Datorövning 2

Regressions- och tidsserieanalys

Syfte

1. Lära sig skapa en korrelationsmatris
2. Lära sig skatta en multipel linjär regressionsmodell samt plotta variablerna mot varandra
3. Lära sig beräkna VIF-värden
4. Lära sig skatta en linjär regressionsmodell med interaktionstermer

Exempel

För att genomföra exemplen nedan betraktar vi exempel 8.1 i "*Applied Regression Analysis and Other Multivariable Methods*" av Kleinbaum et.al, fjärde upplagan. Datamaterialet innehåller 12 observationer på barn med en viss sjukdom. De variabler som är observerade är vikt, längd och ålder. Vi kan läsa in datat på följande sätt:

```
data work.ex81;
input weight height age;
datalines;
64 57 8
71 59 10
53 49 6
67 62 11
55 51 8
58 50 7
77 55 10
57 48 9
56 42 10
51 42 6
76 61 12
68 57 9
;
run;
```

Skapa en korrelationsmatris

För att skapa en korrelationsmatris använder vi samma kod som i Datorövning 1, Regressions- och tidsserieanalys. Vi skriver

```
proc corr data=work.ex81;
var weight height age;
run;
```

Detta ger en utskrift som ser ut så här

The CORR Procedure

3 Variables: weight height age

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
weight	12	62.75000	8.98610	753.00000	51.00000	77.00000
height	12	52.75000	6.82409	633.00000	42.00000	62.00000
age	12	8.83333	1.89896	106.00000	6.00000	12.00000

Pearson Correlation Coefficients, N = 12
Prob > |r| under H0: Rho=0

	weight	height	age
weight	1.00000	0.81426 0.0013	0.76982 0.0034
height	0.81426 0.0013	1.00000	0.61384 0.0337
age	0.76982 0.0034	0.61384 0.0337	1.00000

Vi vill att den beroende variabeln ska ha en stark korrelation med de båda oberoende variablerna. Däremot vill vi inte att de båda oberoende variablerna ska ha allt för stark korrelation sinsemellan. I utskriften ser vi att korrelationen mellan "*weight*" och "*height*" är skattad till 0.81426 samt att korrelationen mellan "*weight*" och "*age*" är skattad till 0.76982. Vi undersöker även korrelationen mellan "*height*" och "*age*". Den är skattad till 0.61384.

Under skattningarna av korrelationen finner vi p-värdena. Vi ser att p-värdet för skattningen av korrelationen mellan "*weight*" och "*height*" och "*weight*" och "*age*" är 0.0013 respektive 0.0034. Dessa värden antyder att vi ska förkasta hypotesen $H_0 : \rho = 0$, vilket är väntat. p-värdet för skattningen av korrelationen mellan "*height*" och "*age*" är 0.0337, vilket kan förkasta nollhypotesen om ingen korrelation på 5% signifikans nivå.

Skatta en multipel linjär regressionsmodell

Vi vill nu skatta en multipel linjär regressionsmodell där "*weight*" är beroende variabel och "*height*" och "*age*" är oberoende variabler. Vi använder "*proc reg*" och skriver

```
proc reg data=work.ex81;
model weight=age height;
run;
quit;
```

Denna kod genererar en utskrift som ser ut som följer

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Number of Observations Read	12
Number of Observations Used	12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.82261	346.41130	15.95	0.0011
Error	9	195.42739	21.71415		
Corrected Total	11	888.25000			

Root MSE	4.65984	R-Square	0.7800
Dependent Mean	62.75000	Adj R-Sq	0.7311
Coeff Var	7.42605		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.55305	10.94483	0.60	0.5641
age	1	2.05013	0.93723	2.19	0.0565
height	1	0.72204	0.26081	2.77	0.0218

Under rubriken "Analysis of Variance" ges en ANOVA tabell. Där kan vi utläsa resultatet av ett omnibus $F-$test, där vi testar

$$H_0 : \beta_1 = \beta_2 = 0$$

mot

$$H_A : \text{ minst en av } \beta_1 \text{ och } \beta_2 \text{ skild från noll}$$

Teststatistikan till testet är

$$F = \frac{SSR/k}{SSE/(n-k-1)} \sim F_{k,n-k-1,1-\alpha}$$

Det observerade $F-$värdet ges i utskriften. Vi ser att $F_{obs} = 15.95$ och att p-värdet till testet är 0.0011. Det går alltså att förkasta H_0 på en signifikansnivå som är större än 0.0011.

