4. ZÁVISLOST MEZI SPOJITÝMI VELIČINAMI

Naším cílem v dnešní lekci je vyšetřit faktory související s hmotností mladého člověka. Za faktory byly vybrány:

- Věk
- Výška
- Pohlaví
- Obvod pasu
- Obvod hrudníku
- Obvod boků
- Obvod krku
- Velikost bot
- Počet hodin, po který se respondent průměrně věnuje týdně sportu
- Počet hodin, který respondent průměrně denně stráví u PC a TV

Průzkum byl prováděn mezi studenty vybrané střední školy. Soubor Data-celek.xls obsahuje údaje o 110 respondentech.

Obsah kapitoly

4.1 Krátký teoretický úvod

Závislost spojitých veličin se vyšetřuje pomocí dvojice metod, a to regrese a korelace. Úkolem regrese je najít vhodný funkční model této závislosti. Úkolem korelace je změřit sílu lineární závislosti. Jinými slovy, regrese popisuje daný vztah a korelace zjišťuje jeho těsnost. Známe dva základní typy regresní analýzy, a to jednoduchou a vícenásobnou.

Cílem jednoduché (simple) regrese je najít model funkční závislosti (spojité) veličiny Y na jedné (spojité) veličině (na tzv. regresoru) X. Tvar funkce často napoví bodový graf dat. Příkladem může být zkoumání závislosti mezi platem a výší úspor či mezi výší exportu a výší HDP.

Cílem vícenásobné (multiple) regrese je najít model funkční závislosti (spojité) veličiny Y na více (spojitých) veličinách (regresorech). Příkladem může být zkoumání závislosti výše úspor na platu, výdajích za potraviny, výdajích za spotřební zboží a výdajích za kulturu.

Při jednoduché regresi můžeme hledat modely různých typů. Mezi nejvíce používané patří:

Lineární model: $y = b_1 x + b_0$ (1)

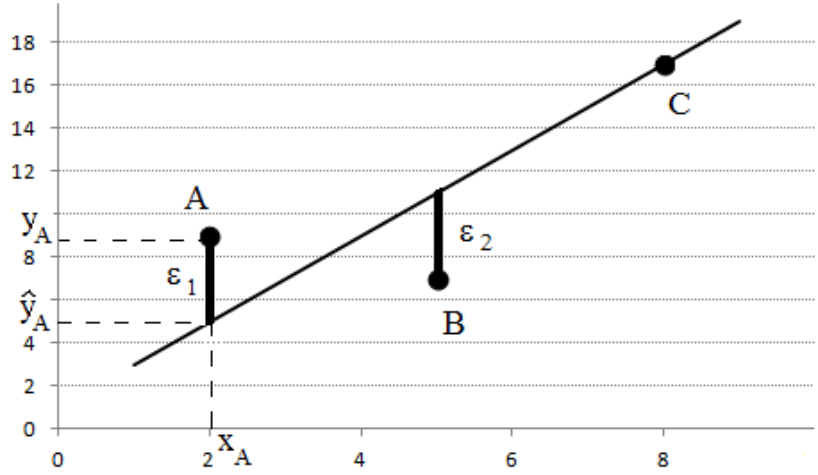
Polynomický model: $y = \sum_{i=0}^{n} b_i x^i$ (2)

Mocninný model: $y = b_0 x^{b_1}$ (3)

Logaritmický model: $y = b_1 \ln x + b_0$ (4)

Hledáním regresního modelu, resp. regresní funkce rozumíme hledání regresních koeficientů b_i, přičemž typ regresního modelu musíme stanovit sami (na základě zkušeností a vzhledu bodových grafů), hodnoty jednotlivých regresních koeficientů pak nalezne metoda (např. implementovaná v SW).

Blíže se nyní seznámíme s nejjednodušším a nejčastěji využívaném typu, a to lineárním modelem, v němž hledáme funkci, jejímž grafem je přímka, viz obr.



Snažili jsme se „proložit“ tři body A, B a C regresní přímkou. Hledáme funkci (přímku) ve tvaru $\hat{y} = b_1 x + b_0$. Vidíme, že platí $y = b_1 x + b_0 + \varepsilon$, neboli že naměřené hodnoty se „o něco“ liší od vypočítaných hodnot odhadu. Tomuto rozdílu říkáme reziduum, značíme ε. V regresním modelu by měla mít rezidua normální rozdělení se střední hodnotou 0.

Hodnoty b_1 a b_0 odhadujeme pomocí Metody nejmenších čtverců, která je založena na principu hledání minima funkce více proměnných pomocí parciálních derivací. Z této metody je možno získat následující vzorce.

$$b_1 = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \tag{5}$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} \tag{6}$$

Směrnice přímky b_1 odpovídá změně závislé proměnné při nárůstu nezávislé proměnné o jednu jednotku.

Výpočty regresních koeficientů pomocí uvedených vzorců jsou poněkud pracné. Proto se většinou v praxi využívá různých pomocníků. V případě lineárních modelů můžeme výpočty provést na kalkulačkách, a to pomocí speciálních funkcí. Ještě efektivnější je využití různých SW, například i velmi rozšířeného Microsoft Excelu, nebo mnoha komerčních statistických SW, jako je STATISTICA či SPSS.

