Matematisk Statistik

SF1911 Statistik för bioteknik: Lp 2 2017
Lab 2 för
CBIOT3

Introduktion

Detta är handledningen till Laboration 2. Läs gärna handledningen två gånger. Försäkra dig om att du förstår hur de MATLAB-kommandon som finns i den bifogade koden fungerar. Laborationen bedöms som godkänd eller ej godkänd. Arbete i grupp är tillåtet med **högst två**, personer per grupp. Godkänd laboration ger 2 poäng till ordinarie tentamenstillfälle.

Syfte och introduktion

Börja laborationen med att ladda ner filen moore.mat från kurshemsidan
`https://www.math.kth.se/matstat/gru/sf1911/`
Se till att de ligger i den mapp du kommer att arbeta i. För att kontrollera att du har lagt filerna rätt, skriv `ls` och se om filerna ovan listas.
Du kan skriva dina kommandon direkt i MATLAB-prompten men det är absolut att föredra att arbeta i editorn. Om den inte är öppen så kan du öppna den och skapa ett nytt dokument genom att skriva `edit lab2.m`. Koden som ges nedan är skriven i celler. En ny cell påbörjas genom att skriva två procenttecken. `Ctrl+Enter` exkeverar innehållet i en cell.

1 Problem 1: Linjär regression

Vi kommer börja med att titta på fenomenet som kallas Moores lag. På hemsidan finns en fil som heter moore.mat, ladda in datat med *load*. Det kan vara så att ni behöver se till att ni befinner er i rätt mapp i MATLAB eller uppge hela adressen till filen. Datat har två variabler, y är antal transistorer/yta medan x är årtal. Det betyder att om vi plottar dem mot varandra så ser vi en plot av utvecklingen över tid av antalet transistorer per yta. I MATLAB kan linjära modeller göras med hjälp av *fitlm*. Ni ska anpassa de följande tre linjära modellerna till datat:

Modell 1: Polynom av grad 1

Modell 2: Polynom av grad 5

Modell 3: Polynom av grad 1 men med log(y) istället för y

Plotta modellernas linjer. Tänk på att modell 3 ger värden i log(y) men att ni ska plotta dem mot y så ni behöver transformera till baka värdena som modellen ger

Vad blir R^2 för modellerna? Vad blir p-värdet för modellerna?

Är residualerna normalfördelade? Kan undersökas med *qqplot*

Vad förutsäger modellerna att antalet transistorer är år 2020?

2 Problem 2 : Skillnaden mellan två populationer

I det här problemet ska vi använda Iris dataset som användes i lab1 och undersöka om det finns någon skillnad mellan *versicolor* och *virginica*. Datasetet finns inbyggt i MATLAB så ni behöver ingen fil utan de räcker med att skriva *load(fisheriris)*. När vi undersökte variablerna i lab 1 så såg de ut att vara normalfördelade så då kan vi använda t-test för att undersöka skillnaden. Mer specifikt så kommer vi att undersöka om det finns någon statistisk signifikant skillnad mellan väntevärdena av de två arterna i någon av de fyra variablerna.

Beräkna 95% konfidensintervall för skillnaden i medelvärde mellan *versicolor* och *virginica* för alla fyra variabler

Undersök hypotesen att skillnaden i medelvärdet för första och andra variablerna med populationerna ovan skiljer sig ifrån noll påsignnifikansnivån 5%. Detta motsvaras av att testa det dubbelsidiga alternativet, $H_0 : \mu_1 - \mu_2 = 0$ mot $H_1 : \mu_1 - \mu_2 \neq 0$. Det testet går att göra i Matlab med

```
ttest2
```

En av parametrarna till `ttest2` bestämmer om det antas vara lika varianser eller inte. Anta olika varians i det här problemet.

Vad blir konfidensintervallen, p-värden och slutsatserna?

I det här problemet har vi gjort fyra olika statistiska tester. För att ta hänsyn till antalet tester vi gjort bör vi därför använda oss av så kallad Bonferroni-korrektion, där vi helt enkelt delar vår önskade signifikansnivå(α) med antalet tester vi gjort. Det nya α blir då 5/4%. Vad blir konfidensintervallen och slutsatserna med den nya nivån?

3 Problem 3: Bootstrap

Ett annat sätt att undersöka skillnaden mellan väntevärdena är genom att använda bootstrap.En kort introduktion till bootstrap ingår t.ex. i slajdsen för föreläsning 9 på `https://www.math.kth.se/matstat/gru/sf1911/aktuellt16.html`

eller i filen pÃ¥ canvass
bootstrapnature.pdf
Bootstrap återsamplar datat och via det går det att få fram osäkherheten i en statistikas fördelning. Bootstrap kräver inte att variablerna är normalfördelade. Genom att ge Matlab kommandot `bootstrp(M, @mean, x)` genereras M stycken bootstrapreplikat av medelvärdet av dina verkliga data x. Med quantile kan sedan ett intervall skapas.

Undersök skillnaden mellan de två arterna i det tidigare problemet med bootstrap genom att beräkna dubbelsidiga konfidensintervall med samma konfidensgrad som användes med Bonferroni-korrektionen.

Jämför intervallet för variablerna ett och två med det som `ttest2` gav. Finns det några skillnader jämfört mot resultatet av t-testen?