Datorövning 1

Regressions- och tidsserieanalys

Syfte

1. Lära sig plotta en beroende variabel mot en oberoende variabel
2. Lära sig skatta en enkel linjär regressionsmodell
3. Lära sig beräkna en skattning av korrelationen, ρ, mellan två variabler
4. Lära sig utföra test om $\rho = 0$
5. Lära sig rita konfidens- och prediktionsband för regressionslinjen
6. Lära sig beräkna konfidensintervall för parameterarna

Exempel

För att illustrera med exempel betraktar vi "*table 5.1*" i "*Applied Regression Analysis and Other Multivariable Methods*" av Kleinbaum et.al. Exemplet innehåller 30 observationer på variabeln "SBP", systolic blood pressure och "Age". Vi kan läsa in datat på följande sätt

```
data work.bloodpressure;
input SBP Age;
datalines;
144 39
220 47
138 45
145 47
162 65
142 46
170 67
124 42
158 67
154 56
162 64
150 56
140 59
110 34
128 42
130 48
135 45
114 17
116 20
124 19
136 36
142 50
120 39
120 21
160 44
158 53
144 63
130 29
125 25
175 69
;
run;
```

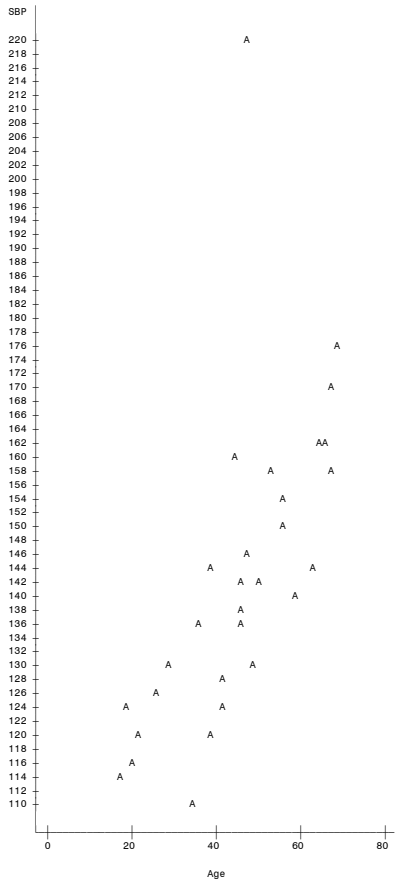
Plotta en beroende variabel mot en oberoende variabel

I exemplet från boken är variabeln blodtryck den beroende variabeln och variabeln ålder är förklarande variabel. Vi vill nu plotta dessa två för att se om det kan föreligga ett linjärt samband mellan dem. Vi använder "*proc*

plot" för att plotta de båda variablerna

```
proc plot data=work.bloodpressure;
plot SBP*Age;
run;
quit;
```

Denna kod genererar en plot som ser ut som följer



Varje observation i plotten är markerad med "A". Vi kan ändra detta genom att skriva

```
proc plot data=work.bloodpressure;
plot SBP*Age='*';
run;
quit;
```

När vi lägger till " ='*' " får vi istället en stjärna som symbol. Det finns flera alternativ, ett annat exempel är att skriva " ='+' " för att få observationerna markerade som ett plus-tecken.

Vi kan också förändra y- och x-axeln genom att skriva

```
proc plot data=work.bloodpressure;
plot SBP*Age='*'  / haxis= 0 to 80 by 10
                    vaxis= 100 to 230 by 10;
run;
quit;
```

I koden ovan anger kommandot "*haxis=*" hur vi vill att *x-axeln* ska se ut, och kommandot "*vaxis=*" anger hur *y-axeln* ska se ut. "*haxis*" står för "horisontal axis" och "*vaxis*" står för "vertical axis". Ovan väljer vi alltså att markeringarna på *x-axeln* ska gå från 0 till 80 i steg om 10, istället för i steg om 20. Vi väljer även att markeringarna på *y-axeln* ska gå från 100 till 230 i steg om 10 istället för i steg om 2.

