y_{it}^0 | \mathbf{x}_{it}, A_i, t]$.[17] Det er forøvrig også et poeng at estimatene er veldig sensitive for målefeil siden det er lite variasjon i behandling for hvert individ. Målefeilene kan derfor fort utgjøre større andel av variasjon enn det gjør i krysseksjon.

10.5.3 Estimering

Si litt high level om dette. I praksis knyttes til estimering av ligningsystem.. stacker matriser og sånn, blir heavy notasjon i lineær algebra. Kan også si noe om forskjell på fixed og random effects.

[17] Vet ikke hvorfor det ikke er $\epsilon_{it} := y_{it} - E[y_{it}|..]$ Må se på dette senere

Fixed effects (within)

Vi vil se på variasjon innad i gruppe. En naturlig måte er å demeane alle variabler innad i hver gruppe. Finner først gjennomsnitt i hver gruppe, der jeg åpner for at antallet medlemmer kan variere.[18]

$$\frac{1}{T(i)}\sum_t y_{it} = \frac{1}{T(i)}\sum_t [\alpha + \beta x_{it} + \alpha_i + \epsilon_{it}] \tag{10.97}$$

$$\bar{y}_i = \alpha + \beta \bar{x}_i + \alpha_i + \bar{\epsilon}_i \tag{10.98}$$

Deretter trekker jeg dette fra hver observasjon

$$\ddot{y}_{it} := y_{it} - \bar{y}_i = \beta \ddot{x}_{it} + \ddot{\epsilon}_{it} \tag{10.99}$$

slik at vi kun trenger at $cov(x_{it}, \epsilon_{it}) = 0$.[19] Alternativt kan jeg bruke dummies til eksplisitt å estimere fast-effektene. Dette illustrerer litt koblingen mellom panel og dummies generelt. Ved å introdusere dummies så definerer man eksplisitt undergrupper og identifiserer effekt av behandling ved å se på variasjon innad i gruppene. Således kan jo regresjon med kjønnsdummy betraktes som panel der hvert kjønn utgjør gruppe.

Pooled OLS

Random effects

Betrakte α_i som tilfeldig variabel. For at det estimator skal være konsistent må vi anta at $cov(x_{it}, \alpha_i) = 0$ slik at det egentlig ikke er noe stort poeng i å bruke panel struktur, bortsett fra at vi kan bruke strukturen til å få mer effektiv estimator enn vanlig OLS. Denne strukturen utnyttes gjennom feasible generalized least square (FGLS). Tror ikke det er veldig stort poeng, bortsett fra at det ikke er kosher.

Det kan derimot være relevant å modellere på denne måten dersom man ikke er interessert i kausal parameter, men kun endring i forvetningsverdi... bruker da MLE til å estimere.

Kan bruke såkalt hausmannstest for å vurdere om $cov(x_{it}, \alpha_i) = 0$. Hvis den er oppfylt vil både F.E og R.E estimatorene være konsistent, men hvis den ikke er oppfylt vil de konvergere mot ulike størrelser. Huasmann gir oss test som sier om differansen er stor nok til at vi kan forkaste nullhypotesen om at $cov(x_{it}, \alpha_i) = 0$.

[18] Såkalt ubalansert panel. Ikke stort problem dersom ikke skyldes systematisk skjevt frafall som er brudd på forutsetning om tilfeldig utvalg.

[19] Tror det er ekvivalent med $cov(\ddot{x}_{it}, \ddot{\epsilon}_{it}) = 0$.. kunne kanskje vist.

Between estimator

En mulig strategi er å kollapse variablene innad i hver gruppe slik at vi bare har gjennomsnittene. Mister variasjon innad i gruppe (tidsdimensjon hvis gjentatt observasjon av individ) og ser bare på forskjeller mellom grupper. Kan identifisere kausaleffekt hvis gjennomsnittlig behandling ikke er korrelert med uobservert heterogenitet α_i, men det vil være en lite effektiv estimatorer siden vi mister mye informasjon,

$$\bar{y}_i = \bar{\mathbf{x}}_i\beta + \bar{u}_i \tag{10.100}$$

GLS som vektet gjennomsnitt av within on between

hm. kan ta resultat og intuisjon, men utledning er ganske fucked. kunne brukt det som anledning til å si noe om GLS generelt, men gjør det et annet sted. Angrist sier at GLS er teit uansett, men kan gi litt intuisjon om 2SLS som GLS på wald estimatorer... eller noe sånt.

10.5.4 Dynamisk panel

Kan ønske å utnytte tidsseriedimensjonen i panel, altså det faktum at vi observerer samme individer på flere tidspunkt, til å modellere dynamikk. For det første kan det være slik at en behandling x_{it} påvirker utfall ikke bare i tidspunkt t men også i fremtidige perioder $t+1, t+2, ...$ Dette kan vi ta hensyn til ved å inkludere laggede uavhengige variabler. Vi kan også ha lyst til å ta med lagged avhengig variabel. Tror at motivasjonen for dette er at vi vil sammenligne likt med likt og derfor må ta med historien til individene. Selv om de ser like ut på et gitt tidspunkt t, så kan informasjon om utfallene deres på tidligere tidspunkt gi informasjon om hva slags type person de er...

Lagged avhengig variabel

Vi kan motivere dette med evaluering av arbeidsmarkedstiltak. Personer på tiltak har ofte hatt tap av inntekt. Hvis de har høyere lønnsvekst etter tiltak så kan dette fange opp at de har høyere *potensiell lønn* enn sammenlignbare personer i kontrollgruppe? Vet ikke helt, men tar litt utledning uansett. Enkleste modell er

$$y_{it} = \gamma y_{i,t-1} + u_{it}, \quad \text{der } u_{it} = \alpha_i + \epsilon it \tag{10.101}$$

der $(\epsilon_{it})_{t\in\mathbb{N}}$ er hvit støy (altså ingen seriekorrelasjon). Skal nå se på mulighet til å estimere den kausale γ. Det forutsetter at den uavhengige variabelen ikke er endogen.[20] Jeg har et

[20] Jeg er ikke 100% komfortabel med begrepet endogen. Brukes vel bare short-hand for å være korrelert med feilledd, men tror det er greit å se litt på ligningssystem for at terminologi skal gi litt mening.

enkelt oppsett for å vurdere konsistens til OLS-estimator i enkel univariat regresjon,

$$\hat{\beta} = \frac{\sum x_n y_n}{\sum_n x_n^2} = \frac{\sum x_n(\beta x_n + u_n)}{\sum_n x_n^2} = \beta + \frac{\sum x_n u_n}{\sum x_n^2} \tag{10.102}$$

Vi kan da vurdere konsistens ved å skalere med $1/N$ og slik at teller konvergerer til $cov(x_n, u_n)$ når $n \to \infty$. Mange estimatorer kan betraktes som OLS på transformerte variabler, så her er det bare å plugge inn. Prøver først med fixed effect der $x_n := \sum_{t=1}(y_{i,t-1} - \bar{y}_{i,-1})$ og $y_n = \sum_{t=1}(y_{i,t} - \bar{y}_)$

$$\hat{\beta} = \frac{\sum_n \sum_{t=1}(y_{i,t-1} - \bar{y}_{i,-1})(y_{i,t} - \bar{y})}{\sum_n \sum_{t=1}(y_{i,t-1} - \bar{y}_{i,-1})^2} \tag{10.103}$$

$$= \gamma + \frac{\sum_n \sum_{t=1}(y_{i,t-1} - \bar{y}_{i,-1})(u_{i,t} - \bar{u}_i)}{\sum_n \sum_{t=1}(y_{i,t-1} - \bar{y}_{i,-1})^2} \tag{10.104}$$

$$\tag{10.105}$$

Det kan sikkert vises at det stemmer, men jeg bare plugger inn de analoge størrelsene fra transformert OLS. Litt usikker på hvordan jeg skal vise hva som er problemet her. Tror problemet er at $y_{i,t-1}$ er positivt korrelert med α_i i feilleddet. Personer som har (uobserverte) konstante egenskaper som gir de høy lønn i forrige periode vil i gjennomsnitt ha høyere feilledd i denne perioden ikke sant.. så hvis høyt utfall i forrige periode er bb-ehandling", så vil behandlingseffekten fange opp dette og vi får skjevt estimat av kausal effekt.

Løsningen på dette er å bruke momentbetingelser implisert av modellen. For at et instrument z skal være gyldig må

1. Relevans, må forklare noe variasjon i behandling: $cov(y_{i,t-1}, z) \neq 0$
2. Eksogenitet, må ikke være korrelert med uobserverte variabler som påvirker utfall: $cov(u_{i,t}, z) = 0$.

Siden prosess er autoregressiv vil utfall i tidligere periode propagere gjennom prosessen. Laggede avhengige variabler er relevant og ikke korrelert med idiosynkratiske delen av feilleddet i hvertfall. Vil være korrelert med uobservert heterogenitet, men dette kan vi bli kvitt ved å ta first difference. Les om Arelleano-Bond hvis dette blir relevant i fremtiden...

10.5.5 Instrument

Selv i med fixed effect må vi anta at $cov(x_{it}, \epsilon_{it}) = 0$, altså at behandling er ukorrelert med idiosynkratisk (?) uobserverte variabler. Kan forsøke å legge inn forklaringsvariabler slik at komponent av ϵ_{it} som ikke er forklart av kontroll-variablene ikke er korrelert med behandling, men i likhet med i krysseksjon kan de fortsatt være relevante utelatte variabler som varierer over tid. En alternativ strategi er å bruke instrument til å identifisere

behandlingseffekt med utgangspunkt i variasjonen av behandling som kan forklares med instrumentet. Tror jeg trenger litt recap på instrument i krysseksjon først.

10.5.6 Panelmetoder på andre datastrukturer (multi-level, hierarki, cluster)

Kan tenke oss at observasjoner er med i overodnede grupper, for eksempel familie, skoleklasse, bedrift eller geografisk område. Vi kan jo også tenke at uobserverte egenskaper på ved dette gruppenivået - som dermed er felles for medlemmene av gruppen - påvirker utfallet til observasjonene. Hvis gruppetilhørighet også påvirker sannsynlighet for eksponering for behandling er $cov(x_{it}, a_i) \neq 0$ og OLS som ikke tar hensyn til gruppetilhørighet gir inkonsistent estimat på kausal effekt. Vi kan håndtere dette med å se på variasjon innad i gruppe med samme metoder som over. Dette forutsetter riktig nok at det er variasjon i eksponering for behandling innad i gruppe; hvis ikke kan vi ikke frikoble behandlingseffekten fra resten av a_i.

Hvis vi antar at $cov(x_{it}, a_i) = 0$ trenger vi ikke ta hensyn til gruppetilhørighet og bare kjøre OLS, men kan oppnå bedre estimat ved å modellere avhengighet i gruppe med random effects. I begge tilfeller vil inferens være korrekt dersom det er en tilfeldig sample. I praksis så er sampling ofte to-delt; man velger først et tilfeldig utvalg av *clusters* (overordnet enhet) og deretter trekker tilfeldig utvalg av observasjoner innad i hver cluster.[21] Denne samplingen er fortsatt tilfeldig slik at estimator er konsistens, men vi må justere standardfeil for å ta hensyn til korrelasjon mellom observasjoner (pga. felles α_i innad i cluster). Vet ikke hvordan det fungerer i praksis.

Finnes stor litteratur på multi-level modelling. Vet ikke helt om de bryr seg om kausalitet.

