Bertil Wegmann
Avdelning statistik och maskininlärning (STIMA)
IDA
Linköpings universitet
bertil.wegmann@liu.se
2019-11-13

Inlämningsuppgift - Bayesiansk statistik, 7.5 hp

Moment 2 - Bayesiansk linjär regression

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 måndagen den 25:e november kl. 23:59**.

Datafil till denna inlämningsuppgift finns på Lisam som **Data_Moment2.RData**. Ladda in data i R och skriv därefter följande kommando: **set.seed(*tal*);y <- rnorm(n,A,B)**, där ***tal*** är den dag då du föddes. Om du t.ex. föddes den 13:e augusti 1995, så skriver du **set.seed(*950813*)**. **Obs! Redovisa den dag då du föddes i labbrapporten.**

Variabeln **y** som skapades med kommandot är lägenhetspriser i tusentals kronor för 57 stycken lägenheter som sålts i en stad. Börja med att transformera om **y** till miljontals kronor med **y/1000**. Datafilen innehåller matrisen **X**, där varje kolumn är en potentiell förklaringsvariabel till lägenhetspris i en linjär regressionsmodell enligt följande:

kolumn 1: x_1 = area i kvadratmeter på lägenheten
kolumn 2: x_2 = antal rum i lägenheten
kolumn 3: x_3 = bostadsavgift per månad i tusentals kronor för lägenheten
kolumn 4: x_4 = antal trappor till lägenheten i bostadshuset
kolumn 5: x_5 = dummyvariabel som är lika med 1 om lägenheten såldes i region City
kolumn 6: x_6 = dummyvariabel som är lika med 1 om lägenheten såldes i region Syd

Antag följande Bayesianska linjära regressionsmodell för responsvariabeln y med högst 6 standardiserade (förutom dummyvariablerna) förklaringsvariabler (lägg till en vektor med 1:or till X för att modellera interceptet β_0):

$$y_i|\mu_i, \sigma, X \overset{iid}{\sim} N(\mu_i, \sigma)$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \beta_6 x_{6i}$$

$$\beta_0 \sim N(3, 10),$$

$$\beta_j \sim N(0, 10), \; j = 1, 2, 3, 4, 5, 6.$$

$$\ln \sigma \sim N(0, 1)$$

Nu till uppgifterna:

1. (6.5 poäng) I denna uppgift ska du med hjälp av kvadratisk approximation anpassa en Bayesiansk linjär regressionsmodell med den beroende variabeln y och **utan** förklaringsvariabler.

   (a) Plotta priorfördelningen för y. Motivera om priorfördelningen för y verkar rimlig utifrån hur y är definierad.

   (b) Använd kvadratisk approximation för att anpassa en Bayesiansk linjär regressionsmodell med den beroende variabeln y och **utan** förklaringsvariabler. Jämför priorfördelningarna för μ och σ med deras respektive posteriorfördelningar. Kommentera skillnaderna utifrån hur informativa priorfördelningarna är jämfört med informationen från data.

   (c) Skapa en tabell med medelvärde, standardavvikelse och 90.9 % kredibilitetsintervall för μ och σ. Tolka kredibilitetsintervallen i ord. Redovisa korrelationsvärdet för μ och σ och kommentera om detta värde verkar rimligt.

   (d) Gör en modellutvärdering genom att replikera data (*in-sample fit*) från den posterior prediktiva fördelningen $p(\tilde{y}|y)$. Använd funktionen *dens* () för att plotta priorfördelningen för y, posterior prediktiva fördelningen $p(\tilde{y}|y)$ och fördelningen för de faktiska värdena på y. Kommentera skillnaderna/likheterna mellan dessa 3 fördelningar och motivera om modellen för data y verkar lämplig.

   (e) Använd nu en uniform prior för μ och $\ln\sigma$, d.v.s.

$$p\left(\mu, \sigma^2\right) \propto \frac{1}{\sigma^2}$$

   Vad är fördelarna/nackdelarna med detta val av priorfördelning? Denna priorfördelning ger dom kända posteriorfördelningarna $p\left(\mu|\sigma^2, y_1, \ldots, y_n\right)$ och $p\left(\sigma^2|y_1, \ldots, y_n\right)$. Plotta dessa posteriorfördelningar mot posteriorfördelningarna från uppgift (b) med funktionen *dens* () och kommentera skillnader/likheter.

2. (5.5 poäng) I denna uppgift ska du med hjälp av kvadratisk approximation anpassa en Bayesiansk linjär regressionsmodell med den beroende variabeln y och **med** den kontinuerliga förklaringsvariabel x (standardiserad) som har högst korrelation med den beroende variabeln y.

   (a) Skapa och tolka ett 90.9 % kredibilitetsintervall för β_1. Hur troligt verkar det vara att förklaringsvariabeln x påverkar den beroende variabeln y på ett linjärt negativt eller positivt sätt?

   (b) Använd en grid av värden på x från det lägsta till det högsta värdet av x. Plotta 95.2 %-iga kredibilitetsintervall för μ som funktion av x (kredibilitetsband) och kommentera vad detta visar.

   (c) Plotta 90.9 %-iga prediktionsintervall för y som funktion av x (prediktionsband) genom att använda samma grid av värden på x som i uppgift (b) och kommentera vad detta visar.

   (d) Beräkna den alternativa Bayesianska förklaringsgraden *Bayesian* R_j^2 för den linjära regressionsmodellen utifrån varje posteriordragning j på μ_{ij} och plotta posteriorfördelningen för denna förklaringsgrad med funktionen *dens* ().

3. (4 poäng) I denna uppgift ska du med hjälp av kvadratisk approximation anpassa en Bayesiansk linjär regressionsmodell med den beroende variabeln y och **alla** standardiserade (förutom dummyvariablerna) förklaringsvariabler.

   (a) Beräkna oddset för positiv eller negativ lutning för respektive lutningsparameter och tolka oddset i ord. Hur troligt är det att respektive förklaringsvariabel, givet de övriga i modellen, påverkar y på ett entydigt linjärt positivt eller negativt sätt? Vilka förklaringsvariabler verkar bidra var för sig till regressionsmodellen utifrån de beräknade oddsen, givet att dom övriga förklaringsvariablerna finns med i modellen?

   (b) Avgör med hjälp av 90.9 % kredibilitetsintervall för respektive lutningsparameter om respektive förklaringsvariabel bidrar på ett entydigt linjärt positivt eller negativt sätt till modellen, givet att dom övriga förklaringsvariablerna finns med i modellen.

   (c) Gör en modellutvärdering genom att replikera data (*in-sample fit*) från den posterior prediktiva fördelningen $p(\tilde{y}|y)$. Använd funktionen *dens* () för att i samma figur

1. plotta den posterior prediktiva fördelningen $p(\tilde{y}|y)$ för modellen i denna uppgift
2. plotta den posterior prediktiva fördelningen $p(\tilde{y}|y)$ från uppgift 1(d)
3. plotta fördelningen för de faktiska värdena på y från uppgift 1(d)

Kommentera skillnader/likheter mellan fördelningarna. Verkar det som att regressionsmodellen med alla förklaringsvariabler anpassar data bättre än regressionsmodellen utan förklaringsvariabler i uppgift 1? Stämmer detta med vad man kan förvänta sig?