

X6: Vysvětlující proměnná, relativní velikost asimilační plochy (listová plocha/hmotnost sušiny).

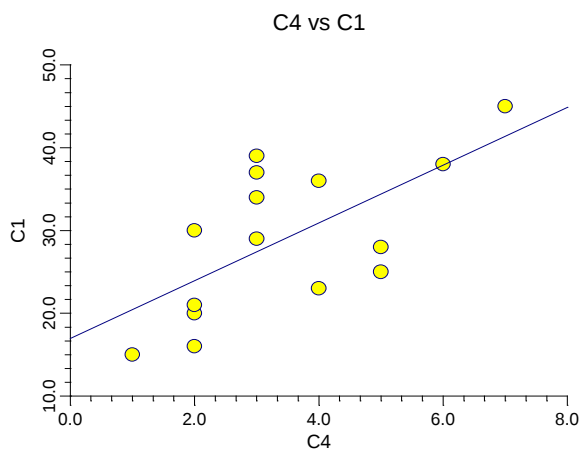
Y4: Vysvětlovaná proměnná, výška rostliny v cm.

Y4	X6
15	1
16	2
20	2
21	2
23	4
25	5
28	5
29	3
30	2
34	3
36	4
37	3
38	6
39	3
45	7

Příklad 16: Vliv odlehklých měření na hodnotu regresního koeficientu.

Použitá data: X6, Y4

Máme dvě proměnné Y4 (kódována jako C1) a X6 (kódována jako C4), jejichž graficky znázorněná závislost vypadá takto.



Jednoduchou lineární regresí zjistíme hodnoty koeficientů a , b a sestavíme regresní rovnici. Koeficient b je významně odlišný od nuly ($p = 0.0085$), ale determinační koeficient je poměrně nízký, $R^2 = 42\%$, neboť relativně velká část variability zůstává tímto regresním modelem nevysvětlená.

Regression Equation Section

Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level	Decision (5%)	Power (5%)
Intercept	16.97987	4.315628	3.9345	0.001711	Reject Ho	0.952580
C4	3.486577	1.126882	3.0940	0.008545	Reject Ho	0.815741
R-Squared	0.424087					

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1	12673.07	12673.07			
Model	1	483.0072	483.0072	9.5729	0.008545	0.815741
Error	13	655.9261	50.45586			
Total(Adjusted)	14	1138.933	81.35238			

Root Mean Square Error	7.103229	R-Squared	0.4241
Mean of Dependent	29.06667	Adj R-Squared	0.3798
Coefficient of Variation	0.2443771	Press Value	831.0445
Sum Press Residuals	100.4082	Press R-Squared	0.2703

Pokud v zadávacím okně regrese (zelená kostka) změníme *Show Residual Report* z *No* na *Yes*, získáme část výsledků označenou *Residual Report*. Ta udává konkrétní hodnoty reziduálů, tj. rozdílů mezi naměřenou a predikovanou hodnotou proměnné Y pro dané X regresním modelem (nebo jinými slovy mezi naměřenou hodnotou a hodnotou ležící na regresní přímce pro dané X). Hodnoty reziduálů si můžeme uložit pro případné pozdější výpočty (zadáme-li toto v dialogovém okně před provedením vlastního výpočtu v oddíle *Data Storage/Residuals* a zadáme příslušný sloupeček, kam chceme hodnoty uložit).

