Laboration 2: Styrkefunktion samt Regression

Syfte

Styrkefunktion

Syftet med dagens laborationen om styrkefunktionen är att du skall:

- bli mer förtrogen med konfidensintervall.
- bli mer förtrogen med hypotestest.
- bli mer förtrogen med styrkefunktion.
- använda styrkefunktionen för att konstruera en enkel försöksplan.

Regression

Syftet med detta laborationsavsittet är att du skall:

- bli mer förtrogen med regressionsanalysen.
- undersöka modellen "enkel linjär regression" och de olika parametrarna i modellen.
- anpassa regressionsmodellerna och använda modellerna för prediktion och kalibrering.
- få förståelse för begreppet residualer.

Kurskompendium:

Olbjer kap 7.1–7.8 samt kap 10, 11.1–11.7. Läs igenom hela handledningen före laborationstillfället. Notera vilka resultat och figurer du behöver till den muntliga redovisningen.
Till laborationens start är det underlättar det om du med dig lösningar till uppgifterna nedan.

Förberedelseuppgifter – Styrkefunktion

1. Givet 10 observationer av $x_i \in \mathsf{N}(\mu, \sigma^2)$, hur konstrueras ett konfidensintervall för μ om
   (a) σ är känd?
   (b) σ är okänd?
2. Hur kan intervallet ovan användas för att testa $H_0\colon\ \mu = 0$ mot $H_1\colon\ \mu \neq 0$?
3. Vilket intervall borde användas om mot-hypotesen är $H_1\colon\ \mu > 0$?
4. Givet n_x observationer av $x_i \in \mathsf{N}(\mu_x, \sigma^2)$ och n_y observationer av $y_i \in \mathsf{N}(\mu_y, \sigma^2)$. Hur konstrueras ett intervall för skillnaden $\mu_y - \mu_x$ om σ är okänd?
5. Hur tolkas signifikansnivån, α, i ett hypotestest?
6. Styrkefunktionen kan ses som en betingad sannolikhet, vilken?

Förberedelseuppgifter – Regression

1. Antag att givet är talpar (x_i, y_i), i=1,. . .,10 där man anser att sambandet mellan x och y är linjärt. Modellen är $y_i = \alpha + \beta x_i + \epsilon_i$ där ϵ_i är oberoende observationer från $\mathsf{N}(0, \sigma^2)$.

   (a) Vad är den grafiska tolkningen av α och β i modellen?

   (b) Residualanalys är ett viktigt instrument vid analys av regressionsmodeller. Hur definieras residualerna i ovanstående modell?

   (c) Skattningen av σ^2 i modellen bygger på residualerna. Hur ser skattningen ut?

   (d) Vad är skillnaden mellan konfidensintervallet och prediktionsintervallet i en regressionsmodell? Använd gärna ett konkret exempel för att klargöra skillnaden.

   (e) Vid kalibrering vill man för ett givet värde på y, y_0, skaffa ett intervall för motsvarande x_0. Visa grafiskt hur detta kan göras utgående från ett prediktionsintervall.

2. Antag att y, responsvariabeln, beror av två oberoende variabler x_1 och x_2. Vid 10 olika försök har man noterat (x_1, x_2, y). Modellen är nu

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i, \qquad i = 1, \ldots, 10 \qquad \epsilon_i \in \mathsf{N}(0, \sigma^2)$$

   (a) Ange matriserna i matrisformuleringen av modellen.

   (b) Vad menas med att x-variablerna är kolinjära?

Styrkefunktion och hypotestest

1 Konfidensintervall för μ

Illustration av konfidensintervall

För att undersöka hur antalet observationer och σ påverkar konfidensintervall, gör uppgift **4.26** och **4.30** i övningshäftet. Jämför resultaten med förberedelseuppgift 1 och 2.

2 Hypotestest för μ

Illustration av styrkefunktion

För att undersöka hur styrkefunktionen beror av n, σ och α, gör uppgift **4.38**. Fundera på sambandet mellan α, sannolikheten att **felaktigt förkasta** H_0, styrkefunktionen och möjligheten att upptäcka en avvikelse från H_0.