Skatta en enkel linjär regressionsmodell

Vi kan se i plotten ovan att datat verkar kunna beskrivas av en enkel linjär regressionsmodell. För att skatta modellen i SAS använder vi "*proc reg*". Vi skriver koden

```
proc reg data=work.bloodpressure;
model SBP=Age;
run;
quit;
```

I koden ovan talar vi om hur vi vill att modellen som vi skattar ska se ut. Vi väljer att skriva "*SBP*" först, då det är vår beroende variabel. Efter den beroende variabeln kommer "*=Age*" eftersom vi vill definiera "*Age*" som oberoende variabel.

Denna kod genererar en utskrift som ser ut så här

```
                    The REG Procedure
                      Model: MODEL1
                  Dependent Variable: SBP

          Number of Observations Read          30
          Number of Observations Used          30


                    Analysis of Variance

                               Sum of          Mean
Source                  DF       Squares        Square   F Value

Model                    1    6394.02269    6394.02269     21.33
Error                   28    8393.44398     299.76586
Corrected Total         29         14787

                    Analysis of Variance

                  Source              Pr > F

                  Model               <.0001
                  Error
                  Corrected Total


      Root MSE             17.31375    R-Square     0.4324
      Dependent Mean      142.53333    Adj R-Sq     0.4121
      Coeff Var            12.14716


                    Parameter Estimates

                  Parameter      Standard
Variable    DF     Estimate         Error   t Value   Pr > |t|

Intercept    1     98.71472      10.00047      9.87     <.0001
Age          1      0.97087       0.21022      4.62     <.0001
```

I utskriften kan vi utläsa antalet observationer, en "ANOVA"-tabell (Analysis of variance) och de skattade regressionsparametrarna med tillhörande mått. De mått som ges är frihetsgraderna, skattningarna av β_0 och β_1 och $s_{\widehat{\beta}_0}$ och $s_{\widehat{\beta}_1}$. Efter kommer t-värdet till testen

$$H_0 : \beta_0 = 0$$

och

$$H_0 : \beta_1 = 0.$$

Teststatistikan till testet där $\beta_1 = 0$ ser ut som följer

$$t = \frac{\widehat{\beta}_1}{s_{\widehat{\beta}_1}} \sim t(n-2).$$

Teststatistikan till testet där $\beta_0 = 0$ har samma fördelning. Vi ser att β_0 är skattad till 98.71472 och att β_1 är skattad till 0.97087. Respektive standardfel

är 10.00047 och 0.21022. Vi kan räkna ut de observerade t-värdena genom att beräkna

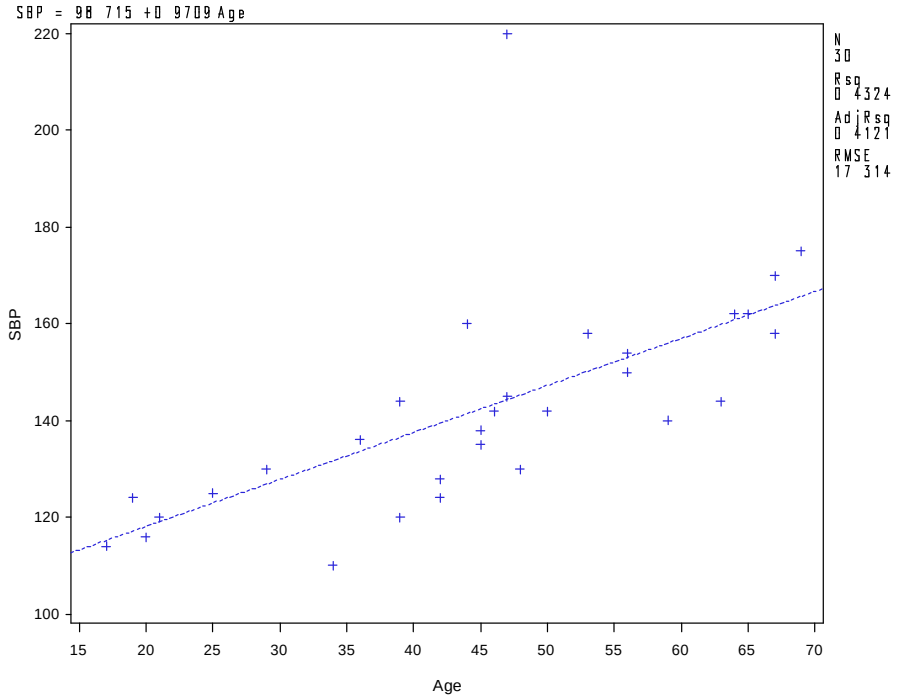
$$\begin{aligned} t_{obs} &= \frac{\widehat{\beta}_1 - \beta_1}{s_{\widehat{\beta}_1}} \\ &= \frac{0.97087 - 0}{0.21022} \\ &= 4.62 \end{aligned}$$

vilket även går att utläsa ur utskriften. p-värdet till detta test är < 0.0001, vi bör alltså förkasta $H_0 : \beta_1 = 0$. t-värdet för testet om β_0 beräknas analogt.

