Truncated data (trunkerad data/censorerad data)

Introduktion

- Trunkerade data kontra censuerade data.
- Trunkerade stokastisk varaibel.
- Trunkerad normalfördelning.
- Moment av trunkerade fördelningar.
- Trunkerad regressionsmodell.

Trunkerade vs censurerade data

Låt oss först diskutera vad trunkering och censorering om:

- **Trunkering handlar om urval/inkludering (selection):** vissa enheter kommer aldrig med i datamängden.
- **Censorering = mätning/registrering (measurement):** enheter är med, men y är **inte exakt observerad** för vissa.

Trunkering (truncation)

Vad händer i datan?

Vi har observationer utanför ett intervall som **saknas helt**. Du ser dem inte, du vet inte hur många de är (utan extern information).

Exempel (inkomst)

Du genomför en studie “inkomster under 30 000 kronor/mån” och **samplar bara** personer under denna nivån. Då finns inga rader i datan med inkomst över 30 000 kronor på grund av att dessa är bortfiltrerade redan vid urvalet.

Konsekvensen av trunkering

Man har egentligen data från en **subpopulation**:

$$Y|(Y \leq c)$$

och likelihood måste därför korrigeras för att urvalet är villkorat:

$$f_{trunk}(y) = \frac{f(y)}{P(Y \leq c)}, \quad y \leq c$$

Censorering (censoring)

Vad händer i datan?

Alla enheter finns med (eller åtminstone kan finnas med), men för vissa observationer vet du bara att värdet ligger **över/under en gräns.**

Exempel (inkomst)

Du samplar alla, men inkomster över 30 000 registreras som “30 000+” eller till och med bara “30 000”. Då finns de personer i datan, men deras sanna inkomst är okänd bortom gränsen.

Vi har två vanliga varianter av dessa:

- **Right censoring (typ 30 000+):** du vet att $Y \geq c$, men inte exakt hur mycket.
- **Top coding (Tobit-lik):** du lagar $Y_{obs} = \min(Y^*, c)$. Det vill säga alla över gränsen får samma observerade värde c.

Konsekvens (likelihood ide)

- För exakta värde: bidrag $f(y_i)$.
- För censurerade (t.ex $Y_i \geq c$) : bidrag $P(Y \geq c) = 1 - F(c)$.

Snabbt sätt att känna igen skillnad med trunkerad och censorerad

Tänk: **finns personer i datan?**

- **Trunkering:** Personer med $Y > c$ finns **inte alls** i datasetet.

- **Censorering:** Personer finns i datasetet men Y är bara känt som $\geq c$ eller satt till c.

Täthetsfunktionen för en stokastisk variabel

Teorin här är mycket komplicerad och väldigt mycket därför kommer jag i följande delar att gå igenom allt långsamt och djupt.

1) Vad betyder trunkering underifrån vid a?

Vad detta betyder är att vi tittar på en slumpvariabel X **givet att** $X > a$**.** Alla observationer med $X \leq a$ är helt borta.

Då vill vi ha en ny täthet som bara lever på intervallet $[a, \infty]$ och som fortfarande integrerar till 1. **Observera att en täthet är definerad då den integreras till 1.**

2) Varför delar man med P(X>a)

Den gamla tätheten $f(x)$ är normaliserad så att $\int_{-\infty}^{\infty} f(x)dx = 1$. Men om vi bara behåller området $x > a$ så blir massan som finns kvar:

$$P(X > a) = \int_{a}^{\infty} f(x)dx$$

Den nya trunkerade tätheten skall bli:

$$f(x|X > a) = \begin{cases} \dfrac{f(x)}{P(X > a)}, & x > a, \\ 0, & x \leq a. \end{cases}$$

Det är bara “samma form” som innan, men uppskalad så att arean över $x > a$ integreras till 1. Man kan se det som normaliseringskonstanten i normalfördelningen vilket gör att fördelningen integreras till 1.

Exempel

Vi har en slumpvariabel $X \sim U(0, 1)$ som är likformigt fördelat på intervallet 0 och 1. Det gäller att:

- $f(x) = 1, \quad 0 \leq x \leq 1$

- annars 0

Om man trunkerar den stokastiska variabeln vid $a = \frac{1}{3}$då fås:

$$P(X > \frac{1}{3}) = \int_{1/3}^{1} 1dx = 1 - 1/3 = 2/3$$

Det vill säga att vi får då:

$$f(x|X > 1/3) = \frac{f(x)}{P(X \geq \frac{1}{3})} = \frac{1}{2/3} = \frac{3}{2}, \quad \frac{1}{3} \leq x \leq 1.$$

Trunkerad normalfördelning

Om $X \sim N(\mu, \sigma^2)$ då gäller:

$$P(X > a) = 1 - \Phi(\frac{a - \mu}{\sigma}) = 1 - \Phi(\alpha)$$

Här är $\alpha = \frac{a-\mu}{\sigma}$ och $\Phi(\cdot)$ är fördelningsfunktionen för en standard normalfördelad variabel.