Pro obecné modely jednoduché regrese můžeme využít maticového vzorce, který rovněž vychází z metody nejmenších čtverců.

$$\vec{b} = (F^T F)^{-1} F^T \vec{y} \,. \tag{7}$$

kde $\vec{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix}$ je vektor regresních koeficientů, $\vec{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ yn \end{pmatrix}$ je vektor hodnot veličiny *y*,

$F = \begin{pmatrix} f_1(x_1), \cdots, f_m(x_1) \\ f_1(x_2), \cdots, f_m(x_2) \\ \vdots \\ f_1(x_n), \cdots, f_m(x_n) \end{pmatrix}$ je regresní matice příslušná danému regresnímu modelu.

Například pro lineární jednoduchou regresi má matice F následující tvar: $F = \begin{pmatrix} 1, x_1 \\ 1, x_2 \\ \vdots \\ 1, x_n \end{pmatrix}$, pro kvadratický model pak: $F = \begin{pmatrix} 1, x_1, x_1^2 \\ 1, x_2, x_2^2 \\ \vdots \\ 1, x_n, x_n^2 \end{pmatrix}$, nebo pro logaritmický model: $F = \begin{pmatrix} 1, \ln(x_1) \\ 1, \ln(x_2) \\ \vdots \\ 1, \ln(x_n) \end{pmatrix}$.

Při vícenásobné regresi převážně využíváme lineárních vztahů. Výpočet regresních koeficientů se provádí opět s využitím vzorce (7). V tomto případě má matice F tvar:

$F = \begin{pmatrix} 1, x_{11}, x_{12}, \cdots, x_{1m} \\ 1, x_{21}, x_{22}, \cdots, x_{2m} \\ \vdots \\ 1, x_{n1}, x_{n2}, \cdots, x_{nm} \end{pmatrix}$, kde x_{ij} znamená *i*-tou hodnotu *j*-tého regresoru.

4.1.1 Kvalita regresních modelů

Jak již bylo výše uvedeno, reziduum ε značí odchylku naměřené hodnoty od hodnoty vypočítané, čili $\varepsilon_i = y_i - \hat{y}_i$.

Ve výpočtech pak z důvodu odstranění znaménka (+, –) pracujeme s druhými mocninami těchto reziduí, neboli s reziduálními čtverci ε_i^2. Metoda nejmenších čtverců hledá minimum tzv. součtu reziduálních čtverců Q_e.

$$Q_e = \sum_{i=1}^{n} \varepsilon_i^2 \tag{8}$$

Kvalitu regresního modelu vyhodnocujeme pomocí následujících charakteristik.

Reziduální rozptyl

$$s_e^2 = \frac{Q_e}{n-p} \quad , \tag{9}$$

kde n je počet měření (bodů) a p je počet parametrů modelu (pro lineární model $p = 2$).
Platí, že $s_e^2 \geq 0$ a dále, že čím větší je hodnota s_e^2, tím hůře model vystihuje data.

Index determinace

$$I^2 = \frac{Q_{\hat{y}}}{Q_y} = 1 - \frac{Q_e}{Q_y}, \tag{10}$$

kde $Q_{\hat{y}} = \sum_{i=1}^{n} (f(x_i) - \bar{y})^2$ a $Q_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

Někdy se můžeme setkat s názvem Koeficient determinace a také s označením R^2.

Platí, že $I^2 \in \langle 0; 1 \rangle$. Hodnotu indexu determinace pro interpretaci převádíme na procenta. Jeho hodnota nám pak říká, kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem. Srozumitelněji a zjednodušeně jinými slovy můžeme říci, z kolika procent model vystihuje daná data. Je tedy zřejmé, že čím vyšší index determinace, tím lepší model.

Nutno ovšem podotknout, že index determinace závisí na počtu vysvětlujících proměnných a s růstem jejich počtu narůstá i jeho hodnota. V důsledku toho index determinace zvýhodňuje složitější modely (tj. modely s více parametry). Toto je nepříjemná vlastnost, která částečně snižuje jeho kvalitu. Pokud tedy využíváme indexu determinace k porovnání dvou modelů s různým počtem parametrů, měli bychom jeho vyhodnocení doplnit i vyhodnocením například pomocí reziduálního rozptylu.

Upravený index determinace

Porovnání modelů s různým počtem parametrů můžeme také provést pomocí upraveného indexu determinace.

$$I_{upr}^2 = 1 - (1 - I^2)\frac{n-1}{n-p} \tag{11}$$

Pearsonův korelační koeficient

$$|r| = \sqrt{I^2} \text{ a } \operatorname{sgn}(r) = \operatorname{sgn}(b_1) \qquad (12)$$

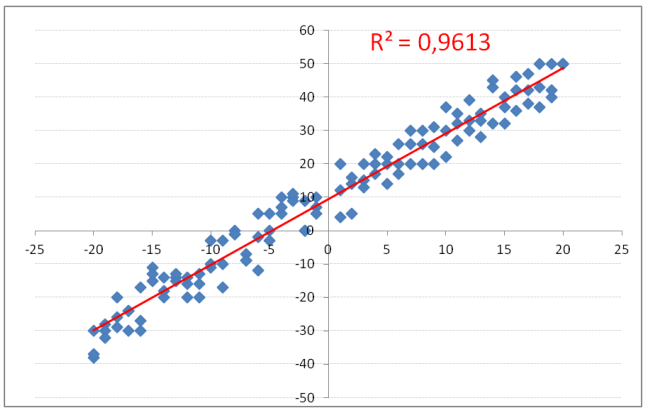
Korelační koeficient má smysl počítat pouze pro lineární model. Korelační koeficient má stejné znaménko jako směrnice regresní přímky.