10.6 Forskjeller i forskjeller

Fixed effect er vel og bra hvis vi har variasjon i eksponering for behandling innad i grupper og føles oss komfortabel med å si at behandling er omtrent betinget uavhengig av potensielle utfall. I praksis er ofte behandling bestemt overnfra og ned i betydningen at noen grupper blir eksponert for tiltak/reform og andre ikke. Hvis vi kun har krysseksjon med utfallet til gruppene etter behandling er det vanskelig å trekke noen konklusjoner siden gruppene nok var forskjellig i utgangspunktet. Hvis vi derimot har informasjon om gruppene over tid kan vi se på forskjell i *endring* for hver gruppe i stedet for nivå. Dette gjør det mulig å identifisere behandlingseffekten under antagelse om at de ville hatt felles trend i fravær av behandling. Dette skal jeg nå vise formelt med to grupper,

[21] Kan også observere alle enheter i cluster, men gjør ingen forskjell.

to tidsperioder og to behandlingsnivåer. Deretter skal jeg utvide til til flere tidsperioder, flere behandlingsnivåer og kontroll for individuelle covariates.

10.6.1 Identifikasjon

Potensielt utfall i periode t:

$$Y_{it} = Y_{it}^0 + D_i(Y_{it}^1 - Y_{it}^0) \tag{10.106}$$

Vi vil estimere average treatment effect on treated (ATT):

$$\delta_1 = E[Y_{i1}^1 - Y_{i1}^0 | D_i = 1] \tag{10.107}$$

Identifiserende antagelse er felles trend i fravær av behandling:

$$E[Y_{i1}^0 | D_i = 1] - E[Y_{i0}^0 | D_i = 1] \tag{10.108}$$
$$= E[Y_{i1}^0 | D_i = 0] - E[Y_{i0}^0 | D_i = 0] \tag{10.109}$$

Antar altså at $(Y_{i1}^0 - Y_{i0}^0) \perp\!\!\!\perp D_i$. Bevis for identifikasjon:

$$\delta_1 = E[Y_{i1}^1 | D_i = 1] - E[Y_{i1}^0 | D_i = 1] \tag{10.110}$$
$$= E[Y_{i1} | D_i = 1] - E[Y_{i1}^0 | D_i = 1] \tag{10.111}$$

Vi observerer $E[Y_{i1}|D_i = 1]$, men $E[Y_{i1}^0|D_i = 1]$ er kontrafaktisk. Kan finne proxy gitt identifiserende antagelse:

$$E[Y_{i1}^0 | D_i = 1] = E[Y_{i0}^0 | D_i = 1] + (E[Y_{i1}^0 | D_i = 1] - E[Y_{i0}^0 | D_i = 1]) \tag{10.112}$$
$$= E[Y_{i0} | D_i = 1] + (E[Y_{i1}^0 | D_i = 0] - E[Y_{i0}^0 | D_i = 0]) \tag{10.113}$$
$$= E[Y_{i0} | D_i = 1] + (E[Y_{i1} | D_i = 0] - E[Y_{i0} | D_i = 0]) \tag{10.114}$$

Alle størrelsene er observerbare og den kausale effekten er identifisert:

$$\delta_1 = E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1] - (E[Y_{i1}|D_i = 0] - E[Y_{i0}|D_i = 0]) \tag{10.115}$$

Merk at antagelsen er ekvivalent med stabil seleksjonsskjevhet:

$$E[Y_{i1}^0 | D_i = 1] - E[Y_{i1}^0 | D_i = 0] \tag{10.116}$$
$$= E[Y_{i0}^0 | D_i = 1] - E[Y_{i0}^0 | D_i = 0] \tag{10.117}$$

10.6.2 Flere grupper og flere tidsperioder

I det enkleste eksempelet trenger vi egentlig bare fire størrelser: gjennomsnitt i hver av gruppene før og etter eksponeringen for behandling. Dersom vi har individdata er det flere fordeler ved å sette opp en regresjonsmodell på form

$$y_{ist} = \gamma_s + \delta_t + \delta D_{st} + \epsilon_{ist} \tag{10.118}$$

10.6.3 Kontinuerlig behandling og individuelle egenskaper

10.7 Limited Dependent Variable

10.7.1 Binært valg

Vi vil ofte forsøke å forstå hvordan valg om å enten gjøre eller ikke gjøre noe avhenger av egenskaper ved valgsituasjonen. Hvilke variabler kan bidra til å forklare om hvorfor personer velger å jobbe eller ikke, reise kollektivt eller ikke, og så videre. Vi kan modellere dette med en latent, uobserverbar kontinuerlig variabel y^*, der valg av y avhenger av terskelverdi av y^* som vi uten tap av generalitet kan sette lik 0. Dette medfører at

$$y^* = \mathbf{x}'\beta + \epsilon \tag{10.119}$$

$$y = I\{y^* > 0\} = I\{\mathbf{x}'\beta + \epsilon > 0\} \tag{10.120}$$

slik at

$$P[y = 1|\mathbf{x}] = P[\mathbf{x}'\beta + \epsilon > 0|\mathbf{x}] \tag{10.121}$$

$$= \mathbb{P}[\epsilon > -\mathbf{x}'\beta|\mathbf{x}] \tag{10.122}$$

$$= F(\mathbf{x}'\beta) \tag{10.123}$$

der $F(\cdot)$ er kumulativ fordeling til $-\epsilon$, som vi antar er symmetrisk slik at det også er cdf til ϵ. Det er tre vanlige valg av $F(\cdot)$ som gir probit, logit og LPM. Kan definere LPM slik at det blir gyldig cdf, men litt usikker på det.

10.7.2 Estimering

For å finne betinget likelihood må jeg si noe om hvordan sannsynlighet for ulike y-verdier avhenger av x-verdi. Denne sammenhengen er beskrevet av ukjent parameter som vi forsøker å lære.

$$P(y|\mathbf{x}) = \begin{cases} F(\mathbf{x}'\beta), & y = 1 \\ 1 - F(\mathbf{x}'\beta), & y = 0 \end{cases} \tag{10.124}$$

som kan skrives som én-linjer. Kan også betrakte F.O.B til score-funksjonen, men vet ikke hvor interessant dette er.

$$\frac{\partial \log L(\beta)}{\partial \beta} = \left[\frac{y - F(\mathbf{x}'\beta)}{F(\mathbf{x}'\beta)(1 - F(\mathbf{x}'\beta)} f(\mathbf{x}'\beta)\right] x \tag{10.125}$$

der $E[score] = 0$ impliserer at den såkalte generaliserte residualen i klammeparantes er ortogonal på $span(\mathbf{x})$.

Tolke koeffisienter

Merk nå at β er fra den underliggende latente modellen og ikke har noen opplagt tolkning. Det vi er interessert i er hvordan sannsynligheten for $P[y = 1|\mathbf{x}]$ avhenger av $\mathbf{x}$. For kontinuerlige variabler kan vi bruke kjerneregel til å derivere uttrykket,

$$\frac{\partial}{\partial x_k} F(\mathbf{x}'\beta) = \frac{\partial F(u)}{\partial u}\beta_k \tag{10.126}$$

$$= f(\mathbf{x}'\beta)\beta_k \tag{10.127}$$

Vi kan merke at effekt partiell effekt på betinget sannsynlighet alltid har samme fortegn som β_k siden $f(\cdot)$ er sannsynlighetstetthet, men størrelsen avhenger av hvor vi evaluerer $\mathbf{x}$. Vi kan betrakte $f(\mathbf{x}'\beta)$ som en skaleringsfaktor. Tre vanlige valg av skaleringer er

1. Plugge inn noen verdier. Interessant dersom vi har har noen få dummies, men i praksis vil vi ofte ha enklere sammendragsmål.

2. Partial effect at average (PEA): Plugger in $\bar{\mathbf{x}}$. Litt problem dersom har transformerte variabler, siden tar gjennomsnitt etter transformasjon.

3. Average partial effect (APE): Evaluerer i hver $\mathbf{x}_n$ som jeg observerer i utvalg og tar gjennomsnitt, $\frac{1}{N}\sum_n f(\mathbf{x}_n'\beta)\beta_k$.[22]

For variabler som er diskret er ikke den deriverte en meningsfull størrelse. Vi tar da differanse i verdi, $(x_k + 1) - x_k$, men det avhenger fortsatt av verdi til andre variabler. Kan bruke samme fremgangsmåter som over, f.eks blir APE:

$$APE(x_k) = \frac{1}{N}\sum_n [f(\mathbf{x}_{n,-k}\beta_{-k} + \beta_k(x_k + 1) - \mathbf{x}_n'\beta] \tag{10.128}$$

der $f(\mathbf{x}_{n,-k}\beta_{-k}$ er vektorene med de resterne $K - 1$ variablene.

[22] Dette blir da hele uttrykket for partial effect og ikke bare skaleringsfaktoren.

10.7.3 Flere kategorier

10.7.4 Multinomial

Ordnet logit

Hvis alternativene kan rangeres kan vi modellerer de fra underliggende latent variabel, bare at vi finner en partisjonering slik at $y = j$ hvis $y^*(\gamma_{j-1}, \gamma_j]$. Tror grenseverdiene blir parametere. Bestemmer at $\gamma_1 = 1$ for de skal være entydig siden $\mathbf{x}\beta$ ikke har noen naturlig skala.

Merk at det er kun for kategoriene i endene der fortegn på koeffisient har entydig effekt på sannsynlighet for at observasjon tilhører kategori.

10.7.5 Sensurert regresjon (tobit)

En annen situasjon - som egentlig er ganske forskjellig! - som analyseres på tilsvarende måte er valg med hjørneløsning. I mange situasjoner er det ikke mulig å velge negativ kvantum slik at det blir en opphopning av verdier i $y = 0$. Jeg tenker at vi da kunne ha modellert sannsynlighet for positivt kvantum med probit og modellert $E[y|x, y > 0]$ med OLS hver for seg. Med Tobit gjør vi begge deler samtidig ved å anta at begge avhenger latent variabel.

Vi begynner som alltid ved å beskrive modell for latent variabel og hvordan den er relatert til observert utfall:

$$y* = \mathbf{x}'\beta_0 + u, \quad \text{der } u|\mathbf{x} \sim N(0, \sigma_0^2) \tag{10.129}$$

$$y = \max\{y^*, 0\} \tag{10.130}$$

Estimering

Jeg vil beskrive hvordan betinget sannsynlighet til y gitt $\mathbf{x}$ avhenger av parametre. Vil har et uttrykk for tetthet til sannsynlighet av y for en gitt $\mathbf{x}$ ikke sant, men dette er en mixed fordeling som punktsannsynlighet i $y = 0$. For å håndtere dette deler jeg opp den betingede fordelingen og behandler hvert tilfelle separat.

$$f(y|\mathbf{x}) = \begin{cases} 0, & y < 0 \\ P(y^* \leq 0|\mathbf{x}), & y = 0 \\ f(y^*|\mathbf{x}), & y > 0 \end{cases} \tag{10.131}$$

Jeg kan nå bruke antagelsene fra den latente modellen til å gi et eksplisitt uttrykk for hver størrelsene, noe som samtidig gir meg likelihood-funksjonen når jeg betrakter det

som funksjon av parameter slik at jeg kan estimere parametre fra observerte data.

$$P(y=0|\mathbf{x}) = P(y^* \leq 0|\mathbf{x}) \tag{10.132}$$
$$= P(\mathbf{x}'\beta_0 + u \leq 0|\mathbf{x}) \tag{10.133}$$
$$= P(u < -\mathbf{x}'\beta_0|\mathbf{x}) \tag{10.134}$$
$$= P\left(\frac{u}{\sigma_0} < -\frac{\mathbf{x}'\beta_0}{\sigma_0}|\mathbf{x}\right) \tag{10.135}$$
$$= \Phi\left(-\frac{\mathbf{x}'\beta_0}{\sigma_0}\right) \tag{10.136}$$
$$= 1 - \Phi\left(\frac{\mathbf{x}'\beta_0}{\sigma_0}\right) \tag{10.137}$$

For de positive verdiene kan vi utlede på samme måte som i vanlig regresjon. Begynner med å ta sannsynlighet for intervall siden punktsannsynlighet til tetthet=0..

$$P(y < y|\mathbf{x}, y > 0) = P(y^* < y|\mathbf{x}, y > 0) \tag{10.138}$$
$$= P(\mathbf{x}'\beta_0 + u < y|\mathbf{x}, y > 0) \tag{10.139}$$
$$= P\left(\frac{u}{\sigma_0} < \frac{y - \mathbf{x}'\beta_0}{\sigma_0}|\mathbf{x}, y > 0\right) \tag{10.140}$$
$$= \Phi\left(\frac{y - \mathbf{x}'\beta_0}{\sigma_0}\right) \tag{10.141}$$

Deriverer for å finne uttrykk for tetthet

$$f(y|\mathbf{x}, y > 0) = \frac{1}{\sigma_0}\phi\left(\frac{y - \mathbf{x}'\beta}{\sigma_0}\right) \tag{10.142}$$

Trenger ikke gjøre det relativt til standardisert fordeling.. skal se om jeg kan omskrive dette senere. Uansett, det gir en representasjon av loglikelihoodfunksjon,

$$logL(\beta, \sigma) = \sum_n I\{y_n = 0\} \log\left(1 - \Phi\left(\frac{\mathbf{x}_n'\beta_0}{\sigma_0}\right)\right) \tag{10.143}$$
$$+ I\{y_n > 0\} \log\left(\frac{1}{\sigma}\phi\left(\frac{y - \mathbf{x}_n'\beta}{\sigma}\right)\right) \tag{10.144}$$

Når vi har estimert parametrene kan vi gi estimerte mål på størrelser vi er interessert i og se hvordan de er relatert til parameter β fra underliggende latent modell.

