Inlämningsuppgift 1

Statistik och dataanalys III, HT. 2023

Mona Sfaxi

Introduktion

Den här inlämningsuppgiften består av två delar. I båda dessa delar ska ni självständigt samarbeta inom era grupper för att lösa uppgifterna. För att bli godkänd måste alla frågor och delfrågor besvaras.

Del 1 av inlämningsuppgiften handlar om hur man kan utnyttja `Inverse transform sampling` för att simulera observationer från en fördelning.

Del 2 handlar om maximum likelihood metoden och hur man kan använda sig av det i Poissonregression.

Instruktioner

Inlämningsuppgiften ska göras i grupper om 2-3 personer. Alla personer i gruppen ska kunna redogöra för alla delar i inlämningsuppgiften.

En qmd-fil samt en html-fil (alternativt en pdf-fil) innehållande välmotiverade svar, relevanta figurer och tillhörande R kod ska lämnas in i mappen `Inlämningsmapp` på Athena. Det ska vara tydligt vilka delar av uppgiften koden avser. Numrera därför varje avsnitt och delavsnitt.

Inlämningstillfällen

Inlämning 1: 29 September 17:00

Komplettering: 7 November 17:00 (obs. gäller för båda inlämningsuppgifter i kursen)

Lägg märke till att om man missar att lämna in vid det första tillfället så kan man fortfarande lämna in vid kompletteringstillfället men det är inte möjligt att komplettera efter det. Kontakta labbansvarig ifall du stöter på problem att ladda upp ditt dokument på Athena innan deadline för inlämning/komplettering.

Del 1 - Transformeringar och simuleringar

Under den här delen av inlämningsuppgiften ska ni finna fördelningen för en variabel U, som är en funktion av en annan stokastisk variabel Y. Därefter ska ni simulera från denna fördelning genom att använda er av Inverse transform sampling.

Låt Y vara en stokastisk variabel med sannolikhetsfördelning

$$f(y) = 5ky + 2, \qquad 0 \leq y \leq 1$$

Med okänt värde k.

1.1 Bestäm konstanten k så att $f(y)$ blir en sannolikhetsfördelning. Visa era uträkningar. Det går bra att bifoga en bild på uträkningarna.

Låt sedan U vara en funktion av Y, där

$$U = 2Y$$

1.2 Finn $f_u(u)$ med en lämplig metod. Glöm inte att skriva ut gränserna för u. Skapa sedan en funktion i R som representerar denna fördelning. Visa uträkningarna. Glöm inte att skriva ut eventuella antaganden för att kunna finna $f(u)$.

1.3 Låt sedan $v = F_u(u)$, där $F_u(u)$ är fördelningsfunktionen för U. Använd dig av Inverse transform sampling genom att först finna inversen: $h^{-1}(v)$. Du bör få två olika lösningar men endast en av dessa är korrekt. Simulera minst 10 000 observationer från en likformig fördelning (se avsnitt 4, datorlabb 2 för en liknande uppgift) och kalla din vektor för v. Skapa en funktion i R som avser $h^{-1}(v)$ och simulera nu fördelningen för $f_u(u)$ genom att använda dig av dina likformigt fördelade observationer i $h^{-1}(v)$.

1.4 Rita den teoretiska fördelningen $f_u(u)$ i en graf tillsammans med dina simulerade värden för $f_u(u)$. Kommentera din graf.

1.5 Beräkna sannolikheten att $u \geq 1.6$ från din simulerade fördelning och jämför med det teroretiska värdet. Kommentera resultatet.

Del 2 - Poissonregression och Maximum likelihood

I det här avsnittet ska ni undersöka sambandet mellan antal läkarbesök och ett par andra variabler i datasetet `doctorvisits`. Variablerna i datasetet är:

`numvisits` som består av antal besök till läkaren som gjorts av tyska kvinnor under en tremånadsperiod. `badh` som är en dummyvariabel, kodad 1 ifall deras hälsa klassificeras som dålig och 0 annars. `Age` representerar kvinnornas åldrar och slutligen `loginc` som är deras logaritmerade inkomster.

