Inlämningsuppgift 1

Statistik och dataanalys III, VT 2024

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 beräkningar av sannolikheter och simuleringar.

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: 20 Februari 17:00

Komplettering: 5 April 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.

Kom ihåg att det är strikt förbjudet att använda sig av AI-genererad kod och textsvar i inlämningsuppgiften.

Del 1 - En Kort uppgift

Uppgift 1.1 En kortlek består av 52 olika kort. Jennifer och Michel har en kortlek var. Jennifer hävdar att sannolikheten att man får 2 ess om man drar 5 kort från en kortlek är 0.05 medan Michel hävdar att sannolikheten är 0.04 (talen är avrundade till två decimaler). Jennifer lägger tillbaka sitt kort och blandar sedan kortleken efter varje dragning medan Michel inte lägger tillbaka något draget kort i sin kortlek. Vem har rätt? Utnyttjar någon utav dem någon fördelning för sina uträkningar? Skriv i så fall vilka parametrar som används och vad dessa värden antar här. Ställ upp deras formler och visa också stegen i deras uträkningar.

Uppgift 1.2 Låt ett lyckat försök vara att man får ett Ess och ett misslyckat försök vara att man inte får ett Ess i en enskild dragning ur en kortlek. Beräkna sannolikheterna för experimenten ovan genom att först låta antal dragningar från kortleken vara lika med 5 och beräkna sannolikheten för antal lyckade försök för varje utfall som går från 0 till 5 för både Jennifer och Michel. Upprepa sedan samma procedur men låt istället antal dragningar vara lika med 15 från kortleken och beräkna sedan sannolikheterna för antal lyckade försök för var och en av utfallen (0-15). Plotta sedan sannolikheterna i samma linjegraf för både Jennifer och Michel. Låt alltså x-axeln i plotten ha värdena 0 till 15 och rita de två linjerna för respektive person i grafen. Vad händer när antal dragningar ökar från 5 till 15 för både Jennifer och Michel?

Uppgift 1.3 Istället för sannolikheten att få Ess ett visst antal gånger är vi nu intresserade av sannolikheten att få något av korten; Ess, Kung, Dam, Knekt ett visst antal gånger. Låt Y beteckna utfallet att man får något av dessa kort vid en dragning från en full kortlek, så att sannolikheten P(Y) är 4/13 ifall utfallet är lyckat (Y = 1) och sannolikheten är 9/13 ifall utfallet är misslyckat (Y = 0). Upprepa experimentet ovan *grafiskt* Låt först antal dragningar ur Jennifers, respektive Michels kortlek vara lika med 5. Upprepa sedan proceduren men med 15 dragningar istället och låt avslutningsvis antal dragningar vara 30. Vad händer och varför tror ni att det blir så? Jämför även resultatet med det ni fick i Uppgift 1.2.

Uppgift 1.4 Simulera Jennifer och Michels experiment från uppgift 1.3 ovan. Låt antal dragningar vara 15. Illustrera simuleringarna grafiskt med hjälp av stolpdiagram.

Uppgift 1.5 Låt Antal dragningar från kortleken vara lika med 15 i samliga fall nedan. Approximera Jennifers experiment med hjälp av normalfördelningen. Rita sedan den teoretiska, den approximativa, och det simulerade experimentet i samma graf och kommentera grafen. Beräkna därefter sannolikheten att hon får 3 eller 4 lyckade försök. Jämför den approximativa sannolikheten med sannolikheten från

det simulerade experimentet och med den teoretiska sannolikheten. Är approximationen bra? Är resultatet konsekvent med grafen? Kan det finnas några problem med att använda sig av en approximation här? Diskutera.

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. `Age` representerar kvinnornas åldrar. `educ` är antal år i skolan 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 en lämplig modell för att prediktera antal läkarbesök. Datasetet går att ladda ner från Athena som en textfil.

