Matematisk Statistik

SF1920/SF1921 Tillämpad statistik, vt 2023
Laboration 2

1 Introduktion

Denna laborationen kommer på redovisningstillfället att bedömas som antingen godkänd eller ej godkänd. De studenter som blir godkända på laborationen behöver inte lösa uppgift 12 på tentamens del I och får dessutom 3 bonuspoäng på tentamens del II. Observera att dessa fördelar endast gäller vid den ordinarie tentamen och det första omtentamenstillfället.

Läs först igenom labbspecifikationen två gånger. Försäkra dig om att du förstår hur de MATLAB-kommandon som finns i den bifogade koden fungerar. Svaren på förberedelseuppgifterna ska kunna redovisas **individuellt**. Arbete i grupp är tillåtet (och uppmuntras) med **högst två** personer per grupp. **Ta med** en utskriven kopia av labbspecifikationen till redovisningstillfället för att kunna använda som kvitto på att laborationen är godkänd.

2 Förberedelseuppgifter

1. En Rayleighfördelad stokastisk variabel X har täthetsfunktionen

$$f_X(x) = \frac{x}{b^2} e^{-\frac{x^2}{2b^2}}.$$

   Antag nu att du har n stycken Rayleighfördelade variabler.

   a) Bestäm ML-skattningen av b.

   b) Bestäm MK-skattningen av b.

2. Beskriv hur du kan ta fram ett approximativt konfidensintervall för parametern b. Motivera varför det är rimligt att göra den approximation som du har gjort. Ledning: Använd MK-skattningen av b.

3. Beskriv idén bakom linjär regression. Beskriv hur man i MATLAB m.h.a. kommandot `regress` kan skatta parametrarna i modellen

$$w = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k \tag{1}$$

3 Nödvändiga filer

Börja med att ladda ner följande filer från kurshemsidan.

- `wave_data.mat`
- `hist_density.m`
- `birth.dat`
- `birth.txt` - beskrivning av datat `birth.dat`
- `moore.dat`

Se till att filerna 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` exekverar innehållet i en cell.

4 Laborationsuppgifter

Problem 1 - Simulering av konfidensintervall

Ett konfidensintervall med konfidensgrad $1-\alpha$ för en (okänd) parameter μ innehåller det sanna μ med sannolikhet $1-\alpha$. Vi ska försöka förstå innebörden av detta begrepp med hjälp av simuleringar. Koden nedan använder $n = 25$ oberoende observationer från $N(2,1)$-fördelningen för att skatta ett konfidensintervall för väntevärdet med konfidensgrad 95%. Detta upprepas 100 gånger vilket ger 100 konfidensintervall. Hur många av dessa intervall kan förväntas innehålla det sanna värdet på μ?

```
%% Problem 1: Simulering av konfidensintervall
    % Parametrar:
    n = 25; %Antal matningar
    mu = 2; %Vantevardet
    sigma = 1; %Standardavvikelsen
    alpha = 0.05;
    %Simulerar n observationer for varje intervall
    x = normrnd(mu, sigma,n,100); %n x 100 matris med varden
    %Skattar mu med medelvardet
    xbar = mean(x); %vektor med 100 medelvarden.
    %Beraknar de undre och ovre granserna
    undre = xbar - norminv(1-alpha/2)*sigma/sqrt(n);
    ovre = xbar + norminv(1-alpha/2)*sigma/sqrt(n);
```

```
%% Problem 1: Simulering av konfidensintervall (forts.)
    %Ritar upp alla intervall
    figure(1)
    hold on
    for k=1:100
        if ovre(k) < mu % Rodmarkerar intervall som missar mu
            plot([undre(k) ovre(k)],[k k],'r')
            elseif undre(k) > mu
            plot([undre(k) ovre(k)],[k k],'r')
        else
            plot([undre(k) ovre(k)],[k k],'b')
        end
    end
    %b1 och b2 ar bara till for att figuren ska se snygg ut.
    b1 = min(xbar - norminv(1 - alpha/2)*sigma/sqrt(n));
    b2 = max(xbar + norminv(1 - alpha/2)*sigma/sqrt(n));
    axis([b1 b2 0 101]) %Tar bort outnyttjat utrymme i figuren
    %Ritar ut det sanna vardet
    plot([mu mu],[0 101],'g')
    hold off
```

Vad visar de horisontella strecken och det vertikala strecket? Hur många av de 100 intervallen innehåller det sanna värdet på μ? Stämmer resultatet med dina förväntningar? Kör simuleringarna flera gånger.

