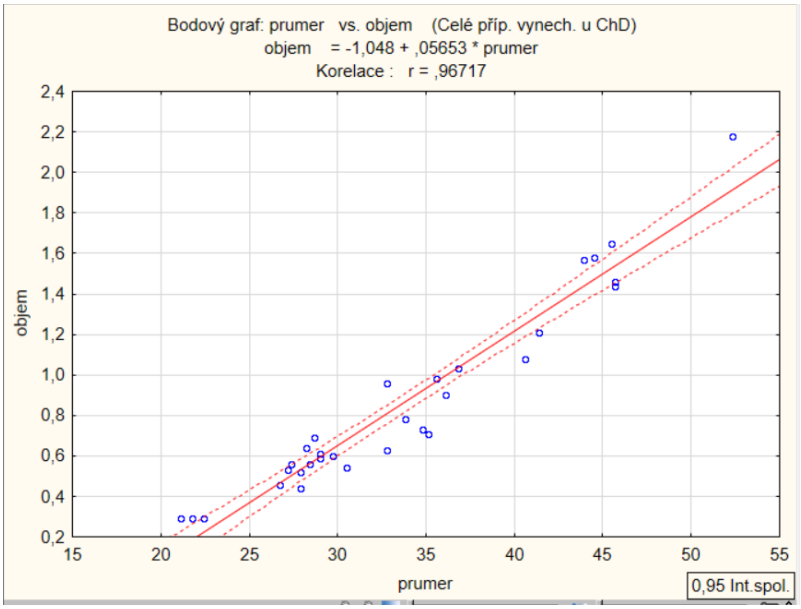
Lineární regrese – řešení.

Datový soubor TREŠNĚ. Máme údaje o 31 třešňových stromech: průměr kmene v prsní výšce [cm] a odhad objemu dřevní hmoty [m^3]. Hledáme model, který by popsal lineární závislost objemu dřevní hmoty na průměru kmene.

Tento problém **řeší lineární regresní analýza**: mám dvě kvantitativní proměnné a chci popsat, jak hodnoty průměru kmene mohou předpovídat hodnoty objemu dřevní hmoty.

Rovnice modelu: **OBJEM = $\beta_0 + \beta_1$*PRŮMĚR + E**
Regresní analýzou odhaduji hodnotu regresních koeficientů β_0 a β_1.

Situace graficky: body jsou dvojice měření na jednom stromě [x = průměr, y = objem], plná čára je hledaný lineární model, přerušované čáry jsou konfidenční intervaly odhadů středních hodnot objemu pro všechny hodnoty průměrů kmene ze zobrazeného intervalu (zde cca 22 až 55 cm).
V tomto bodovém grafu vidíme, že závislost mezi hodnotami existuje a je poměrně těsná.



Zadání ve STATISTICA:

a) Grafy → Bodový graf
b) Statistiky → Základní statistiky → Korelační matice → 2 seznamy → *Základní výsledky*: Grafy nebo *Detailní výsledky*: 2D bodové grafy.

Výsledky analýzy:

Výsledky regrese se závislou proměnnou : objem (tresne)
R= ,96717330 R2= ,93542419 Upravené R2= ,93319743
F(1,29)=420,08 p<,00000 Směrod. chyba odhadu : ,12038

N=31	b*	Sm.chyba z b*	b	Sm.chyba z b	t(29)	p-hodn.
Abs.člen			-1,04782	0,095315	-10,9932	0,000000
prumer	0,967173	0,047188	0,05653	0,002758	20,4960	0,000000

Zadání ve STATISTICA:

Statistiky → Vícenásobná regrese → *zadat proměnné (nepopleťte závislou a nezávislou!)* → OK → *Základní výsledky*: „Výpočet: výsledky regrese“.

Rovnice modelu s odhadem koeficientů (sloupeček **b**)
OBJEM = -1,048 + 0,0565*PRŮMĚR + E

Jsou **regresní koeficienty** (alias parametry) rovnice statisticky významné? → Pomocí t-testu testujeme hypotézu, že skutečná hodnota koeficientu je nula, **H_0: $\beta_1 = 0$**. Totéž pro **β_0**, ale pro hodnotu tohoto koeficientu většinou nemáme smysluplnou interpretaci, alespoň v biologii. V tomto příkladu zamítáme hypotézu o nulovosti regresního koeficientu, testová statistika = 20,496, p-hodnota < 0,001 (poslední dva sloupečky). Znamená to, že sklon regresní přímky je průkazně nenulový, že existuje (statistická) závislost mezi průměrem kmene a jeho objemem, zamítáme možnost nezávislosti průměru a objemu.

