Datorövning 3

Regressions- och tidsserieanalys

Syfte

1. Lära sig granska om modellantaganden uppfylls, för både en enkel och en multipel linjär regressionsmodell

2. Lära sig identifiera outliers

Exempel

Vi går igenom de antaganden som gäller för en multipel linjär regressionsmodell. De antaganden som gäller för en enkel linjär regressionsmodell är liknande, men för endast ett X. Antagandena är:

1. För varje kombination av värden på de oberoende variablerna $X_1, X_2, ..., X_k$ betraktas Y som en stokastisk variabel med en viss fördelning. Denna fördelning har ett ändligt väntevärde och en ändlig varians.

2. Observationerna på den stokastiska variabeln Y är statistiskt oberoende av varandra.

3. Väntevärdet för Y, $\mu_{Y|X_1,X_2,...,X_k}$ för varje specifik kombination av värden på de oberoende variablerna $X_1, X_2, ..., X_k$ är en linjär funktion av $\beta_0, \beta_1, ..., \beta_k$. Detta kan skrivas som

$$\begin{aligned} \mu_{Y|X_1,X_2,...,X_k} &= \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \Rightarrow \\ Y &= \beta_0 + \beta_1 X + ... + \beta_k X_k + E \end{aligned}$$

   där E är feltermen som visar på skillnaden mellan det individuella observerade värdet på Y och det sanna medelvärdet, $\mu_{Y|X_1,X_2,...,X_k}$.

4. Variansen för Y är densamma för alla kombinationer av värden på de oberoende variablerna $X_1, X_2, ..., X_k$, det vill säga homoskedastisk varians. Detta kan skrivas som

$$\sigma^2_{Y|X_1,X_2,...,X_k} = Var(Y \mid X_1, X_2, ..., X_k) \equiv \sigma^2$$

5. För varje kombination av värden på de oberoende variablerna $X_1, X_2, ..., X_k$ är Y normalfördelad. Detta kan skrivas som

$$Y \sim N(\mu_{Y|X_1,X_2,...,X_k}, \sigma^2),$$

eller

$$E \sim N(0, \sigma^2),$$

vilket är ekvivalent.

Antagande 2, 4 och 5 kan sammanfattas med antagandet

$$E \sim iidN(0, \sigma^2).$$

Det vill säga, vi antar att feltermen, E, har en normalfördelning med väntevärde 0 och konstant varians. Vi antar dessutom att varje E_i, där $i = 1, ..., n$ är oberoende. *iid* står för "independent, identically distributed".

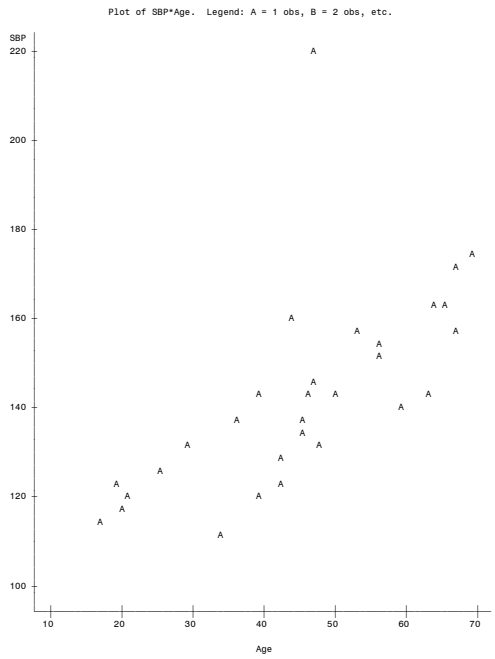
Granska modellantaganden och identifiera outliers