Vi kan undersöka huruvida någon av de enskilda parameterskattningarna är signifikanta. Under rubriken "Parameter Estimates" hittar vi skattningarna av parametrarna, standardfelen, observerade $t-$värden samt tillhörande p-värden. I utskriften ser vi att β_0- och $\beta_1$skattningarna inte är signifikanta på signifikansnivån 5%. β_2 skattningen är däremot signifikant på signifikansnivån 5%, men inte på någon signifikansnivå mindre än 0.0218.

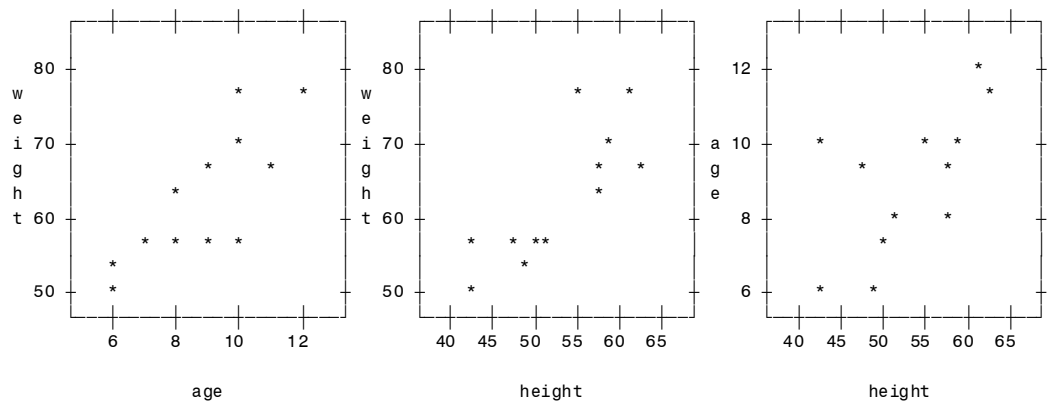
Vi ser också i utskriften att förklaringsgraden, R^2, är 0.78 och att den justerade förklaringsgraden R^2_{adj} är 0.7311. Denna utskrift går även att hitta i kursboken.

När vi skattar modellen kan vi även lägga till koden "plot" som vi gjorde i koden för den enkla linjära regressionsmodellen. För att få plottarna bredvid varandra och på ett lättöverskådligt sätt kan vi skriva

```
proc reg data=work.ex81 lineprinter;
model weight=age height;
plot weight*(age height) age*height/vplots=3 hplots=3 symbol='*';
run;
quit;
```

I översta raden skriver vi "*lineprinter*" för att vi vill ha utskriften i "output"-fönstret. Vi skriver "*plot weight*(age height)*" för att plotta weight mot de båda oberoende variablerna. Vi skriver "*age*height*" för att plotta de båda oberoende variablerna mot varandra. Efter snedstrecket kommer kod som ändrar utseendet på plottarna. Vi specificerar här att det är 3 plottar vi genererar samt att vi vill att observationerna ska vara markerade med

stjärnor. Plottarna ser ut så här



Vi ser att ett linjärt samband verkar föreligga mellan den beroende variabeln och de båda oberoende variablerna, men också att ett linjärt samband kan föreligga mellan de båda oberoende variablerna.

Det går även bra att skapa plottarna utifrån anvisningarna i Datorövning 1, Regressions- och tidsserieanalys.

Beräkna VIF-värden

VIF står för "*variance inflation factor*" och används för att mäta kollinjäritet i en multipel regressionsmodell. Kursboken tar upp VIF-värden i kapitel 14. För att beräkna VIF-värden i SAS skriver vi koden

```
proc reg data=work.ex81;
model weight=age height / vif;
run;
quit;
```

Vi gör alltså ett tillval i raden där vi specificerar modellen, vi skriver "/vif". Detta genererar en utskrift som ovan med tillägget "*Variance Inflation*"

längst ner till höger.

The REG Procedure
Model: MODEL1
Dependent Variable: weight

Number of Observations Read 12
Number of Observations Used 12

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	692.82261	346.41130	15.95	0.0011
Error	9	195.42739	21.71415		
Corrected Total	11	888.25000			

Root MSE	4.65984	R-Square	0.7800
Dependent Mean	62.75000	Adj R-Sq	0.7311
Coeff Var	7.42605		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6.55305	10.94483	0.60	0.5641	0
age	1	2.05013	0.93723	2.19	0.0565	1.60462
height	1	0.72204	0.26081	2.77	0.0218	1.60462

Inget av VIF-värdena är för högt, det vill säga över 10.