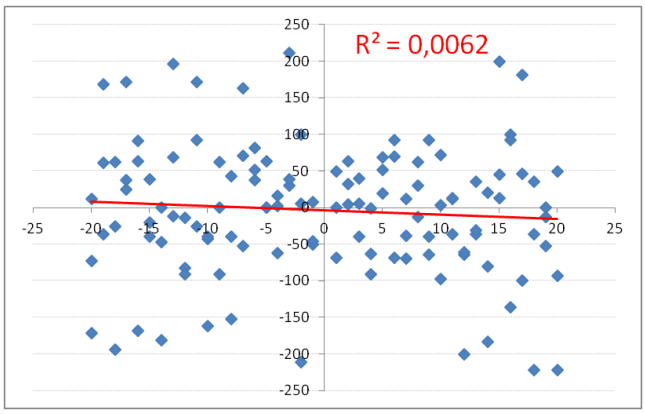
Tento výpočet korelačního koeficientu (12) je velmi zdlouhavý, proto se více využívá následujícího upraveného vzorce.

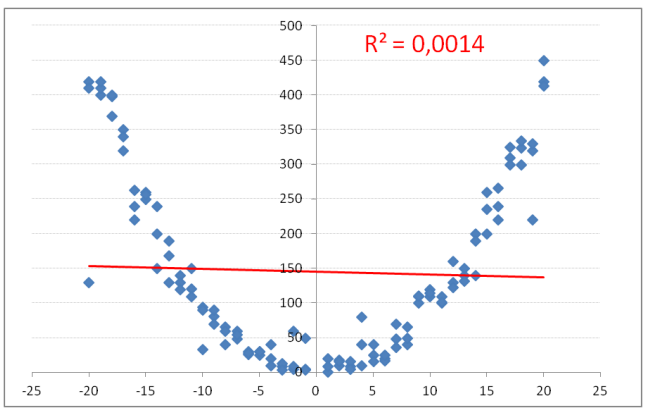
$$r = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2) \cdot (\overline{y^2} - \bar{y}^2)}} \qquad (13)$$

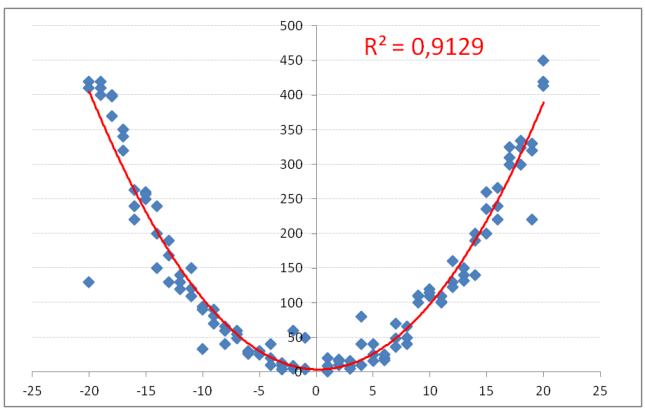
Platí, že $r \in \langle -1; 1 \rangle$. Čím blíže je jeho hodnota ke krajním hodnotám tohoto rozmezí, tím je lepší model. Pro vyhodnocení hodnot korelačního koeficientu existuje speciální test hypotéz. Zjednodušeně lze však říci, že pokud je jeho hodnota blízká 1 (pro dostatečné množství dat se většinou uvádí podmínka větší než 0,8), pak mluvíme o silné přímé lineární závislosti. Je-li jeho hodnota blízko -1 (menší než -0,8), pak mluvíme o silné nepřímé lineární závislosti. Pokud je jeho hodnota blízko 0 (v rozmezí od -0,3 do +0,3), pak říkáme, že není lineární závislost. Slovo lineární v poslední větě je velmi důležité. Je nutno si uvědomit, že neexistence lineární závislosti nevylučuje existenci funkční závislosti jiného druhu (kvadratické, logaritmické, …)

Příklady vybraných situací v datech









Spearmanův koeficient pořadové korelace

Pokud data nesplňují předpoklady rozložení dat (jiné než normální rozložení proměnných, nelinearita vztahu, data obsahující odlehlá pozorování, ordinální data) je vhodnější použít neparametrický ekvivalent, a to Spearmanův koeficient pořadové korelace. Jsou-li hodnoty proměnných x_i a y_i seřazeny vzestupně do dvou řad a každé hodnotě je přiděleno pořadí, pak koeficient pořadové korelace je dán vztahem:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)}, \qquad (14)$$

kde D_i je rozdíl pořadí hodnot x_i a y_i.

Pokud se vyskytuje v souboru více stejných hodnot, pak všechny z nich obdrží hodnotu pořadí vypočítanou jako průměr z hodnot jednotlivých pozic (př. pokud se v souboru vyskytují na pozicích 5-8 čtyři stejné hodnoty, pak všechny tyto hodnoty obdrží pořadí $\frac{5+6+7+8}{4} = 6{,}5$.)

