Matematisk Statistik

SF1910 Tillämpad statistik, HT 2018
Laboration 2 för CSAMH

1 Introduktion

Denna laboration är poänggivande och godkänd laboration kan ge 4 bonuspoäng vid ordinarie tentamen och första omtentamen. Laborationen bedöms som godkänd eller ej godkänd. Läs först 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. Definiera likelihood och log-likelihood samt förklara sambandet mellan dessa begrepp. Beskriv idén bakom Minsta-kvadratmetoden (MK) respektive Maximum-likelihoodmetoden (ML).

2. 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.

3. 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.

4. Beskriv idén bakom linjär regression. Förklara vad polynomregression är.

5. 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 Syfte och vidare introduktion

Börja med att ladda ner följande filer från kurshemsidan.

- `wave_data.mat`
- `moore.mat`
- `poly.mat`
- `birth.dat`
- `birth.txt` - beskrivning av datat `birth.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 *.*at` 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 lab3.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 - Deskriptiv statistik

I denna uppgift studerar vi skillnaden i väntevärden hos två populationer. I synnerhet undersöker vi 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. Ni kan skapa två variabler `x` och `y` för födelsevikter hörande till icke-rökande respektive rökande mammor enligt

```
>> 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 8: Deskriptiv statistik
    load lab2data/birth.dat
    x = birth(birth(:, 20) < 3, 3);
    y = birth(birth(:, 20) == 3, 3);
    subplot(2,2,1)
    boxplot(x)
    axis([0 2 500 5000])
    subplot(2,2,2)
    boxplot(y)
    axis([0 2 500 5000])
    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 ni dra?

Problem 2 - Centrala Gränsvärdessatsen

Koden nedan simulerar exponentialfördelade slumptal och summerar sedan dessa. Studera koden och fundera ut vad `N` representerar.

```
%% Problem 2: Centrala gransvardessatsen
    M = 1e3;
    N = 4;
    mu = 5;
    X = exprnd(mu, M, N);
    S = cumsum(X, 2);
    for k = 1:N
        hist(S(:, k), 30)
        xlabel(num2str(k))
        pause(0.1)
    end
```

Justera `N`, vad händer när du ökar respektive minskar värdet? Varför? Vid vilket `N` ser det ut som att det inte gör någon skillnad att öka `N`? Vilken fördelning verkar summorna ha? Varför har de denna fördelning?

Problem 3 - Monte Carlo-skattning av väntevärden

På förra laborationen såg vi att väntevärdet av en stokastisk variabel kunde skattas numeriskt med hjälp av Monte Carlo-simulering, dvs. genom att beräkna det aritmetiska medelvärdet av ett stort antal simulerade slumptal som dras från samma fördelning som den stokastiska variabeln.

Fördelen med Monte Carlo-simulering av väntevärden är att det kan användas även för väntevärden som är svåra att beräkna exakt. Låt exempelvis X och Y vara oberoende stokastiska variabler där $X \in \text{Exp}(4)$ och $Y \in N(0,1)$. Väntevärdet $E[e^{X\cos(Y)}]$ ges då av

$$\begin{aligned} E[e^{X\cos(Y)}] &= \int_0^\infty \int_{-\infty}^\infty e^{x\cos(y)} f_{X,Y}(x,y)dydx \\ &= \int_0^\infty \int_{-\infty}^\infty e^{x\cos(y)} 4e^{-4x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dydx, \end{aligned}$$

vilket är en rätt knepig integral.

Skriv en egen MATLAB-kod som beräknar väntevärdet $E[(e^{X\cos(Y)}]$ med Monte Carlo-simulering. Upprepa simuleringen av väntevärdet och se hur resultatet varierar. Prova också att variera antalet termer i det aritmetiska medelvärdet och se hur det påverkar skattningen av väntevärdet.

Problem 4 - Maximum likelihood/Minsta kvadrat

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

```
%% Problem 3: 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 4: Maximum likelihood/Minsta kvadrat (forts.)
    plot(0:0.1:6, raylpdf(0:0.1:6, my_est_ml), 'r')
    hold off
```

Problem 5 - 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 kan förväntas innehålla det sanna värdet på μ?

```
%% Problem 4: 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);
    %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 6 - Konfidensintervall för Rayleighfördelning

Vi ska nu undersöka en Rayleigh-fördelad 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 5: Konfidensintervall for Rayleighfordelning
    load wave_data.mat
    subplot(211), plot(y(1:100))
    subplot(212), hist_density(y)
```

Om du ändrar `y(1:100)` till `y(1:end)` så kan du se hela signalen. Skatta parametern på datat på samma sätt som i uppgift 4. 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 5: 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 5: 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 7 - 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.mat` på samma sätt som tidigare. I datat så är `y` antalet transistorer/yta 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 yta. Inför modellen

$$w_i = \log(y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i. \tag{2}$$

Skatta β_0 och β_1 med hjälp av MATLABs funktion `regress`. Om du skattar parametrar m.h.a. data från 1971 till 2011, vad är då din prediktion för antalet transistorer år 2020?

Problem 8 - Polynomregression

Regression kan användas även när sambandet mellan variablerna inte är linjärt utan ges av någon annan funktion, exempelvis ett polynom med grad större än ett. Vi ska undersöka detta i denna uppgift. Börja med att ladda in filen `poly.mat`. Plotta `y1`, `y2` respektive `y3`, var för sig mot `x1`, `x2`, respektive `x3`. Ser de ut att kunna beskrivas av polynom?

Inför nu modellen

$$y_k = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n. \tag{3}$$

Bilda för var och en av de tre datamängderna en X-matris på lämpligt vis. Alltså studera plottarna och designa sedan ett X sådant att det kan representera ett polynom av den grad som du tror passar. I fallet för modellen (**??**) ovan så ser X ut så här:

$$X = \begin{bmatrix} 1 & x & x^2 & \dots & x^n \\ 1 & x & x^2 & \dots & x^n \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & x & x^2 & \dots & x^n \end{bmatrix}. \tag{4}$$

Ta sedan fram din skattning av $\hat{\beta}$ med hjälp av `regress` och plotta din skattade modell

$$\hat{y} = X\hat{\beta}, \tag{5}$$

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

```
%% Problem 7: Regression
    res =  X*beta_hat - y1;
    subplot(211), normplot(res)
    subplot(212), hist(res)
```

Vilken fördelning ser de ut att komma från? Vad kan du dra för slutsatser om modellen?

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.