Enkel linjär regressionsmodell Vi använder data-setet "*work.bloodpressure*" från Datorövning 1, Regressions- och tidsserieanalys (Table 5.1 i kursboken) när vi genomför exemplet. Vi använder inte det utökade data-setet (innehållande kön) utan det ursprungliga. 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;
```

För att granska **antagande 3** plottar vi Y mot X för att se om en linjär regressionsmodell kan beskriva relationen mellan Y och X. Vi gjorde detta på "Datorövning 1, regressions- och tidsserieanalys", och plotten ser ut som

följer



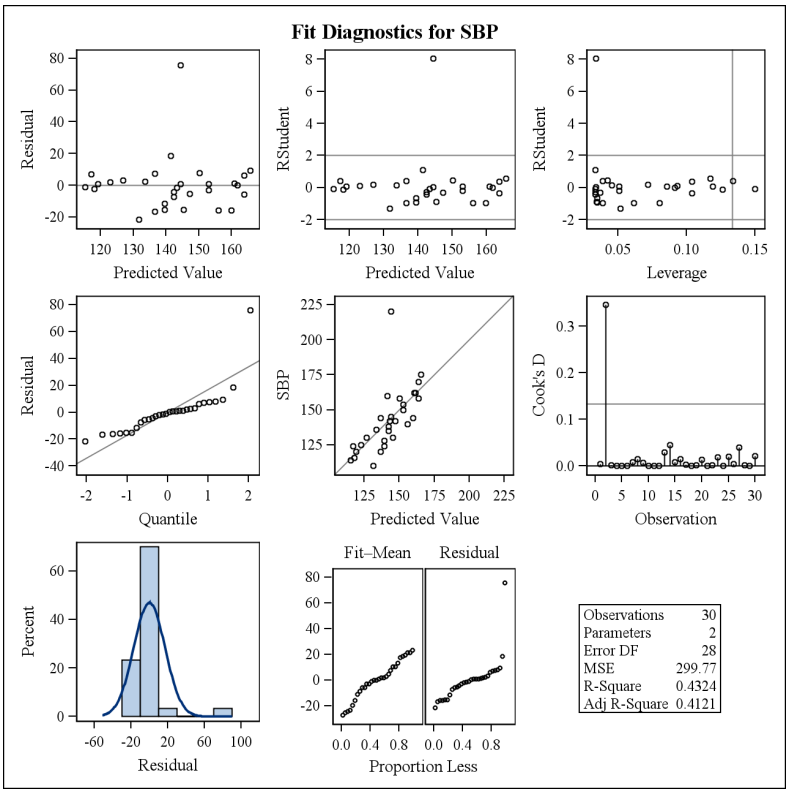
Här ser vi att ett linjärt samband verkar föreligga mellan "SBP" och "Age". Vi ser även att det finns en outlier i data-setet. Övertyga er om var den ligger i grafen.

Vi kan även granska linjäritetsantagandet genom att plotta residualerna mot $\widehat{Y}$, de skattade $Y-$värdena. Detta ger samma information som plotten ovan. Det mönster som då förväntas föreliggea om Y beror linjärt på X, är att residualernas värden ska variera runt 0 (se subfigur längst upp till vänster på nästa sida).

För att granska **antagande 2** och **4** plottar vi residualerna mot $\widehat{Y}$. För att kontrollera **antagande 5** kan vi dels titta på histogrammet för residualerna, men också rita en $Q-Q-$plot. Vi ritar dessa plottar i SAS genom att skriva koden

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

Ovan lägger vi till ett tillval, nämligen "*plots*". Vi skriver "*plots (only) = diagnostics*". När vi skriver "*diagnostics*" får vi en sammanfattning av diagnostiska plottar. (Vi skulle även kunna skriva "*all*" och få ut alla plottar som går att skapa i SAS för vår modell.) När vi lägger till "*(only)*" genererar SAS endast de plottar vi ber om, och på så sätt utesluts de plottar som genereras "*by default*". Detta ger utskriften



Den första och andra plotten i ordningen är plottar med residualerna respektive "jackknife residuals", plottade mot $\widehat{Y}$ (residualerna kallas "*rstudent*" i SAS, men "*jackknife*" i boken). Dessa plottar används för att kontrollera antagande 2 och 4, men även för att upptäcka outliers. I plot nummer 2 bör observationerna ligga inom intervallet $[-2, 2]$, för att inte klassas som outliers. Vi ser i plotten att det finns en outlier. Bortsett från den observationen ser det ut som att vi kan anta att $Y-$värdena är oberoende av varandra, då de inte uppvisar något systematiskt mönster, samt att det föreligger konstant varians. Vi ser också att väntevärdet hos feltermerna verkar vara 0, vilket följer från minsta-kvadrat metoden av parameterskattningarna.

Det finns vissa typer av beroende som inte kan upptäckas i plottarna ovan. Om vi har upprepade mätningar, på exempelvis samma individ, föreligger beroende. Denna typ av beroende är inte självklart att man ser genom att plotta residualerna mot $\widehat{Y}$.

Även beroende över tiden är en typ av beroende som inte upptäcks i residualplotten. För att upptäcka tidsberoende ska residualerna plottas mot tiden.

Plot nummer 4 och 7 i ordningen används för att kontrollera antagandet om normalfördelning. Plot nummer 4 kallas en $Q-Q-$plot, och om feltermerna följer en normalfördelning ska observationerna ligga längs med linjen. Även här sticker observationen som klassas som en outlier ut, vilket medför att datat avviker från en normalfördelning. I plot nummer 7 ser vi ett histogram för residualerna. Här ser vi att residualerna inte följer en normalfördelning.

För att identifiera observationen som tycks vara en outlier skriver vi