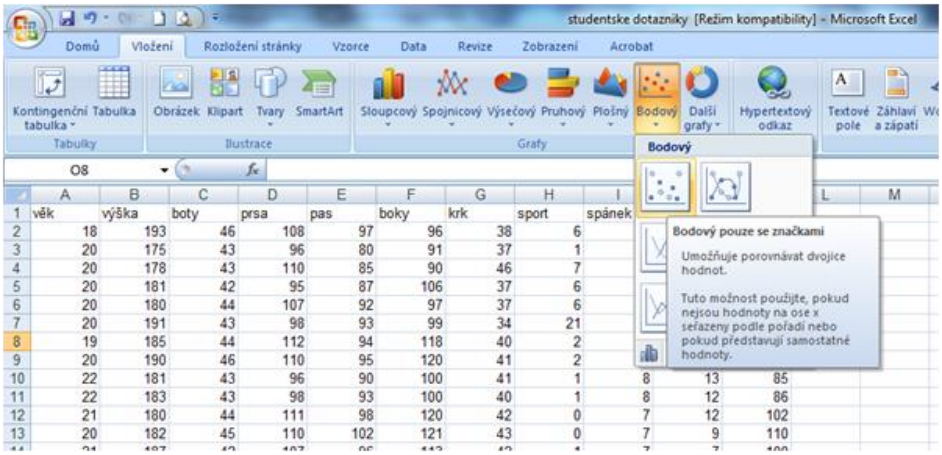
Spearmanův koeficient pořadové korelace nabývá stejně jako korelační koeficient hodnot z intervalu $r_s \in \langle -1; 1 \rangle$. Vyhodnocení síly závislosti pak probíhá obdobně jako u korelačního koeficientu.

Kromě Spearmanova korelačního koeficientu existují i další neparametrické korelační koeficienty jako např. Kendallovo τ.

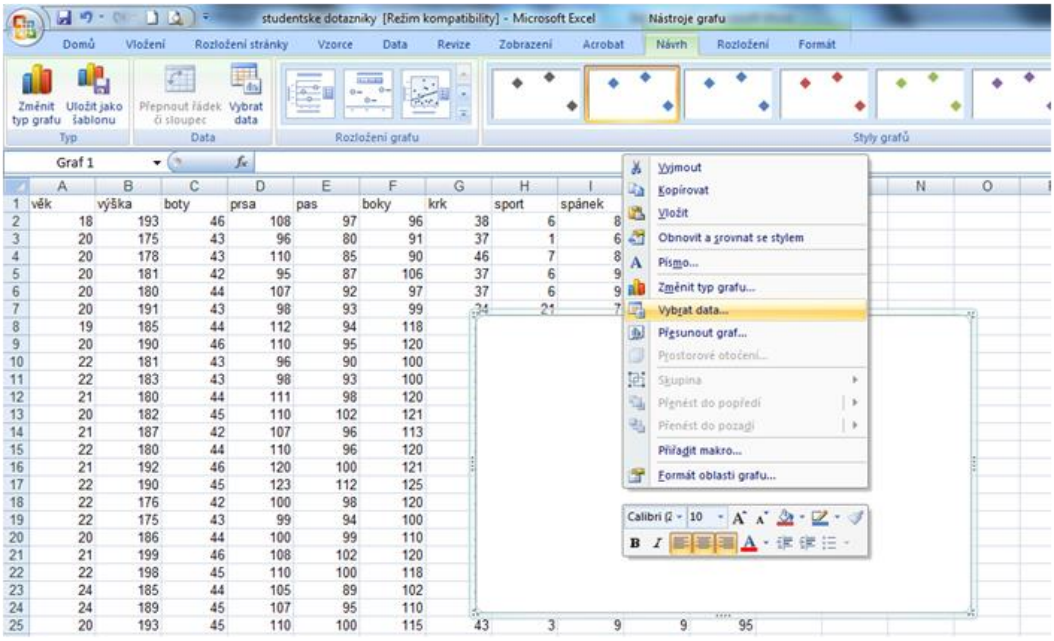
4.2 Regresní analýza pomocí SW Excel

4.2.1 Grafické znázornění

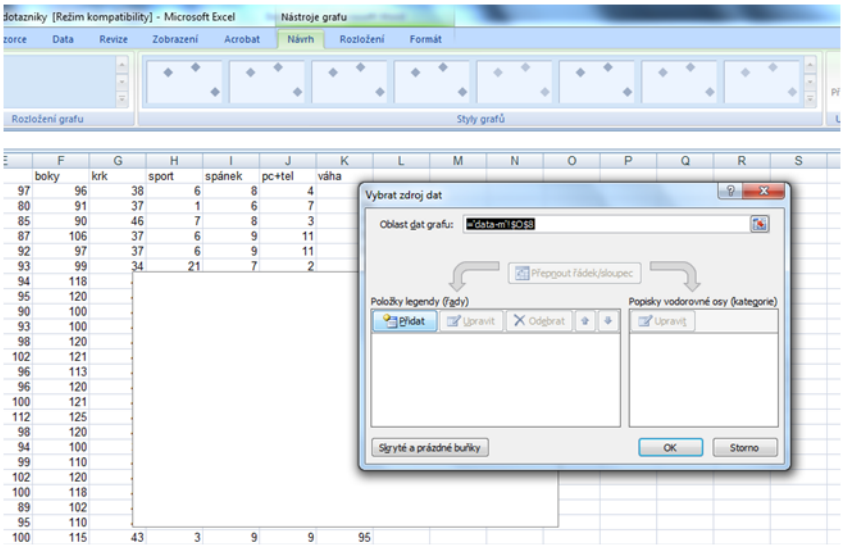
Pokud chceme provést regresní analýzu, vytvoříme si nejprve graf. V Excelu je pro tyto účely nejvhodnější bodový graf. V nabídce zvolíme *Vložit – Bodový graf (pouze se značkami)*.