Slumpvariabeln Xs täthetsfunktion kan skrivas som:

$$f(x) = \frac{dF(x)}{dx} = \frac{d\Phi(\frac{x-\mu}{\sigma})}{dx} = \frac{1}{\sigma}\phi(\frac{x - \mu}{\sigma})$$

Detta ger den **trunkerade normalfördelningen:**

$$f(x|X > a) = \frac{f(x)}{1 - \Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{1 - \Phi(\alpha)}$$

Här är $\phi(\cdot)$ är täthetsfunktionen för en standard normalfördelad variabel.

Moment av trunkerade fördelningar

Innan vi går vidare måste man först ha en förståelse av vad ett **moment** är. Ett moment är bara ett “sammanfattande mått” som man får genom att ta ett medelvärde av någon funktion av X (som är en slumpvariabel).

Första momentet

$$\mathbb{E}[X] = \mu$$

Det vill säga första momentet är medelvärdet.

Andra momentet

Andra momentet hänger ihop med variansen; låt mig visa.

$$\mathbb{E}[X^2]$$

Variansen defineras som:

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

Nu när vi vet om att $E[X] = \mu$ kan vi formulera om variansen med få den genom andra momentet:

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

Så man kan notera att för att få fram variansen krävs både första momentet och andra momentet.

Medelvärdet efter trunkering

När när vi har en förståelse av vad ett moment är kan vi definera det efter trunkeringen. Detta innebär **att man kollar endast på värden som är större än a.** Detta måste även defineras för momentet!

$$\mathbb{E}[X[X > a] = \int_a^\infty x f(x|X > a) dx$$

Vad innebär denna formeln?

Denna formeln innebär att man tar den trunkerade tätheten $f(x|X > a)$ (som bara gället för $x > a$ och summerar/integrerar till 1 där) och räknar medelvärdet på vanligt sätt: “x gånger sannolikhetstäthet”, integrerat över de tilltåna värdena a till ∞.

Exempel

Det trunkerade medelvärdet i föregående exempel blir:

$$\mathbb{E}[X|X > \frac{1}{3}] = \int_{\frac{1}{3}}^{1} x \cdot \frac{3}{2} dx = \frac{2}{3}$$

Variansen för en likformig fördelad variabel på intervallet (a, b) är $\frac{(b-a)^2}{12}$ vlket ger oss:

$$Var[X|X > \frac{1}{3}] = \frac{1}{27}$$

Väntevärde och varians för den icke-trunkerade fördelningen är $\frac{1}{2}$ och $\frac{1}{12}$ respektive.

Vad händer om trunkeringen sker underifrån?

- Om trunkeringen sker underifrån, så är det trunkerade medelvärdet större än det ursprungliga medelvärdet.
- Den trunkerade variansen är längre än variansen i den icke trunkerade fördelningen.

Moment av den trunkerade normalfördelningen

Om vi har en slumpvariabel $Y \sim N(\mu, \sigma^2)$ och a är en konstant. Då gäller:

$$\begin{gathered}\mathbb{E}[Y|\text{trunkering}] = \mu + \sigma\lambda(\alpha) \\ Var[Y|\text{trunkering}] = \sigma^2(1 - \delta(\alpha)) \\ \text{Där, } \alpha = \frac{a - \mu}{\sigma}\end{gathered}$$

Vidare gäller:

$$\begin{gathered}\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)}, \quad \text{om} \quad y > a \\ \lambda(\alpha) = -\frac{\phi(\alpha)}{\Phi(\alpha)}, \quad \text{om} \quad y < a \\ \delta(\alpha) = \lambda(\alpha)(\lambda(\alpha) - \alpha)\end{gathered}$$

Observera:

- $\phi(\alpha)$ = tätheten för $N(0, 1)$.
- $\Phi(\alpha) =$ fördelningsfunktionen för $N(0, 1)$.

Vi har $0 > \delta(\alpha) < 1$, vilket medför att $Var[Y|\text{ trunkering}] < \sigma^2$.

Funktionen $\lambda(\alpha)$ kallas för den **inversa Millskvoten** och även fördelningens **hazardfunktion.**

Förklaring och motivering till ovan

När man säger att en normalfördelad variabel $Y \sim (\mu, \sigma^2)$ är trunkerad vid a betyder det att vi inte längre tillåter alla möjliga värden, utan vi tittar på Y givet att den hamnar på ena sidan om a. Antingen studerar man:'

- $Y|Y > a$ (vi behåller bara värden över a).
- $Y|Y < a$ (vi behåller bara värden under a).