```
proc reg data=work.bloodpressure;
model SBP=Age ;
output out=work.res rstudent=jackknife;
run;
quit;
```

Med denna kod skapar vi ett nytt data-set som heter "*work.res*". I det data-setet kommer våra ursprungliga observationer sparas samt ytterligare en variabel som heter "*jackknife*" där "*jackknife"* residualerna anges. (Kommandot "*student=*" ger "*studentized"* residualer.) För att titta på datat skriver vi

```
proc print data=work.res;
run;
```

och får en utskrift som ser ut så här

Obs	SBP	Age	jackknife
1	144	39	0.43082
2	220	47	8.04826
3	138	45	-0.25435
4	145	47	0.03776
5	162	65	0.01063
6	142	46	-0.07932
7	170	67	0.37464
8	124	42	-0.90785
9	158	67	-0.34604
10	154	56	0.05336
11	162	64	0.06820
12	150	56	-0.17960
13	140	59	-0.95219
14	110	34	-1.30447
15	128	42	-0.66886
16	130	48	-0.89719
17	135	45	-0.42855
18	114	17	-0.07503
19	116	20	-0.12942
20	124	19	0.41815
21	136	36	0.13555
22	142	50	-0.30440
23	120	39	-0.97589
24	120	21	0.05421
25	160	44	1.09468
26	158	53	0.45552
27	144	63	-0.95491
28	130	29	0.18437
29	125	25	0.11995
30	175	69	0.56443

I utskriften ser vi att observation 2 har ett värde på "jackknife" som är 8.04826. Eftersom det är större än 2 kan man dra slutsatsen att observation 2 är en outlier.

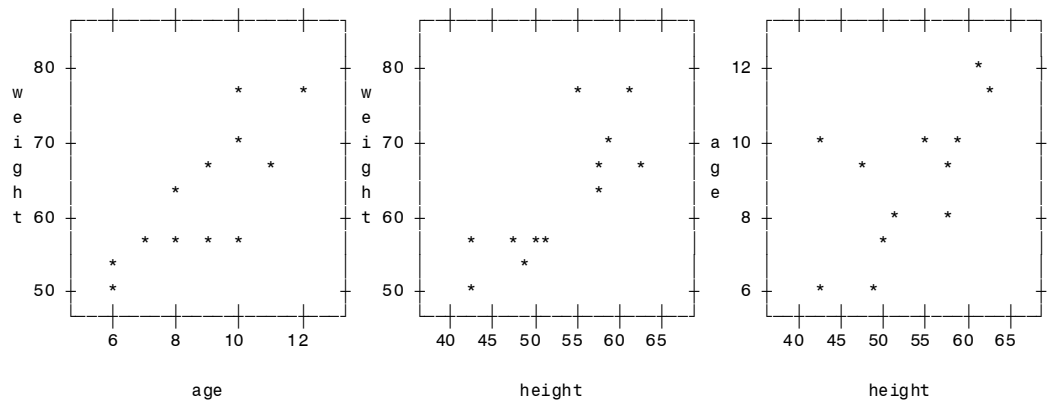
Vi kan modifiera koden ovan ytterligare om vi även vill spara "Cook distance" och "leverage" i det nya data-setet. Vi skriver då