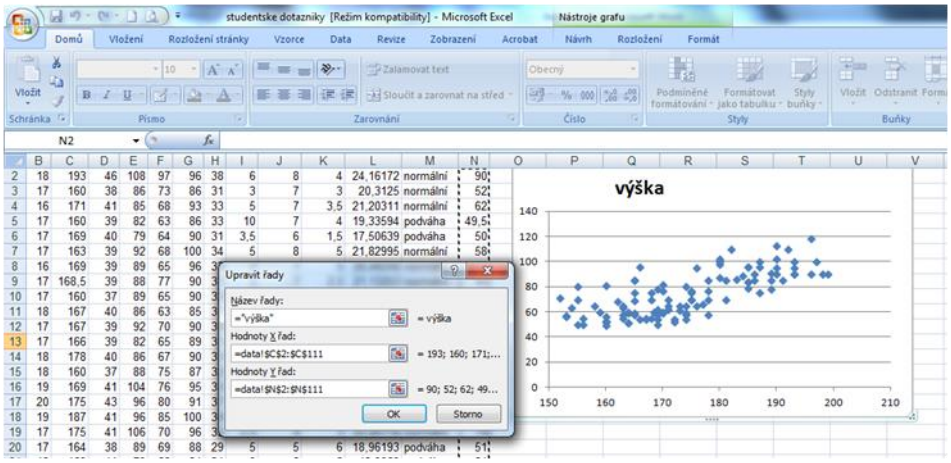
Po této volbě nám Excel vytvoří prázdnou plochu pro graf. V menu pro tuto oblast vybereme možnost *Vybrat data*.



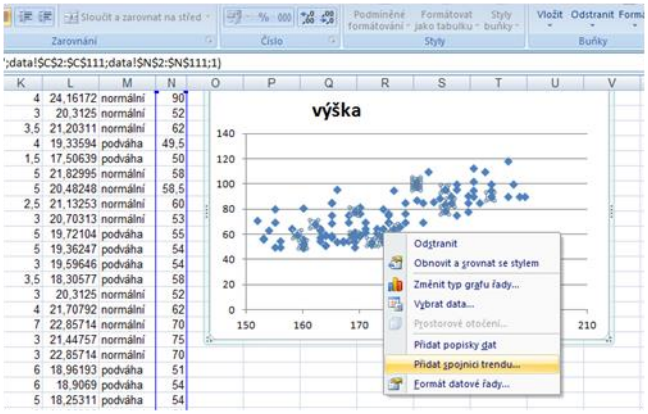
V následujícím dialogovém okně vybereme možnost *Přidat Položky legendy (řady)*.



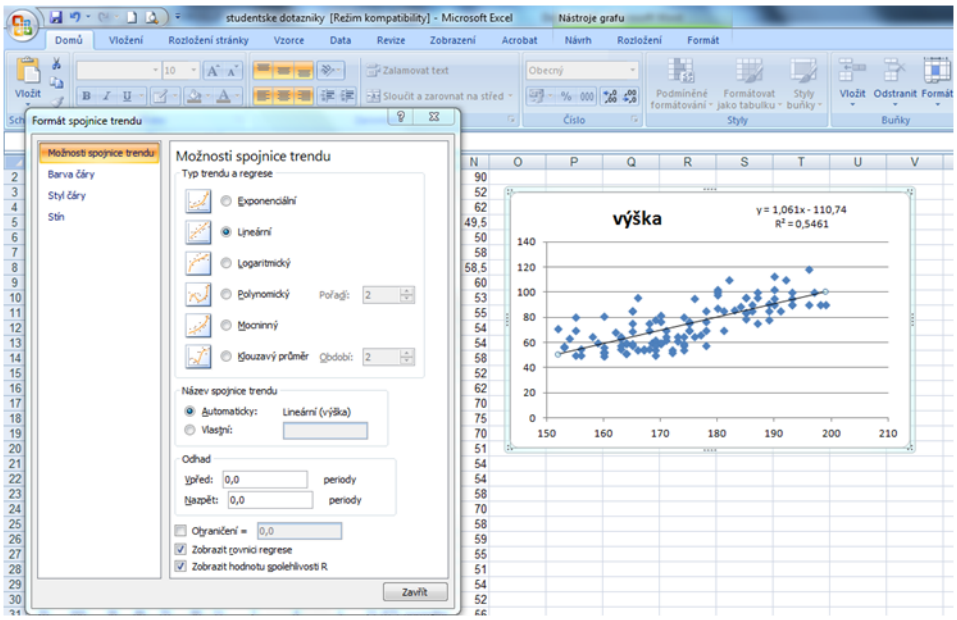
Následně vyplníme požadovaný název řady a tažením myší označíme sloupce s daty.