Tolke koeffisient

Når vi har fått estimert β så ligner output litt på vanlig regresjon og det kan være fristende å tolke det på vanlig måte. Men β er parameter i $\mathbb{E}[y^*|\mathbf{x}] = \mathbf{x}'\beta$. Tror dette kan ha direkte tolkning hvis data er sensuert, men ikke dersom vi modellerer hjørneløsning. Da vil vi

istedet bruke latent modell til å beskrive sentraltendens i det faktiske utfallet.

$$\mathbb{E}[y|\mathbf{x}] = \mathbb{E}[y|\mathbf{x}, y > 0]P(y > 0|x) + 0 \tag{10.145}$$

$$= \mathbb{E}[y|\mathbf{x}, y > 0]\Phi\left(\frac{\mathbf{x}'\beta_0}{\sigma_0}\right) \tag{10.146}$$

der jeg har brukt at $P(y > 0|\mathbf{x}) = 1 - P(y = 0|\mathbf{x})$. Det er ganske rimelig at vi kan evaluere forventingsverdi ved å ta vektet sannsynlighet av betingede forventninger på partisjonering, men litt usikker på hvordan jeg viser det formelt. Se lov om iterated expectations. Vil finne et uttrykk for betinget sannsynlighet av utfall gitt at det er positivt.

$$\mathbb{E}[y|\mathbf{x}, y > 0] = \mathbb{E}[y^*|\mathbf{x}, y > 0] \tag{10.147}$$

$$= \mathbb{E}[\mathbf{x}'\beta_0 + u|\mathbf{x}, y > 0] \tag{10.148}$$

$$= \mathbf{x}'\beta_0 + \sigma_0\mathbb{E}\left[\frac{u}{\sigma_0}|\mathbf{x}'\beta_0 + u > 0\right] \tag{10.149}$$

$$= \mathbf{x}'\beta_0 + \sigma_0\mathbb{E}\left[\frac{u}{\sigma_0}|\frac{u}{\sigma_0} < -\frac{\mathbf{x}'\beta_0}{\sigma_0}\right] \tag{10.150}$$

$$= \mathbf{x}'\beta_0 + \sigma_0\frac{\phi(-c)}{1 - \Phi(-c)} \tag{10.151}$$

$$= \mathbf{x}'\beta_0 + \sigma_0\frac{\phi(c)}{\Phi(c)} \tag{10.152}$$

$$= \mathbf{x}'\beta_0 + \sigma_0\lambda(c) \tag{10.153}$$

der $c := \frac{\mathbf{x}'\beta_0}{\sigma_0}$ og $\lambda(c) := \frac{\phi(c)}{\Phi(c)}$ betegnes som den inverse mills ratioen. Dette gir oss et uttrykk for $\mathbb{E}[y|\mathbf{x}, y > 0]$ som er en størrelse vi kunne forsøkt å estimere ved å avgrense til kun observasjoner med positivt utfall. Kan se at det består både av helning til forventningsverdi av latentverdi og et ekstra ledd. Litt usikker på hvordan vi tolker, men det har noe sammenheng med at observasjoner med positivt utfall har større verdi av uobservert variabel u i latent modell. Dersom vi er interessert i parameter β_0 fra latent modell vil vi derfor få skjevt estimat dersom vi avgrenser til observasjon med positivt utfall.. Uansett, kan nå også plugge inn størrelsen og få et uttrykk

$$\mathbb{E}[y|\mathbf{x}] = \mathbb{E}[y|\mathbf{x}, y > 0]P(y > 0|x) \tag{10.154}$$

$$= \left[\mathbf{x}'\beta_0 + \sigma_0\lambda(c)\right]\Phi(c) \tag{10.155}$$

$$= \mathbf{x}'\beta_0\Phi(c) + \sigma_0\phi(c) \tag{10.156}$$

der $c := \frac{\mathbf{x}'\beta_0}{\sigma_0}$. Når jeg har estimert parametrene er estimatene av $\mathbb{E}[y|\mathbf{x}]$ og $\mathbb{E}[y|\mathbf{x}, y > 0]$ bare to deterministiske funksjoner av $\mathbf{x}$. Disse funksjonene er en forenklet representasjon av egenskaper ved virkeligheten og er dessuten upresise fordi utvalget gir begrenset infor-

masjon om den *sanne* prosessen[23] som genererer data. Uansett, det er ihvertfall objekter jeg kan jobbe med og bruke til å svare på spørsmål. Vi begynner med å se på hvordan $\mathbb{E}[y|\mathbf{x}, y > 0]$ endres når vi endrer en variabel x_j.

$$\frac{\partial}{\partial x_j}\mathbb{E}[y|\mathbf{x}, y > 0] = \frac{\partial}{\partial x_j}(\mathbf{x}'\beta_+\sigma_\lambda(c)) \tag{10.157}$$

$$= \beta_j \ \sigma\frac{\partial}{\partial c}\lambda(c)\frac{\beta_j}{\sigma} \tag{10.158}$$

Må finne $\frac{\partial}{\partial c}\lambda(c)$. Bruker at $\partial\phi(c)/\partial c = -c\phi(c)$.[24]

$$\frac{\partial}{\partial c}\frac{\phi(c)}{\Phi(c)} = \frac{-\mathbf{x}'\beta\phi(c)\Phi(c) - \phi(c)^2}{\Phi(c)^2} \tag{10.159}$$

$$= \frac{-\phi(c)[c\Phi(c) + \phi(c)]}{\Phi(c)^2} \tag{10.160}$$

$$= -\lambda(c)\frac{c\Phi(c) + \phi(c)}{\Phi(c)} \tag{10.161}$$

$$= -\lambda(c)[c + \phi(c)] \tag{10.162}$$

slik at

$$\frac{\partial}{\partial x_j}\mathbb{E}[y|\mathbf{x}, y > 0] = \beta_j\{1 - \lambda(c)[c + \phi(c)]\} \tag{10.163}$$

Siden $\mathbb{E}[y|\mathbf{x}] = \mathbb{E}[y|\mathbf{x}, y > 0]P(y > 0|x)$ og er $P(y > 0|x) = \Phi(c)$, kan vi bruke produktregel og plugge inn

$$\frac{\partial}{\partial x_j}\mathbb{E}[y|\mathbf{x}] = \beta_j\{1 - \lambda(c)[c + \phi(c)]\}\Phi(c) + [c + \sigma\lambda(c)]\frac{\beta_j}{\sigma}\phi(c) \tag{10.164}$$

$$= \beta_j\{\Phi(c) - \phi(c)[c + \phi(c)]\} + [c + \sigma\lambda(c)]\frac{\beta_j}{\sigma}\phi(c) \tag{10.165}$$

$$= \beta_j\Phi(c) - \beta_j\phi(c)c - \beta_j\phi(c)^2 + c\frac{\beta_j}{\sigma}\phi(c) + \lambda(c)\beta_j\phi(c) \tag{10.166}$$

$$\tag{10.167}$$

wtf. Bruker istedet

$$\mathbb{E}[y|\mathbf{x}] = \mathbf{x}'\beta\Phi(c) + \sigma\phi(c) \tag{10.168}$$

$$\frac{\partial}{\partial x_j}\mathbb{E}[y|\mathbf{x}] = \beta_j\Phi(c) + \mathbf{x}'\beta\phi(c)\frac{\beta_j}{\sigma} - c\sigma\phi(c)\frac{\beta_j}{\sigma} \tag{10.169}$$

$$= \beta_j\Phi(c) + c\phi(c)\beta_j - c\phi(c)\beta_j \tag{10.170}$$

$$= \beta_j\Phi(c) \tag{10.171}$$

[23] Eller vår representasjon av den..

[24] hvorfor?

der jeg igjen har brukt $\partial/\partial c\phi(c) = -c\phi(c)$.

Partial effect avhenger av hvor vi evaluerer $\mathbf{x}$. Kan bruke average partial effect (APE) på samme måte som i probit. Merk også at derivert gir en lokal lineær tilnærming. Kan få eksakt endring ved én enhets endring ved å plugge verdier inn i $g(\cdot) := \mathbb{E}[y|\cdot]$...

Kausalitet i Tobit

Så langt har jeg sett på Tobit som fremgangsmåte for å modellere $E[y|x]$ og $E[y|x, y > 0]$ for utfall som er begrenset til å være positiv gjennom en lineær latent modell. Har sett at $E[y|x]$ består av to komponenter: sannsynlighet for å ha positivt utfall og verdi gitt positiv. Hvis vi vil undersøke kausal effekt av binær tilfeldig behandling D_i på et slikt begrenset utfall kan det være fristende å bruke Tobit for å beregne de ulike effektene.[25] Hvis vi plotter histogram av betinget fordeling av utfall for behandling og kontroll grupper så vil de ha tyngdepunkt i 0 og en eller anne fordeling for positive verdier. Vi kan føle at forskjellen i gjennomsnittet ikke fanger alle aspekter ved behandlingseffekten, og det er forsåvidt sant. Vi kan bruke Tobit til å dekomponere behandlingseffekt,

$$
\begin{aligned}
&E[y_i|D_i = 1] - E[y_i|D_i = 0] && (10.172)\\
&= E[y_i|D_i = 1, y_i > 0]P[y_i > 0|D_i = 1] && (10.173)\\
&- E[y_i|D_0 = 1, y_i > 0]P[y_i > 0|D_i = 0] && (10.174)\\
&= E[y_i|D_i = 1, y_i > 0]P[y_i > 0|D_i = 1] - E[y_i|D_i = 1]P[y_i > 0|D_i = 0] && (10.175)\\
&+ E[y_i|D_i = 1]P[y_i > 0|D_i = 0] - E[y_i|D_0 = 1, y_i > 0]P[y_i > 0|D_i = 0] && (10.176)\\
&= \{P[y_i > 0|D_i = 1] - P[y_i > 0|D_i = 0]\}E[y_i|D_i = 1, y_i > 0] && (10.177)\\
&+ \{E[y_i|D_i = 1, y_i > 0]E - [y_i|D_0 = 1, y_i > 0]\}P[y_i > 0|D_i = 0] && (10.178)
\end{aligned}
$$

Hmm... litt usikker på om jeg skal beholde det. Poeng at *conditional on positive* (y>0) er å betinge på utfall av behandling. Gruppene med positivt utfall i behandling og kontroll er systematisk forskjellig på andre måter slik at forskjellene ikke fanger kausal effekt. Noe å tenke på dersom man bruker Tobit til å analysere kausal effekt...

