EKSEMPEL: RESIDUALPLOTT OG TRANSFORMASJON AV PREDIKTOR

Simulering av data

De 20 observasjonene nedenfor er simulert fra modellen

y = 1 + log x1 + x2 + epsilon

der epsilon er normalfordelt med forventing 0 og st.avvik 0.02,
x1 er uniformt fordelt på intervallet (1,3),
x2 er normalfordelt med forventning 1 og st.avvik. 0.5.
log=naturlig logaritme

Data Display

Row	y	x1	log x1	x2
1	1,31195	1,71892	0,54169	-0,22539
2	1,92490	1,29874	0,26139	0,64553
3	2,15363	1,50675	0,40996	0,74746
4	1,50252	1,88131	0,63197	-0,10866
5	3,27578	2,73379	1,00569	1,26352
6	3,50639	2,75716	1,01420	1,46375
7	2,30543	1,41297	0,34569	0,96261
8	2,37145	1,46043	0,37873	0,96975
9	3,26009	2,85981	1,05076	1,23181
10	3,02505	2,05531	0,72043	1,29565
11	3,58540	2,70916	0,99664	1,59818
12	1,95918	1,18627	0,17082	0,76620
13	2,44863	1,62882	0,48786	0,93981
14	2,77606	1,43178	0,35892	1,38495
15	3,23609	2,15803	0,76920	1,47332
16	2,44614	2,77940	1,02223	0,41908
17	2,13068	1,70870	0,53574	0,56905
18	2,99815	1,48511	0,39549	1,57772
19	2,90921	1,96982	0,67794	1,22693
20	2,80508	2,81991	1,03671	0,78403

Den gale modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (1)$$

er så tilpasset i MINITAB:

Regression Analysis: y versus x1; x2

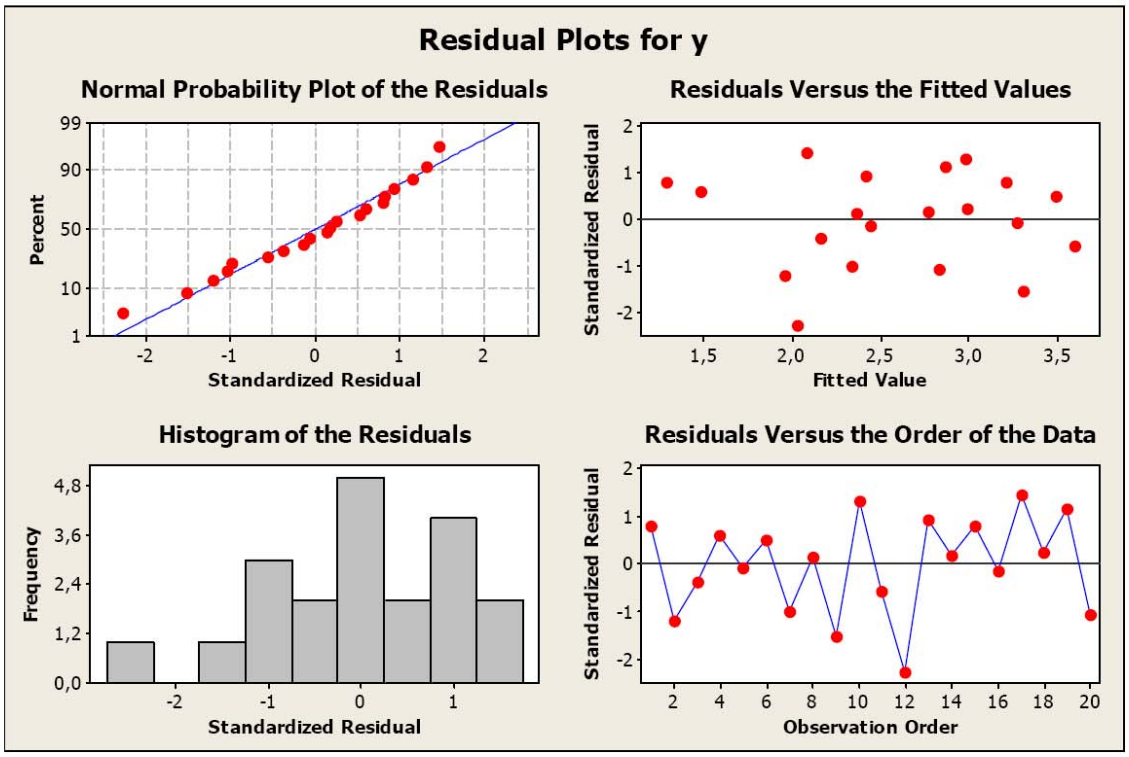
The regression equation is
y = 0,686 + 0,483 x1 + 1,01 x2

Predictor	Coef	SE Coef	T	P
Constant	0,68582	0,02829	24,24	0,000
x1	0,48315	0,01344	35,94	0,000
x2	1,00608	0,01537	65,45	0,000