Dále do grafu vložíme křivku regresní funkce, a to tak, že si otevřeme menu řady (např. zmačknutím pravého tlačítka myši, pokud kurzor ukazuje na libovolný bod řady). V tomto menu vybereme možnost *Přidat spojnici trendu*.

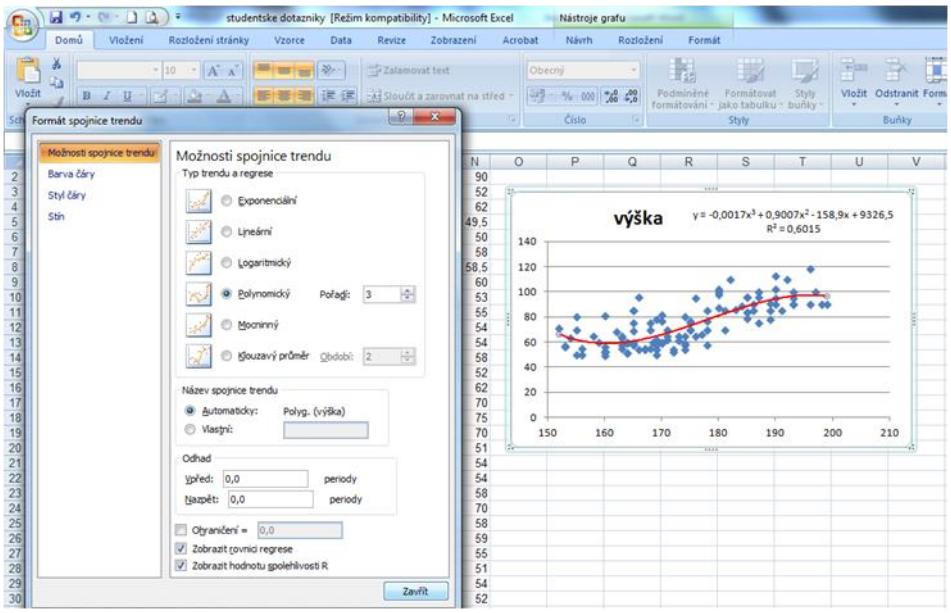


Tato volba nám do grafu vloží regresní přímku a zároveň se nám otevře dialogové okno pro úpravu této přímky. Ve spodní části dialogového okna zaškrtneme možnosti *Zobrazit rovnici regrese* a *Zobrazit hodnotu spolehlivosti R* (= hodnotu indexu determinace).

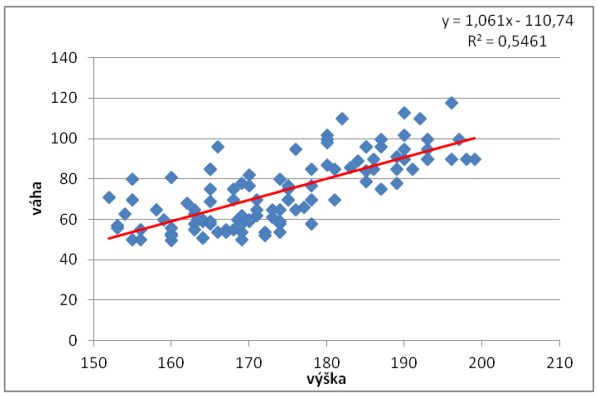


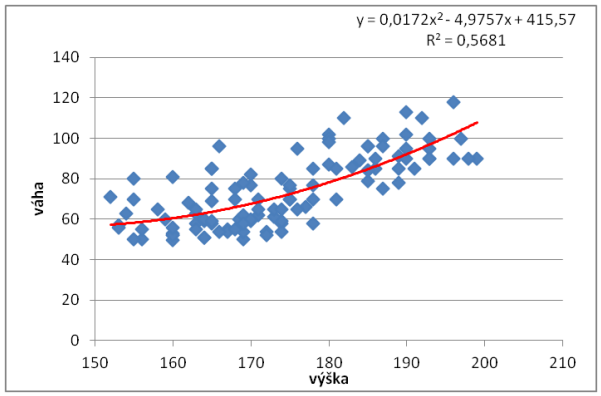
Dále můžeme pomocí tohoto okna volit typ regresní funkce. Na výběr máme lineární model (implicitní možnost), dále model exponenciální, logaritmický, mocninný a různé modely polynomické, a to až do stupně 6.

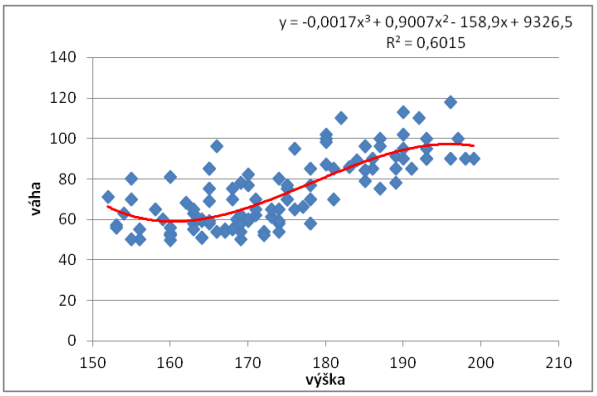
V levé části tohoto okna můžeme volit různé možnosti pro úpravu vzhledu křivky regresní funkce, a to jak barvu, styl i tloušťku čáry.