Významnost celého modelu: hodnota F(1,29) v záhlaví tabulky (třetí řádek). Je to testová statistika k testu hypotézy, že variabilita vysvětlená modelem je nulová. Testujeme F-testem, tedy porovnáváme variabilitu (odhad rozptylu) reziduálů předpovězených hodnot (tj. předpověď objemu minus průměrný objem) a

variabilitu reziduálů v modelu (tj. naměřený objem minus předpovězený objem). Odhady těchto rozptylů jsou dobře vidět v tabulce ve sloupci „Průměr čtverců“. Tedy rozptyl reziduálů kolem modelové přímky je malý (= 0,0145), model funguje dobře; rozptyl reziduálů předpovězených hodnot je velký (6,087), to znamená, že modelovat tato data pouhým průměrným objemem by byla chyba, a také to říká, že jsme modelem vysvětlili 6,087 ze 6,507 dílů variability.

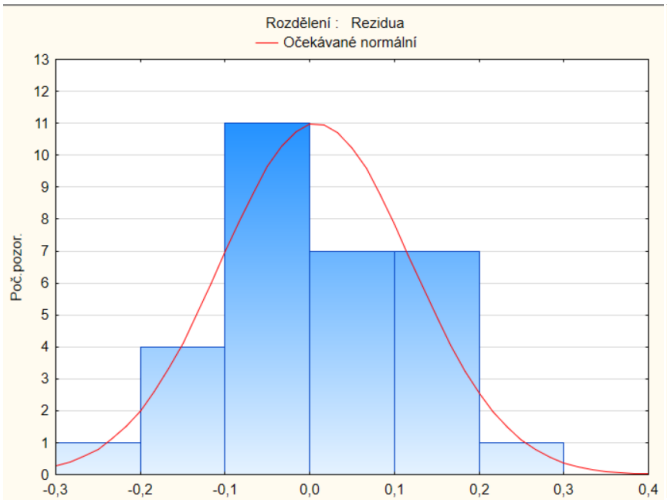
Efekt	Analýza rozptylu (01 tresne) Součet čtverců	sv	Průměr čtverců	F	p-hodn.
Regres.	6,087155	1	6,087155	420,0846	0,000000
Rezid.	0,420219	29	0,014490		
Celk.	6,507374				

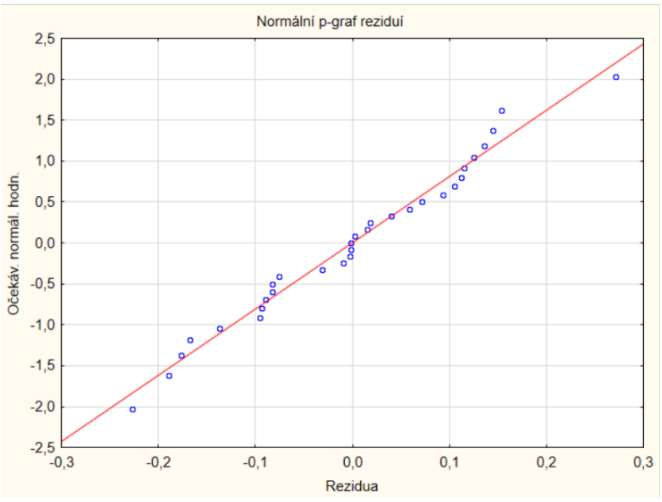
Zadání ve STATISTICA:
Detailní výsledky:
ANOVA (Celk. vhodnost modelu).

<u>Podíl variability vysvětlené modelem</u>: Je to právě těch 6,087 ze 6,507 dílů variability, tedy 93,54 %. Toto číslo označujeme jako koeficient determinace, R^2, a v první výsledkové tabulce ho najdeme v záhlaví na druhém řádku: R2 = 0,9354.
R (bez mocnění) = 0,9672 je korelační koeficient (platí ale jen v jednoduché lineární regresi).
Upravené R2 = 0,9332 používáme, když máme více vysvětlujících proměnných nebo když máme jen málo pozorování.

