Bayesiansk statistik, 732g43, 7.5 hp

Moment 4 - Logistisk regression, Binomialregression, Poissonregression, Multilevelmodeller

Bertil Wegmann

STIMA, IDA, Linköpings universitet

Översikt moment 4

- Bayesiansk logistisk regression
- Bayesiansk binomialregression
- Bayesiansk Poissonregression
- Bayesianska multilevelmodeller

Bayesiansk logistisk regression

- Den beroende variabeln Y är binär, d.v.s. antingen är $Y = 0$ eller $Y = 1$. Vi vill anpassa en modell där $P(Y = 1)$ beror på olika förklaringsvariabler i en regressionsmodell.
- Modellen kan skrivas som

$$y_i \overset{iid}{\sim} bin(1, p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}(p_i) = \boldsymbol{\beta}\boldsymbol{x}_i',$$

där $\boldsymbol{\beta} = (\beta_0\, \beta_1\, \ldots\, \beta_k)$ är en radvektor med parametrar och $\boldsymbol{x} = (1\, x_1\, \ldots\, x_k)$ en radvektor med förklaringsvariabler.

- Oberoende priors för $\boldsymbol{\beta}$ (boken):

$$\beta_j \sim N(0, 10),$$

där $j = 0, 1, \ldots, k$.

Exempel: Bayesiansk logistisk regression med eBay data

- Wegmann och Villani (2011) analyserade förpackningar av samlarmynt från 1000 auktioner på auktionssajten eBay. Datamaterialet "eBaydata" finner ni på kurshemsidan och består av
  - beroende variabel $y = \textit{UnOpen}$: en dummyvariabel med värdet 1 om myntförpackningen levereras sluten och oöppnad i dess originalförpackning.
  - förklaringsvariabel $x_1 = \textit{PowerSeller}$: en dummyvariabel med värdet 1 om säljaren är rankad bland de mest lyckosamma försäljarna på eBay.
  - förklaringsvariabel $x_2 = \textit{LogBookValue}$: naturliga logaritmen av priset på myntförpackningen från den stora mynthandlaren Golden Eagle Coins på Internet (http://www.goldeneaglecoin.com).

Exempel: Bayesiansk logistisk regression med eBay data

- Jämför resultaten mellan följande 4 modeller:
  1. Modell_utan: Bayesiansk logistisk regressionsanalys utan förklaringsvariabler.
  2. Modell_x1: Bayesiansk logistisk regressionsanalys med endast förklaringsvariabeln x_1.
  3. Modell_x2: Bayesiansk logistisk regressionsanalys med endast förklaringsvariabeln x_2.
  4. Modell_x1x2: Bayesiansk logistisk regressionsanalys med båda förklaringsvariabler.

Effekt från förklaringsvariablerna i Bayesianska logit

- Log-oddset förändras med $m\beta_j$ då en förklaringsvariabel x_j ökar med m enheter, givet att $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$ hålls konstanta.
- Oddset förändras med en faktor $e^{m\beta_j}$ då en förklaringsvariabel x_j ökar med m enheter, givet att $(x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k)$ hålls konstanta.
- Marginell effekt på $P(Y = 1)$ från förklaringsvariabeln x_j:

$$\frac{\partial P(Y=1)}{\partial x_j} = \frac{\partial \left(\frac{\exp(\boldsymbol{\beta x'})}{1+\exp(\boldsymbol{\beta x'})} \right)}{\partial x_j} = \frac{e^{\beta \mathbf{x}'}}{\left(1 + e^{\beta \mathbf{x}'}\right)^2} \cdot \beta_j$$

$$= P(Y=1) \cdot \frac{\beta_j}{1 + e^{\beta \mathbf{x}'}}$$

Bayesiansk binomialregression

- Den beroende variabeln Y är binomialfördelad. Vi vill anpassa en modell där sannolikheten för lyckat försök för observation i, p_i, beror på olika förklaringsvariabler i en regressionsmodell.
- Modellen kan skrivas som

$$y_i \overset{iid}{\sim} bin(n_i, p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}\,(p_i) = \boldsymbol{\beta}\boldsymbol{x}_i',$$

där $\boldsymbol{\beta} = (\beta_0\, \beta_1\, \ldots\, \beta_k)$ är en radvektor med parametrar och $\boldsymbol{x} = (1\, x_1\, \ldots\, x_k)$ en radvektor med förklaringsvariabler.