Variera nu μ, σ, n och α (en i taget) och ser hur de olika parametrarna påverkar resultatet.

Problem 2 - Maximum likelihoodskattning och minsta kvadratskattning

I denna uppgift ska vi undersöka två olika punktskattningar av värdet på parametern i en Rayleighfördelning. Koden nedan genererar en samling Rayleighfördelade stokastiska variabler med parametervärde 4 och plottar sedan skattningarna `my_est_ml` och `my_est_mk`. Använd dina två skattningar från förberedelseuppgift **??**.

```
%% Problem 2: Maximum likelihood/Minsta kvadrat
    M = 1e4;
    b = 4;
    x = raylrnd(b, M, 1);
    hist_density(x, 40)
    hold on
    my_est_ml = % Skriv in din ML-skattning har
    my_est_mk =  % Skriv in din MK-skattning har
    plot(my_est_ml, 0, 'r*')
    plot(my_est_mk, 0, 'g*')
    plot(b, 0, 'ro')
    hold off
```

Ser din skattning bra ut? Kontrollera hur täthetsfunktionen ser ut genom att plotta den med din skattning:

```
%% Problem 2: Maximum likelihood/Minsta kvadrat (forts.)
    plot(0:0.1:6, raylpdf(0:0.1:6, my_est_ml), 'r')
    hold off
```

Problem 3 - Konfidensintervall för Rayleighfördelning

Vi ska nu undersöka en Rayleighfördelad signal, bestämma en punktskattning av parametervärdet samt ta fram ett konfidensintervall för parametern. Ladda in data genom att skriva `load wave_data.mat`. Filen innehåller en signal som du kan plotta genom att skriva följande kod.

```
%% Problem 3: Konfidensintervall for Rayleighfordelning
    load wave_data.mat
    subplot(2,1,1), plot(y(1:100))
    subplot(2,1,2), hist_density(y)
```

Om du ändrar `y(1:100)` till `y(1:end)` så kan du se hela signalen. Skatta parametern i datat på samma sätt som i Problem 2. Spara din skattning som `my_est`. Ta fram ett konfidensintervall för skattningen och spara övre respektive undre värdet som `upper_bound` respektive `lower_bound`. Plotta nu intervallet för din skattning av parametern

```
%% Problem 3: Konfidensintervall (forts.)
    hold on     % Gor sa att ploten halls kvar
    plot(lower_bound, 0, 'g*')
    plot(upper_bound, 0, 'g*')
```

Kontrollera hur täthetsfunktionen ser ut genom att plotta den med din skattning på samma vis som i föregående avsnitt:

```
%% Problem 3: Konfidensintervall (forts.)
    plot(0:0.1:6, raylpdf(0:0.1:6, my_est), 'r')
    hold off
```

Ser fördelningen ut att passa bra?

Rayleighfördelningen kan t.ex. användas för att beskriva hur en radiosignal avtar. Experimentella mätningar på Manhattan har visat att Rayleighfördelningen beskriver radiosignalers fädning (engelska: fading) på ett bra sätt i den sortens stadsmiljö [**?**].

Problem 4 - Jämförelse av fördelningar hos olika populationer

I denna uppgift undersöker vi en datamängd visuellt med hjälp av MATLAB för att se om vi kan göra några intressanta iakttagelser. Filen `birth.dat` innehåller data om 747 förstföderskor i Malmö under åren 1991-1993. Filen innehåller olika 26 variabler, varav några är numeriska (såsom längd och vikt hos modern) och andra kategoriska, dvs antar ett av 2-3 fixa värden (exempelvis "1" om barnet var planerat och "2" om det inte var planerat). Använd informationen i `birth.txt` och MATLAB-funktionerna `subplot` och `hist_density.m` för att generera en figur med fyra olika histogram som visar fördningarna för barnets födelsevikt, moderns ålder, moderns längd respektive moderns vikt.