Skatta en regressionsmodell med interaktionstermer

För att skatta en modell med interaktionstermer använder vi datat "*work.bloodpressure*" från Datorövning 1, regressions- och tidsserieanalys. Vi läste då in datat med 30 observationer. Det första vi ska göra är att ta bort observation 2 som kan anses vara en outlier. De resterande 29 observationerna ska ses som observationer på variabeln "*SBP*" och "*age*" hos kvinnor. Vi ska alltså modifiera data-setet genom att lägga till en variabel, som exempelvis kan heta "*gender*". Vi ska dessutom sätta en 1 på varje rad, där 1 representerar kvinna. Sedan ska vi utöka data-setet med 40 observationer där vi har värden på variablerna "*SBP*", "*age*" och "*gender*". Dessa ska ha värde 0 för variabeln "*gender*", då dessa 40 observationer är gjorda på män. Vi kan läsa om exemplet och hitta datat i kapitel 11.3.3 i kursboken. Vi kan

läsa in datat så här:

```
data work.bloodpressure;
input SBP Age gender @@;
datalines;
144 39 1 138 45 1 145 47 1
162 65 1 142 46 1 170 67 1
124 42 1 158 67 1 154 56 1
162 64 1 150 56 1 140 59 1
110 34 1 128 42 1 130 48 1
135 45 1 114 17 1 116 20 1
124 19 1 136 36 1 142 50 1
120 39 1 120 21 1 160 44 1
158 53 1 144 63 1 130 29 1
125 25 1 175 69 1 158 41 0
185 60 0 152 41 0 159 47 0
176 66 0 156 47 0 184 68 0
138 43 0 172 68 0 168 57 0
176 65 0 164 57 0 154 61 0
124 36 0 142 44 0 144 50 0
149 47 0 128 19 0 130 22 0
138 21 0 150 38 0 156 52 0
134 41 0 134 18 0 174 51 0
174 55 0 158 65 0 144 33 0
139 23 0 180 70 0 165 56 0
172 62 0 160 51 0 157 48 0
170 59 0 153 40 0 148 35 0
140 33 0 132 26 0 169 61 0
;
run;
```

där kommandot "@@" innebär att data läses in radvis.

Vi vill nu skatta regressionsmodellen

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$$

där

$$\begin{aligned} X &= \text{age} \\ Z &= \text{gender} \\ XZ &= \text{interaktion mellan age och gender} \end{aligned}$$

Vill vi skatta denna modell med "*proc reg*" i SAS måste vi manipulera data-

setet ytterligare. I data-steget måste vi då skriva

```
data work.bloodpressure;
input SBP Age gender @@;
age_gender=Age*gender;
datalines;
144 39 1 138 45 1 145 47 1
162 65 1 142 46 1 170 67 1
124 42 1 158 67 1 154 56 1
162 64 1 150 56 1 140 59 1
110 34 1 128 42 1 130 48 1
135 45 1 114 17 1 116 20 1
124 19 1 136 36 1 142 50 1
120 39 1 120 21 1 160 44 1
158 53 1 144 63 1 130 29 1
125 25 1 175 69 1 158 41 0
185 60 0 152 41 0 159 47 0
176 66 0 156 47 0 184 68 0
138 43 0 172 68 0 168 57 0
176 65 0 164 57 0 154 61 0
124 36 0 142 44 0 144 50 0
149 47 0 128 19 0 130 22 0
138 21 0 150 38 0 156 52 0
134 41 0 134 18 0 174 51 0
174 55 0 158 65 0 144 33 0
139 23 0 180 70 0 165 56 0
172 62 0 160 51 0 157 48 0
170 59 0 153 40 0 148 35 0
140 33 0 132 26 0 169 61 0
;
run;
```

Vi måste alltså lägga till en rad där vi definierar en ny variabel som heter "*age_gender*" och som skrivs $Age \times gender$, det vill säga som representerar interaktionseffekten.