- Oberoende priors för $\boldsymbol{\beta}$ (boken):

$$\beta_j \sim N\,(0, 10)\,,$$

där $j = 0, 1, \ldots, k$.

Bayesiansk binomial regression - school admissions

- Graduate school ansökningar för män och kvinnor till 6 olika institutioner på UC Berkley (ett av bokens datamaterial). Data kolumner innehåller för respektive kön och institution antal ansökningar (*applications*) samt antal beviljade (*admit*) och icke-beviljade (*reject*) ansökningar.

- Totalt antal ansökningar är 4526, varav 1755 stycken eller 38.8 % beviljades.

- Mål med binomialregression: avgöra om kön kan förklara sannolikheten att en ansökan beviljas eller ej.

Bayesiansk binomial regression - school admissions

- 2 modeller med $y =$ antal beviljade per kön och institution och med eller utan dummyvariabeln $x_1 = 1$ om ansökan kom från en man och 0 annars:

$$y_i \overset{iid}{\sim} bin(n_i, p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}\,(p_i) = \beta_0 + \beta_1 x_{1i},$$

  där n_i är antalet ansökningar (applications) och p_i är sannolikheten för att en ansökan beviljas för en kombination i av kön och institution.

- Resultatet visar på att sannolikheten för en beviljad ansökan är i genomsnitt över institutioner betydligt högre för män än kvinnor.

- Vad kan detta resultat bero på?

Bayesiansk binomial regression - school admissions

- Posterior prediktiv check visar att modellen är dålig på att skatta andelen beviljade ansökningar för män och kvinnor per institution.
- Kvinnor har en högre andel beviljade ansökningar på flera institutioner.
- Skillnaderna mellan män och kvinnors andel beviljade ansökningar bottnar i att kvinnor har i högre grad ansökt till institutioner som beviljar en betydligt lägre andel ansökningar.
- Omformulering av mål med binomialregression: avgöra hur stor den genomsnittliga skillnaden är i sannolikhet för beviljad ansökan mellan könen per institution?

Bayesiansk binomial regression: olika intercept

- Modell:

$$y_i \overset{iid}{\sim} bin(n_i, p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}\,(p_i) = \beta_{0j} + \beta_1 x_{1i},$$

$$\beta_{0j} \overset{iid}{\sim} N\,(0, 10)\,,$$

$$\beta_1 \sim N\,(0, 10)\,,$$

där index $j = 1, \ldots, 6$, är för varje institution j.

Bayesiansk binomial regression: olika intercept - resultat

- Akaikevikter för de 4 modellerna med WAIC: med/utan dummyvariabel för män och med/utan olika intercept för institutioner.
- Akaikevikterna visar på att det är ungefär fifty-fifty att modellerna med olika intercept för institutioner med/utan dummyvariabel för män är bäst på att göra prediktioner.
- Oddset för beviljad ansökan förväntas förändras med en faktor $e^{-0.1} = 0.9$, där -0.1 är posterior medelvärdet för parametern till dummyvariabeln x_1, från en kvinna till en man som ansöker.
- Fördelningen för faktorn som oddset förändras med visar att det är mer troligt att en ansökan från en kvinna blir beviljad per institution, men skillnaden är liten.
- Posterior prediktiv check visar att modellen är nu mycket bättre på att skatta andelen beviljade ansökningar för män och kvinnor per institution.

Poisson regression

- Modell:

$$y_i \overset{iid}{\sim} Poisson(\lambda_i)$$

$$\log \lambda_i = \boldsymbol{\beta} \boldsymbol{x}_i',$$

där $\boldsymbol{\beta} = (\beta_0\, \beta_1\, \ldots\, \beta_k)$ är en radvektor med parametrar och $\boldsymbol{x} = (1\, x_1\, \ldots\, x_k)$ en radvektor med förklaringsvariabler.