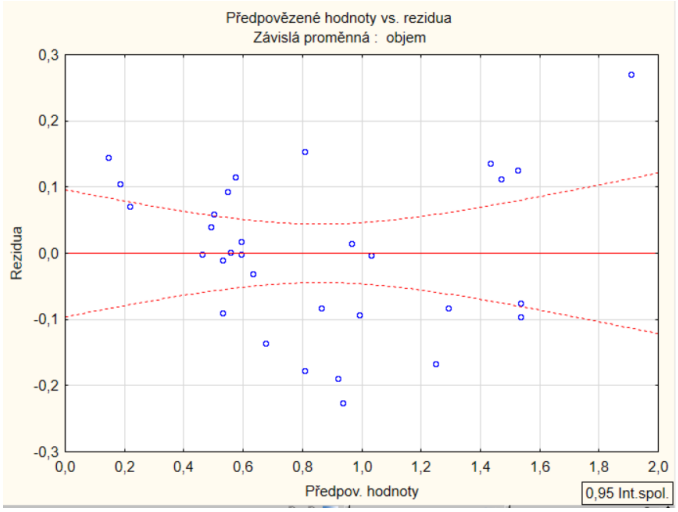
<u>Kontrola předpokladů</u> (záložka *Rezidua/předpoklady/předpovědi* → Reziduální analýza. Dále je to podrobně v přednáškových slidech):

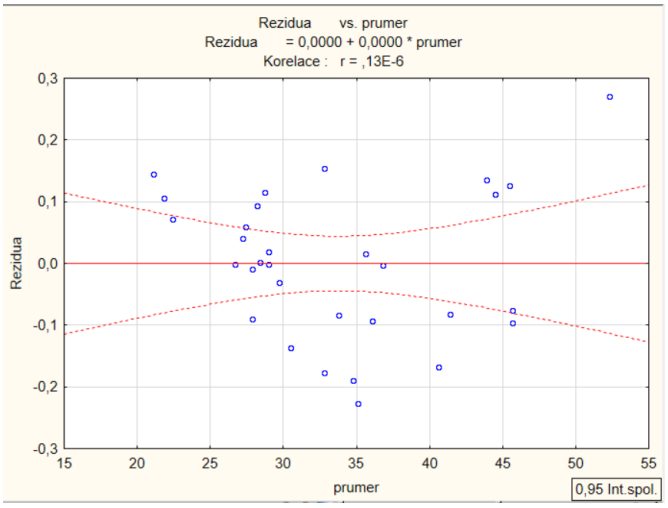
1) **Rezidua modelu mají normální rozdělení** – splněno.





2) **Rozptyl těchto reziduí se nemění s hodnotou nezávislé (vysvětlující) proměnné.** Body jsou uspořádány do jakési misky → rýsuje se zde kvadravická závislost, dalo by se tedy zkoušet do modelu přidat člen + β_2*PRŮMĚR^2. Totéž můžeme usuzovat i z dalšího grafu.





3) **Střední hodnota závislé proměnné (EY) je lineární funkcí nezávisle proměnné.** Jinými slovy: jestliže v našem modelu chybí nějaký další vysvětlující člen (např. výška stromu nebo průměr^2), budou naše předpovědi vychýlené. To se projeví právě na reziduálech – nebudou uspořádány rovnoměrně kolem nuly, ale budou nějak „zahnuté“. V tomto případě právě do tvaru misky, což signalizuje, že ve členu EY je schovaná ještě nějaká „sudá mocnina“. V tomto případě je to skutečně průměr^2. Můžete si vytvořit v datové tabulce sloupeček s napočítanou druhou mocninou průměru a tuto novou proměnnou přidat do modelu ☺ V našem případě modelu s jednou vysvětlující (nezávislou) proměnnou je graf totožný s předchozím grafem (rezidua na průměru).

Celá analýza pak dopadne takto (tohle už nemusíte předvádět u zkoušky!!):

Rozšířený model: OBJEM = $\beta_0 + \beta_1$*PRŮMĚR + β_2*PRŮMĚR^2 + E
Koeficient b_2 pro průměr^2 je průkazně nenulový (t = 4,33, p = 0,00017), ale na samotný průměr už nezbyla žádná práce, nezamítám hypotézu, že $b_1 = 0$. Samotný průměr tedy mohu z modelu vypustit (smazat).