Konstruera en försöksplan

En viktig fråga i medicinska prövningar är hur många patienter som behöver ingå i en studie för att upptäcka en läkemedels-effekt. Gör övningsuppgift **4.37** och undersöka hur många patienter som behövs för att upptäcka en biverkning.

Regression

3 Enkel linjär regression

Illustration av modell:

I ett enkelt simuleringsexperiment ska du undersöka hur värdet på σ påverkar modellen och de slutsatser man kan dra från data. (För att ge illustrativa bilder ges fullständiga Matlab kommandon i denna del av laborationen.) Skapa en vektor x med värden 1, 2, ..., 10 och en variabel y som erhålls genom det teoretiska linjära sambandet y=α+βx, där α och β är kända. Välj t ex y=10+2x. Till variabeln y adderas två uppsättningar av normalfördelade mätfel $\mathsf{N}(0, \sigma^2)$ med olika värden på σ, förslagsvis σ=1 och σ=5.

```
>> x=[1:10]'
>> y1=10+2*x+normrnd(0,1,10,1);
>> y2=10+2*x+normrnd(0,5,10,1);
```

Vektorn y1 består alltså nu av 10 observationer från $\mathsf{N}(10 + 2x, 1^2)$ medan y2 består av 10 observationer från $\mathsf{N}(10 + 2x, 5^2)$.

- Titta på data i samma diagram och jämför.

```
>> plot(x,10+2*x)
>> hold on
>> plot(x,y1,'x')
>> plot(x,y2,'o')
```

- För att skatta regressionslinjen och titta på residualerna utnyttjar vi den specialskrivna m-filen `reggui`.

```
>> reggui(x,y1)
>> reggui(x,y2)
```

- Titta på residualerna för de båda linjerna. Hur påverkas de av värdet på σ?

- I figurerna som alstras av `reggui` ges även skattningar och konfidensintervall för modellens parametrar. Jämför de erhållna intervallen med de sanna värdena på α och β; täcker intervallen över parametrarna?

Matlabs egen inbyggda regressionsrutin

I Matlab finns en inbyggd funktion för regressionsanalys, `regress`, som kan användas vid multipel linjär regression (och därmed förstås även vid enkel linjär regression). Observera att `reggui` endast kan användas vid enkel linjär regression samt vid polynomregression som är ett specialfall av multipel linjär regression.
Pröva hjälpkommandot `help regress` för att ta reda på hur in- och utargumenten ser ut.

- Använd `regress` för att skatta en av de två regressionslinjerna ovan. Då måste vi först bilda matrisen X som är en (10 × 2) matris med första kolumnen enbart ettor och andra kolumnen bestående av x-värdena.

```
>> X=[ones(10,1) x]
>> [b bint r]=regress(y1,X,0.05)
```

Utargumentet `bint` ger konfidensintervall för parametrarna α och β (med konfidensgrad 0.95 här ovan) och `r` är residualerna från regressionen.

- Kontrollera att de erhållna skattningarna och intervallen stämmer med de du fick från `reggui`.

4 Kalibreringskurva

Man vill göra en kalibreringskurva för en kalorimetrisk analys av fluorjoner i vatten och mäter därför transmittansen två oberoende gånger för ett antal kända koncentrationer av fluorjoner. Resultat finns i filen `kalibrer.mat`.
Eftersom vi har två mätningar per koncentration (x-värde) måste koncentrations vektorn replikeras innan vi kan göra en regression:

```
>> x = [Konc'; Konc'];
>> y = [Trans(1,:)' ; Trans(2,:)'];
```

- Pröva att anpassa en enkel linjär regressionsmodell till data med hjälp av `reggui` (observera att `reggui` behöver ej den inledande kolumnen av ettor)
- Verifiera att modellen är rimlig genom att titta på residualerna.
- Prediktion: Vad är den förväntade transmittansen då fluorkoncentrationen är 5.0? Vad är motsvarande 95% prediktionsintervall?
- Kalibrering (invers prediktion): Då man i framtiden ska använda linjen som kalibreringskurva, vill man till ett värde y bestämma ett intervall som med 95% sannolikhet täcker provets verkliga halt. Skatta ett 95% intervall för fluorkoncentrationen då man för ett prov med okänd koncentration avläst trans=82.8.

