Laboration 4
Matematisk statistik för K, TMA074

1 Introduktion

Syftet med den här laborationen är att få en djupare förståelse för linjär regression samt lära oss hur vi kan använda Matlab för regression.

2 Linjär regression

Vid linjär regression har vi en *responsvariabel*, y, som antas vara en linjär funktion av en *förklarande variabel* x. Förutsättningen för linjär regression är att vi kan välja värdena på x och att dessa kan bestämmas utan fel. Detta gör att vi kan betrakta x-värdena som fixa konstanter. För varje värde på x har motsvarande värde på y en viss variation, till exempel på grund av mätfel. Vi väljer nu ett antal värden $x_1, \ldots, x_n$ och mäter responsvariabeln för dessa värden. Vi får då mätvärden $y_1, \ldots, y_n$, där y_1 är värdet som är uppmätt då den förklarande variabeln är x_1 och så vidare. Vi har alltså talpar (x_i, Y_i) där x_i är ett fixt värde och Y_i är en slumpvariabel. Modellen vi ansätter för Y_i är

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

där ε_i är oberoende $\mathsf{N}(0, \sigma^2)$ slumpvariabler som beskriver mätfelet och β_0 och β_1 är okända parametrar som beskriver det linjära sambandet.

Vi använder minsta-kvadratmetoden för att skatta parametrarna, vilket ger

$$\beta_1^* = S_{xy}/S_{xx}$$
$$\beta_0^* = \bar{y} - \beta_1^* \bar{x}$$

Variansen σ^2 beskriver spridningen kring linjen och skattas som

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \beta_0^* - \beta_1^* x_i)^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

Här är

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$
$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}$$

Dessa uttryck för parameterskattningarna är anpassade för att göra beräkningarna för hand. Om vi istället använder dator för beräkningarna finns ofta färdiga program som gör själva regressionen. Annars kan man enkelt beräkna skattningarna genom att först skapa matriserna

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Skattningarna av β_0 och β_1 ges sen av

$$\begin{pmatrix}\beta_0^* \\ \beta_1^*\end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

och variansen skattas sen med den vanliga kvadratsumman, vilket är ett betydligt enklare sätt att göra skattningen om vi har tillgång till en dator.

3 Simulerad data

Vi börjar med att göra ett enkelt simuleringsexempel för att undersöka hur värdet på σ påverkar modellen och de slutsatser man kan dra från data.

Skapa en vektor x med värden $1, 2, \ldots, 10$ och en variabel y som erhålls genom den teoretiska regressionsmodellen (1) där β_0 och β_1 är kända. Vi sätter $\beta_0 = 1$ och $\beta_1 = 2$ och simulerar två olika datamängder där vi i den första har $\sigma = 1$ och i den andra har $\sigma = 5$:

```
>> beta0 = 1;
>> beta1 = 2;
>> n = 10;
>> x = [1:n]';
>> y1 = beta0 + beta1*x + 1*randn(n,1);
>> y2 = beta0 + beta1*x + 5*randn(n,1);
```

Titta på det linjära sambandet och de två datamängderna:

```
>> figure;
>> subplot(121)
>> plot(x,beta0+beta1*x,'k')
>> hold on
>> plot(x,y1,'o')
>> subplot(122)
>> plot(x,beta0+beta1*x,'k')
>> hold on
>> plot(x,y2,'o')
```

Vi skattar nu parametrarna och ritar upp den skattade regressionslinjen

```
>> X = [ones(n,1) x];
>> p1 = (X'*X)\(X'*y1);
>> s21 = sum((y1-X*p1).^2)/(n-2);
>> subplot(121)
>> plot(x,p1(1) + p1(2)*x)
```

- Skatta parametrarna baserat på den andra datamängden också och rita upp den skattade regressionslinjen.
- Rita upp residualerna $e_i = y_i - \beta_0^* - \beta_1^* x_i$ för de två skattningarna. Hur påverkas de av värdet på σ?