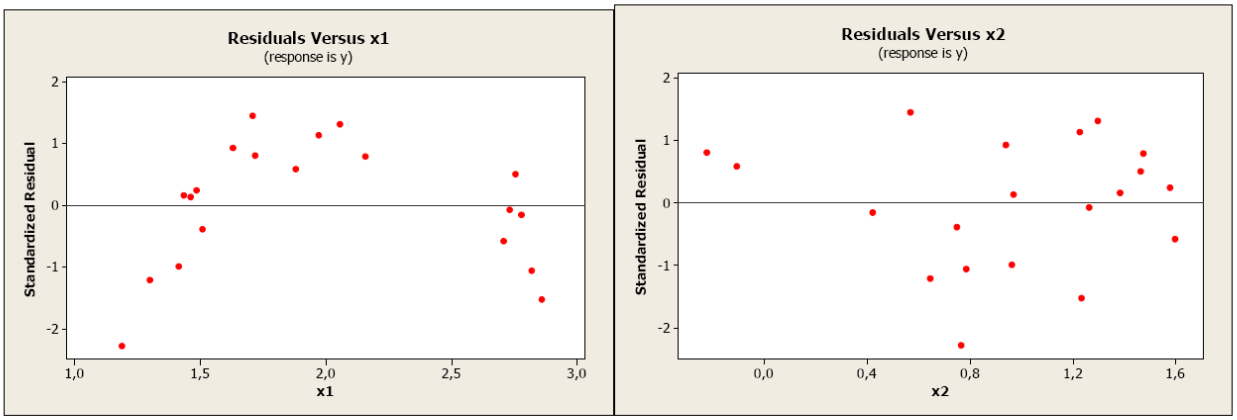
S = 0,0336068 R-Sq = 99,8% R-Sq(adj) = 99,7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	7,9130	3,9565	3503,14	0,000
Residual Error	17	0,0192	0,0011		
Total	19	7,9322			



Normalplottet kan virke OK ved første øyekast, men har en systematisk konveks krumning. Histogrammet nederst til venstre tyder på at residualenes fordeling er noe skjev mot venstre. Residuals Versus the Fitted Values ser OK ut.



Her ser vi en tydelig "omvendt U"-fasong på residualene mot x1. Det tyder på at vår modell gitt ved (1) ikke er riktig, Residualene mot x2 ser mer OK ut, men kanskje er det noe som skurrer her også?

Transformasjon av x1

Residualplottet ovenfor for x1 er en indikasjon på at en transformasjon av variabelen x1 vil gi bedre resultat. Siden vi her vet hva den underliggende modellen er, er det lett å foreslå at vi erstatter prediktoren x1 i modellen med log x1. Vi tilpasser dermed følgende modell i MINITAB:

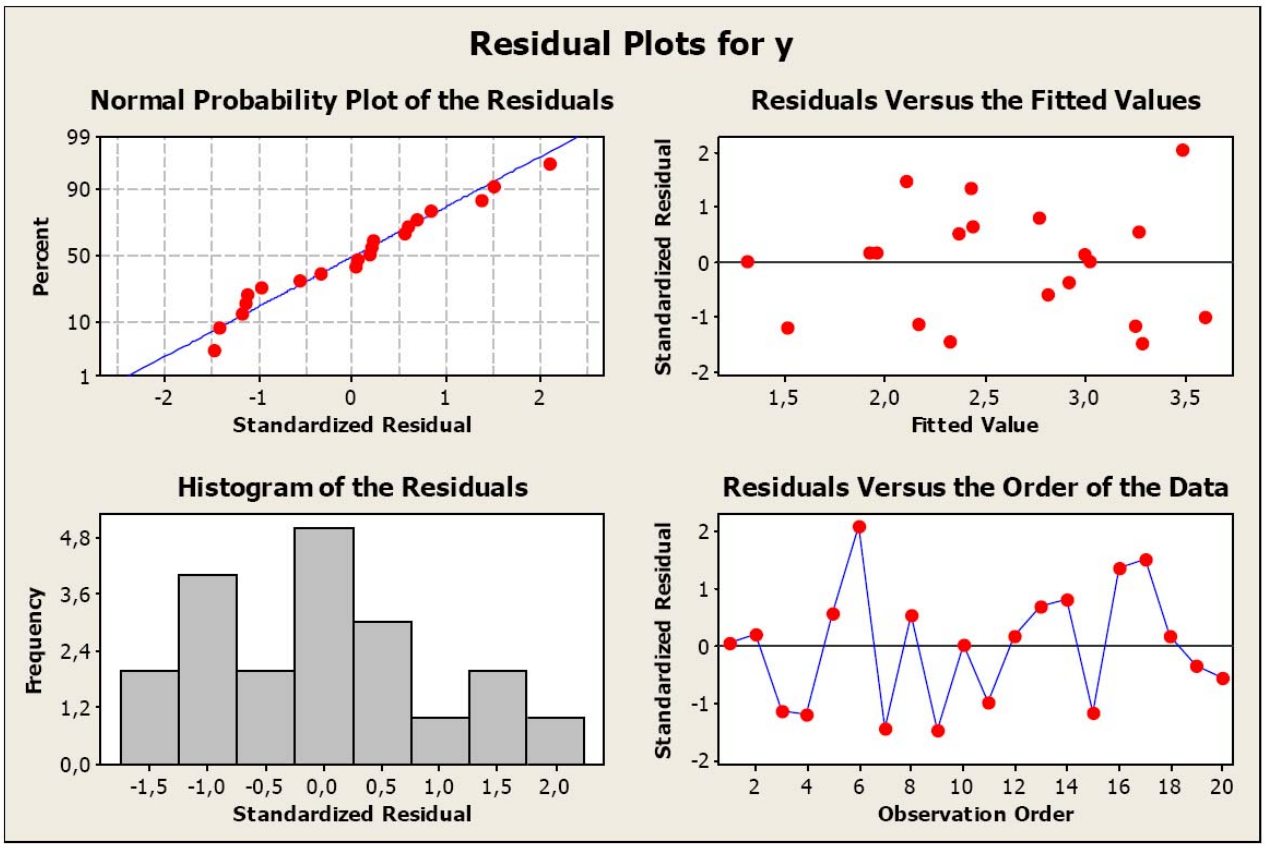
Y = beta0 + beta1 log x1 + beta2 x2 + epsilon