10.7.6 Heltallsverdier (Poisson-regresjon)

I mange sammenhenger kan utfallsvariabelen kun ta ikke-negative heltallsverdier. Et eksempel er antall barn en kvinne føder eller antall dager før person havner tilbake i fengsel. Hvis vi bruker MLE må vi modellere hele den betingede fordelingen, men vi kan også avgrense oss til kun å modellere sentraltendensen $\mathbb{E}[y|\mathbf{x}] = g(\mathbf{x}, \beta)$. Siden y ikke kan

[25] Merk at binær behandling så er CEF, $E[y|D]$ nødvendigvis lineær slik at det er god grunn til å bruke vanlig regresjon. Mer generelt er de gjerne ikke-linære med LDV slik at det blir større motivasjon for å beregne form med ikke-lineær kurve... Det vil uansett gjerne være slik at det ikke er så stor forskjell i gjennomsnittlig marginal effekt.

ta negative verdier gir det lite mening om denne funksjonen gjør det. Et mulig valg er $g(\mathbf{x}, \beta) = \exp\{\mathbf{x}'\beta\}$. Denne funksjonen er ikke lineær i parametre, så vi kan ikke bruke vanlig closed form OLS. Vi kan derimot bruke ikke-linær OLS som er enkel generalisering og løse det numerisk.

I praksis kan det være greit å påføre mer struktur ved å modellere den betingede fordelingen. Hvis vi velger en parametrisk fordeling så kan vi estimere den betingede sannsynligheten for de ulike utfallene i stedet for bare å ha sentraltendens og mål på spredning. Et vanlig valg er poisson-fordeling.

$$p(y; \lambda) = \frac{e^{-\lambda}}{y!}\lambda^y \tag{10.179}$$

$$p(y|\mathbf{x}; \beta) = \exp(-\exp(\mathbf{x}'\beta))\exp(\mathbf{x}'\beta)^y/y! \tag{10.180}$$

$$LogL_n(\beta) = y_n\mathbf{x}'\beta - \exp(\mathbf{x}'\beta) \tag{10.181}$$

Poisson-fordelingen har egenskapen at $\mathbb{E}[y|\mathbf{x}] = \mathbb{V}[y|\mathbf{x}] = \lambda(\mathbf{x}, \beta) = \exp(\mathbf{x}'\beta)$. I praksis kan dette stemme dårlig med den gitte fordelingen vi observerer og dette bør vi ta hensyn til. MLE-estimatoren kan gi konsistent estimat på $\mathbb{E}[y|\mathbf{x}]$ i en større klasse av fordelinger, men standardfeilen som vi utleder fra informasjonsmatrisen vil være feil. Kan innføre ny parameter σ der $\mathbb{V}[y|\mathbf{x}] = \sigma\lambda(\mathbf{x}, \beta)$. Bruker som vanlig en sandwich til å skalere, men i dette tilfelle er estimator en skalar, slik at

$$\widehat{se}_{QMLE} = \hat{\sigma}\widehat{se}_{MLE} \tag{10.182}$$

må finne ut av dette senere.

10.8 Modellere seleksjon

Så langt har vi antatt at vi har et tilfeldig utvalg slik at det er representativt for populasjonen. Av ulike grunner så kan det være slik at vi ikke observerer alle egenskaper vi er interessert i for alle enheter. Noen kan la være å svare på alle spørsmål, noen kan la vær å svare overhodet, noen kan falle fra i forskningsopplegget slik at vi ikke ser utfallet. Selv med observasjonsdata kan det være systematiske skjevheter i hvilke variabler vi observerer for ulike personer. Man må for eksempel ha jobb for at vi skal observere lønn. Med slike ufullstendige utvalg kan ikke nødvendigvis konklusjon fra utvalg generaliseres til populasjon.

Jeg skal nå utvikle et rammeverk for å betrakte ufullstendige utvalg. Jeg vil se på egenskapene til estimatoren med ulike former for seleksjon og se på mulighet til å korrigere for eventuell skjevhet.

10.8.1 Rammeverk

Poenget er at vi kan konstruere en bineær tilfeldig variabel $s_n := I\{observerer\ hele(\mathbf{x}_n, y_n)\}$. Det er en tilfeldig variabel med betinget fordeling. Vi kan modellere denne og forsøke å undersøke hva som påvirker om vi observerer.

Vi kan anta at prosessen $(\mathbf{x}_n, y_n)_{n\in\mathbb{N}}$ oppfyller alle gode egenskaper, men istedet for tilfeldig utvalg $\{(\mathbf{x}_n, y_n) : n = 1, \ldots, N\}$ observerer vi $\{s_n(\mathbf{x}_n, y_n) : n = 1, \ldots, N\}$. Setter alle verdier lik null dersom noen mangler.

10.8.2 Avkortet (truncated) regression

I sensurert regression kan vi observere et tilfeldig utvalg av enheter i populasjon og deres karakteristikker, med unntak av utfall kan være sensuert i endepunktene. I avkortet regresjon blir observasjoner med gitte verdier av utfall ekskludert fra utvalget. Det er ikke lenger et tilfeldig (representativt) utvalg. Det er ingenting i veien for å estimere størrelser betinget av eksklusjonskriteriet, men dette kan avvike fra egenskaper ved underliggende latent modell som kan være det vi egentlig er interessert i. Begynner som alltid med å beskrive latent modell og relasjon mellom latente utfall og observerte utfall.

$$y^* = \mathbf{x}'\beta_0 + u, \quad u|\mathbf{x} \sim N(0, \sigma_0^2) \tag{10.183}$$

$$y = \begin{cases} y^*, & y^* > 0 \\ (y, x) \text{ ikke observert}, & y^* \leq 0 \end{cases} \tag{10.184}$$

10.8.3 Heckit..

Kapittel 11

Kalkulus

Jeg trenger litt greier for optimering. Ta recap på basic theorem, derivasjon og integral, optimering i $\mathbb{R}$ og $\mathbb{R}^2$ (både betinget og ubetinget), generalisering til flere dimensjoner, konkavitet, gradient, jacobi, hesse, lagrange, koblinger til lineær algebra (kvadratisk form, (lokalt) lineære transformasjoner,...)

11.1 Litt bakgrunn

11.1.1 Mengder

En menge er en kolleksjon av distinkte objekter. Objektene kalles elementer av mengden. Mengder betegnes ofte på formen

$$A = \{x \in S : P(x)\} \tag{11.1}$$

der S er en annen mengde som består av universet av objekter vi betrakter og $P(\cdot)$ er et medlemskriterium. Et objekt i univseret er element i mengden A dersom påstanden $P(\cdot)$ er sann når det blir evaluert for det objektet. En mengde B er er en delmengde av A hvis alle elementene i B også er element i A

$$B \subset A \iff x \in B \implies x \in A \tag{11.2}$$

De reelle tallene $\mathbb{R}$ er mengden av alle ikke-imaginære tall. Intervaller utgjør viktige delmengder. Eksempler på intervall er

$$A = \{x | a < x < b\} = (a, b) \subset \mathbb{R} \tag{11.3}$$

$$B = \{x | a \leq x \leq b\} = [a, b] \subset \mathbb{R} \tag{11.4}$$

der det første er åpnet og det andre er lukket. Det eksisterer en presis definisjon på om en mengde er åpen eller lukket som avhenger av om den inkluderer endepunktene.

Må inkludere dette senere. Mengder er veldig grunnleggende størrelser og vi vil beskrive relasjon mellom element i ulike mengder. La A og B være delmengder av mengden S, der S er universet vi betrakter

- Union: $A \cup B = \{x \in S | x \in A \text{ eller } x \in B\}$
- Interseksjon eller snitt: $A \cap B = \{x \in S | x \in A \text{ og } x \in B\}$
- Differanse: $A \setminus B = \{x \in S | x \in A \text{ og } x \notin B\}$
- Komplement: $A^C = \{x \in S | x \notin A\}$

Merk at det finnes ulike måter å gi ekvivalente representasjoner av samme utrykk. For eksempel er $A \setminus B = \{x \in S | x \in A \text{ og } x \notin B\} = x \in S \wedge x \in A \wedge x \notin B$. Tror jeg.. de har samme sannhetsmengde i hvertfall..Vi kan analysere den logiske formen til utrykk om mengder og betrakte sammenhengen mellom reglene for operasjoner på mengder og reglene for ekvivalens mellom utrykk. Eksempel

$$x \in A \setminus (B \cap C) \tag{11.5}$$
$$P \wedge \neg(Q \wedge R) \tag{11.6}$$
$$P \wedge (\neg Q \vee \neg R) \tag{11.7}$$
$$(P \wedge \neg Q) \vee (P \wedge \neg R) \tag{11.8}$$
$$x \in A \setminus B \cup A \setminus C \tag{11.9}$$

der jeg brukte at $x \in A$ (osv.) er påstander og dermed kan representeres med bokstav. Sannhetsverdi avhenger av variabel x så jeg kunne også ha betegnet det med $P(x)$. Eksempler på vanlige univers er

- $\mathbb{R}$, mengden av reelle tall, det vil si alle tall på tallinjen.
- $\mathbb{Z}$, menden av heltal $\{..., -1, 0, 1, ...\}$
- $\mathbb{Q}$, mengden av rasjonelle tall, det vil si tall som kan skrives som brøk av heltall.
- $\mathbb{N}$, mengden av naturlige tall, $\{0, 1, ...\}$. Merk at noen ikke inkluderer 0 som naturlig tall.

Det er også vanlig å avgrense disse mengdene, for eksempel ved å kun betrakte positive reelle tall. Det kan betegnes som $\mathbb{R}^+$. Elementene i mengder trenger ikke være tall. Det kan i prinsippet være alle mulige typer objekter, inkludert andre mengder. Mengder av mengder betegnes ofte som en familie. Et eksempel på dette er *power set* til en mengde A som består av alle delmengdene til A.

$$\mathscr{P}(A) = \{x : x \subseteq A\} \tag{11.10}$$

Ettersom mengder er så fleksible vil vi også ha en mer fleksibel måte å konstruere de. En alternativ måte er å bruke elementer fra en annen mengde I som indeks når vi konstruerer

$$P = \{p_i : i \in I\} \tag{11.11}$$

Ofte vil vi betrakte samling av element der hvert element består av komponenter som kommer fra ulike mengder. For å håndtere dette definerer vi en tuple som en endelig følge av distinkte objekt. Et eksempel på en tuple er $(a_1, ..., a_N)$. Mengder av tupler blir ofte konstruert av kartesisk produkt av mengder som består av alle tuplene som kan konstruerer fra de mengdene

$$A_1 \times ... \times A_N = \{(a_1, ..., a_N) | a_n \in A_n \text{ for } n = 1, ..., N\} \tag{11.12}$$

I praksis bruker vi ofte kryssprodukt av mengder av reelle tall der hvert element er en vektor. Vektorrommet $\mathbb{R}^N = \mathbb{R} \times ... \times \mathbb{R}$ der $\mathbf{x} = (x_1, ..., x_N) \in \mathbb{R}^N$. Husker at intervall er viktige delmengder av $\mathbb{R}$. Dette kan generaliseres til $\mathbb{R}^N$ som rektangler som består av kartesisk produkt av intervall

$$I = \times_{n=1}^{N} [a_n, b_n) = \{(x_1, ..., x_N) | a_n \leq x_n < b_n \text{ for } n = 1, ...N\} \tag{11.13}$$

11.1.2 Algebra

Det mulig å uttrykke påstander der sannhet avhenger av verdi til variabel x. En vanlig problemstilling med praktisk anvendelse er å finne sannhetsmengde til påstand. Delmengder av tallinjen er intervall eller kombinasjon av intervall (snitt/union). Avstand mellom to tall er absoluttverdi. Eksempler på påstander og deres sannhetsmengder:

$$|x| = D \iff x \in \{-D, D\} \iff x = -D \vee x = D \tag{11.14}$$

$$|x - a| = D \iff x = a - D \vee x = a + D \tag{11.15}$$

$$(2 - x)^{-1} < 3 \iff x \in (-\infty, \frac{5}{3}) \cap (-2, \infty) \tag{11.16}$$

Det tredje eksempel var mest sammensatt. Kan generelt løse ved å omskrive på form $\frac{a \cdot b}{c \cdot d} < 0$, finne ut for hvilke x-verdier hver av faktorene er negative og dermed finne sannhetsmengde til hele uttrykket. En annen vanlig problemstilling er å finne røttene til en funksjon, altså verdiene av x der $f(x) = 0$. En grunn til at dette dukker opp så ofte er at vi alltid kan omskrive $g(x) = h(x) \iff g(x) - h(x) = 0 \iff f(x) = 0$, der $f := g - h$.[1]. En mulig fremgangsmåte her er å faktorisere $f(x) = a \cdot b \cdot c \cdot ..$ og finne får hvilke x-verdier hver av faktorene er lik null.