Det är av medicinskt intresse att bestämma riskfaktorer som ökar sannolikheten för att en barn föds med alltför låg födelsevikt. Låg födelsevikt definieras som en födelsevikt under 2500 g, mycket låg födelsevikt som en födelsevikt under 1500 g och extremt låg födelsevikt som en födelsevikt under 1000 g. Vi kan användra den givna datamängden för att försöka dra slutsatser om riskfaktorer för låg födelsevikt genom att jämföra viktfördelningen för barn vars mödrar har en viss riskfaktor med viktfördelning för barn vars mödrar saknar riskfaktorn.

En känd riskfaktor för låg födelsevikt hos nyfödda är rökning. Vi undersöker därför skillnaden i födelsevikt mellan barn vars mammor röker respektive inte röker under graviditeten. I filen `birth.txt` ser man att kolonn 20 i `birth.dat` innehåller rökvanor och att värdena 1 och 2 betyder att mamman inte röker under graviditeten, medan värdet 3 betyder att hon gör det. Skapa två variabler `x` och `y` för födelsevikter hörande till icke-rökande respektive rökande mammor enligt

```
%% Problem 4: Fordelningar av givna data
    load birth.dat
    x = birth(birth(:, 20) < 3, 3);
    y = birth(birth(:, 20) == 3, 3);
```

Vad som händer här är att `birth(:, 20) < 3` returnerar en vektor av "sant" och "falskt" och att bara de rader av kolonn 3 (födelsevikterna) i `birth` för vilka jämförelsen är sann, väljs ut. Använd funktionen `length` eller kommandot `whos` för att se storleken på vektorerna `x` och `y`. Använd koden nedan för att visuellt inspektera datat.

```
%% Problem 4: Fordelningar av givna data (forts.)
    subplot(2,2,1), boxplot(x),
    axis([0 2 500 5000])
    subplot(2,2,2), boxplot(y),
    axis([0 2 500 5000])
```

```
%% Problem 4: Fordelningar av givna data (forts.)
    subplot(2,2,3:4), ksdensity(x),
    hold on
    [fy, ty] = ksdensity(y);
    plot(ty, fy, 'r')
    hold off
```

Vad betyder plotarna? Vilka slutsatser kan du dra om rökande mödrars påverkan på barns födelsevikt?

Välj nu ut en annan av de kategoriska variablerna i datat som du misstänker kan påverka födelsevikten och undersök med samma metod om det förefaller föreligga ett samband mellan variabeln och födelsevikten. Observera att om du väljer en kategorisk variabel med tre olika värden, så behöver du göra om den till en variabel med två olika värden på samma sätt som för rökvanevariabeln ovan.

Problem 5 - Test av normalitet

Många statistiska metoder baseras på antagandet att datat är normalfördelat. Det är därför av intresse att kunna avgöra om en given datamängd är normalfördelad eller ej. Två metoder för att visuellt undersöka om en datamängd är normalfördelad ges av de båda MATLAB-kommandona `normplot` och `qqplot` som båda jämför den empiriska datamängdens kvantiler med kvantilerna för en normalfördelning. Undersök med ett av dessa kommandon om variablerna för barnets födelsevikt, moderns ålder, moderns längd och/eller moderns vikt kan sägas vara normalfördelade och om inte på vilket sätt de olika variablerna avviker från normalfördelning slumpvariabler.

Metoderna `normplot` och `qqplot` bygger på att man gör en visuell bedömning av en graf och innehåller därför ett visst mått av subjektivitet. Det finns dock statistiska test för att avgöra normalitet. Ett sådant är Jarque-Beras test av normalitet som bygger på en jämförelse mellan det empiriska datats skevhet och kurtosis och motsvarande storheter för en normalfördelning. För en slumpvariabel X med väntevärde μ och standardavvikelse σ definieras skevheten γ och kurtosisen κ som

$$\gamma = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] \quad \text{respektive} \quad \kappa = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right].$$

I Jarque-Beras test av normaltitet används testvariabeln

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right),$$

där n är antalet observationer och S och K är skattningar av datats skevhet och kurtosis. Under nollhypotesen att datat är normalfördelat, så är testvariabeln approximativt χ^2-fördelad med två frihetsgrader.

