

EKSEMPEL: RESIDUALPLOTT OG TRANSFORMASJON AV PREDIKTOR

Simulering av data

De 20 observasjonene nedenfor er simulert fra modellen

$$y = 1 + \log x_1 + x_2 + \text{epsilon}$$

der epsilon er normalfordelt med forventning 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

$$Y = \text{beta}_0 + \text{beta}_1 x_1 + \text{beta}_2 x_2 + \text{epsilon} \quad (1)$$

er så tilpasset i MINITAB:

Regression Analysis: y versus x1; x2

The regression equation is
 $y = 0,686 + 0,483 x_1 + 1,01 x_2$

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

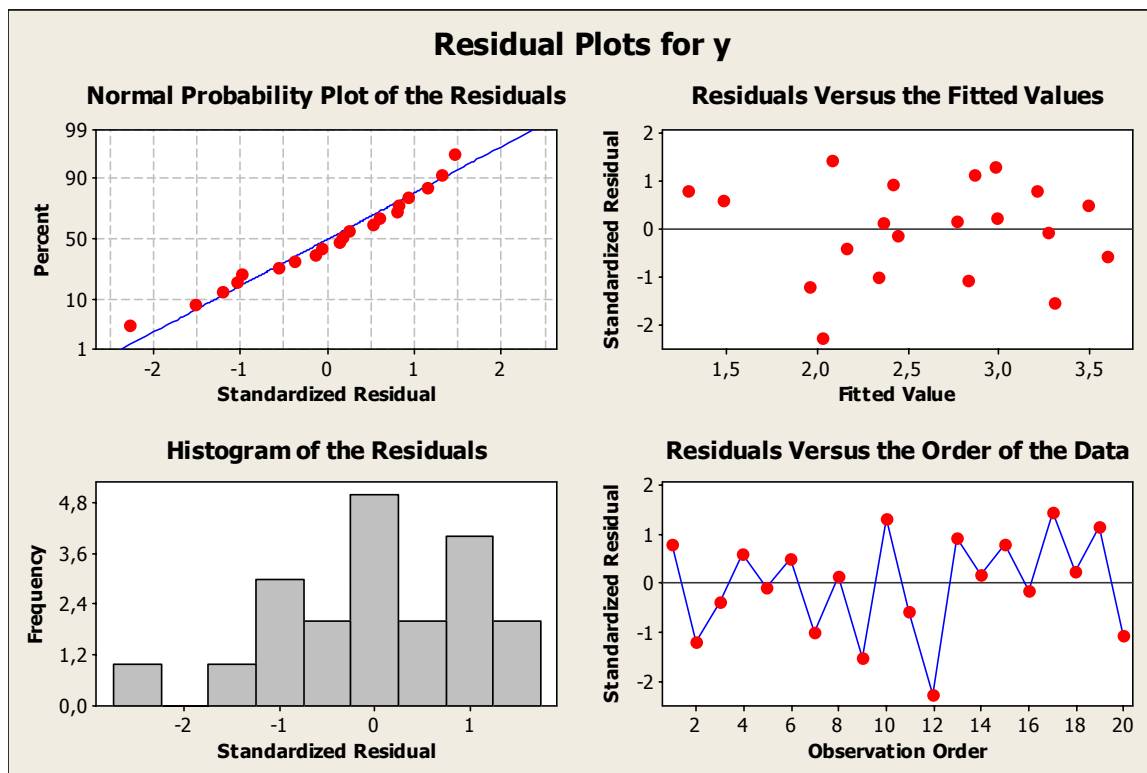
Source	DF	SS	MS	F	P
Regression	2	7,9130	3,9565	3503,14	0,000
Residual Error	17	0,0192	0,0011		
Total	19	7,9322			

Source	DF	Seq SS
x1	1	3,0749
x2	1	4,8381

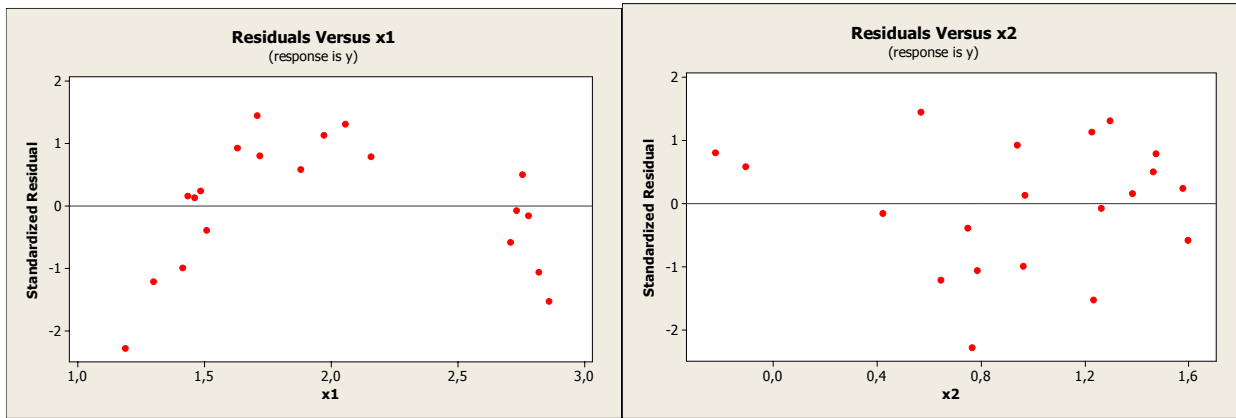
Unusual Observations

Obs	x1	y	Fit	SE Fit	Residual	St Resid
12	1,19	1,95918	2,02984	0,01281	-0,07066	-2,27R

R denotes an observation with a large standardized residual.



