DATORÖVNING 4
MATEMATISK STATISTIK FÖR E — FMSF20

Syfte:

Syftet med den här laborationen är att du skall bli mer förtrogen med

- Enkel linjär regression
- Multipel linjär regression

Specialrutiner och data finns att hämta på kursens hemsida:
`www.maths.lu.se/kurshemsida/fmsf45masb03/`

1 Förberedelseuppgifter

1. Repetera normalfördelningsdiagram, läsa igenom hela regressionsdelen i statistikkompendiet (avsnitt 4-6) och hela laborationshandledningen.
2. Ange modellen för enkel linjär regression med normalfördelade fel.
   (a) **Mozquizto 1:** Hur skattar vi α, β och σ^2?
   (b) Vilken fördelning får α^* och β^*?
   (c) Hur gör vi konfidensintervall för α, β?
   (d) **Mozquizto 2:** Hur kan vi testa huruvida linjens lutning är 0?
3. **Mozquizto 3:**
   (a) Hur ser ett konfidensintervall för $\mu_0 = \alpha + \beta x_0$ ut?
   (b) Vad är ett prediktionsintervall och hur räknas det ut?
   (c) Vad är ett kalibreringsintervall och hur kan det konstrueras?
4. Residualanalys är ett centralt moment i all regressionsanalys. Hur bör residualerna se ut vid en korrekt regressionsanalys? Ange några tekniker för att kontrollera detta.
5. Ange modellen för multipel linjär regression på matrisform.
6. **Mozquizto 4:** Hur ser normalekvationerna ut och hur löser vi dessa? Vad blir kovariansmatrisen för $\boldsymbol{\beta}^*$?
7. Lös uppgift ST35.

2 Enkel linjär regression

Vid enkel linjär regression söker anpassas en rät linje till datamaterialet, dvs modellen är

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

där ε_i är oberoende likafördelade störningar med väntevärdet 0 och variansen σ^2.
Vi kommer i den följande framställningen att arbeta med matrisformuleringen av modellen,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

där de ingående matriserna har följande form:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \quad \text{och} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Vi skall använda MATLAB-funktionen `regress` som skattar parametrar, beräknar konfidensintervall för dem, beräknar residualer och litet till. Gör `help regress` för att se vad funktionen gör.

2.1 Uppgift ST35

Använd `regress` för att räkna uppgift ST35:

```
>> x = [1:8]';                                % En kolumn med x-värden.
>> y = [1.5 2.3 1.7 2.0 2.5 1.9 2.2 2.4]';  % En kolumn med y-värden.
>> X = [ones(size(x)) x];     % En kolumn med ettor och en med
                              % x-värdena.
>> [b,bint] = regress(y,X)    % Skatta alfa och beta samt
                              % konfidensintervall för dem.
>> mu = X*b;                  % Beräkna den skattade linjen.
>> plot(x,y,'*',x,mu,'-')     % Rita obervationer och skattad linje.
>> refline(b(2), b(1))        % Alternativt kan vi addera en reference linje
```

Identifiera β^* och I_β i `b` och `bint` och jämför med dina tidigare beräkningar i förberedelseuppgift 7. Vill vi bara skatta parametrarna kan vi snabbt göra detta utan regress med `X\y`.

Uppgift: Stämmer parameterskattningarna med beräkningar i förberedelseuppgift 7?

Det finns en specialskriven funktion, `reggui`, som, förutom att skatta modellparametrarna, också ritar upp data, skattad linje, residualer och konfidens- och prediktionsintervall. Använd den för att lösa uppgift ST35 igen:

```
>> help reggui
>> reggui(x,y)
```

Uppgift: Identifiera β^* och I_β i figuren och jämför med dina tidigare beräkningar.