Vi kan nu skatta modellen med "*proc reg*". Vi skriver

```
proc reg data=work.bloodpressure;
model SBP = age gender age_gender;
run;
quit;
```

Vi får då en utskrift som ser ut så här

The REG Procedure
Model: MODEL1
Dependent Variable: SBP

Number of Observations Read	69
Number of Observations Used	69

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	18010	6003.44290	75.02	<.0001
Error	65	5201.43940	80.02214		
Corrected Total	68	23212			

Root MSE	8.94551	R-Square	0.7759
Dependent Mean	148.72464	Adj R-Sq	0.7656
Coeff Var	6.01481		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	110.03853	4.73610	23.23	<.0001
Age	1	0.96135	0.09632	9.98	<.0001
gender	1	-12.96144	7.01172	-1.85	0.0691
age_gender	1	-0.01203	0.14519	-0.08	0.9342

Vi ser att skattningen av β_2 är -12.96144. Detta betyder att blodtrycket för kvinnorna förväntas vara 12.96144 enheter lägre för kvinnor än för män, då de övriga oberoende variablerna hålls konstant. Vi tolkar β_2 utifrån kvinnorna eftersom det är för kvinnorna som variabeln "*gender*" antar värdet 1. Vi ser också att varken β_2 eller β_3 skattningen är signifikant.

Det finns ett annat sätt att skatta samma modell i SAS, och det är genom att använda "*proc GLM*" (GLM står för "*generalized linear models*"). Denna kod är enklare att använda eftersom vi då inte behöver definiera den nya variabeln "*age_gender*" i data-steget. Vi kan istället direkt skriva

```
proc GLM data=work.bloodpressure;
model SBP = age gender age*gender;
run;
quit;
```

Det räcker alltså med att sätta en stjärna mellan "*age*" och "*gender*" i raden med "*model*" kommandot. Detta går inte att göra i "*proc reg*". Utskriften

vi får är mycket lik den som genereras med "*proc reg*"

The GLM Procedure

Dependent Variable: SBP

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	18010.32871	6003.44290	75.02	<.0001
Error	65	5201.43940	80.02214		
Corrected Total	68	23211.76812			

R-Square	Coeff Var	Root MSE	SBP Mean
0.775914	6.014814	8.945510	148.7246

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Age	1	14951.25461	14951.25461	186.84	<.0001
gender	1	3058.52475	3058.52475	38.22	<.0001
Age*gender	1	0.54936	0.54936	0.01	0.9342

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Age	1	7971.007140	7971.007140	99.61	<.0001
gender	1	273.443297	273.443297	3.42	0.0691
Age*gender	1	0.549356	0.549356	0.01	0.9342

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	110.0385285	4.73610350	23.23	<.0001
Age	0.9613526	0.09632327	9.98	<.0001
gender	-12.9614443	7.01172459	-1.85	0.0691
Age*gender	-0.0120301	0.14519328	-0.08	0.9342

Det kan vara en fördel att använda "*proc GLM*" när man skattar en regressionsmodell med många interaktionseffekter. Vi slipper då definiera dessa variabler i data-steget utan kan skriva in dem direkt i "*proc GLM*" koden.

Uppgifter

Basuppgifter

1. Betrakta det utökade data-setet "work.bloodpressure".

   (a) Plotta variablerna mot varandra för att undersöka sambandet mellan dem. Hur ser det ut? Är det vettigt att tolka alla plottar?

   (b) Tag fram korrelationsmatrisen.

   (c) Tag fram VIF-värden

(d) Utifrån utskrifterna i exemplet samt plottarna, korrelationsmatrisen och VIF-värdena, skatta en linjär regressionsmodell som beskriver datat bättre. (Vilka variabler borde ingå i modellen? Vilka variabler bidrar inte med någon information?)

2. Utgå från övning 3, kapitel 8 i kursboken. Det finns 25 observationer på variablerna "Cholesterol" (Y),"Weight" (X_1) och "Age" (X_2). Vi läser in datat

```
data work.problem83;
input Cholesterol Weight Age;
datalines;
354 84 46
190 73 20
405 65 52
263 70 30
451 76 57
302 69 25
288 63 28
385 72 36
402 79 57
365 75 44
209 27 24
290 89 31
346 65 52
254 57 23
395 59 60
434 69 48
220 60 34
374 79 51
308 75 50
220 82 34
311 59 46
181 67 23
274 85 37
303 55 40
244 63 30
;
run;
```

I boken anges tre modeller

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E \\ Y &= \beta_0 + \beta_1 X_1 + E \\ Y &= \beta_0 + \beta_1 X_2 + E \end{aligned}$$

Vilken modell beskriver datat bäst? Undersök detta genom att

- plotta variablerna mot varandra
- bestämma korrelationsmatrisen
- beräkna $t-$ och $p-$värden för parameterskattningarna
- beräkna förklaringsgraden