- Oberoende priors för $\boldsymbol{\beta}$:

$$\beta_j \sim N\left(0, 10\right),$$

där $j = 0, 1, \ldots, k$.

Poisson regression - awards för studenter på high school

- Antalet awards för studenter på ett high school, y, kan bero på antalet skrivningspoäng på ett matematiktest ($x_1 = \textit{math}$ i hundratals) och vilket av följande tre studieprogram som studenten går på: akademiskt (academic), allmänt (general) eller yrkesprogram (vocation).

- I exemplet på

  https://stats.idre.ucla.edu/r/dae/poisson-regression/

  har data samlats in för variablerna ovan.

- En Bayesiansk Poisson regression ska anpassas till datamaterialet.

Poisson regression - awards för studenter på high school

- Studieprogrammen kodas om till
  $1 = \textit{academic}$, $2 = \textit{general}$, $3 = \textit{vocational}$.

- Poisson regression med olika intercept för program:

$$y_i \overset{iid}{\sim} Poisson(\lambda_i)$$

$$\log(\lambda_i) = \beta_{0j} + \beta_1 x_{1i},$$

$$\beta_{0j} \overset{iid}{\sim} N(0, 10),$$

$$\beta_1 \sim N(0, 10),$$

där index $j = 1, 2, 3$ för varje program j.

Poisson regression: resultat

- Akaikevikter för 4 modeller med WAIC: med/utan förklaringsvariabel $x_1 = \textit{math}$ och med/utan olika intercept för program.
- Akaikevikterna visar på att förklaringsvariabeln *math* ska vara med i modellen tillsammans med/utan intercept för program.
- Sannolikheten är ungefär 99 % att modellen med *math* tillsammans med intercept för program är bäst på att göra prediktioner jämfört med sannolikheten på 1 % för modellen med endast *math*.
- 90.9 % kredibilitetsintervall för λ och y kan skapas som funktion av *math*.
- Posterior prediktiv check med kredibilitetsintervallen ovan visar att modellen skattar antalet antalet awards på ett ok sätt, men modellen verkar ha problem med ökande varians i antalet awards då *math* ökar.

Bayesiansk multilevelmodell för barns intelligens

- Wänström och Wegmann (2017) analyserade hur antalet yngre och äldre syskon påverkar barns intelligensnivå.
- Datamaterialet *Individual Development and Adaptation (IDA)* inkluderar alla skolbarn (n=665) i årskurs 3 på lågstadiet i normala skolor i Örebro, 1965. Barnen följdes även upp i vuxen ålder omkring år 2000 för att bland annat studera karriärval och inkomstnivå.
- Effekt på intelligensnivå kan också förklaras med hjälp av andra variabler. Från datamaterialet extraherades följande bakgrundsvariabler:
  - socioekonomisk status (SES)
  - föräldrars ålder (M.age och F.age)
  - kön (Gender)

Bayesiansk multilevelmodell för barns intelligens

- Datamaterialet innehåller också information om vilken klass respektive elev (barn) tillhörde. Det är naturligt att ta hänsyn till olika klassmiljöer som kan innebära olika skillnader i intelligensnivå mellan elever i olika klasser.
- Skolklasserna är ett urval från populationen med alla skolklasser i årskurs 3 på lågstadiet i Sverige. Därför vill man kunna generalisera effekterna till populationen barn i alla skolklasser. Därmed valdes en Bayesiansk multilevelmodell (random effects model) med skolklasser som gruppvariabel (klustervariabel).
- Enbart generalisera resultaten till barnen i studien - fixed effects model med dummyvariabler för klasstillhörighet kan väljas. Detta kan dock ge osäkra skattningar i små klasser med få antalet barn.
- Multilevelmodellen som kan användas kallas för Bayesiansk varierande-intercept modell, eftersom den modellerar olika intercept för olika skolklasser.