[1] Jeg tror addisjon og subtraksjon av funksjoner er veldefinert, men er ikke helt sikker.

11.2 Litt analyse

Det sies at kalkulus er studie av endring. Vi mapper tall, så det er jo litt interessant å se hvor mye verdien av output endres når vi gjør en liten endring i input. Historisk har folk brukt konspetet om *infinitesmall* endringer, men i siste halvdel av 1800-tallet kom teorien på litt tryggere grunn med følger og grenser (som man kan lære mer om i reell analyse).

11.2.1 Følger

Vi kan betrakte en følge $(x_1, x_2, ...x_n, ...) := (x(n) : n \in \mathbb{N}) = (x_n)_{n\in\mathbb{N}}$ som en uendelig tuple der hvert positive heltall blir mappet til et objekt $x(n) := x_n$. Dette er en ganske generell datastruktur, men i kalkulus betrakter vi tilfelle der $x_n \in \mathbb{R}^N$. Vi er interessert i om følger konvergerer til et tall $\mathbf{r} \in \mathbb{R}^N$. Vi sier at $\lim x_n = \mathbf{r}$ eller $x_n \to \mathbf{r}$ dersom det er slik at uansett hvor lite vi gjør et nabolag om $\mathbf{r}$, så vil vi kunne finne en index N slik at alle elementene i følgen med høyere indeks befinner seg i nabolaget. Vi beskriver nabolaget med en såkalt ϵ-ball om $\mathbf{r}$,

$$B_\epsilon(\mathbf{r}) := \{\mathbf{x} : d(\mathbf{x}, \mathbf{r}) < \epsilon\} \tag{11.17}$$

der $d(\cdot)$ er en norm, for eksempel den eukledisk:$d(\mathbf{x}, \mathbf{r}) = \sqrt{(\mathbf{x}-\mathbf{r})'(\mathbf{x}-\mathbf{r})} := \|\mathbf{x}-\mathbf{r}\|$.

11.2.2 Rekker

En rekke er summen av leddene i en følge. Kan være entent endelig eller uendelig. Vi er interessert i om en uendelig rekke konvergerer. Tror vi kan betrakte summen av de n første leddene som element i en følge, $x(n) = \sum_{i=1}^{n} y_i$, og undersøke om $(x_n)_{n\in\mathbb{N}}$ konvergerer.

11.2.3 Grenser og kontinuitet

Bruker epsilon-delta til å definere grense.

$$\lim_{x\to a} f(x) = f(a) \iff \exists\delta|x-a| < \delta \implies |f(x) - f(a)| < \epsilon, \quad \forall\epsilon > 0 \tag{11.18}$$

Hvis grensen til f eksister i a så sier vi at funksjonen er kontinuerlig i a. Hvis funksjonen er kontinuerlig for alle $a \in I$ så er den kontinuerlig på intervallet I.[2] Det er også venstre- og høyrekontinuitet som jeg er litt usikker på hvordan man definerer formelt.

$$\lim_{x\to a} f(x) = f(a) \iff \lim_{x\to a^-} f(x) = f(a) \wedge \lim_{x\to a^+} f(x) = f(a) \tag{11.19}$$

[2]Sikkert noe tekniske greier om endepunktene.

11.2.4 Topologi

Vil innføre noen definisjoner til å beskrive mengden en funksjon er definert på.

Åpne mengder

En mengde er *åpen* dersom den ikke inkluderer grenseverdiene. Mer formelt er en mengde S åpen hvis det for hver $x \in S$ eksisterer en $\epsilon > 0$ slik at $B_\epsilon(x) \subset S$. Funksjoner på åpne mengder har ikke nødvendigvis noen ekstremverdier så de kan være vanskelig å optimere. Dette motiverer også bruk av inf og sup

Lukket mengde

En mengde er lukket dersom den inneholder grenseverdi til alle konvergente følger av elementer i mengden, $x_n \to x \implies x \in S$ dersom $x_n \in S \forall n \in \mathbb{N}$.

Begrenset mengde

En mengde er begrenset dersom det eksisterer et tall B slik at $\|x\| < B$ for alle $x \in S$.

Kompakt mengde

En mengde er kompakt dersom den er både lukket og begrenset.

11.3 Terminologi for funksjoner

Funksjoner er den nest mest grunnleggende størrelsen. En funksjon $f : A \to B$ er en regel som knytter hvert element i A til et unikt element i B. Mengden A er domenet til f og mengden B som den mapper til er *codomain*. Hvis $f(a) = c$ er a preimage av c under f. Verdimengden til funksjonen er delmengden av codomenet B der

$$\{b \in B | f(a) = b \; for \; noen \; a \in A\} \tag{11.20}$$

Denne mengden kalles $rng(f)$. En funksjon er injective (én-til-én) hvis distinkte element i A alltid mapper til distinkte element i B. Altså hvis

$$a \in A, a' \in A \text{ og } a \neq a' \implies f(a) \neq f(a') \tag{11.21}$$

Funksjonen $f : A \to B$ er *onto* hvis $rng(f) = B$. En funksjon er en *bijection* hvis den er både onto og én-til-én. Hvis en funksjon er én-til-én eksisterer det alltid en invers funksjon $f^{-1} : B \to A$ der $f^{-1}(b) = a \iff f(a) = b$.

11.3.1 Real valued functions

En funksjon f er en real valuedfunksjon hvis den mapper til tallinjen; $f : A \to \mathbb{R}$. Vi er ofte interessert i å finne for hvilke element i A at funksjonen tar sin høyeste verdi.

- Maximizer av f på A er $\{a^* \in A | f(a^*) \geqslant f(a), \forall a \in A\}$
- Maksimumsverdi er da $f(a^*)$

Det er en utfordring at maxmizers ikke alltid eksisterer. Dette gjelder for eksempel for monotont voksende funksjoner på åpne intervall. For å håndtere dette kan vi definerer en supremum $s := supA$ der (1) $a \leq s$, $\forall a \in A$ og (2) det eksisterer en følge $(x_n), x_n \in A$, slik at $x_n \to s$

11.3.2 Inverse funksjoner

Hvis $f'(x) > 0$ for alle $x \in I$ er funksjonen strengt monotont voksende på intervallet slik at $b > a \implies f(b) > f(a)$. Da eksisterer det en funksjon f^{-1} med definisjonsmengde $\{f(x) : x \in I\}$, verdimengde I og der $f^{-1}(f(x)) = x$. Det er en ny funksjon som tar output og mapper til input under f. Funksjonen f må være strengt monoton fordi ellers vil flere input mappe til samme output og dermed vil ikke den inverse tilfredstille definisjonen av en funksjon og er dermed ikke definert.

11.4 Lineær tilnærming av funksjoner

Funksjoner kan gjøres vilkårlig kompliserte og de kan potensielt være vanskelige å beskrive og manipulere. Mye av kalkulus handler om å finne en enklere (eg. lineær) representasjon av funksjonen som gir tilnærming av funksjonen i et omegn om $x^* \in S$.

11.4.1 Derivasjon

Hvis funksjonen er kontinuerlig så kan vi tegne punktene $\{(x, f(x) : x \in I\}$ uten å løfte blyanten. Dermed kan vi tenke oss å lage en sekant som er en rett linje mellom to punkter på grafen $(x, f(x)), (x + h, f(x + h))$ og se hva som skjer med helningen når h går mot 0. Dette er den deriverte til funksjonen i x.

$$f'(x) := \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \tag{11.22}$$

for at den deriverte skal være definert må grenseverdien eksistere. Da er kontinuitet en nødvendig, men ikke tilstrekkelig betingelse. $f'(x)$ er også en funksjon av x og denne må være kontinuerlig i a for at $f'(a)$ skal være definert. Uansett, fra definisjonen over kan

man utlede derivasjonsregler for alle (?) funksjonstyper og kombinasjoner av disse, så det er forholdsvis enkelt å derivere. Lite utvalg:

- Kjerneregel: $\frac{d}{dx}f(g(x)) = \frac{d}{dx}f(u) = \frac{df}{du}\frac{du}{dx}$
- Generalisering av kjerneregel:
- Noe om inverse funksjoner?

Når jeg jeg deriverer en funksjon $f : \mathbb{R} \to \mathbb{R}$ får jeg ut en ny funksjon $f' : \mathbb{R} \to \mathbb{R}$. Hvis jeg evalurer den i et element av inputmengden får jeg $f'(x^*) = a \in: \mathbb{R}$; et tall som angir grenseverdien av helningen til f i x^*. Jeg kan bruke dette til å beskrive funksjonen som gir verdier av tangentlinjen

$$g : s \mapsto f(x^*) + f'(x^*)(s - x^*) \approx f(x^* + s) \tag{11.23}$$

der jeg har bruke s istedet for Δx. Dette er en affine funksjon, men vi kan se på lineær representasjon ved å se direkte på endring av funksjonsverdi i stedet for nivå.

$$h : s \mapsto f'(x^*)(s - x^*) \approx f(x^* + s) - f(x^*) \tag{11.24}$$

Generelt kan vi skrive $h(s) = DF(s^*)x$. Hvis funksjonen $f : \mathbb{R}^N \to \mathbb{R}$ får vi lineær tilnærming

$$h : (s_1, ..., s_N) \mapsto \sum \frac{\partial}{\partial x_n} F(\mathbf{x}^*)s_n = DF(\mathbf{x}^*)\mathbf{s} \tag{11.25}$$

der $DF(\mathbf{x}^*) := \left[\frac{\partial}{\partial x_1}F(x^*), ..., \frac{\partial}{\partial x_N}F(x^*)\right]$ som også kalles Jacobi-matrisen til F i $\mathbf{x}^*$. Dette skiller seg fra gradienten som er en kolonnevektor. Vi kan enkelt utvide til en funksjon $F : \mathbb{R}^N \to \mathbb{R}^M$ ved å observere at vi kan behandle hver av komponentene separat.[3]

$$F(\mathbf{x}) = [F_1(\mathbf{x}), ..., F_M(\mathbf{x})]' \in \mathbb{R}^M \tag{11.26}$$

For å finne lineær tilnærming er det bare å derivere hver av de M komponent funksjonene med hensyn på de N inputvariablene

$$DF(\mathbf{x}^*) = [DF_1(\mathbf{x}^*), ..., DF_M(\mathbf{x}^*)]' \tag{11.27}$$

11.4.2 Taylor-tilnærming

Funksjoner kan være komplisert å arbeide med analytisk. Vi kan forsøke å finne en funksjon p der $p(x) \approx f(x)$ for x i et nabolag til a. Det kan være rimelig å kreve at $p(a) = f(a)$ og

[3] Husk at det finnes ulike måter å betrakte samme transformasjon. Vi kan for eksempel tenke på $\cos(x^2) := f(x)$ eller som $g(h(x))$ der $h : x \mapsto x^2$ og $g : y \mapsto \cos(y)$.

at $p'(x)|_a = f'(x)|_a$. Dette gir

$$f(x) \approx p(x) := f(a) + f'(a)(x-a) \tag{11.28}$$

Vi kan bruke dette til å beregne endring i y for små endringer i x

$$dy := p(x+dx) - p(x) \approx \Delta y := f(x+dx) - p(x) \tag{11.29}$$

Dette er en lineær tilnærming av funksjonen, også kalt første ordens taylor tilnærming. Hvis den n'te deriverte til f er definert i a kan vi generelt definere n'te ordens taylor tilnærming av f i a som

$$p(x) = f(a) + \sum_n^N \frac{1}{n!} f^(n)(a)(x-a)^n \approx g(x) \tag{11.30}$$

det er mulig å vise at differansen $g(x) - p(x)$ er gitt ved lagranges feilledd

$$R_{n+1}(x) = \frac{1}{(n+1)!} f^{(n+1)}(c) x^{n+1} \tag{11.31}$$

Størrelsen på leddet avhenger av c som er vanskelig å finne. Vi kan velge c som gir størst absoluttverdi av $f^{(n+1)}(c)$ for å finne en øvre begrensing på feilen i intervallet vi tilnærmer funksjon.