3 Polynomregression

Datamaterialet som du skall arbeta med i detta avsnitt är koldioxidhalter uppmätta vid Mauna Loa[1] varje månad under 32 år, totalt finns $32 \cdot 12 = 384$ mätvärden. Materialet finns i filen `co2.dat`, och den kan laddas in i MATLAB med kommandot `load co2.dat`. Mätvärdena hamnar då i en vektor med namnet `co2`. Plotta mätvärdena.

```
>> figure      % Ny figur
>> plot(co2)
```

[1] `www.co2.earth/monthly-co2`

Det finns uppenbarligen en kraftig periodicitet (årsvariation) i mätningarna, och en sådan låter sig inte så lätt fångas med en polynomiell regressionsfunktion. Detta problem kan lösas på flera sätt men det enklaste är att medelvärdesbilda över varje år.
Först gör vi om vektorn co2 till en matris med 12 rader (en rad per månad och en kolonn för varje år).

```
>> z = reshape(co2, 12, []);
```

Notera att vi bara behöver ange antalet rader i `reshape`; `[]` markerar att antalet columer ska beräknas.
Funktionen `mean` beräknar sen medel i varje column (d.v.s. årsmedelvärdena), och resultatet transponeras till en kolonn-vektor med `'`.

```
>> y = mean(z)'
```

Vi skapar även en vektor med den förklarande variabeln (årtalet, räknat från lämplig nollpunkt).

```
>> x = (1:length(y))';
```

(Utrycket `a:s:b` skapar en radvektor med värden från `a` till `b` i steg om `s` och `length(y)` ger antalet värden i `y`) Plotta mätvärdena.

```
>> plot(x, y, 'o')
```

Uppgift: Vad kan vi säga om periodiciteten?

Vi skall nu göra polynomregression på materialet, dvs vår modell är

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i, \quad i = 1, \ldots, n,$$

där ε_i är oberoende likafördelade störningar med väntevärdet 0 och variansen σ^2. För använda matrisformuleringen inför vi de nya förklarande variablerna $x_{ij} = x_i^j$ för $j = 1, \ldots, k$, $i = 1, \ldots, n$, och kan skriva skriva

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Vi ska nu använda funktionen `reggui` för att anpassa polynom av olika grad till datamaterialet.

```
>> reggui(x, y)
```

1. Börja med en linjär funktion ($k = 1$). Verkar en rät linje vara en tillfredsställande regressionsmodell?

2. Nästa steg är att försöka anpassa en *kvadratisk* funktion till mätvärdena, dvs vi använder ordningstalet $k = 2$ för regressionspolynomet (använd knappen ”degree” i reggui).

3. Pröva även med tredje (eller högre) grad på polynomet. Leder det till några varningsmeddelanden i kommandofönstret (testa $k \geq 5$)? Vad kan det i så fall bero på?

För att avgöra vilken modell som är lämpligt undersöker vi

- Om residualerna är oberoende (d.v.s. saknar mönster)?
- Om residualerna är normalfördelade?
- De 95 %-iga konfidensintervallen för β_k (vad är det vi vill testa?)

Mozquizto 5: Vad verkar vara ett lämpligt gradtal på polynomet för CO_2 datan?

Mozquizto 6: Vad är β-koefficienterna för detta polynom?

4 Multipel regression

Multipel linjär regression följer samma matrisform som linjär regression där vi utökar matrisen $\mathbf{X}$ med en kolonn för varje ny förklarande variabel.

4.1 Cementdata

I ett klassiska experiment mättes, i 13 försök, värmeutvecklingen i stelnande cement som funktion av viktprocenten av några ingående ämnen. I filen `cement` finns följande variabler kolonnvis:

cem1	viktprocent av $3CaO \cdot Al_2O_3$
cem2	viktprocent av $3CaO \cdot SiO_2$
cem3	viktprocent av $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$
cem4	viktprocent av $2CaO \cdot SiO_2$
värme	utvecklad värme i kalorier per gram cement

```
>> load cement.dat
>> cement
```

Vissa av de fyra cementvariablerna samvarierar kraftigt med varandra vilket påverkar regressionsanalysen. Utnyttja gärna `corrcoef`, som räknar ut korrelationsmatrisen. Plotta de olika cementvariablerna mot värme och även de olika cementvariablerna mot varandra.

```
>> corrcoef(cement)
>> x = cement(:,1:4);
>> Y = cement(:,5);
>> help plotmatrix
>> plotmatrix(cement)
```

Mozquizto 7: Vilka variabler verkar samvariera?

