X4, X5, X6: Vysvětlující proměnné, např. počet dní od vytvoření prvního listu, relativní příkon radiace (suma záření od předchozího měření), relativní velikost asimilační plochy (listová plocha/hmotnost sušiny).
Y4: Vysvětlovaná proměnná, např. výška rostliny v cm.

Y4	X6	X5	X4
15	1	1	1
16	2	1	2
20	2	2	3
21	2	1	4
23	4	3	5
25	5	5	6
28	5	5	8
29	3	2	9
30	2	0	12
34	3	1	13
36	4	2	14
37	3	0	15
38	6	1	16
39	3	0	17
45	7	2	20

Příklad 18: Mnohonásobná lineární regrese.

Použitá data: Y4 (kódována jako C1), X4 (C2), X5 (C3), X6 (C4)

V případě, že máme několik vysvětlujících proměnných a jednu proměnnou vysvětlovanou, můžeme jejich vztah analyzovat za použití mnohonásobné lineární regrese. Cílem této regrese je stanovit hodnoty parciálních regresních koeficientů b_i, pro každou vysvětlující proměnnou jeden, které určují podobu regresní rovnice. Mnohonásobná lineární regrese se v NCSS opět provádí v **Analysis/Regression/Correlation/Multiple Regression**.

Výpočtem mnohonásobné regrese uvedených proměnných zjistíme, že v případě X4 a X5 je parciální regresní koeficient významně odlišný od nuly (p = 0, p = 0.042), ne však již v případě X6 (p = 0.90), tj. není průkazná závislost mezi X6 a Y4. Podle vypočítaných parciálních regresním koeficientů by tedy regresní rovnice měla podobu $\mathbf{y = 13.3 + 1.53x_4 + 0.62x_5 - 0.0374x_6}$, ovšem s tím, že vliv proměnné X6 je zřejmě minimální. ANOVA pro výsledný regresní model je nicméně vysoce průkazná, přičemž model vysvětluje zhruba 99% variability.

Multiple Regression Report

Regression Equation Section

Independent Variable	Regression Coefficient	Standard Error	T-Value (Ho: B=0)	Prob Level	Decision (5%)	Power (5%)
Intercept	13.34089	0.5466015	24.4070	0.000000	Reject Ho	1.000000
C2	1.529524	7.608472E-02	20.1029	0.000000	Reject Ho	1.000000
C3	0.6173783	0.2682373	2.3016	0.041913	Reject Ho	0.555238
C4	-3.742499E-02	0.2986096	-0.1253	0.902523	Accept Ho	0.051511
R-Squared	0.993393					

Regression Coefficient Section

Independent Variable	Regression Coefficient	Standard Error	Lower 95% C.L.	Upper 95% C.L.	Standardized Coefficient
Intercept	13.34089	0.5466015	12.13782	14.54395	0.0000
C2	1.529524	7.608472E-02	1.362063	1.696985	1.0268
C3	0.6173783	0.2682373	2.699205E-02	1.207765	0.1081
C4	-3.742499E-02	0.2986096	-0.6946602	0.6198102	-0.0070
T-Critical	2.200985				

Analysis of Variance Section

Source	DF	Sum of Squares	Mean Square	F-Ratio	Prob Level	Power (5%)
Intercept	1	12673.07	12673.07			
Model	3	1131.409	377.1362	551.3236	0.000000	1.000000
Error	11	7.524617	0.6840561			

Total(Adjusted)	14	1138.933	81.35238	
Root Mean Square Error		0.8270769	R-Squared	0.9934
Mean of Dependent		29.06667	Adj R-Squared	0.9916
Coefficient of Variation		2.845448E-02	Press Value	12.69629
Sum |Press Residuals|		11.5552	Press R-Squared	0.9889

Problémem mnohonásobné regrese je ale možnost, že vysvětlující proměnné jsou mezi sebou navzájem závislé. Pokud tomu tak opravdu je, jednotlivé nezávislé proměnné postihují (částečně nebo úplně) stejnou část variability závisle proměnné. Může se nakonec stát, že vysvětlující potenciál jedné nezávisle proměnné je obsažen v jiné nebo v kombinaci jiných nezávislých proměnných X a tato proměnná je tedy pro daný regresní model zcela nadbytečná. V našem případě je velmi pravděpodobné, že proměnná X6, jejíž parciální koeficient není průkazně odlišný od nuly, není pro regresní model nezbytná.