11.5 Noen vanlige funksjoner

11.5.1 Polynomial

11.5.2 Eksponential og logaritmer

Det er mange størrelser som vokser med % av egen verdi. Populasjoner, penger i banken, mm. Dette kan beskrives med en eksponetialfunksjon $f(x) = ca^x$ der

- $f(x+1)/f(x) = ca^{x+1}/ca^x = a \iff f(x+1) = f(x)a$
- $f(3) = c \cdot a \cdot a \cdot a$
- $f(0) = c$.
- $f(\frac{m}{n}) = ca^{m/n}$
- $f(-x) = ca^{-x} = \frac{c}{a^x}$

derivert? valg av grunntall

Det er en monoton funksjon som er strengt voksende for $a > 1$ og strengt avtagende for $a < 1$. Hvis $a = 1$ er $f(x) = c$ for alle x. Med unntak av dette tilfelle har eksponentialfunksjoner en invers som kalles logaritmen

egenskaper, litt usikker på hvordan utleder

- $\log(xy) = ln(x) + ln(y)$
- $\log(\frac{x}{y}) = ln(x) - ln(y)$
- $\log(\frac{1}{x}) = 1 - ln(x)$
- $\log(x^r) = r \cdot ln(x)$

Naturlig logaritme

One base to rule them all.

$$a = e^{ln(a)} \iff a^x = e^{ln(a)x} = e^{\lambda x}, \quad \text{der } \lambda = ln(a) \tag{11.32}$$

Log-lineær

Tror dette er når sammenhengen mellom variabler i lineær i logartimen, eks:

$$y = Ax^a \implies \log y = \log[Ax^a] = \log A + \log x^a = \log A + a \log x \tag{11.33}$$

11.5.3 Trigonometriske funksjoner

11.5.4 Sammensatte funksjoner

11.6 Kort om integral

11.7 Flervariable funksjoner

Betrakt en funksjon

$$f : \mathbb{R}^d \to \mathbb{R} \tag{11.34}$$

$$: \mathbf{x} = (x_1, ..., x_d) \mapsto f(\mathbf{x}) \tag{11.35}$$

Generaliseringen av den deriverte til funksjonen er gradienten som angir den partiellderiverte med hensyn på hver av variablene i inputvektoren.[4]

$$\nabla f : \mathbb{R}^d \to \mathbb{R}^d \tag{11.36}$$

$$: \mathbf{x} \mapsto \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, ... \frac{\partial f(\mathbf{x})}{\partial x_d} \right)' \tag{11.37}$$

Gradienten er en vektor som lever i inputspace. Eller i utgangspunktet er det en vektor av funksjoner. Det blir vektor av tall når vi angir spesifikk verdi av $\mathbf{x}$. Uansett hvor vi evaluere vil vektoren peke i retningen der funksjonen vokser raskest. Lengden på vektoren sier noe om hvor raskt den vokser.

Generaliseringen av den andrederiverte til funksjonen er hesse-matrisen

$$\mathbf{H}f : \mathbb{R}^d \to \mathbb{R}^{d \times d} \tag{11.38}$$

$$: \mathbf{x} \mapsto \mathbf{H}f(\mathbf{x}) \tag{11.39}$$

der

$$\mathbf{H}f(\mathbf{x})_{i,j} = \frac{\partial}{\partial x_j} \left(\frac{\partial f(\mathbf{x})}{\partial x_i} \right) \tag{11.40}$$

Funksjonen er strengt konkav hvis[5]:

$$\mathbf{x}' \left[\mathbf{H}f(\mathbf{x}) \right] \mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0} \tag{11.41}$$

11.8 Ubetinget optimering

11.8.1 Gradient descent

I maskinlæring har vi en kostnadsfunksjon $C : \Theta \to \mathbb{R}$ som angir sum av tap til hver av observasjonene i treningsdata for kandidatparameter θ. Jeg vil velge kandidaten som minimerer kostnaden. En mulighet er å finne gradientent $\nabla_\theta C$ og løse $\nabla_\theta C(\theta) = 0$. For å finne kandidater til optimum. Utfordringen er at $C(\cdot)$ kan være vilkårlig komplisert og avhenge av mange variabler slik at vi ikke klarer å løse det ikke-lineære ligningssystemet analytisk. Alternativet er da å gå frem numerisk. Hvis vi har et eksplisitt uttrykk for gradienten kan vi evaluere den i vilkårlig θ_{start}. Gradienten er vektor i parameterrommet som peker i retning der kostnaden vokser raskest og lengden avhenger av hvor bratt

[4] Tror vel egentilig generaliseringen er jacobimatrisen som er et lineær map. Må avklare forhold mellom jacobi og gradient..

[5] Merk at den kvadratiske formen er generaliseringen for om en matrise er *positiv* eller *negative*. Sier forhåpentligvis mer om dette i delen om lineær algebra.

funksjonen er. Vi velger derfor å gå i helt motsatt retning,

$$\theta_{ny} = \theta_{start} - \eta \nabla_\theta C(\theta_{start}) \tag{11.42}$$

der η er den såkalte læringsraten som skalerer gradienten og påvirker hvor raskt vi beveger oss i parameterrommet. Det er viktig at den ikke er for høy slik at vi overskyter bunnpunktet og begynner å divergere, men bør heller ikke være for lav slik at konvergens tar for lang tid. Bestemmer oss et treshold som avslutter algoritmen når $\|\nabla_\theta\| < k$ siden vi i praksis ikke for den eksakt lik null.

Merk at dette er en lokal metode som bare kjenner helning til funksjon akkurat der den blir evaluert. Sikrer kun at det lokale minimum som den konvergerer mot også tilsvarer det globale minimum dersom funksjonen er konveks. Skal nå se på alternativ fremgangsmåte som kan håndtere funksjoner med platå og flere lokale minimum.

Stokastisk gradient descent

Denne fremgangsmåten bruker kun subset av treningsdata når den evaluerer hvordan kostnaden blir påvirker av valg av parameter. Når vi evaluerer i ulike subsets får vi ulike kostnadsfunksjoner slik at gradient hopper litt rundt om kring, men i gjennomsnitt så vil den konvergere mot global minimum. Kan være effektiv fordi den bruker mindre data og dermed kjører raskere, selv om det tar flere steg og en litt mindre ryddig vei mot målet. Også fordel at den kan hoppe ut at av blindveier.

11.9 Betinget optimering

Kapittel 12

Lineær algebra

12.1 Vektorer

En vektor er en tuple med reelle tall, $\mathbf{x} = (x_1, ..., x_N)$, der $x_n \in \mathbb{R}$ for $n = 1, ..., N$. De er to grunnleggende operasjoner som er definert på vektorer: skalering og summering. I tillegg er det definert noen andre konsept som ikke lager nye vektorer

- Indre produkt: $\langle \mathbf{x}, \mathbf{y} \rangle = \sum x_n y_n$
- Eukledisk norm : $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$

Det er noen grunnleggende ulikheter som gjelder for disse

- Triangelulikheten: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
- Cauchy-Schwarz-ulikheten: $|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|$

Det er mulig å uttrykke nye vektorer som lineær kombinasjon av eksistererende. Hvis vi har en mengde av vektorer $X = \{\mathbf{x_1}, ...\mathbf{x_K}\} \subset \mathbb{R}^N$ så finnes det en mengde

$$\{y \in \mathbb{R}^N | y = \sum \alpha_k x_k \text{ der } \alpha_k \in \mathbb{R} \text{ og } \mathbf{x_k} \in X \text{ for } k = 1, ..K\} \tag{12.1}$$

Denne mengden av vektorer som kan uttrykkes som lineær kombinasjon av vektorene i X betegnes som $span(X)$. Som vi skal se er dette vesentlig for å vurdere eksistens av løsning på lineære ligningssystem, altså om det eksisterer en $\mathbf{x}$ slik at $y = \sum \alpha_k x_k$ for en gitt y. Løsningen eksister hvis og bare hvis $y \in span(X)$.

En mengde av vektorer er lineært uavhengige hvis

$$\sum \alpha_k x_k = \mathbf{0} \implies \mathbf{a} = \mathbf{0} \tag{12.2}$$

Dette impliserer også at ingen vektorer i mengden kan uttrykkes som en lineær kombinasjon av de resterende vektorene. Det er vesentlig for unikhet av løsning siden hvis X er

en lineær uavhenig mengde vil

$$y = \sum \alpha_k x_k = y = \sum \alpha'_k x_k \implies y = \sum (\alpha_k - \alpha'_k) x_k = \mathbf{0} \implies \alpha_k = \alpha'_k \tag{12.3}$$

En mengde av lineært uavhengige vektorer Z utgjør en basis for $span(X)$ hvis $span(Z) = span(X)$. For øvring utgjør $span(X)$ et underrom av $\mathbb{R}^N$ siden det er lukket under skalering og addisjon. Generelt er en mengde S et underrom av $\mathbb{R}^N$ hvis det for alle $\alpha \in \mathbb{R}$

- $\mathbf{x} \in S \implies \alpha\mathbf{x} \in S$
- $\mathbf{x}, \mathbf{y} \in S \implies \mathbf{x} + \mathbf{y} \in S$

og dimensjonen til et underrom er antall vektorer i basisen.