Konfidensintervall för parametrarna ges av:

$$I_{\beta_0} = \left(\beta_0^* \pm t_{\alpha/2}(n-2)s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\right)$$

$$I_{\beta_1} = \left(\beta_1^* \pm t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right)$$

Vi kan till exempel beräkna ett konfidensintervall för β_0 som

```
>> Sxx = sum((x-mean(x)).^2);
>> cl = p1(1) - tinv(1-0.025,n-2)*sqrt(s21*(1/n + mean(x)^2/Sxx));
>> cu = p1(1) + tinv(1-0.025,n-2)*sqrt(s21*(1/n + mean(x)^2/Sxx));
```

Alternativt kan vi också använda funktionen `regress` för att skatta parametrarna och beräkna konfidensintervallen. Se `help regress` för att förstå vad funktionen gör.

- Beräkna parameterskattningarna och konfidensintervallen med `regress` och jämför med skattningarna vi beräknade ovan.
- Jämför parameterskattningarna med de sanna parametervärdena. Innehåller konfidensintervallen de sanna värdena?

Ett konfidensintervall för $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$ ges av

$$I_{\mu_Y(x_0)} = \left(\beta_0^* + \beta_1^* x_0 \pm t_{\alpha/2}(n-2) s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \right)$$

I Matlab kan vi beräkna detta för alla värden x_0 som:

```
>> cu = p1(1) + p1(2)*x0 + tinv(1-0.025,n-2)*sqrt(s21)*sqrt(1/n + (x0-mean(x)).^2/Sxx);
>> cl = p1(1) + p1(2)*x0 - tinv(1-0.025,n-2)*sqrt(s21)*sqrt(1/n + (x0-mean(x)).^2/Sxx);
```

Vi ritar ut intervallet i samma figur:

```
>> plot(x0,cu,'r')
>> plot(x0,cl,'r')
```

- Skatta konfidensintervallet för den andra datan också. Är regressionslinjerna innehållna i konfidensintervallen?
- Skatta och rita ut motsvarande prediktionsintervall. Vad är skillnaden på konfidensintervallen och prediktionsintervallen?

4 Modellvalidering

En mycket viktig komponent i en regressionsanalys är validering av modellen, vilket betyder att vi måste försäkra oss om att det är lämpligt att ansätta en enkel regressionsmodell. Det vanligaste sättet att göra detta på är att beräkna residualerna $e_i = y_i - \beta_0^* - \beta_1^* x_i$. Om modellen är korrekt bör dessa residualer

- vara ungefär normalfördelade med väntevärde noll.
- inte uppvisa någon speciell struktur, som till exempel att de först är negativa, sen positiva, och sen negativa igen.
- ha ungefär samma variation för alla olika värden på x, vi får till exempel inte ha att variansen verkar vara större för stora värden på x.

Låt oss simulera en ny datamängd, skatta parametrarna och beräkna residualerna:

```
>> n = 20;
>> x = [1:n]';
>> y = 1 + 2*x + normrnd(0,1,n,1);
>> p = regress(y,[ones(size(x)),x]);
>> e = y - p(1) - p(2)*x
```

Det enklaste sättet att undersöka om villkoren i modellen är uppfyllda är att rita upp residualerna och visuellt undersöka dem. Vi plottar datan, den skattade linjen, residualerna som funktion av x, samt ett normalfördelningsdiagram för residualerna:

```
>> subplot(131)
>> plot(x,y,'.')
>> hold on
>> plot(x,p(1) + p(2)*x)
>> subplot(132)
>> plot(x,e,'.')
>> subplot(133)
>> normplot(e)
```

Undersök nu vad som händer med dessa figurer om vi ändrar modellen vi simulerar från så att modellantagandena inte är uppfyllda. Ni kan till exempel testa simulera data med en kvadratisk trend

```
>> y = 1 + 2*x.^2 + 10*normrnd(0,1,n,1);
```

byta till någon annan fördelning på mätfelet

```
>> y = 1 + 2*x + 10*trnd(2,n,1);
```

eller låta variansen på mätfelet bero på x

```
>> y = 1 + 2*x + x.*normrnd(0,1,n,1);
```

Kan ni upptäcka från residualplottarna att datan är simulerad från fel modell?

5 Projektuppgift

Utför den sista delen på projektet.