

## 75520 STATISTIKK 2

### EKSEMPEL 8 (H1997)

I dette eksemplet skal vi studere virkningen av multikollinearitet.

Vi har simulert 20 observasjoner av  $X_1$  som er  $N(0,1)$  og latt  $X_2 = X_1 + V$  der  $V$  er  $N(0,0.01)$ . Korrelasjonskoeffisienten mellom  $X_1$  og  $X_2$  blir dermed 0.9995. Til slutt har vi simulert

$$Y = 10 + X_1 + E$$

der "feilen"  $E$  er  $N(0,0.5)$ .

#### Data Display

Merk her at kolonnene for  $x_1$  og  $x_2$  er nesten identiske, og dermed nesten lineært avhengige.

Row	y	x1	x2
1	11.0433	1.57445	1.58114
2	9.9307	0.09271	0.09397
3	10.3324	0.53145	0.54462
4	9.6987	-0.44827	-0.46366
5	9.5924	-0.10146	-0.08361
6	9.7023	-0.30631	-0.30229
7	10.6160	0.66720	0.67605
8	9.1374	0.08235	0.06326
9	10.3527	0.35854	0.36576
10	10.1546	-0.00850	-0.00350
11	10.8577	1.09870	1.09889
12	11.6884	0.78880	0.77879
13	7.0046	-2.47167	-2.46897
14	10.7838	1.00506	0.98438
15	8.9609	-0.40892	-0.40325
16	8.6036	-0.86300	-0.86482
17	8.5149	-1.55303	-1.55902
18	9.3103	-0.28791	-0.27311
19	10.8328	0.56694	0.55574
20	9.6344	0.05502	0.04370

#### Regression Analysis (med Y som respons og $X_1$ , $X_2$ som uavhengige variable)

The regression equation is  
 $y = 9.82 + 5.06 x_1 - 3.98 x_2$

Predictor	Coef	StDev	T	P
Constant	9.81589	0.09196	106.74	0.000
$x_1$	5.057	8.415	0.60	0.556
$x_2$	-3.978	8.420	-0.47	0.643

S = 0.4109      R-Sq = 87.0%      R-Sq(adj) = 85.4%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	19.1348	9.5674	56.67	0.000
Error	17	2.8702	0.1688		
Total	19	22.0050			

Source	DF	Seq SS
$x_1$	1	19.0971
$x_2$	1	0.0377

#### Unusual Observations

Obs	$x_1$	y	Fit	StDev Fit	Residual	St Resid
8	0.08	9.1374	9.9807	0.1821	-0.8433	-2.29R
12	0.79	11.6884	10.7070	0.1433	0.9814	2.55R

R denotes an observation with a large standardized residual

#### Prediksjon

Vi bruker nå modellen til å predikere  $Y$  for et nytt sett av  $x_1, x_2$ , nemlig  $x_1=1, x_2=-1$ .

Fit	StDev Fit	95.0% CI	95.0% PI
18.8510	16.8314	(-16.6690, 54.3709)	(-16.6796, 54.3815) XX

X denotes a row with X values away from the center  
 XX denotes a row with very extreme X values

Som vi ser, er prediksjonen meget daarlig.

-1

#### Matrisen ( $X'X$ )

0.050	-0.167	0.166
-0.167	419.450	-419.628
0.166	-419.628	419.868

Vi ser at denne har store verdier, noe som skyldes at de to siste kolonnene i  $X$  er nesten lineært avhengige.

#### Best Subsets Regression

Response is y

Vars	R-Sq	R-Sq (adj)	C-p	S	x x
1	86.8	86.1	1.2	0.40193	X
1	86.7	85.9	1.4	0.40354	X
2	87.0	85.4	3.0	0.41089	X X

#### Regression Analysis (med kun $X_1$ med i modellen)

The regression equation is  
 $y = 9.82 + 1.08 x_1$

Predictor	Coef	StDev	T	P
Constant	9.81746	0.08989	109.21	0.000
$x_1$	1.08146	0.09947	10.87	0.000

S = 0.4019      R-Sq = 86.8%      R-Sq(adj) = 86.1%

#### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	19.097	19.097	118.21	0.000
Error	18	2.908	0.162		
Total	19	22.005			

Unusual Observations	Obs	x1	y	Fit	StDev Fit	Residual	St Resid
	12	0.79	11.6884	10.6705	0.1181	1.0179	2.65R
	13	-2.47	7.0046	7.1445	0.2635	-0.1399	-0.46 X

R denotes an observation with a large standardized residual  
 X denotes an observation whose X value gives it large influence.

Vi merker oss at koeffisienten til x1 naa er hoyst signifikant!

Fit	StDev Fit	95.0% CI	95.0% PI
10.8989	0.1327	( 10.6201, 11.1778)	( 10.0095, 11.7884)

Legg merke til at prediksjonen naa er mye bedre enn i tilfellet da baade x1 og x2 var med i modellen.