Antal besök till läkaren är alltså en numerisk, diskret variabel. Ni ska här undersöka ifall Poissonregression kan vara lämplig för att prediktera antal läkarbesök. Datasetet är uppdelat i två delar; `dv_train` och `dv_test` och finns att ladda ner på Athena. `dv_train` ska användas för att skatta regressionskoefficienterna för modellen. `dv_test` ska användas för att utvärdera hur bra modellen presterar när den träffar på data som den tidigare inte har sett.

> **Uppgift 2.1** Börja med att ladda ner `dv_test` och `dv_train` från Athena, och spara den i samma mapp som din Quarto-fil, förslagsvis med hjälp av koden:

```
dv_test <- read.delim("dv_test.txt", sep = " ")
dv_train <- read.delim("dv_train.txt", sep = " ")
```

Antag att observationerna i `numvisit` kan modelleras med en Poissonregression med okänd parameter λ_i. Dvs låt;

$$y_i|x_i \sim Poisson(\lambda_i)$$

Där

$$\lambda_i = \exp(\beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \beta_4 x_{i3}) \tag{1}$$

> **Uppgift 2.2** Ni ska nu använda Poissonregression och sedan optim för att skatta dessa β-parameter för ert träningsdata `dv_train`. Låt övriga tre variabler i datasetet vara de oberoende variablerna i er regression. Det går bra att följa denna tutorial steg för steg när det kommer till multipel Poissonregression men tänk på att ni har 3 oberoende variabler och inte 2 som i exemplet så ni vill sammantaget skatta fyra parametrar (interceptet och koefficienterna för varje oberoende variabel) med hjälp av `optim()`

> **Uppgift 2.3** Använd sedan de skattade parametrarna och de oberoende variablerna från `dv_test` för att prediktera antal läkarbesök.

> **Uppgift 2.4** Beräkna därefter RMSE för testdata, som ges av:

$$RMSE_{test} = \sqrt{\frac{\sum(y_{test} - \hat{y}_{test})^2}{n_{test}}}$$

Där y_{test} är varabeln `numvisit` från `dv_test` och n_{test} är antal observationer i ert testdata.

Uppgift 2.5 Skatta nu en vanlig linjär regressionsmodell i R, med samma variabler, med funktionen `lm()` för ert träningsdata. Prediktera sedan antal läkarbesök med ert testdata. Tips: använd funktionen `predict(min_lm_modell, new_data)`. Beräkna sedan $RMSE_{test}$ för den linjära regressionsmodellen och jämför resultatet med modellen ovan. Vilken modell föredrar ni?

Tips

Nedan följer några tips för Poissonregression ifall man fastnar:

- ☐ Skapa en separat dataframe som består av alla oberoende variabler från `dv_train`. Kalla denna dataframe för något passande såsom exempelvis `X_train`. Skapa även en vektor `y_train` och låt den vara lika med `numvisit` från datasetet `dv_train`.
- ☐ Skapa en funktion i R som avser log-likelihood funktionen för en Poissonfördelning. Ge den ett passande namn såsom exempelvis `loglik_Poisreg`. Låt sedan funktionen ta argumenten `Beta`, `y` och era `x`-variabler. Här är det absolut nödvändigt att argumentet `Beta` kommer först i ordningen.
- ☐ Definera sedan en variabel `lambda` innuti din funktion och låt denna vara lika med exponenten av en linjär regressionsmodell likt Ekvation 1.
- ☐ Använd `optim()` för att få fram ML-skattningarna av β-parametrarna från funktionen `loglik_Poisreg`.
- ☐ Använd de skattade parametrarna för att skapa prediktioner, $\hat{y}$ för testdatan. Kom ihåg att:

$$\hat{y}_{test} = \exp(b_1 + b2 \cdot x_{1\ test} + b3 \cdot x_{2\ test} + b4 \cdot x_{3\ test})$$

Där $x_{1,test}$ är den första oberoende variabeln från `dv_test`, $x_{2,test}$ är den andra oberoende variabeln från `dv_test` och $x_{3,test}$ är den tredje oberoende variabeln från `dv_test`.