Oddíl výsledků označený jako *Multicollinearity Section* nás informuje o tom, zda-li mezi vysvětlujícími proměnnými X existuje lineární závislost. Hodnota R^2 vs ostatním X (*R-Squared Vs Other X's*) udává podíl variability dané nezávislé proměnné vysvětlené kombinací zbývajících nezávislých proměnných (tj. provádí se mnohonásobná lineární regrese jednotlivých proměnných X na zbývajících), *Variance Inflation* (*Factor*) je převrácená hodnota *Tolerance* (přičemž *Tolerance* $= 1 - R^2$ *vs ostatním X*). Čím vyšší hodnota *Variance Inflation* a R^2 *vs ostatním X*, tím více je daná proměnná X korelována s ostatními a tím nižší je tedy její vysvětlující síla (z toho následně vyplývá, že tím méně je potřebná pro vytvoření regresního modelu). Výsledky ukazují, že nejmenší vysvětlující sílu má opravdu proměnná X6 (C4).

Multicollinearity Section

Independent Variable	Variance Inflation	R-Squared Vs Other X's	Tolerance	Diagonal of X'X Inverse
C2	4.344128	0.769804	0.230196	8.462587E-03
C3	3.674401	0.727847	0.272153	0.1051832
C4	5.179296	0.806924	0.193076	0.1303514

V části *Eigenvalues of Centered Correlations* získáme obdobnou informaci o vysvětlující síle jednotlivých nezávislých proměnných X. Čím vyšší je hodnota *Eigenvalue* (v podstatě míra variance), tím větší je vysvětlující síla dané proměnné. V dalších dvou sloupečcích je pak tato hodnota vyjádřená procentuálně jako příspěvek k celkové variabilitě. Opět je vidět, že nejnižší vysvětlující sílu má proměnná X6 (řádek č. 3).

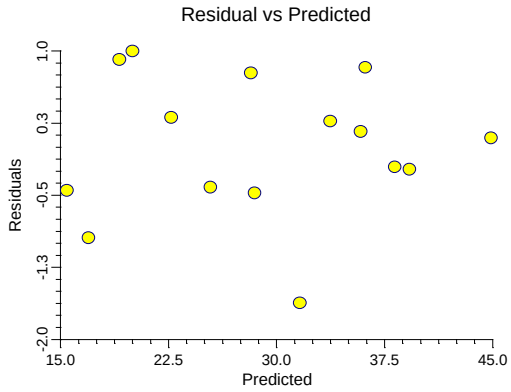
Eigenvalues of Centered Correlations

No.	Eigenvalue	Incremental Percent	Cumulative Percent	Condition Number
1	1.637380	54.58	54.58	1.00
2	1.277907	42.60	97.18	1.28
3	0.084713	2.82	100.00	19.33

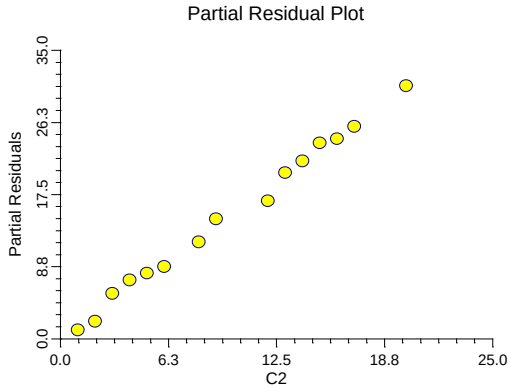
All Condition Numbers less than 100. Multicollinearity is NOT a problem.

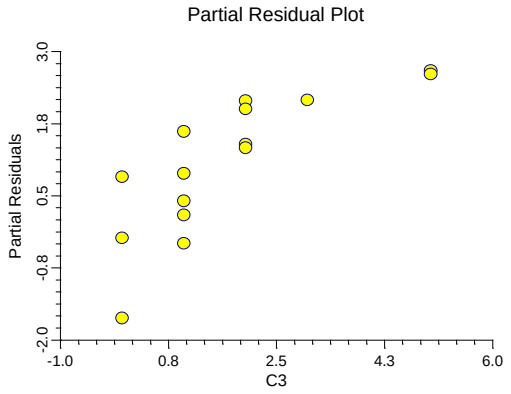
Rozložení reziduálů ukazuje, že použitý lineární model je v pořádku a odpovídá závislosti studovaných proměnných.