```
proc reg data=work.bloodpressure;
model SBP=Age ;
output out=work.res rstudent=jackknife Cookd=Cookdistance h=leverage;
run;
quit;
```

Här har vi skapat ytterligare två variabler; "*Cookdistance*" och "*leverage*".

Multipel linjär regressionsmodell Vi använder data-setet "*work.ex81*" från Datorövning 2, regressions- och tidsserieanalys.

Vi börjar med att plotta variablerna mot varandra för att se om linjära samband verkar föreligga samt för att se om det finns någon outlier. Vi skapade dessa plottar på Datorövning 2, regressions- och tidsserieanalys och de ser ut som följer.



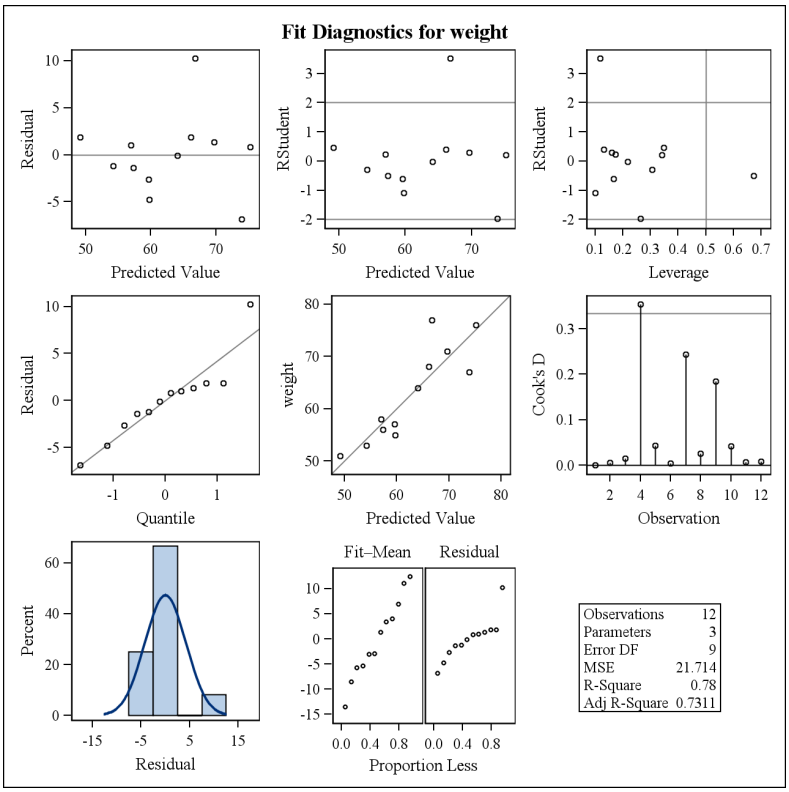
Ett linjärt samband verkar föreligga mellan variablerna. Det är dock svårt att upptäcka någon outlier.

Även i fallet med den multipla linjära regressionsmodellen kan vi granska linjäritetsantagandet genom att undersöka residualplotten. Den information vi behöver samlas då i en plot, istället för i flera plottar för variablerna.

Vi skapar nu diagnostiska plottar för att kontrollera antagande 2, 4 och 5. Vi skriver

```
ods rtf;
ods graphics on;
proc reg data=work.ex81 plots (only)=diagnostics;
model weight=age height;
run;
quit;
ods graphics off;
ods rtf close;
```

Detta genererar plottarna



I plot nummer 2 ser vi att en observation faller utanför intervallet $[-2, 2]$ och att en observation ligger på gränsen. Dessa två observationer kan vara outliers. Värdena är dock inte så höga, utan ligger relativt nära intervallgränserna. Det kan vara svårt i detta skede att avgöra om de borde klassas som outliers eller ej. Bortsett från dessa observationer verkar variansen vara konstant och observationerna oberoende.

Plottarna som används för att granska normalfördelningsantagandet visar att normalfördelning inte kan antas. Kanske kan det vara så att det är de två observationerna som klassas som outliers som gör att detta antagande inte uppfylls? För att identifiera *vilka* av observationerna som är outliers skriver

vi samma kod som förut, enligt

```
proc reg data=work.ex81;
model weight=age height;
output out=work.res rstudent=jackknife;
run;
quit;
```

Även här skapar vi ett nytt data-set som heter "*work.res*" (det går bra att döpa data-setet till vad man vill) med en ny variabel som heter "*jackknife*" (även här går det bra att döpa variabeln till vad man vill). Vi använder "*proc print*" och får en utskrift som ser ut så här

Obs	weight	height	age	jackknife
1	64	57	8	-0.02520
2	71	59	10	0.29866
3	53	49	6	-0.30116
4	67	62	11	-1.97756
5	55	51	8	-1.09296
6	58	50	7	0.22191
7	77	55	10	3.51929
8	57	48	9	-0.60305
9	56	42	10	-0.49623
10	51	42	6	0.46266
11	76	61	12	0.20019
12	68	57	9	0.40338

I utskriften kan vi se att observation 4 och observation 7 har relativt höga värden på "*jackknife*".

Uppgifter

Basuppgifter

1. Använd data-setet "work.bloodpressure". Avlägsna observationen som är en outlier och granska återigen modellantagandena. Granska även vilka värden "*jackknife*" residualerna antar genom att spara dem i ett nytt data-set. Blir det någon skillnad när outliern tas bort?

2. Nedan anges data för en budfirma där Y = restid, X_1 = mil och X_2 = antal leveranser under resan.

Y	X_1	X_2
9.3	100	4
4.8	50	3
8.9	100	4
6.5	100	2
4.2	50	2
6.2	80	2
7.4	75	3
6.0	65	4
7.6	90	3
6.1	90	2

   Läs in datat och anpassa en multipel linjär regressionsmodell. Utför modellkontroll.

Extra uppgifter

1. Läs in datat nedan och skatta en enkel linjär regressionsmodell

Y	X
0.60	80
6.70	220
5.30	140
4.00	120
6.55	180
2.15	100
6.60	200
5.75	160

(a) Utför modellkontroll.

(b) Verkar det som att datat kan förklaras av någon annan modell än den enkla linjära regressionsmodellen? Om så är fallet, skatta den samt granska återigen modellantagandena.