Residual Report

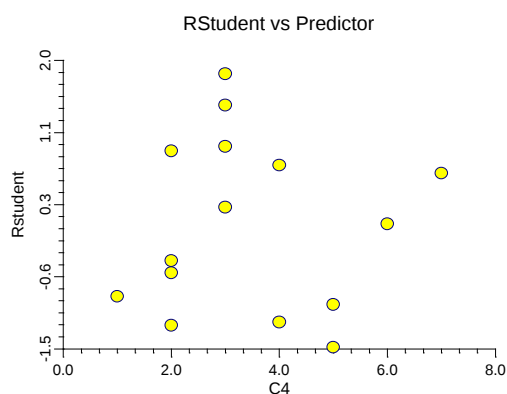
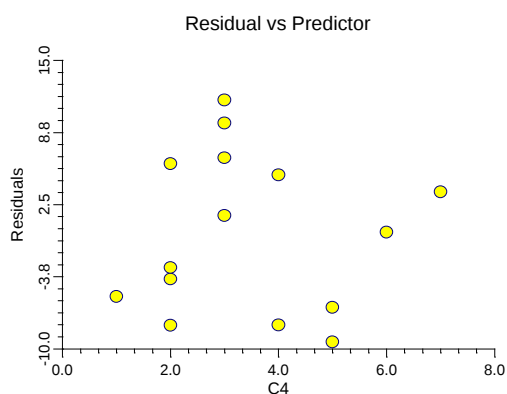
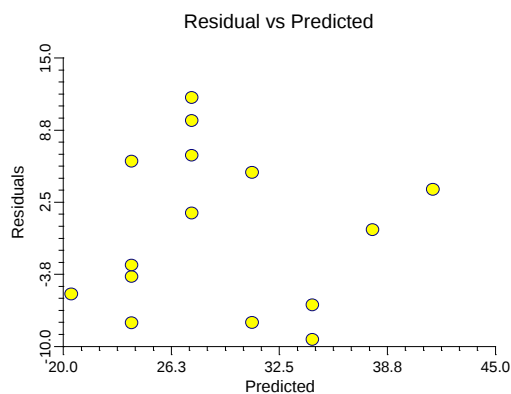
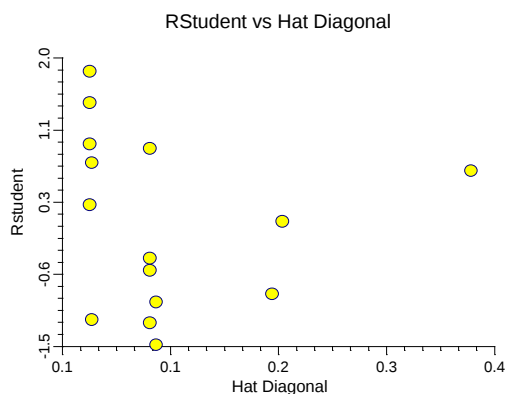
Row	Actual	Predicted	Residual	Percent Error	MSEi
1	15	20.46644	-5.466443	36.44	51.46882
2	16	23.95302	-7.95302	49.71	48.66539
3	20	23.95302	-3.95302	19.77	53.17939
4	21	23.95302	-2.95302	14.06	53.83397
5	23	30.92617	-7.926175	34.46	49.00785
6	25	34.41275	-9.412752	37.65	46.21433
7	28	34.41275	-6.412752	22.90	50.74024
8	29	27.4396	1.560403	5.38	54.44183
9	30	23.95302	6.04698	20.16	51.19466
10	34	27.4396	6.560403	19.30	50.79506
11	36	30.92617	5.073825	14.09	52.3442
12	37	27.4396	9.560403	25.84	46.45148
13	38	37.89933	0.1006711	0.26	54.65942
14	39	27.4396	11.5604	29.64	42.65762
15	45	41.38591	3.614094	8.03	52.90244

Tato část regresní diagnostiky (opět je třeba před výpočtem v *Show Regression-Diagnostics Report* požadovat *Yes*) nám umožní identifikovat případná odlehlá měření, která mohou výrazně ovlivňovat hodnoty stanovených koeficientů, především parametru *b*, a tedy celého regresního modelu. První dva sloupce (*Studentized Residual*, *Rstudent*) hodnotí odlehlost měření ve směru osy y, třetí sloupec (*Hat Diagonal*) hodnotí odlehlost měření ve směru osy x, čtvrtý sloupec (*Cook's D*) pak kombinuje oboje předchozí a udává míru vlivu každého měření na výslednou hodnotu regresního koeficientu *b*; vzorce pro výpočet těchto parametrů jsou uvedeny v nápovědě programu. Obecně, při pohledu na hodnoty, žádné z měření se neodchyluje od ostatních, snad jen 15-té měření se zdá být poněkud odlehlé ve směru osy x, *Rstudent* = 0.38 (viz také grafy níže). Ovšem odlehlost tohoto měření ve směru osy y není nijak výrazná (*Rstudent* = 0.63, což je plně v mezích ostatních hodnot) a tudíž ani jeho vliv na hodnotu regresního koeficientu (*Cook's D*) není výrazně vyšší než u jiných měření. Konstatujeme, že žádné z měření není hodnoceno jako odlehlé, aby ovlivnilo nežádoucím způsobem výsledný regresní model.