5 Exponentiella samband

Provfiske har genomförs i Bolmen sedan 1967, vid provfiskning undersöks förekomsten av olika arter, deras vikt och längd samt halten av olika miljögifter i fisken (kvicksilver, kadmium, PCB, dioxiner, etc). Filen `BolmenGadda.mat` innehåller längd (cm) och vikt (g) för 183 gäddor från Bolmen. Vi vill undersöka om det finns ett samband mellan längd och vikt.

- Börja med att plotta vikt som funktion av längd hos gäddorna

  ```
  >> plot(Langd,Vikt,'.')
  >> xlabel('Langd (cm)')
  >> ylabel('Vikt (g)')
  ```

- Använd `reggui` för att undersöka om modellen

  $$\text{Vikt} = \alpha + \beta\text{Langd} + \epsilon \qquad\qquad \epsilon \in \mathsf{N}(0,\, \sigma^2)$$

  är rimlig för data.

Ett bättre alternativ kan vara att använda ett log-log samband

$$\begin{aligned} \text{Vikt} &= a\text{Langd}^k \cdot \epsilon \\ \underbrace{\log \text{Vikt}}_{y} &= \underbrace{\log a}_{\alpha} + \underbrace{k}_{\beta}\underbrace{\log \text{Langd}}_{x} + \log \epsilon \qquad\qquad \log\epsilon \in \mathsf{N}(0,\, \sigma^2) \end{aligned}$$

- För att undersöka sambandet kan man plotta data i ett loglog-diagram

  ```
  >> loglog(Langd,Vikt,'.')
  >> xlabel('Langd (cm)')
  >> ylabel('Vikt (g)')
  ```

- Använd `reggui` för att undersöka den nya modellens lämplighet.
- Beräkna (med `reggui`'s hjälp) den predikterade vikten och ett 95% procentigt prediktionsintervall för vikten hos fiskar som är 58 cm långa. Spara värdena i två lämpliga variabler (ersätt ? nedan med lämpliga värden)

  ```
  >> yhat58 = ?;
  >> predI58 = [? ?];
  ```

- För att kunna jämföra modellen med data i sin naturliga skala sparar vi parametrarna från `reggui`, konstruerar en vektor med längder från den minsta till den största fisken

  ```
  >> v = linspace(min(Langd), max(Langd), 100);
  ```

  och plottar både data och den anpassade modellen (ersätt ? med det skattade sambandet mellan längd och vikt)

  ```
  >> plot(Langd,Vikt,'.', v, ?, 'r')
  >> xlabel('Langd (cm)')
  >> ylabel('Vikt (g)')
  ```

- Genom att plotta prediktionen för vikten då Langd = 58 i samma figur kan vi undersöka hur intervallbredden stämmer med spridningen i data.

  ```
  >> hold on %l\"{a}gg till fler linjer i en existerande figur
  >> plot(58, yhat58, '*r') %har du kommit ih\aa{}g att transformera?
  >> plot(58, predI58, '+r')
  ```

6 Muntlig Redovisning

Diskutera följande frågor med labhandledaren

6.1 Styrkefunktion

1. Givet 100 st 95%-konfidensintervall för μ. Hur många av intervallen kan förväntas **inte** innehålla μ?
2. Varför används den större t-kvantil istället för en normal-kvantil när σ är okänt?
3. Hur ser en "ideal styrkefunktion" ut?
4. Hur påverkas styrkefunktionen av n, σ och α?
5. Föklara avvägningen mellan testets styrka och α.
6. I **4.37**, hur många personer måste man mäta på för att upptäcka den nedsatta salivproduktionen?

6.2 Regression

1. Hur påverkar σ residualerna och konfidensintervallen för α och β i simuleringsexperimentet?
2. Vad blev 95%-intervall för fluorkoncentrationen då man för ett prov med okänd koncentration avläst trans=82.8?
3. Hur ser modellen i naturlig skala ut för sambandet mellan längd och vikt hos gäddorna från Bolmen (visa figuren)?
4. Hur stämmer prediktionsintervallet för vikt då Langd = 58 med de observerade vikterna (zooma i figuren)? faktorförsök. Använd figurerna från avsnittet "`Illustration av modell`".