Bayesiansk multilevelmodell för barns intelligens

- Multilevelmodell med 2 nivåer:

$$y_i \overset{iid}{\sim} N(\mu_i, \sigma_y)$$

$$\mu_i = \alpha_j + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i},$$

$$\alpha_j \sim N(\alpha, \sigma_\alpha),$$

$$\alpha \sim N(0, 10),$$

$$\sigma_\alpha \sim \mathit{halfcauchy}(0, 1),$$

$$\beta_k \sim N(0, 10),$$

$$\sigma_y \sim \mathit{halfcauchy}(0, 1),$$

där $y = IQ$, x_1 är antalet äldre syskon (Older), x_2 är antalet yngre syskon (Young), $x_3 = 1$ för pojke och 0 för flicka, $x_4 = SES$, x_5 och x_6 är åldersklass för respektive moderns och faderns ålder vid födseln (M.age och P.age). Index $j = 1, \ldots, 46$ är för skolklass j och index $k = 1, \ldots, 6$ är för förklaringsvariabel k.

Bayesiansk multilevelmodell för barns intelligens: resultat

- Kvadratisk approximation fungerar inte för MCMC, eftersom modellen består av 2 nivåer och därmed inte ger en enhetlig funktion för log av posteriorn $\rightarrow$MCMC till vår tjänst med hjälp av funktionen *map2stan* ().
- Priorn för intercepten α_j kallas för *hyperprior*, eftersom *hyperpriorn* beror på priorn för dess *hyperparametrar* α och σ_α.
- Posterior resultaten visar att det är väldigt troligt att barn med fler antalet äldre syskon har, i genomsnitt, lägre IQ än barn med färre antalet äldre syskon.
- Posterior resultaten visar att det är väldigt troligt att barn med högre SES har, i genomsnitt, högre IQ än barn med lägre SES.
- MCMC diagnostik visar på konvergens till posteriorn utifrån $n_{eff} > 100$, $\hat{R} < 1.1$, zick-zack mönster i MCMC trace plott och stabilisering av posterior medelvärdet för varje parameter över MCMC iterationerna.

Bayesiansk multilevelmodell: fördelar

- Om observationerna tillhör olika grupper, så bör man givetvis ta hänsyn till den gruppspecifika strukturen. Särskilt viktigt om det finns mycket heterogenitet mellan homogena grupper.
- Modellen lär sig om varje enskild grupp (kluster) samtidigt som den lär sig om populationen av grupper (kluster). Modellen poolar information över grupperna.
- En grupp med lite data påverkas mer av den poolade informationen. Naturligt, eftersom lite data för en grupp innebär lite information om den specifika gruppen.
- Den poolade informationen över grupperna minskar risken för dominerande inferens från en grupp med mycket data.
- Priorn för varje intercept krymper alla intercept mot deras gemensamma medelvärde α (*shrinkage effect*). Priorn är en regulariserande prior för intercepten som uppdateras från data.

Bayesiansk multilevelmodell: fördelar

- Modellen tar hänsyn till all variation inom och mellan grupper jämfört med en modell utan varierande intercept som väger alla observationer lika mycket till ett enda intercept.
- Modellen skattar variationen inom grupper med σ_y och variationen mellan grupper med σ_α. Intraklasskorrelationen mäter andel variation mellan grupperna som

$$\rho_{ICC} = \frac{\sigma_\alpha}{\sigma_y + \sigma_\alpha}.$$

- Posteriorfördelningar för variation inom och mellan grupper ger posteriorfördelningen för intraklasskorrelationen.
- Intraklasskorrelationen är ett mått på hur starkt sambandet är mellan observationer inom samma grupp.
- Wänström och Wegmann (2017) visar en genomsnittlig intraklasskorrelation på 0.44 i en modell utan förklaringsvariabler, vilket innebär en ganska hög andel variation i *IQ* mellan grupperna.