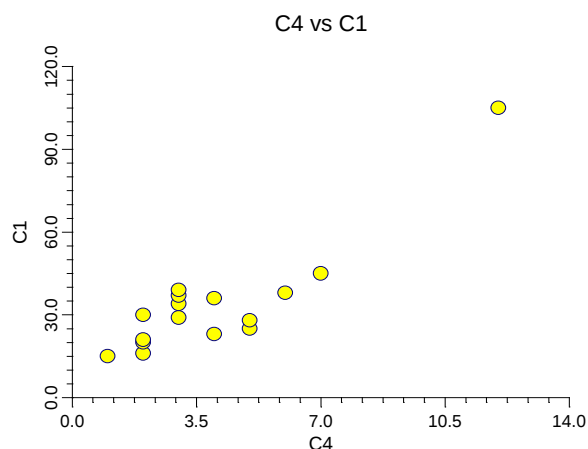
Regression Diagnostics Section

Row	Studentized Residual	Rstudent	Hat Diagonal	Cook's D	Dffits	Covratio
1	-0.871256	-0.862640	0.219799	0.106925	-0.457866	1.333701
2	-1.194081	-1.215848	0.120805	0.097958	-0.450692	1.058113
3	-0.593514	-0.578116	0.120805	0.024201	-0.214297	1.263509
4	-0.443372	-0.429236	0.120805	0.013505	-0.159110	1.294806
5	-1.159475	-1.176480	0.073826	0.053580	-0.332155	1.018627
6	-1.417311	-1.480923	0.125839	0.144585	-0.561881	0.959707
7	-0.965590	-0.962881	0.125839	0.067109	-0.365329	1.156886
8	0.228056	0.219549	0.072148	0.002022	0.061221	1.254768
9	0.907905	0.901330	0.120805	0.056631	0.334106	1.170957
10	0.958816	0.955609	0.072148	0.035742	0.266472	1.092297
11	0.742221	0.728710	0.073826	0.021956	0.205737	1.162040
12	1.397272	1.456254	0.072148	0.075906	0.406077	0.913476
13	0.016132	0.015499	0.228188	0.000038	0.008428	1.520531
14	1.689576	1.837533	0.072148	0.110986	0.512397	0.770356
15	0.646627	0.631497	0.380872	0.128611	0.495303	1.775612

Plots Section



Vliv odlehlých hodnot na parametry regresního modelu demonstruje dodatečné měření ($x = 12$, $y = 105$), které je velmi výrazně odlehlé ve směrou obou os, viz grafická závislost.



Hodnoty koeficientů získaných jednoduchou lineární regresí se znatelně změní. Podstatně se zvýší signifikance stanovení regresního koeficientu b a zdvojnásobí se hodnota determinačního koeficientu ($R^2 = 79\%$) – tedy celkový regresní model se oproti předchozímu případu znatelně zlepšil.

Multiple Regression Report

Regression Equation Section

Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level	Decision (5%)	Power (5%)
Intercept	6.18287	4.575707	1.3512	0.198052	Accept Ho	0.242492

C4	6.907407	0.9593289	7.2002	0.000005	Reject Ho	0.999999
R-Squared	0.787375					

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1	18292.56	18292.56			
Model	1	5152.926	5152.926	51.8436	0.000005	0.999999
Error	14	1391.512	99.39368			
Total(Adjusted)	15	6544.438	436.2958			
Root Mean Square Error		9.969638	R-Squared	0.7874		
Mean of Dependent		33.8125	Adj R-Squared	0.7722		
Coefficient of Variation		0.2948507	Press Value	3485.558		
Sum Press Residuals		165.2449	Press R-Squared	0.4674		

Přestože se jedná o velmi odlehlé měření, vlastní hodnota příslušného reziduálu (měření č. 16) se vůči dalším reziduálům moc neliší a nijak výrazně nenaznačuje extremitu tohoto měření.