12.2 Matriser

Kan bruke matriser til å *stacke* linære ligningssystem,

$$a_{11}x_1 + a_{12}x_2 = y_1 \tag{12.4}$$
$$a_{21}x_1 + a_{22}x_2 = y_2 \tag{12.5}$$

tilsvarer

$$\begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{12.6}$$
$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{12.7}$$
$$\mathbf{A}\mathbf{x} = \mathbf{y} \tag{12.8}$$

Vi bruker $\mathbf{A}_{ij}$ til å referere til komponent fra i'te rekke og j'te kolonne. Ofte har vi lyst til å gi en representasjon av $\mathbf{A}$ uten å beskrive alle de individuelle komponentene. Vi kan da gruppere de inn i rekke- og kolonnevektorer. Jeg er litt usikker på hvilken notasjon jeg bruker.[1] Jeg kan da gi alternativ representasjon av matrisemultiplikasjon

[1] En mulighet er å bruke $\mathbf{A}_i$ til å betegne rekke og $\mathbf{A}_j$ for kolonne. Et mulig problem er at det er ambigiuøst om jeg betrakter $\mathbf{A}_i$ som en kolonnevektor; altså transponerte av rekken i matrisen. Også problem om jeg vil referere til spesifikk tall. Kan brukе$\mathbf{a}_{\bullet 1}$ og $\mathbf{a}_{1\bullet}$ til å referere til henholdsvis første kolonne og rekke. hmm

med utgangspunkt i disse blokkene,

$$\mathbf{A}\mathbf{x} = \begin{bmatrix}\mathbf{a}_1 & \mathbf{a}_2\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix}\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 \tag{12.9}$$

$$= \begin{bmatrix}\mathbf{a}_{1\bullet} \\ \mathbf{a}_{2\bullet}\end{bmatrix}\mathbf{x} = \begin{bmatrix}\mathbf{a}_{1\bullet}\mathbf{x} \\ \mathbf{a}_{2\bullet}\mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{a}_1'\mathbf{x} \\ \mathbf{a}_2'\mathbf{x}\end{bmatrix}, \quad \mathbf{A} := \begin{bmatrix}\mathbf{a}_1' \\ \mathbf{a}_2'\end{bmatrix} \tag{12.10}$$

der jeg i siste likhet lar $\mathbf{a}_1$ være rekke i matrisen på kolonneform.[2]. Det er også poeng at man kan blokkpartisjonere matriser på andre måter og gjøre operasjon på blokkene så lenge de er kompatible, men det må bli en annen gang.

$$A = \begin{bmatrix}a & \dots \\ \vdots & \end{bmatrix} \tag{12.11}$$

12.2.1 Derivasjon (flytte?)

Jeg vil nå begynne å derivere med hensyn på vektor. Det er litt som å ta partiell derivert med hensyn på hver av komponentene i en matrise og stacke de oppå hverandre,

$$\frac{d}{d\mathbf{x}}f(\mathbf{x}) = \begin{bmatrix}\frac{\partial}{\partial x_1}f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_K}f(\mathbf{x})\end{bmatrix} \tag{12.12}$$

Litt usikker på dimensjonene. Går ut i fra at den deriverte har samme dimensjon som $\mathbf{x}$..Noen regneregler [3]

$f(\mathbf{x})$	$\frac{d}{\mathbf{x}}f(\mathbf{x})$
$\mathbf{A}\mathbf{x}$	$\mathbf{A}$
$\mathbf{a}'\mathbf{x}$	$\mathbf{a}$
$\mathbf{x}'\mathbf{a}$	$\mathbf{a}$
$\mathbf{x}'\mathbf{x}$	$2\mathbf{x}$
$\mathbf{x}'\mathbf{A}\mathbf{x}$	$2\mathbf{A}\mathbf{x}$

Eksempel: minste kvadrat

Har tapsfunksjon

$$L = \frac{1}{N}\sum(y_n - \mathbf{x}_n'\mathbf{b})^2 \tag{12.13}$$

[2]Notasjon kan bli ganske forvirrende

[3]Merk at de resuserer til vanlige derivasjonsregler dersom matriser og vektor bare har én komponent

Vil ha en vektor der n'te komponent er $\mathbf{x}_n'\mathbf{b}$. Tilsvarer $\mathbf{Xb}$ der $\mathbf{X}_{n\bullet} = \mathbf{x}_n'$. Kan da skrive det på matriseform,

$$L = \frac{1}{N}(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) \tag{12.14}$$
$$= \frac{1}{N}(\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{Xb}) \tag{12.15}$$
$$= \frac{1}{N}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{Xb} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \tag{12.16}$$
$$= \frac{1}{N}(\mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{Xb} + \mathbf{b}'\mathbf{X}'\mathbf{Xb} \tag{12.17}$$

Ser bort i fra skaleringen $\frac{1}{N}$ og deriverer med hensyn på $\mathbf{b}$,

$$\frac{dL}{d\mathbf{b}} = 2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{Xb} = \mathbf{0} \tag{12.18}$$
$$\implies \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{12.19}$$

gitt at $\mathbf{X}'\mathbf{X}$ er invertibel.

12.3 Lineære transformasjoner

En funksjon $T : \mathbb{R}^K \to \mathbb{R}^N$ er lineær transformasjon hvis

$$T(\alpha\mathbf{x} + \beta\mathbf{y}) = \alpha T\mathbf{x} + \beta T\mathbf{y} \tag{12.20}$$

for alle $\mathbf{x}, \mathbf{y} \in \mathbb{R}^K$ og $\alpha, \beta \in \mathbb{R}$. Vi skriver funksjonen med stor bokstav og uten parantes fordi den oppfører seg som en matrise. Skal vise senere at det er én-til-én korrespondanse mellom matriser og lineære trasnformasjoner mellom vektorrom. Dette er en veldig grei egenskap til lineære transformasjoner som gjør de er mye brukt i anvendt matematikk. Definisjonen over kan generaliseres til

$$T\left(\sum \alpha_k \mathbf{x}_k\right) = \sum \alpha_k T\mathbf{x}_k \tag{12.21}$$

dette impliserer at

$$T\mathbf{x} = \sum \alpha_k T\mathbf{e}_k \tag{12.22}$$

slik at $rng(T) = span(V)$, der $V = \{T\mathbf{e}_1, ..., T\mathbf{e}_K\}$. For $T : \mathbb{R}^N \to \mathbb{R}^N$ så er det ekvivalnes mellom ikke-singularitet og masse greier. Det er en ideell situasjon siden det alltid eksisterer en unik løsning. I praksis må vi ofte finne tilnærmet løsning fordi hvis $T : \mathbb{R}^K \to \mathbb{R}^N$, der $K < N$, så kan det være slik at $\mathbf{y} \notin rng(T)$. Altså finnes det ingen $\mathbf{x}$ slik at $T\mathbf{x} = \mathbf{y}$. Vår beste løsning da er å finne $\mathbf{x}'$ som minimerer $\|\mathbf{y} - T\mathbf{x}'\|$. For å minimere avstand mel-

lom to en vektor og et underrom får vi bruk for ortogonale projeksjoner, fordi løsningen er å gå den strakeste vegen.

12.4 Ortogonale projeksjoner

To vektorer er ortogonale hvis $\langle \mathbf{x}, \mathbf{y} \rangle = 0 \iff \mathbf{x} \perp \mathbf{y}$. Dette konseptet generaliserer også til andre objekt som vi kan definere indre produkt på, som f.eks. tilfeldige variabler. En vektor kan også være ortogonal på en mengde S

$$\mathbf{x} \perp S \iff \mathbf{x} \perp \mathbf{z} \text{ for alle } \mathbf{z} \in S \tag{12.23}$$

Mengden av alle vektorer som er ortogonal på mengden S utgjør dets ortogonale komplement

$$S^{\perp} = \{\mathbf{x} | \mathbf{x} \perp \mathbf{z}, \text{ for alle } \mathbf{z} \in S\} \tag{12.24}$$

En mengde av vektorer X er ortogonale hvis vektorene er parvise ortogonale $\mathbf{x}_j \perp \mathbf{x}_k$, $j \neq k$. Den er i tillegg ortonormal hvis $\|\mathbf{x}\| = 1$ for alle $\mathbf{x} \in X$. Dette er en veldig grei egenskap siden det gjør det enkelt å finne vektene i lineære kombinasjoner

$$\mathbf{y} = \sum \alpha_k \mathbf{x}_k = \sum \langle \mathbf{y}, \mathbf{x}_k \rangle \mathbf{x}_k \tag{12.25}$$

Dette gjør de velegnet som basiser for vektorrom. Det eksisterer alltid ortonormal basiser og vi kan bruke algoritmer (eg. Gram-Schmidt) for å konstruere.

Uansett, det store resultatet er ortogonale projeksjons theoremet! La S være et underrom av $\mathbb{R}^N$ og $\mathbf{y} \in \mathbb{R}^N$. Vi vil finne

$$\hat{\mathbf{y}} = \underset{\tilde{\mathbf{y}} \in S}{\arg\min} \|\tilde{\mathbf{y}} - \mathbf{y}\| \tag{12.26}$$

Theoremet sier da at dette har en unik løsning der $\mathbf{y} - \hat{\mathbf{y}} \perp S$. Merk at for alle andre $\mathbf{z} \in S$ så er

$$\|\mathbf{y} - \mathbf{z}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{z}\|^2. \tag{12.27}$$

Som et eksempel kan vi projektere $\mathbf{y}$ på $\mathbf{1}$. Vil finne $\hat{\mathbf{y}} \in span(\mathbf{1})$ som minimerer avstand til $\mathbf{y}$ og vet at $\langle \mathbf{y} - \alpha \mathbf{1}, \mathbf{1} \rangle = 0$. Kan da finne $\alpha = \bar{y}_N$. Mer generelt kan vi betrakte projeksjon

på et vektorrom S med flere dimensjoner. La $S = span(X)$, der $X := \{\mathbf{x}_1, ..., \mathbf{x}_K\}$.

$$\mathbf{y} - \hat{\mathbf{y}} \perp \mathbf{x}_k, \quad k = 1, ...K \tag{12.28}$$

$$\implies \mathbf{x}_k'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}, \quad k = 1, ..., K \tag{12.29}$$

$$\implies \mathbf{X}'(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0} \tag{12.30}$$

$$\implies \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y \tag{12.31}$$

Det eksisterer en lineær transformasjon som utfører projeksjonen, $\mathbf{P} : \mathbf{y} \mapsto \hat{\mathbf{y}} = proj_S\mathbf{y}$. Fra utledningen over følger det at $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$. Denne transformasjonen har en del egenskaper som følger ganske intuitivt fra at det er en projeksjon. Ofte er vi også interessert i residualen, som vi kan finne med $\mathbf{M} := \mathbf{I} - \mathbf{P}$.

La nå S ha en ortonormal basis U. Merk at $\mathbf{u}_j'\mathbf{u}_k = 0, j \neq k$ og lik 1 hvis $j = k$. Det følger da at $\mathbf{U}'\mathbf{U} = \mathbf{I}$ slik at $\mathbf{P} = \mathbf{U}\mathbf{U}'$ og $\mathbf{P}\mathbf{y} = \sum \mathbf{u}_k\langle\mathbf{u}_k, \mathbf{y}\rangle$. Merk generelt at $\mathbf{y} \in span(X)$ alltid kan skrives som $\mathbf{X}\mathbf{b}$ for noen $\mathbf{b}$, men det kan være vanskelig å finne vektene i den lineære kombinasjonen. Med ortonormal basis blir dette enklere fordi vektene ikke bidrar"i samme retninger og enhetslengde gjør det enkelt å finne riktig skalering... Har ikke helt intuisjonen på dette, men gir sånn omtrent mening.

Kapittel 13

Appendix

13.1 Logikk

En *conjecture* (formodning) er noe som vi tror kan være sant. Hvis vi kan bevise at det er sant kan vi kalle det er theorem. Et eksempel på conjecture er at $x^2 > x$ for alle $x > 1$. Hvordan kan vi bevise at dette er sant? Det er ikke tilstrekkelig å finne eksempelverdier x der det er sant fordi det ikke gir noen garanti for at det sant for alle $x > 1$.

Målet er å bevise formodninger (og eventuelt avsløre om de ikke er sanne). For å klare dette må jeg først innføre et rammeverk for å konstruere gyldige argumenter. De primitive størrelsene i argument er påstander som vi betegner med stor bokstav samt *connectiver* som binder påstander sammen i såkalte formler.[1]. Eksempel på påstand: P = Sverre er kul". Påstander er enten sanne eller usanne. Fra påstander kan vi konstruere premisser og konklusjoner.[2] Et argument er gyldig dersom konklusjonen alltid er sann når premissene er sanne. Et eksempel på gyldig argument er

$$\begin{array}{l} P \vee Q \\ \neg P \\ \hline \therefore Q \end{array}$$

Vi har i hovedsak tre måter å kombinere påstander; $(\wedge), (\vee), (\neg)$. Merk at vi ikke oversetter direkte fra symbol til ord når vi bruker formler til å konstruere setninger (og omvendt). Det er en liten kunst å bevege seg mellom matematikske symbol og ord.