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:

$$Y = \text{beta0} + \text{beta1} \log x1 + \text{beta2} x2 + \text{epsilon}$$

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%

Analysis of Variance

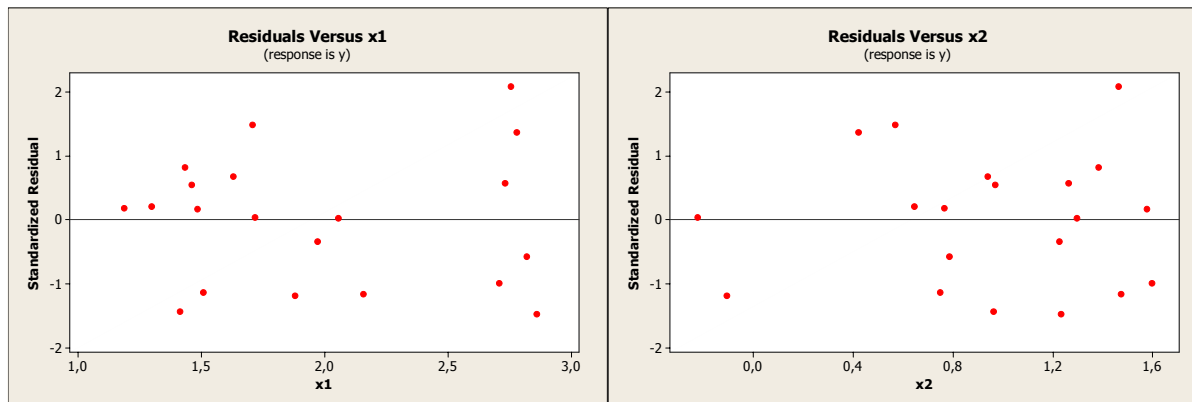
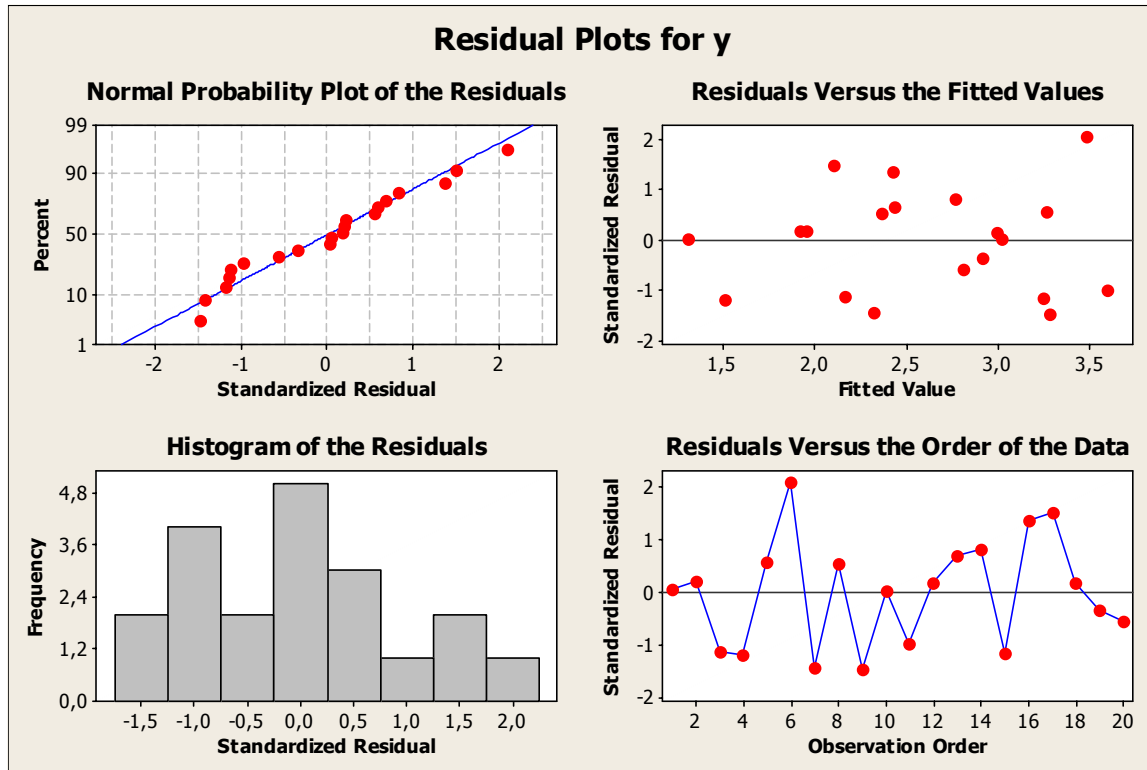
Source	DF	SS	MS	F	P
Regression	2	7,9287	3,9644	19312,92	0,000
Residual Error	17	0,0035	0,0002		
Total	19	7,9322			

Source	DF	Seq SS
log x1	1	2,9988
x2	1	4,9299

Unusual Observations

Obs	log x1	y	Fit	SE Fit	Residual	St Resid
6	1,01	3,50639	3,47904	0,00580	0,02735	2,09R

R denotes an observation with a large standardized residual.



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.