Výsledky regrese se závislou proměnnou : objem (01 tresne)
R= ,98046365 R2= ,96130896 Upravené R2= ,95854532
F(2,28)=347,84 p<0,0000 Směrod. chyba odhadu : ,09483

N=31	b*	Sm.chyba z b*	b	Sm.chyba z b	t(28)	p-hodn.
Abs.člen			0,295997	0,319436	0,92662	0,362041
prumer	-0,390374	0,315855	-0,022819	0,018463	-1,23593	0,226754
prum2	1,367048	0,315855	0,001111	0,000257	4,32809	0,000173

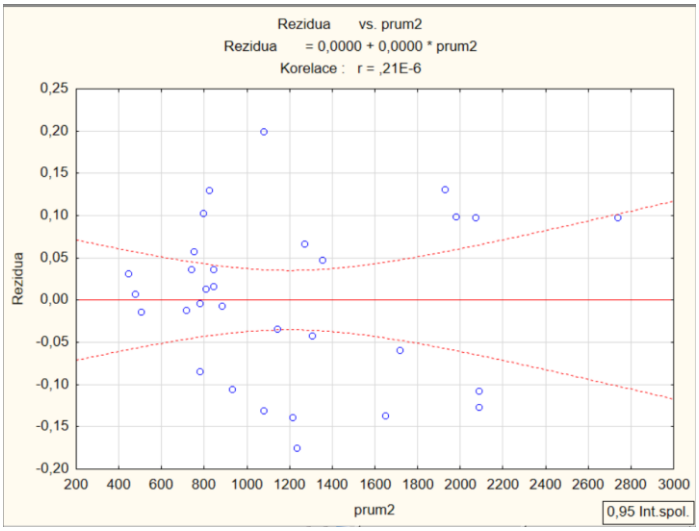
Další verze modelu: OBJEM = $\beta_0 + \beta_2$*PRŮMĚR^2 + E
Nyní oba koeficienty průkazné, hurá! Celý model ještě významnější (F = 681,75, p < 0,001), R2 = 0,959.

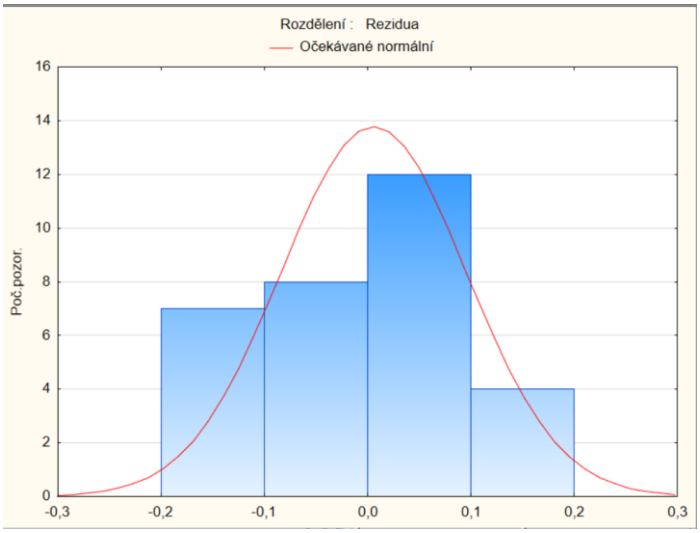
Výsledky regrese se závislou proměnnou : objem (01 tresne)
R= ,97938665 R2= ,95919820 Upravené R2= ,95779124
F(1,29)=681,75 p<0,0000 Směrod. chyba odhadu : ,09568

N=31	b*	Sm.chyba z b*	b	Sm.chyba z b	t(29)	p-hodn.
Abs.člen			-0,095712	0,040258	-2,37750	0,024242
prum2	0,979387	0,037509	0,000796	0,000030	26,11040	0,000000

Nenechejte se zmást malou hodnotou koeficientu $b_2 = 0,000796$, je skutečně průkazně nenulový. Uvažte, že se násobí se čtvercem průměru v centimetrech, což jsou dost velká čísla. Pokud bychom zadali průměr^2 v metrech čtverečních, dostali bychom $b_2 = 7,96$.

Také regresní diagnostika vypadá v pořádku:





Výsledný model tedy je: OBJEM = -0,096 + 0,000796*PRŮMĚR^2.