Börja med att bestämma en full regressionsmodell med värme som responsvariabel och samtliga fyra cementvariabler som förklarande variabler. Först konstruerar vi `X` och beräknar β^*

```
>> X = [ones(size(Y)) x]
>> beta = X\Y
```

Residualer och variansskattning

```
>> res = Y-X*beta
>> [n, c] = size(X)
>> f = n-c
>> s2 = sum(res.^2)/f
```

Varians för β^* och residualer

```
>> Vbeta = s2*inv(X'*X)
>> plot(res,'o')
```

Över och under gränser i konfidens intervall för β kan nu beräknas med (där $t_{0.05/2}(f)$-kvantiler kan fås med `tinv(1-0.05/2,f)`)

```
>> kvantil = tinv(1-0.05/2, f)
>> IbetaL = beta - kvantil*sqrt(diag(Vbeta))
>> IbetaH = beta + kvantil*sqrt(diag(Vbeta))
```

Mozquizto 8: Vilka regressionskoefficienter är signifikant skilda från noll (på 5% nivån)?

Givetvis kunde vi också använt funktionen `regress` direkt

```
>> [beta,Ibeta,res,resint,stats] = regress(Y, X, 0.05)
```

Förmodligen är du inte alls nöjd med den fulla regressionsmodellen du just bestämt för cementdata, t ex samvarierade några av de förklarande variablerna kraftigt och kanske skall inte alla vara med. Försök komma fram till en bra regressionsmodell, vilket ju inte är helt lätt...
Funktionen `stepwise` kan vara till stor hjälp vid modellvalet

```
>> help stepwise
>> stepwise(x,Y)
```

Notera att `stepwise` **alltid** inkluderar ett intercept, därför använder vi `x` istället för `X` i anroppet. Vanligen tar inkluderas intercept även om det inte är signifikant om det inte finns mycket starka skäl att ta bort ett intercept.

Mozquizto 9: Vilken modell kom du fram till?

Använd `regress` för att skatta den nya modellen (ersätt `?` med index för variablerna, kom ihåg interceptet!)

```
>> [betaN, IbetaN] = regress(Y, X(:,[?]), 0.05)
```

Mozquizto 10: Vad blir skattningarna av β för modellen från `stepwise`?

5 Kalibrering av flödesmätare

5.1 Bakgrund

Kalibrering av en flödesmätare genomförs i en speciell kalibreringsrigg där mätningarna från en referensmätare eller referensmetod kan jämföras med mätaren som håller på att kalibreras.

5.2 Metod

Vi har nu tillgång till data från en kalibrering av en ultraljudsflödesmätare. Datamaterialet, som kommer från institutionen för värme- och kraftteknik, omfattar 71 mätningar och är lagrat i filen `flow.mat` (`load flow.mat`); variabeln `fx2` avser referensflödesmätningar från kalibreringsriggen och `fy2` avser motsvarande flöden uppmätta med den testade ultraljudsflödesmätaren (flödeshastigheterna givna i enheten m/s).
Tanken är här att vi med hjälp av de gjorda mätningarna med givare och referens skall skatta parametrarna i en enkel linjär regressionsmodell. Vi antar att referensmätningarnas fel kan försummas i jämförelse med ultraljudsgivarens (varför måste vi bekymra oss om detta?) och att ultraljudsgivarens fel är oberoende, likafördelade och har väntevärdet noll.
För att studera detta datamaterial ska vi använda funktionen `reggui`. Observera att du automatiskt kan rita ut konfidensintervall och/eller prediktionsintervall genom att markera rutan `mark ints`; om du klickar under regressionslinjen fås intervall för μ och y medan klick över linjen ger intervall för x.
För att bilden skall bli tydligare börjar vi med att studera en liten delmängd av materialet, 10 talpar av flödesmätningar som ges i variablerna `fx1` och `fy1`.

```
>> load flow.mat
>> reggui(fx1,fy1)
```