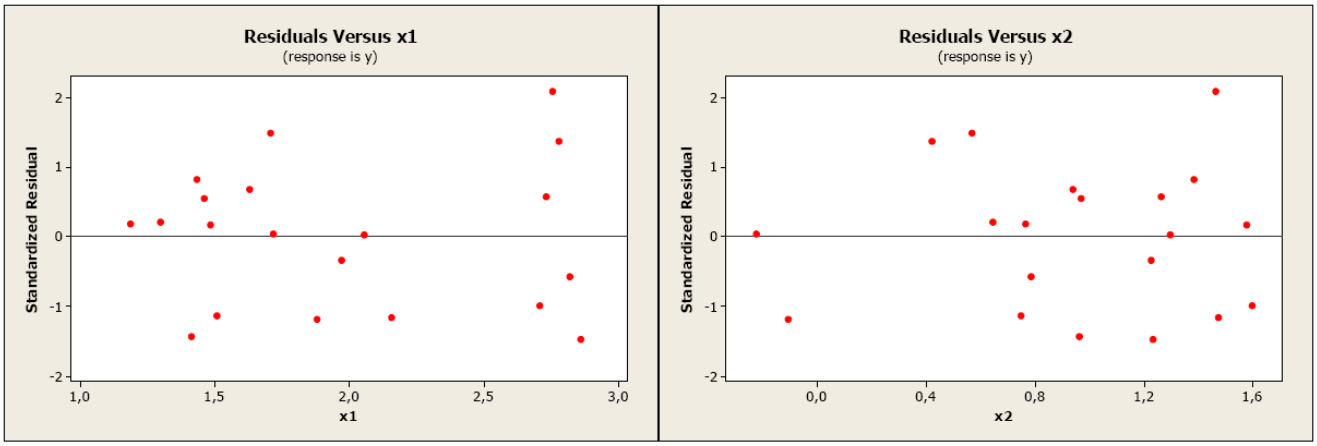
Regression Analysis: y versus log x1; x2

```
The regression equation is
y = 1,02 + 0,967 log x1 + 1,01 x2


Predictor      Coef  SE Coef       T      P
Constant   1,01554  0,00910  111,66  0,000
log x1     0,96748  0,01141   84,76  0,000
x2         1,01266  0,00653  154,97  0,000


S = 0,0143272   R-Sq = 100,0%   R-Sq(adj) = 100,0%
```



Den korrekte modellen er nå tilpasset med MINITAB. Koeffisientene i den estimerte regresjonsfunksjon er nå tilnærmet korrekte (alle beta-ene var 1 i vår simulering av data). Plottene har heller ikke slike systematiske avvik som plottene for den galt spesifiserte modellen.

Example: Multicollinearity

I dette eksemplet skal vi studere virkningen av multikollinearitet.

Vi har simulert 20 observasjoner av X1 som er N(0,1) og latt X2 = X1 + V der V er N(0,0.01).
Korrelasjonskoeffisienten mellom X1 og X2 blir dermed 0.9995. Til slutt har vi simulert

$$Y = 10 + X1 + E$$

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

Data Display

Merk her at kolonnene for x1 og x2 er nesten identiske, og dermed nesten lineart 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 X1, X2 som uavhengige variable)

The regression equation is
y = 9.82 + 5.06 x1 - 3.98 x2

Predictor	Coef	StDev	T	P
Constant	9.81589	0.09196	106.74	0.000
x1	5.057	8.415	0.60	0.556
x2	-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			

Prediksjon

Vi bruker naa modellen til aa predikere Y for et nytt sett av x1,x2, nemlig x1=1, x2=-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.

Matrisen $(X'X)^{-1}$

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 lineaert avhengige.

Best Subsets Regression

Response is y

Vars	R-Sq	R-Sq (adj)	C-p	S	x 1	x 2
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 X1 med i modellen)

The regression equation is
y = 9.82 + 1.08 x1

Predictor	Coef	StDev	T	P
Constant	9.81746	0.08989	109.21	0.000
x1	1.08146	0.09947	10.87	0.000

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

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			

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.