Sannhetsverdi til formler avhenger av sannhetsverdiene til påstandene de består av. For å undersøke i hvilke tilfeller formlene er sanne kan vi konstruere sannhetstabeller som angir sannhetsverdi til formler for alle mulige kombinasjoner av sannhetsverdi til påstander (2^N for N ulike påstander). Dette kan brukes til å undersøke om konklusjon alltid er sann i tilfellene der premissene er sanne. Vi kan merke oss at en formel er en

[1] Liker ikke order formel". Kan kanskje kalle det for utrykk. Det er sammensatt av påstander, men tror også jeg kan betrakte hele utrykket som en påstand. På en annen side er det kanskje lurt å ha distinksjon

[2] Utrykk som er sammensatt av påstander og dermed entent sann eller usann.

tautologi dersom den alltid er sann og en selvmotsigelse dersom den aldri er sann.

Dersom to formler har samme sannhetsverdi for alle kombinasjoner av påstander er de ekvivalente. Analogt med algebra er det operasjoner som vi kan utføre på formler som bevarer ekvivalens. Noen eksempler:

1. $\neg(P \wedge Q) \iff \neg P \vee \neg Q$
2. $\neg(P \vee Q) \iff \neg P \wedge \neg Q$
3. $P \wedge (Q \vee R) \iff (P \wedge Q) \vee (P \wedge R)$
4. $P \vee (Q \wedge R) \iff (P \vee Q) \wedge (P \vee R)$

De ulike reglene kan vi bruke til å forenkle og omformulere uttrykk. Det er også en betinget connectiv; hvis P så Q. Dette er det samme som å si at P er tilstrekkelig for Q. Det betyr at P ikke kan være sann samtidig med at Q er usann.

$$(P \implies Q) \iff \neg(P \wedge \neg Q) \iff (\neg P \vee Q) \tag{13.1}$$

Over har jeg også brukt $\iff$ der $P \iff Q$ betyr $P \implies Q \wedge P \impliedby Q$, altså at P både er nødvendig og tilstrekkelig betingelse for Q. Det betyr at de er ekvivalente i betydningen at de har samme sannhetsverdier i tabellen.

Påstander kan også avhenger av variabler, som er objekt som kan ta ulike størrelser og er betegnet med en bokstav. For å håndtere dette utvider vi notasjonen for påstand til $P(x)$. Hvorvidt denne påstanden er sann avhenger av verdi til variabel. Mengden av verdier der påstanden er sann kalles sannhetsmengden til $P(x)$ og betegnes som $\{x : P(x)\}$.

Vi vil ofte si noe om innholdet i sannhetsmengden til påstander. Vi har to såkalte *quantifiers*:

1. Universal: $\forall x P(x)$, betyr at påstand er sann for alle x i universet vi betrakter
2. Eksistens: $\exists x P(x)$, betyr at det eksisterer minst én x der påstand er sann

I tillegg brukes $\exists! x P(x)$ som betyr at det eksisterer én og bare én x som gjør påstand sann. Dette er mer en notasjonell konvensjon. Vi kan også omformulere utrykk med quantifiers for å finne ekvivalente representasjoner. Merk at

1. $\forall x P(x) \Leftrightarrow \neg\exists x \neg P(x)$. Hvis det er sant for alle x kan det ikke eksistere en x der det ikke er sant. Medfører $\neg\forall x P(x) \Leftrightarrow \exists x \neg P(x)$.
2. $\exists x P(x) \Leftrightarrow \neg\forall x \neg P(x)$. Hvis det finnes minst én x der det er sant, så kan det ikke være usant for all x. Medfører $\neg\exists x P(x) \Leftrightarrow \forall x \neg P(x)$.

Eksempel: Negate uttrykket *everyone has a relative he doesn't like* og utrykk det positivt

$$\neg\forall x\exists y[R(x,y) \wedge \neg L(x,y)] \tag{13.2}$$

$$\exists x\neg\exists y[R(x,y) \wedge \neg L(x,y)] \tag{13.3}$$

$$\exists x\forall y\neg[R(x,y) \wedge \neg L(x,y)] \tag{13.4}$$

$$\exists x\forall y[\neg R(x,y) \vee L(x,y)] \tag{13.5}$$

$$\exists x\forall y[R(x,y) \Rightarrow L(x,y)] \tag{13.6}$$

Ikke alle har en slektning som de ikke liker. Det er ekvivalent med at det eksisterer noen som liker alle slektningene sine. For å omformulere det til et positivt uttrykk må vi flytte negasjonen frem til påstandene og deretter distribuere. Merk at rekkefølgen til *quantifiers* har betydning dersom de er ulike, men dersom de er like kan vi lese $\exists x\exists y$ som at det eksisterer x og y slik at (...), og rekkefølgen har ikke betydning. Merk også at dersom vi bruker quantifiers om to størrelser fra samme univers må vi spesifisere eksplisitt dersom $x \neq y$.

Kan også betegne begrensede quantifiers som jeg tror bare er notasjon, men litt usikker på dette

$$\forall x \in SP(x) \Leftrightarrow \forall x(x \in S \Rightarrow P(x)) \tag{13.7}$$

13.2 Bevise theorem [Må omskrives]

Theorem er argument som har formen at gitt premissene (hypotesene) er sanne, så er også konklusjonen sann. Premissene inkluderer ofte frie variabler og ved sett in verdier av disse får jeg et *instance* av theoremet. Den fremgangsmåten vil ikke være tilstrekkelig siden jeg vil vise at det er sant for *alle* instancer. På en annen side kan fremgangsmåten brukes til å vise at at theorem ikke er sant, fordi en eneste motsigelse viser at konklusjonen ikke med nødvendighet er sann dersom premissene er sanne.

Jeg skal nå begynne å utvikle strategier for å bevise theorem. Det betyr å vise at argumentet er gyldig så jeg får bruk for greiene om logikk. I praksis gjøres bevis i flere steg. Stegene er ikke lineære. Det har en nested struktur for vi manipulerer strukturen, gjør ting og deretter rekonstruerer struktur slik at vi beviser opprinnelig påstand. Vi får bruk for begrepene *givens* og *goals* som er henholdsvis tingene vi antar er sanne og ting vi vil vise er sanne i ulike deler av beviset. Merk at premiss og konklusjon er henholdsvis given og goal på begynnelsen. Merk også at det er en viktig distinksjon mellom å *anta* at P sant og å *hevde* (assert) det. Hvis vi hevder at noe er sant må vi bevise det, mens vi kan anta basically hva som helst. Hvis vi deretter viser at $P \Rightarrow Q$ har vi bevist Q dersom vi hevder at P er sant, men vi kan ikke si noe om sannhetsverdi til Q dersom vi kun har antatt P.

13.2.1 Strategier

Jeg skal nå begynne å utvikle ulike strategier for å bevise theorem. Eksempel: Bevis $P \Rightarrow Q$

```
Anta at P
[Bevis at Q]
Derfor vist at P medfører Q
```

En alternativ og veldig fleksibel strategi er bevis ved motsigelse.

```
Anta at P
Anta at ikke Q
[Bevis at ikke P]
Derfor vist at P medfører Q
```

Goals:

- $P \Rightarrow Q$.
    1. Legg P til i given og bevis Q.
    2. Legg $\neg Q$ til i given og bevis $\neg P$
- $\neg P$
    1. Prøve å finne en ekvivalent for som ikke er negert
    2. Legg til P i given og forsøk å finne en selvmotigelse av ting vi har antatt er sanne.
- $\forall x P(x)$
    1. La x være arbritrær og bevis P(x). Ofte er det implisitt at x er arbitrær, men vi må passe på å ikke gjøre noen antagelser. Vi tar beviset for en spesifikk x, men siden den er vilkårlig gjelder det for alle x.
- $\exists x P(x)$
    1. Vise at det er sant for en bestemt x=a som jeg finner, f.eks. ved å løse ligning.
- $P \wedge Q$
    1. Bevise P og Q separat.
- $P \Leftrightarrow Q$
    1. Det er bare et spesialtilfelle der vi ha $P \Rightarrow Q \wedge Q \Rightarrow P$

- $P \vee Q$
    1. Kan anta $\neg P$ og bevise Q eller omvendt.

Har også strategier for hva jeg kan gjøre med givens

- $P \Rightarrow Q$.
    1. Enten vise at P eller $\neg Q$
- $\neg P$
    1. Prøve å finne en ekvivalent for som ikke er negert
    2. Hvis målet er å finne selvmotigelse kan jeg flytte P over til mål.
- $\forall x P(x)$
    1. Plugge inn en verdi a. Finner spesifikke verdi fra de resten av beviset...
- $\exists x P(x)$
    1. Innfører en x_0 som er en verdi der $P(x_0)$.
- $P \wedge Q$
    1. Behandle som to separate givens.
- $P \Leftrightarrow Q$
    1. Det er bare et spesialtilfelle der vi ha $P \Rightarrow Q \wedge Q \Rightarrow P$
- $P \vee Q$
    1. Bevise separat, én med P og én med Q som given.

Intuisjonen er at vi har en liste med givens. Hvis vi med utgangspunkt i disse kan lage en motsigelse betyr det at minste én av givens ikke er sanne, men siden vi har antatt at de er sanne må den usanne være påstanden som vi la til... Utfordringen er at vi flytter negasjon av goal inn i givens, som gjør at vi ikke har et opplagt goal å jobbe mot.

13.2.2 Eksempler på bevis

Mye av arbeidet skjer inne i brakkeparentes. Ulike ting man kan gjøre er å skrive ut definisjonene til de matematiske utrykkene i premissene og eventuell forsøke å utlede nye ting som er sant og som vi kan bruke. Vi kan også manipulere både givens og goals til

ekvivalente representasjoner. Direkte bevis manipulerer givens slik at de ligner på goal. Eksempel: Bevis at summen av oddetall og partall er oddetall.

$$x = 2z, \exists z \in \mathbb{Z}, y = 2w - 1, \exists w \in \mathbb{Z} \iff x + y = 2z + 2w - 1 = 2v - 1 \tag{13.8}$$

der $v = z + w$ $in\mathbb{Z}$. Vi bruker logikk og symboler når vi utleder beviset, men det endelige resultatet skal bestå av vanlig tekst. Skal nå forsøke å skrive ut beviset

Theorem 1. *La a være et partall og la b være et oddetall. Da er $z = a + b$ er oddetall.*

Bevis. Siden a er partall eksisterer det et heltall n slik at $a = 2n$ og siden b er oddetall eksisterer det et heltall m slik at $b = 2m - 1$. Det medfører at summen av tallene kan skrives som $2m + 2n - 1$. Det medfører at det eksisterer et heltall $x = 2(m + n)$ slik at $a + b = 2x - 1$ som er et oddetall. □

Theorem 2. *La $A \setminus B$ være disjunkt fra C og la $x \in A$. Hvis $x \in C$ er da $x \in B$.*

Bevis. Anta at $x \in C$. Jeg beviser med motsigelse og antar derfor at $x \notin B$. Da er $x \in A \setminus B$ og $x \in C$. Disse mengdene er disjunkte og dette er derfor en motsigelse. Jeg konkluderer derfor med at $x \in B$. Det følger derfor at hvis $x \in C$ er $x \in B$. □

Theorem 3. *Hvis $\exists x(P(x) \Rightarrow Q(x))$ så er $\forall x P(x) \Rightarrow \exists x Q(x)$.*

Bevis. Anta at $P(x) \Rightarrow Q(x))$ er sant for $x = x_0$. Anta $P(x)$ er sann for alle x. Da er $P(x_0)$ sann. Siden $P(x_0)$ impliserer $Q(x_0))$ er $Q(x_0)$ sann. Det medfører at det eksisterer en x slik at $Q(x)$ er sann. Det medfører at for alle x vil P(x) implisere at det eksister en x slik at $Q(x)$. □

TODO: Tror det er god idé å ta litt flere eksempler på bevise og se litt på hvordan de kan formuleres på en gode måte..