Använd nu funktionen interaktivt för att göra följande beräkningar:

1. Beräkna det förväntade värdet enligt ultraljudsmätaren, då flödet enligt kalibreringsriggen är $0.40 m/s$. Beräkna också ett 95%-igt konfidensintervall för detta förväntade värde. Beräkna dessutom ett 95%-igt prediktionsintervall för en framtida observation från ultraljudsmätaren, då kalibreringsriggen ger mätvärdet $0.40 m/s$.

2. Vid användning av den kalibrerade ultraljudsmätaren, behöver vi ”läsa baklänges” i kalibreringskurvan. Antag att vi med ultraljudsmätaren får mätvärdet $0.48 m/s$. Beräkna ett 95%-igt konfidensintervall för den ”sanna” flödeshastigheten (dvs det värde som kalibreringsriggen skulle ge).

3. **Mozquizto 11:** Notera värdena på de tre intervallen eftersom du ska använda dem senare i laborationen.

   $\mu_0^*(0.40) \quad =$

   $I_{\mu_0(0.40)} \quad =$

   $I_{y_0(0.40)} \quad =$

   $I_{x_0(0.48)} \quad =$

4. När vi enligt det ovanstående beräknat olika konfidens- och prediktionsintervall har vi förutsatt att mätfelen är normalfördelade med konstant varians. Var i beräkningarna utnyttjas detta antagande?

Om vi vill använda kalibreringskurvan i seriösa sammanhang måste vi utföra en modellvalidering, d.v.s. vi måste kontrollera att den linjära regressionsmodellen ger en adekvat beskrivning av sambandet. Vi kan bland annat validera modellen genom en grafisk residualanalys. Vid en sådan residualanalys får följande tre diagram, som alla kan fås i `reggui`, anses vara standard:

- Residualer gentemot predikterade y-värden.

- Residualer gentemot den oberoende variabelns värden.
- Residualer i normalfördelningsdiagram.

```
>> reggui(fx2,fy2)
```

Upprepa nu beräkningarna från första frågepunkten ovan.

Mozquizto 12: Hur ser intervall och skattningar ut för det större data materialet?

$\mu_0(0.40)^*$ =

$I_{\mu_0(0.40)}$ =

$I_{y_0(0.40)}$ =

$I_{x_0(0.48)}$ =

Jämför intervallbredderna baserade på de 10 mätningarna med motsvarande intervallbredder för den modell som är anpassad till alla de 71 mätpunkterna, Nu är det inte säkert att du lyckats pricka in precis samma x-värde i de två fallen, men vissa allmänna iakttagelser bör ändå vara möjliga.

Uppgift: Jämför de två konfidensintervallen. Skiljer de sig väsentligt åt (eller inte)? Hur kan det förklaras?

Uppgift: Jämför de två prediktionsintervallen. Skiljer de sig väsentligt åt (eller inte)? Hur kan det förklaras?

Uppgift: Ser residualerna rimliga ut?

För exakta beräkningar kan vi enkelt beräkna relevanta kvadrat-summor

```
>> n = length(fx2)
>> Sxx = sum( (fx2-mean(fx2)).^2 )
>> Syy = sum( (fy2-mean(fy2)).^2 )
>> Sxy = sum( (fx2-mean(fx2)).*(fy2-mean(fy2)) )
```

Mozquizto 13: Använd dessa för att beräkna (istället för att läsa av i `reggui`) $\mu_0^*(0.40)$, $I_{\mu_0(0.40)}$, $I_{y_0(0.40)}$ och $I_{x_0(0.48)}$.