Residual Report

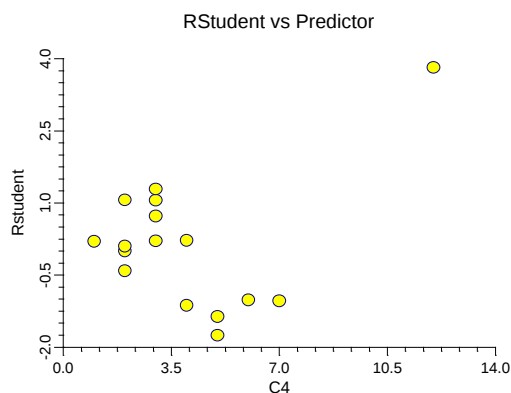
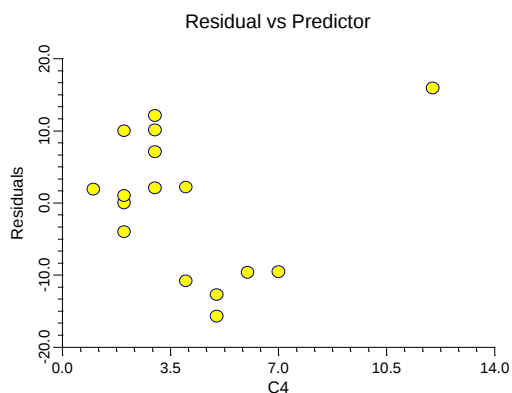
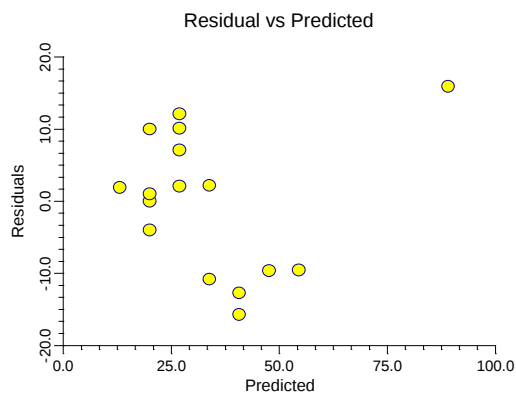
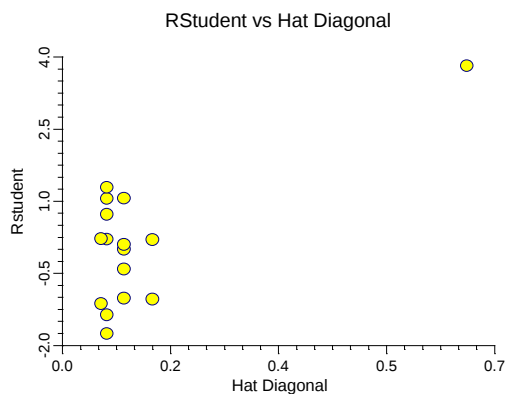
Row	Actual	Predicted	Residual	Percent Error	MSEi
1	15	13.09028	1.909722	12.73	106.7109
2	16	19.99768	-3.997685	24.99	105.6741
3	20	19.99768	2.314815E-03	0.01	107.0394
4	21	19.99768	1.002315	4.77	106.9535
5	23	33.8125	-10.8125	47.01	97.44672
6	25	40.71991	-15.71991	62.88	86.56095
7	28	40.71991	-12.71991	45.43	93.63136
8	29	26.90509	2.094908	7.22	106.6757
9	30	19.99768	10.00231	33.34	98.49278
10	34	26.90509	7.094907	20.87	102.8679
11	36	33.8125	2.1875	6.08	106.6467
12	37	26.90509	10.09491	27.28	98.59433
13	38	47.62732	-9.627315	25.34	99.12161
14	39	26.90509	12.09491	31.01	94.9166
15	45	54.53472	-9.534722	21.19	98.85225
16	105	89.07176	15.92824	15.17	50.45586

Inspekci hodnot v části *Regression Diagnostics* však vidíme, že přidané 16-té měření je velmi výrazně odlehlé od ostatních ve směru obou os. Hodnota *Cook's D* je vyšší než 7, přičemž hodnoty vyšší než 1 jsou považovány za výrazně ovlivňující regresní koeficient (viz nápověda programu). Tato jediná, avšak extrémní hodnota "fixuje" regresní přímku v levé horní části grafu a díky tomu se markantně zvýšila výpovědní síla celého regresního modelu (ze 42% vysvětlené variability na 79%). Odlehlost posledního měření je evidentní také z prohlídky grafického zobrazení příslušných parametrů (především *Rstudent vs. Hat Diagonal*).

Regression Diagnostics Section

Row	Studentized Residual	Rstudent	Hat Diagonal	Cook's D	Dffits	Covratio
1	0.207262	0.200030	0.145833	0.003667	0.082652	1.349452
2	-0.422568	-0.409818	0.099537	0.009869	-0.136254	1.255318
3	0.000245	0.000236	0.099537	0.000000	0.000078	1.287963
4	0.105948	0.102135	0.099537	0.000620	0.033957	1.285899
5	-1.120111	-1.131245	0.062500	0.041822	-0.292086	1.025287
6	-1.636591	-1.753714	0.071759	0.103530	-0.487604	0.817082
7	-1.324263	-1.364404	0.071759	0.067785	-0.379360	0.956014
8	0.218100	0.210524	0.071759	0.001839	0.058534	1.240945
9	1.057276	1.062100	0.099537	0.061782	0.353122	1.090499
10	0.738647	0.726067	0.071759	0.021089	0.201876	1.153935
11	0.226612	0.218770	0.062500	0.001712	0.056486	1.228022
12	1.050976	1.055227	0.071759	0.042695	0.293396	1.060048
13	-1.017637	-1.019033	0.099537	0.057237	-0.338803	1.104468
14	1.259195	1.288550	0.071759	0.061288	0.358269	0.982440
15	-1.034802	-1.037632	0.145833	0.091411	-0.428746	1.158012
16	2.720428	3.818218	0.655093	7.028206	5.262121	0.747142

Plots Section



Je vidět, že jediné měření může mít na celkový regresní model velmi podstatný vliv. Toto odlehlé měření může být zcela v pořádku s podstatou studované závislosti. Přesto by bylo více než vhodné jej před učiněním definitivního závěru a přijetím regresního modelu ověřit. Mohlo například dojít k banální chybě při zadávání dat do počítače (k našemu štěstí), nebo k chybnému odečtu hodnot v terénu (což často lze opakovaným měřením zjistit).