Bayesiansk multilevel logitmodell för familjestorlek

- Multilevelmodell med 2 nivåer:

$$y_i \overset{iid}{\sim} bin(1, p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = \text{logit}\left(p_i\right) = \alpha_j + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i},$$

$$\alpha_j \sim N\left(\alpha, \sigma_\alpha\right),$$

$$\alpha \sim N\left(0, 10\right),$$

$$\sigma_\alpha \sim halfcauchy\left(0, 1\right),$$

$$\beta_k \sim N\left(0, 10\right),$$

där $y = 1$ om det är minst 4 barn i familjen, $x_1 = SES$, x_2 och x_3 är åldersklass för respektive moderns och faderns ålder vid födseln (M.age och P.age). Index $j = 1, \ldots, 46$ är för skolklass j och index $k = 1, \ldots, 3$ är för förklaringsvariabel k.

Pooling av information mellan grupper

- Man kan dela upp information från pooling mellan grupper i 3 kategorier:
  - fullständig pooling: modellen innehåller endast ett intercept för alla grupper.
  - ingen pooling: $J - 1$ dummyvariabler för J grupper. Information från data för en grupp är irrelevant information till alla andra grupper.
  - delvis pooling: adaptiv regulariserande prior för gruppernas intercept i en multilevelmodell.

Pooling av information mellan grupper

- Fullständig pooling underanpassar data genom att ge precis information om endast ett intercept för alla observationer och som kan skilja sig mycket från intercepten i grupperna då intraklasskorrelationen är stor.

- Ingen pooling ger icke-precis information om interceptet i varje grupp, eftersom det är endast data i respektive grupp som uppskattar sitt intercept. Detta medför särskilt stor osäkerhet i skattning av interceptet i grupper med lite data, vilket innebär att modellen överanpassar data med lite information i dessa grupper.

- Delvis pooling i multilevelmodeller med regulariserande prior ger skattningar av gruppspecifika intercept som underanpassar mindre i förhållande till fullständig pooling och överanpassar data mindre för grupper med lite data, eftersom intercepten för dessa grupper lånar styrka från andra grupper och krymper mer mot genomsnittliga interceptet för alla grupper.

Multilevel posterior prediktioner

- Informationskriterier som DIC och WAIC kan användas som förut för att jämföra modellers prediktionsförmåga på nya data (out-of-sample deviance).
- Den adaptiva regularisernade priorn för intercepten i multilevelmodeller minskar överanpassning av data för att ge bättre prediktionsförmåga - multilevelmodeller kan därför aldrig replikera data exakt, eftersom regulariseringen ger en sämre modellanpassning till data.
- Skilj på om modellens prediktiva förmåga ska utvärderas för de specifika grupperna i data (fixed effects) eller om utvärderingen ska kunna generaliseras till populationen med alla grupper (random effects).
- Oavsett modell med fixed eller random effects: Bayesianska modeller är generativa, så simulering från modellen med samplade posteriorvärden för parametrarna ger prediktioner i form av simulerad data. Posteriorfördelningen för prediktioner kan sedan sammanfattas.

Multilevel posterior prediktioner

- Posterior prediktioner för specifika grupper i data som användes till skattning av modellen: specificera vilken grupp och därmed vilket intercept man ska använda för att generera prediktioner i modellen.
- Posterior prediktioner för nya grupper som inte fanns i den data som användes till skattning av modellen: fördelningen för de varierande intercepten är intressant.
- Posterior prediktioner för en ny grupp som antas vara en genomsnittlig grupp: genomsnittliga interceptet α används för att generera prediktioner i modellen.
- Posterior prediktioner för en ny specifik grupp $J+1$ som inte antas vara genomsnittlig: för varje posteriordragning s av α och σ_α dras α_{J+1} från $\alpha_{J+1,s} \sim N(\alpha_s, \sigma_{\alpha,s})$ och för varje intercept $\alpha_{J+1,s}$ genereras en prediktion i modellen. Prediktionerna genereras alltså från lika många genererade intercept $\alpha_{J+1,s}$ som antalet posteriordragningar i MCMC:n.