Pokud vyzkoušíme všechny volby typu regresního modelu, vidíme, že tři z nich by mohly být dobrými modely pro vyjádření vztahu mezi veličinami Výška a Váha. Těmito modely jsou lineární model, kvadratický model a polynomický model 3. stupně.

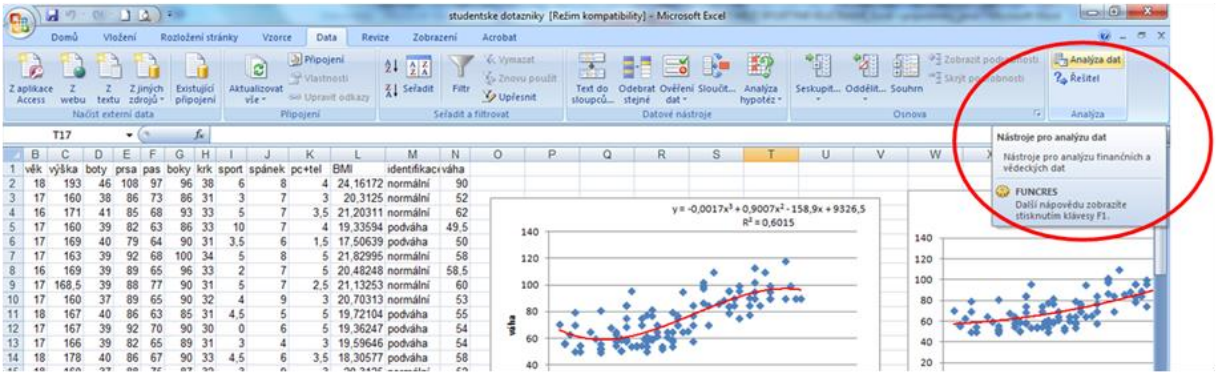




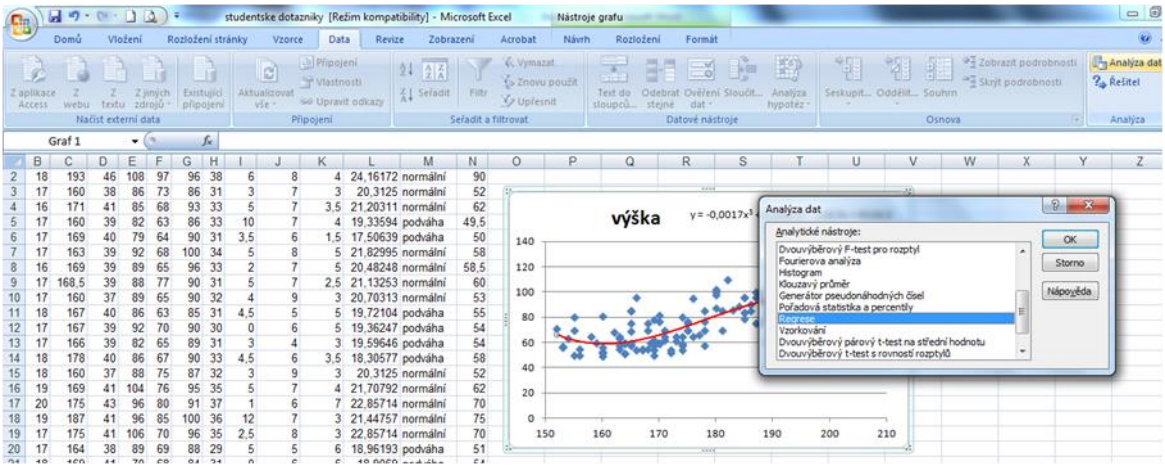


4.2.2 Jednoduchá regresní analýza pomocí doplňku prostředku Excel – lineární model

Abychom však správně vybrali z těchto modelů ten nejlepší, nemůžeme se spolehnout pouze na porovnání indexů determinace jednotlivých modelů, ale potřebujeme podrobnější regresní analýzu. K tomu musíme využít speciálního doplňku Excelu, a to *Analýzu dat*. Nejprve si ukážeme využití tohoto prostředku na jednoduché lineární regresi. Výsledkem by měla být regresní přímka s rovnicí, která je vidět na výše uvedeném obrázku.