Då blir fördelningen inte "en vanlig normal" längre, utan en normal som är omskalad så att **sannolikheten på den förbjudna sidan tas bort**.

För att göra formlerna enklare standardiserar man: $Z = (Y - \mu)/\sigma \sim N(0, 1)$. Gränserna a blir då $\alpha = (a - \mu)/\sigma$. Det är alltså α som säger hur många standardavvikelser a ligger från medelvärdet. Händelsen $Y > a$ motsvarar $Z > \alpha$ och $Y < a$ motsvarar $Z < \alpha$. Eftersom att $Y = \mu + \sigma Z$ räcker det att kunna medelvärdet och variansen för Z efter trunkering, och sen "skala tillbaka" med μ och σ.

När man trunkerar flyttas medelvärdet bort från den avklippta delen. Om vi kräver $Y > a$ så tvingar vi variabeln att liggai högersvansen, och då blir medelvärdet större än μ. Exakt hur mycket större bestäms av funktionen $\lambda(\alpha)$. För trunkeringen över a är:

$$\lambda(\alpha) = \frac{\phi(\alpha)}{1 - \Phi(\alpha)},$$

Där ϕ är tätheten för standardnormalen och Φ är fördelningsfunktionen. Det här talet mäter i praktiken "hur mycket svans som finns kvar" relativt täthetens storlek vid gränsen. Då blir:

$$\mathbb{E}[Y|Y > a] = \mu + \sigma\lambda(\alpha)$$

Om man istället trunkerar under a (dvs $Y < a$) blir medelvärdet mindre än μ, och då används motsvarande form:

$$\lambda(\alpha) = -\frac{\phi(\alpha)}{\Phi(\alpha)}$$

Det är samma ide: medelvärdet flyttas bort från den del som inte längre får vara med.

Variansen blir alltid mindre efter trunkering, eftersom du har tagit bort en del av spridningen och tvingar variabeln att ligga i ett smalare område. Det fångas av funktionen $\delta(\alpha)$ som defineras som:

$$\delta(\alpha) = \lambda(\alpha)(\lambda(\alpha) - \alpha)$$

Den hamnar mellan 0 och 1, alltså:

$$0 > \delta(\alpha) < 1$$

På grund av detta blir därför variansen

$$Var[Y \mid \text{trunkering}] = \sigma^2(1 - \delta(\alpha)) < \sigma^2$$

Så sammanfattningen är: α talar om var gränsen ligger i standardiserad skala, $\lambda(\alpha)$ säger hur mycket medelvärdet flyttas när man bara tillåter enda sidan, och $\delta(\alpha)$ säger hur mycket variansen krymper. Funktionen $\lambda(\alpha)$ kallas också Millskvoten (och är nära kopplad till hazard funktionen) eftersom den har formen “täthete divideras med svans-sannolikhet”, vilket är precis det som dyker upp när man vill beskriva beteendet i en svans.

Den trunkerade regressionsmodellen

Den trunkerade regressionsmodellen säger $Y_i \sim^{iid} N(\mu_i.\sigma^2)$ och antag att $\mu_i = x_i\beta$ vilket ger följande modell:

$$y_i = x_i\beta + \epsilon_i, \quad \text{Där } \epsilon_i \sim^{iid} N(0, \sigma^2)$$

Så detta innebär att:

$$y_i|x_i \sim N(x_i\beta, \sigma^2)$$

I den trunkerade regressionsmodellen behöver dock fördelningen för y_i givet att y_i är större än trunkeringspunkten a. Detta är precis den trunkerade

normalfördelningen som vi studeras nyss **men** nu **ersätter** μ **med** $x_i\beta$ överallt ovan.

Marginella effekten

Den marginella effekten på $\mathbb{E}[y_i|y_i > a] = x_i\beta + \sigma\lambda(\alpha_i)$ där $\alpha_i = \frac{a - x_i\beta}{\sigma}$, vilket blir:

$$\frac{\partial\, \mathbb{E}[y_i \mid y_i > a]}{\partial x_{ij}} = \beta_j \big(1 - \delta(\alpha_i)\big).$$

Marginella effekten från en förklaringsvariabel på det trunkerade medelvärdet är mindre än det ursprungliga medelvärdet, eftersom att $0 < (1 - \delta(\alpha_i)) < 1$.

Om studien syftar till att studera subpopulationen då $y > a$, så är vi intresserade av ovanstående marginella effekt. Däremot om hela populationen ska studera, så är koefficienterna β det faktiska intresset.