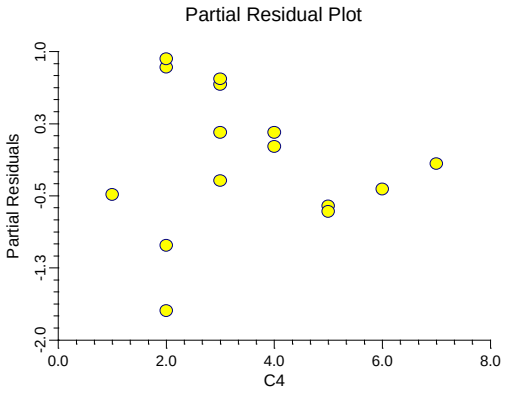
Plots Section



Graficky se o významnosti jednotlivých nezávislých proměnných pro celkový regresní model přesvědčíme také v části *Partial Residual Plot*. Zde jsou vyneseny hodnoty tzv. parciálních reziduálů jednotlivých proměnných na sebe samé – parciální reziduály dané nezávisle proměnné se získají při "odfiltrování" vlivu zbývajících nezávisle proměnných. Čím více se tato závistlost blíží lineární, tím větší je vysvětlující síla dané proměnné. Zatímco lineární závislost je evidentní u proměnné X4 a poměrně zřetelná u X5, rozdělení reziduálů X6 je více méně náhodné, což ukazuje na minimální vysvětlující sílu této proměnné (zcela v souladu s výsledky v předchozích oddílech).







Na základě dosavadních poznatků před námi stojí otázka, jestli je nezbytně nutné v regresním modelu zahrnout všechny tři nezávisle proměnné X. Volba *All possible regressions* nám spočítá regresní modely, resp. příslušné determinační koeficienty R^2, závisle proměnné Y na všech možných kombinacích nezávisle proměnných X. Smyslem je porovnat sílu regresních modelů při zahrnutí různých nezávislých proměnných a následně vybrat model nejvhodnější (tzn. model vysvětlující nejvíce variability závisle proměnné Y). Jednoznačně největší individuální vysvětlující sílu má X4 (C2, A; $R^2 = 0.98$), za ní následuje X6 (C4, C; $R^2 = 0.42$) a na závěr je X5 (C3, B; $R^2 = 0.04$) s minimálním vysvětlujícím potenciálem. V kombinaci dvou nezávisle proměnných však největší sílu vykazuje model zahrnující X4 a X5 ($R^2 = 0.993$), což je přece jen o

něco málo více než model s X4 a X6 ($R^2 = 0.990$). Model se všemi třemi proměnnými pak vysvětluje jen nepatrně více variability ($R^2 = 0.99339$) než model s X4 a X5.

All Possible Regression Report

All Possible Results Section

Model Size	R-Squared	Root MSE	Cp	Model
1	0.983609	1.198338	16.290425	A (C2)
1	0.424087	7.103229	947.877752	C (C4)
1	0.035786	9.191031	1594.388174	B (C3)
2	0.993384	0.792431	2.015708	AB
2	0.990212	0.9638616	7.297421	AC
2	0.750670	4.864583	406.126859	BC
3	0.993393	0.8270769	4.000000	ABC

Nyní stojí otázka takto, je postačující regresní model s proměnnými X4 a X5, nebo zahrnutím proměnné X6 získá regresní model signifikantně vyšší vysvětlující sílu? Odpověď se získá testováním přírůstku ΔR^2 vůči původní hodnotě R^2. Tuto proceduru počítáme v oddíle *Stepwise Regression* (postupná regrese). Můžeme zvolit několik možností výběru proměnných, *stepwise* (předdefinovaná možnost), *forward* nebo *backward*, přičemž *stepwise* mnohonásobná regrese je zřejmě nejčastější volbou. Výsledek této procedury ukazuje, že přidáním X6 k proměnným X4 a X5 již výrazně sílu celého regresního modelu nezvýšíme. Závěrem tedy spočítáme mnohonásobnou regresi při zahrnutí právě těchto dvou proměnných, X4 a X5, čímž získáme příslušné parciální regresní koeficienty.

Stepwise Regression Report

Iteration Detail Section

Iter. No.	Action	Variable	R-Squared	Sqrt(MSE)	Max R-Squared Other X's
0	Unchanged		0.000000	9.019555	0.000000
1	Added	C2	0.983609	1.198338	0.000000
2	Unchanged		0.983609	1.198338	0.000000
3	Added	C3	0.993384	0.792431	0.081943
4	Unchanged		0.993384	0.792431	0.081943

Poznámka závěrem: Může se zdát poněkud paradoxní, že proměnná X6, která sama o sobě vysvětluje cca 42% variability vzhledem k Y4 (viz výše *All possible regressions* nebo Příklad 16), nakonec není pro výsledný regresní model významná. Naopak X5, která sama vysvětluje jen zhruba 4%, je nakonec do regrese zahrnuta. Tato skutečnost vyplývá ze vztahu mezi X4 a X6, který je více méně lineární (viz graf dole). Proměnná X4 již v sobě zahrnuje téměř veškerou vysvětlující informaci (vzhledem k závisle proměnné Y4), která je obsažena v X6. Proto za přítomnosti X4 v regresním modelu se přidáním X6 jeho celková síla prakticky nezvýší a tato vysvětlující proměnná je nadbytečná.