Använd MATLAB-funktionen `jbtest` för att avgöra om variablerna för barnets födelsevikt, moderns ålder, moderns längd och/eller moderns vikt är normalfördelade på signifikansnivån 5%.

Problem 6 - Enkel linjär regression

Linjär regression utvecklades under sent 1700-tal av en ung Gauss. Metoden fick ett genomslag när den förutspådde banan för den genom tiderna först upptäckta asteroiden Ceres. Linjär regression används än flitigare idag med tillämpningar inom i stort sett all vetenskap som behandlar data. Fördjupning i ämnet ges i kursen "Regressionsanalys".

I denna uppgift ska vi undersöka fenomenet Moores lag. Ladda in datat `moore.dat` på samma sätt som tidigare. I datat så är `y` antalet transistorer/ytenhet medan `x` representerar årtalet. Det betyder att om vi plottar dessa variabler mot varandra så ser vi en plot av utvecklingen över tid av antalet transistorer per ytenhet. Antalet transistorer per ytenhet förefaller öka exponentiellt, vilket gör det rimligt att anta att logaritmen av antalet transistorer per ytenhet ökar linjärt med tiden. Inför modellen

$$w_i = \log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

Bilda en matris X som har formen

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},$$

där n är antalet observationer i datat och ta sedan fram en skattning av $\hat{\beta} = (\beta_0, \beta_1)^T$ med hjälp av MATLABs funktion `regress`. Plotta din skattade modell

$$\log(\hat{y}) = X\hat{\beta},$$

genom att jämföra $\hat{y}$ med datat y. Plotta sedan residualerna på följande sätt.

```
%% Problem 6: Regression
    res =  w-X*beta_hat;
    subplot(2,1,1), normplot(res)
    subplot(2,1,2), hist(res)
```

Vilken fördelning ser de ut att komma från? Använd även funktionen `regress` för att bestämma storheten R^2, vilken är ett mått på hur stor andel av variationen i datat som förklaras av modellen. Om du skattar $\hat{\beta}$ m.h.a. data från 1972 till 2019, vad är då din prediktion för antalet transistorer år 2025?

Problem 7 - Multipel linjär regression

Regression kan användas även när en variabel kan antas bero på flera olika förklaringsvariabler, säg x, w och z. Vi kan då ställa upp en multipel linjär regressionsmodell

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \beta_3 z_i + \varepsilon_i.$$

Använd datat i `birth.dat` för att för att först sätta upp en enkel linjär regressionsmodell för hur barnets födelsevikt beror på moderns längd. Sätt sedan upp en multipel linjär regressionsmodell där ni som förklaringsvariabler använder moderns vikt, moderns rökvanor och den ytterligare kategoriska variabel med två möjliga värden som ni undersökte i Problem 4. Notera att de två värdena på de kategoriska variablerna måste väljas till 0 respektive 1. Använd `regress` för att bestämma konfidensintervall för parametrarna i den multipla regressionsmodellen och avgör med hjälp av dessa konfidensintervall om de undersökta variablerna har en signifikant påverkan påbarnets födelsevikt. Plotta även residualerna i den multipla regressionsmodellen med `normplot` och tolka resultatet.

Referenser

[1] Dmitry Chizhik, Jonathan Ling, Peter W. Wolniansky, Reinaldo A. Valenzuela, Nelson Costa, and Kris Huber (2003). Multiple-input-multiple-output measurements and modeling in Manhattan *IEEE Journal on Selected Areas in Communications*, Vol **21**, p. 321-331.

[2] Blom, G., Enger, J., Englund, G., Grandell, J., och Holst, L., (2005). Sannolikhetsteori och statistikteori med tillämpningar.