Bertil Wegmann
Avdelning statistik och maskininlärning (STIMA)
IDA
Linköpings universitet
bertil.wegmann@liu.se
2019-12-12

**Inlämningsuppgift - Bayesiansk statistik, 7.5 hp**

**Moment 4 - Logistisk regression, Binomialregression, Poissonregression, Multilevelmodeller**

**Inlämningsuppgiften skall lösas individuellt**. Valfritt program får användas för att lösa uppgifterna nedan, men mina instruktioner och ledtrådar gäller för programmeringsspråket R. **Bifoga all kod tillsammans med lösningen för varje deluppgift**. **Labbrapport** med lösningar på nedanstående uppgifter samt relevanta datorutskrifter **lämnas in på Lisam senast torsdagen den 9:e januari 2020 kl. 23:59**.

1. (8 poäng) I denna uppgift ska du använda ett datamaterial om vaccinerade barn i ett latinamerikanskt land som kan laddas ned från kursrummet på Lisam. Syftet med uppgiften är att förklara sannolikheten för att ett barn är vaccinerat mot en sjukdom. Icke-informativa priorfördelningar ska väljas på egen hand. Responsvariabeln $y_i = 1$ om ett barn $i$ är vaccinerat mot sjukdomen och 0 annars. Följande förklaringsvariabler finns att tillgå för att förklara sannolikheten att ett barn är vaccinerat:

   $x_1 = 1$ om barnet är minst 2 år, 0 annars.

   $x_2 = 1$ om barnet bor på landsbygden, 0 annars.

   $x_3 = 1$ om modern till barnet har genomgått endast första steget av elementär grundskoleutbildning, 0 annars.

   $x_4 = 1$ om modern till barnet har genomgått både det första och andra steget av elementär grundskoleutbildning, 0 annars (andra steget förutsätter att det första steget är genomgått).

   $x_5 =$ andel procent infödda i det samhälle som barnet bor i.

   (a) Anpassa först en Bayesiansk logistisk regression med $y$ som responsvariabel och förklaringsvariablerna $x_1$ och $x_2$ med hjälp av funktionen *map2stan* (). Utvärdera och dra slutsats om MCMC dragningarna lett till att MCMC algoritmen konvergerat till posteriorn med hjälp av $n_{eff}$, $\hat{R}$, *traceplottar* och plottar för varje parameters ackumulerade posteriormedelvärde över MCMC dragningarna.

   (b) Anpassa motsvarigheten till den Bayesianska logistiska regressionen i (a) med en Bayesiansk binomial regression. Verifiera att den logistiska och binomiala regressionen ger samma resultat.

   (c) Utvärdera hur bra modellen i (a) anpassar sannolikheten $p$ för att ett barn är vaccinerat mot sjukdomen enligt följande: beräkna ett 95.2 % kredibilitetsintervall för $p$ för varje kombination av värden på $x_1$ och $x_2$ (d.v.s. totalt 4 kredibilitetsintervall) och jämför respektive kredibilitetsintervall med motsvarande andel barn i data som är vaccinerade mot sjukdomen. Visa en figur för jämförelsen och kommentera hur bra modellen verkar anpassa data.

   (d) Plotta posteriorfördelningen för den faktor som oddset för vaccinerat barn förändras med då man går från ett barn under 2 år till ett barn på minst 2 år. Kommentera utförligt vad posteriorfördelningen visar.

   (e) Använd informationskriteriet DIC för att avgöra om det är bättre att lägga till förklaringsvariablerna $x_3$ och $x_4$ om moderns utbildningsnivå till modellen i (a). Motivera. Avgör också om posteriorfördelningarna för parametrarna till $x_1$ och $x_2$ förändras nämnvärt av att $x_3$ och $x_4$ läggs till i modellen. Verkar förklaringsvariablerna för moderns utbildningsnivå vara confounding variabler? Motivera.

(f) Anpassa nu en Bayesiansk logistisk regression med $y$ som responsvariabel och med alla förklaringsvariabler $x_1$ till $x_5$ i uppgiften. Avgör utifrån informationskriterierna DIC och WAIC vilken av modellerna i (a), (e) och (f) som är bäst och tolka vad den bästa modellen anses vara bäst på utifrån informationskriterierna. Verkar förklaringsvariabeln $x_5$ introducera confounding i modellen och i så fall på vilket sätt? Beräkna 95.2 % kredibilitetsintervall (kredibilitetsband) för sannolikheten $p$ att ett barn är vaccinerat mot sjukdomen som funktion av förklaringsvariabeln $x_5$ genom att använda typvärdet för respektive $x_1$, $x_2$, $x_3$ och $x_4$ i datamaterialet. Redovisa kredibilitetsbandet i en figur och kommentera vad figuren visar i ord.

2. (3 poäng) I datamaterialet som introducerades i uppgift 1 finns det också information om följande variabel:

   $z_1$ = index för vilket samhälle som barnet tillhör.

   (a) Tag hänsyn till skillnader i sannolikhet för vaccination utifrån vilket samhälle som barnet tillhör. Anpassa därför en Bayesiansk logistisk regression med olika intercept för olika samhällen, där $y$ är responsvariabel och med $x_1, x_2, x_3, x_4$ som förklaringsvariabler. Notera att en multilevelmodell med gemensam fördelning för intercepten inte ska anpassas här, utan endast att man modellerar olika intercept för olika samhällen. Tar denna modell hänsyn till pooling av information från data mellan samhällena för att skatta intercepten? Varför eller varför inte? Utvärdera om MCMC dragningarna lett till att MCMC algoritmen konvergerat till posteriorn med hjälp av $n_{eff}$ och $\hat{R}$. Om konvergens inte uppnås, öka antalet MCMC dragningar tills konvergens uppnås utifrån $n_{eff}$ och $\hat{R}$. Undersök sedan om denna modell är bättre än motsvarande modell utan olika intercept som anpassades i uppgift 1(e) utifrån informationskriterierna DIC och WAIC.

   (b) Tag återigen hänsyn till skillnader i sannolikhet för vaccination utifrån vilket samhälle som barnet tillhör. Anpassa nu en Bayesiansk multilevel logistisk regression med olika intercept för olika samhällen, där $y$ är responsvariabel och med $x_1, x_2, x_3, x_4$ som förklaringsvariabler. Vad är det för skillnad på denna modell jämfört med modellen i uppgift 2(a)? Använd tillräckligt med MCMC dragningar tills konvergens har uppnåtts utifrån $n_{eff}$ och $\hat{R}$. Undersök sedan om denna modell är bättre än modellerna i uppgift 1(e) och 2(a) utifrån informationskriterierna DIC och WAIC.

3. (2 poäng) Använd datamaterialet om antalet flygbombsträffar i London från föreläsningen på moment 1.

   (a) Använd en icke-informativ **konjugerad prior** för medelvärdet $\theta$ och redovisa parametervärdena för priorn. Simulera sedan fram 10000 värden från posteriorfördelningen med hjälp av kvadratisk approximation. Beräkna posteriorfördelningens medelvärde och standardavvikelse. Stämmer dessa överens med dom ungefärliga värdena från föreläsningen på moment 1? Varför eller varför inte?

   (b) Använd modellen för Poisson regression från föreläsningen på moment 4, men **utan** förklaringsvariabler. Anpassa modellen med hjälp av funktionen *map2stan* (). Beräkna följande posteriorsannolikhet för medelvärdet $\lambda$: $P(\lambda > 1 \mid y)$.