Vi kan även ange ett "*plot*" kommando i "*proc reg*" koden för att rita den skattade regressionslinjen. Vi skriver då

```
proc reg data=work.bloodpressure;
model SBP=Age;
plot SBP*Age;
run;
quit;
```

I koden ovan har vi lagt till en rad; "*plot SBP*Age*". Denna rad genererar en utskrift som ser ut så här



Överst till vänster ser vi ekvationen för den skattade linjen: SBP $= 98.715 + 0.9709$ Age. Till höger om plotten ser vi antalet observationer, R^2 (förklaringsgraden), R^2_{adj} och den skattade standardavvikelsen (Root MSE) hos residualerna.

Beräkna korrelationen mellan två variabler samt utföra test om $\rho = 0$

För att beräkna korrelationen mellan två variabler använder vi "*proc corr*". Vi skriver koden

```
proc corr data=work.bloodpressure;
var SBP Age;
run;
quit;
```

Vi skriver "*var SBP Age*" för att beräkna korrelationen mellan dessa två variabler. (Det går givetvis bra att skriva flera variabler.) Koden genererar en utskrift som ser ut så här

The CORR Procedure

2 Variables: SBP Age

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
SBP	30	142.53333	22.58125	4276	110.00000	220.00000
Age	30	45.13333	15.29420	1354	17.00000	69.00000

Pearson Correlation Coefficients, N = 30
Prob > |r| under H0: Rho=0

	SBP	Age
SBP	1.00000	0.65757 <.0001
Age	0.65757 <.0001	1.00000

Längst ner i utskriften, under rubriken "*Pearson Correlation Coefficients*", får vi mått på korrelationerna. Vi ser att korrelationen mellan "*SBP*" och "*SBP*" (självfallet) är 1. Vi ser också att korrelationen mellan "*SBP*" och "*Age*" är skattad till 0.65757. Vi kan själva beräkna det skattade värdet av korrelationen genom formeln

$$r = \frac{s_X}{s_Y}\widehat{\beta}_1.$$

Vi hittar standardavvikelsen för Y respektive X i utskriften ovan. Vi får då

$$r = \frac{15.2942}{22.58125} \times 0.97087 = 0.65757.$$

Under siffran 0.65757 finns värdet < 0.0001. Detta värdet anger p-värdet till testet

$$H_0 : \rho = 0.$$

Eftersom p-värdet är mycket litet bör vi förkasta H_0 och anta att korrelationen mellan "*SBP*" och "*Age*" är skild från noll.

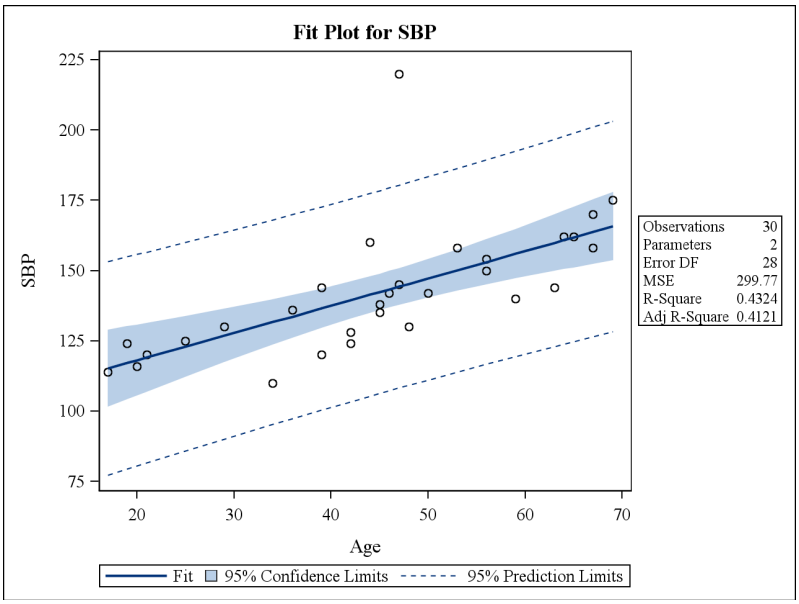
Rita konfidens- och prediktionsband för regressionslinjen