Uppgift 2.1 Börja med att läsa in datastet `doctorvisits` från Athena, förslagsvis med hjälp av koden:

```
df <- read.delim("doctorvisits.txt", sep = " ")
```

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

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

Där

$$\lambda_i = \exp(\beta_0 + \beta_1 x_{i1}) \tag{1}$$

Uppgift 2.2 Välj ut *en* utav variablerna `age`, `educ` eller `loginc` som sedan ska användas i en Poissonregression där `numvisits` är responsvariabel. Målet är alltså att skatta parametrarna i Ekvation 1. Motivera varför ni valt just er specifika förklaringsvariabel.

Uppgift 2.3 Koda fram loglikelihoodfunktionen för en Poissonregression i R. Det går bra att följa denna tutorial steg för steg när det kommer till Poissonregression. Tänk på att ni endast ska ha en förklaringsvariabel i regressionen så ni vill sammantaget skatta två parametrar - β_0 och β_1.

Uppgift 2.4 Skapa sedan en grid av värden för β_0 respektive β_1 (se stycket nedan för mer detaljer). Rita loglikelihoodfunktionen i en 3D-graf med β_0- och β_1-värdena från respektive grid på x- och y-xlarna. Grafen ska alltså visa funktionsvärden för loglikelihoodfunktionen givet olika värden på β_0 och β_1. Var ungefär verkar maximumpunkten ligga och vilket funktionsvärde motsvarar detta?

> Om ni valde `age` som förklaringsvariabel: Låt griden för β_0 gå från 0.2 till 0.7 och för β_1 gå från -0.15 till 0.1.
> Om ni valde `education` som förklaringsvariabel: Låt griden för β_0 gå från 0 till 2 och för β_1 gå från -0.15 till 0.07.
> Om ni valde `loginc` som förklaringsvariabel: Låt griden för β_0 gå från -0.5 till 1.5 och för β_1 gå från -0.05 till 0.2.

Uppgift 2.5 Använd funktionen `optim()` för att skatta de två β-parametrarna. Vilka värden får ni på $\hat{\beta}_{0,ML}$ och $\hat{\beta}_{1,ML}$? Är det konsekvent med figuren ovan i `uppgift 2.4`?

Uppgift 2.6 Beräkna ett 95%-igt konfidensintervall för varje parameter och ge en tolkning av intervallet. (Tips: använd hessianen ni får ut från optim).

Uppgift 2.7 Den nionde personen i datasetet hade en ålder på 57 år, en utbildning som var 10 år och en logaritmerad inkomst på 7.77203. Prediktera hur många läkarbesök hon skulle ha gjort utifrån er skattade modell (använd endast ett av dessa värden utifrån förklaringsvariabeln som ni valde i uppgift 2.2). Beräkna därefter residualen, dvs. hur mycket det predikterade värdet skiljer sig från det faktiska antalet läkarbesök hon gjort.

Uppgift 2.8 Låt $f(\hat{\beta}_{0,ML}, \hat{\beta}_{1,ML})$ beteckna en approximativ bivariat sannolikhetsfördelning för $\hat{\beta}_{0,ML}$ och $\hat{\beta}_{1,ML}$ Simulera minst 10 000 observationer från denna fördelning. Illustrera den sedan i en passande graf och beräkna därefter sannolikheten nedan:

> Om ni valde förklaringsvariabeln `age`, beräkna:
> $P(0.4 \leq \hat{\beta}_{0,ML} \leq 0.5,\ 0.006 \leq \hat{\beta}_{1,ML} \leq 0.009)$
>
> Om ni valde förklaringsvariabeln `education`, beräkna:
> $P(0.78 \leq \hat{\beta}_{0,ML} \leq 0.85,\ -0.01 \leq \hat{\beta}_{1,ML} \leq 0.02)$
>
> Om ni valde förklaringsvariabeln `loginc`, beräkna:
> $P(0 \leq \hat{\beta}_{0,ML} \leq 1,\ 0 \leq \hat{\beta}_{1,ML} \leq 0.12)$