För att rita konfidensband för regressionslinjen anger vi ett tillval. Vi skriver koden

```
ods rtf;
ods graphics on;
proc reg data=work.bloodpressure plots (only) =  fit  ;
model SBP=Age ;
run;
quit;
ods graphics off;
ods rtf close;
```

Vi lägger alltså till "*plots (only) = fit*" för att rita konfidensbanden. Det finns andra kommandon för fler plottar, men detta går vi igenom längre fram. När vi lägger till "*(only)*" genererar SAS *endast* de plottar vi ber om, och på så sätt utesluts de plottar som genereras "by default". När vi använder tillvalet "*plots*" måste vi omsluta koden med "*ods*" kommandon. Det går givetvis bra att välja ett annat format än "*rtf*", exempelvis "*pdf*".

Detta genererar bilden



Beräkna konfidensintervall för parametrarna

För att beräkna ett 95%-igt konfidensintervall för parametrarna, β_0 och β_1, i SAS skriver vi koden

```
proc reg data=work.bloodpressure;
model SBP=Age / clb alpha=0.05;
run;
quit;
```

Vi skriver "*clb*" (confidence limits beta) för att specificera att vi vill ha ett konfidensintervall för respektive β. Vi skriver även "*alpha=0.05*" för att tala om att vi vill göra ett 95%-igt konfidensintervall.

Detta genererar en utskrift som ser ut så här

The REG Procedure
Model: MODEL1
Dependent Variable: SBP

Number of Observations Read	30
Number of Observations Used	30

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6394.02269	6394.02269	21.33	<.0001
Error	28	8393.44398	299.76586		
Corrected Total	29	14787			

Root MSE	17.31375	R-Square	0.4324
Dependent Mean	142.53333	Adj R-Sq	0.4121
Coeff Var	12.14716		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	98.71472	10.00047	9.87	<.0001	78.22969	119.19975
Age	1	0.97087	0.21022	4.62	<.0001	0.54026	1.4014

Utskriften ser ut som den vi fick ovan, med tillägget att vi får konfidensintervall gränserna för β_0 och β_1 längst ner till höger.

Uppgifter

Basuppgift

1. Beakta uppgift 6 i kursboken där 13 barn har undersökts. Hos barnen har medelvärdet av den totala tiden de sover registrerats (ATST) samt

barnens ålder. Värdena är

ATST min/24 h	Age
586.00	4.4
461.75	14.0
491.00	10.1
565.00	6.7
462.00	11.5
532.10	9.6
477.60	12.4
515.20	8.9
493.00	11.1
528.30	7.75
575.90	5.5
532.50	8.6
530.50	7.2

Läs in datat.

(a) Plotta ATST mot ålder för att se om en enkel linjär regressionsmodell kan beskriva datat

(b) Skatta en enkel linjär regressionsmodell, samt plotta den

(c) Beräkna korrelationen mellan den beroende- och den oberoende variablen

(d) Beräkna ett 95 %-igt konfidensintervall för β_1

(e) Testa om $\beta_1 = 0$. Verkar testet överensstämma med konfidensintervallet?

(f) Rita konfidensband för regressionslinjen

Extrauppgift

1. I tabellen nedan redovisas observationer på hastighet (km/h) och bensin-

förbrukning (liter/mil) för en Ford Escort.

Hastighet km/h	Förbrukning liter/mil	Hastighet km/h	Förbrukning liter/mil
10	2.10	90	0.76
20	1.30	100	0.83
30	1.00	110	0.90
40	0.80	120	0.99
50	0.70	130	1.08
60	0.59	140	1.18
70	0.63	150	1.28
80	0.70		

(a) Plotta "*hastighet*" mot "*förbrukning*" (Vilken är beroende variabel?). Ser datat ut att kunna beskrivas av en enkel linjär regressionsmodel? Om inte vilket typ av samband verkar föreligga?

(b) Beräkna korrelationen samt utför testet $H_0 : \rho = 0$. Är resultatet av testet konsistent med vad som kunde utläsas ur plotten?