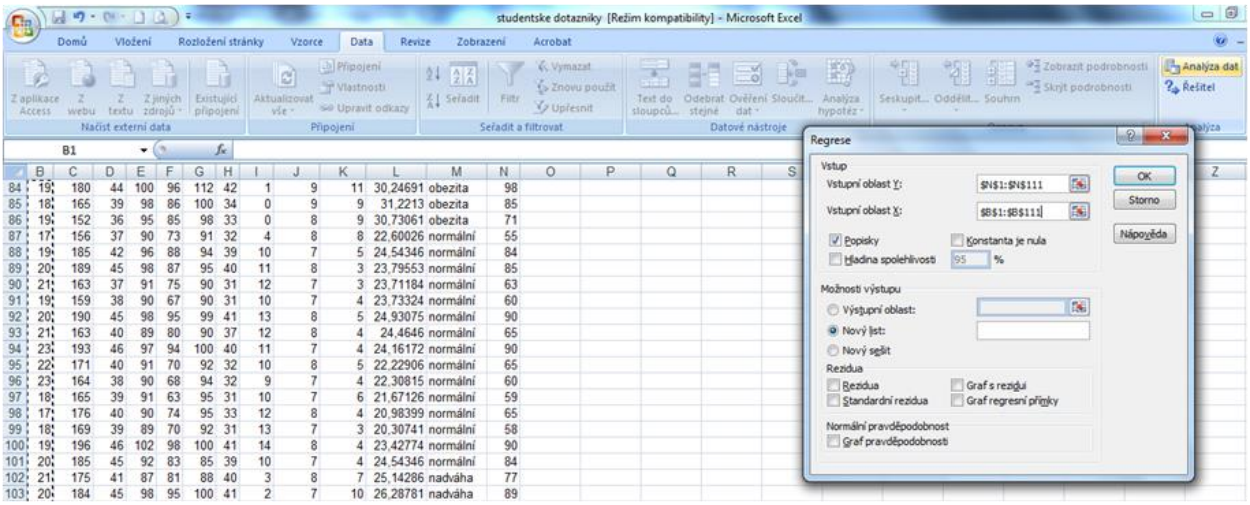
V dialogovém okně, které se nám otevře po výběru tohoto doplňku, vybereme možnost *Regrese*.



Následně vyplníme potřebné údaje v dialogovém okně. Povinně vybereme oblasti obsahující data. Doporučujeme do výběru zahrnout i hlavičky daných sloupců, ulehčí nám to následnou orientaci ve výstupu. V tomto případě však musíme zaškrtnout pole *Popisky*. Někdy můžeme mít potřebu upravit *Hladinu spolehlivosti*. Pokud bychom potřebovali velmi přesné a spolehlivé výsledky, můžeme tuto hodnotu navýšit až na 99 %, v případě nedostatku dat můžeme naopak tuto hodnotu snížit až na 90 %.

Dále si můžeme vybrat, kam nám má Excel umístit výstup. Tuto možnost však doporučujeme nevyužít a ponechat implicitní nastavení, které výstup umístí do nového listu.

Někdy můžeme potřebovat i informace o hodnotách reziduí, případně jejich grafické znázornění, v takovém případě zaškrtneme vybraná požadovaná pole.



Pro potvrzení zadání zvolíme *OK* a poté nám Excel vloží do nového listu následující tabulku.

VÝSLEDEK

Regresní statistika	
Násobné R	0,738967
Hodnota spolehlivosti R	0,546073
Nastavená hodnota spolehlivosti R	0,54187
Chyba stř. hodnoty	11,66689
Pozorování	110

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	1	17684,71	17684,71	129,9236	3,09E-20
Rezidua	108	14700,55	136,1162		
Celkem	109	32385,26			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	-110,739	16,2265	-6,82459	5,34E-10	-142,903	-78,5756
výška	1,061048	0,093087	11,39841	3,09E-20	0,876533	1,245564

Ve výstupu jsme obarvili pár důležitých polí. Žluté pole (Násobné R) nám udává hodnotu korelačního koeficientu. Hodnota 0,739 značí středně silnou lineární závislost mezi veličinami Výška a Váha.

Oranžové pole (Hodnota spolehlivosti R) udává hodnotu indexu determinace (můžeme zkontrolovat s hodnotou, kterou uvedl Excel v případě práce s grafy – viz výše). Uvědomme si, že platí $0{,}739^2 = 0{,}546$ (index determinace je druhou mocninou korelačního koeficientu). Hodnotu 0,546 můžeme interpretovat: „Variabilita veličiny Váha je z 54,6 % popsána veličinou Výška. Popis zbytku, neboli 45,4 % variability Váhy, je nutno hledat jinde.“ Zjednodušeně bychom také mohli říci, že model vystihuje data z 54,6 %.

Hnědé pole (Nastavená hodnota spolehlivosti R) udává hodnotu upraveného indexu determinace, který slouží k porovnání modelů s rozdílným počtem parametrů (regresních koeficientů).

Zelené pole udává hodnotu reziduálního rozptylu (tj. Q_e).

Vidíme, že obě růžová pole mají stejnou hodnotu. Toto platí pouze v případě lineární jednoduché regrese. Jejich hodnota vypovídá o významnosti modelu, respektive regresního koeficientu. Horní růžové pole „*Významnost F*“ se týká významnosti modelu jako celku (testuje se nulová hypotéza, že všechny regresní koeficienty kromě absolutního členu jsou nulové). Spodní růžové pole „Hodnota P“ se týká významnosti pouze jednoho regresního koeficientu, v našem případě koeficientu pro Výšku (testuje se nulová hypotéza, že tento koeficient je roven nule). Pokud je tato hodnota menší než zvolená hladina významnosti (v našem případě 0,05), tj. doplňková hodnota k hladině spolehlivosti, pak je model významný, respektive regresní model je statisticky významně odlišný od 0. V opačném případě (p-hodnota je větší než 0,05) je model nevýznamný. V našem případě je p-hodnota velmi malá ($3{,}09*10^{-20}$), tedy model významný je.

Modrá pole udávají hodnoty regresních koeficientů (možno opět zkontrolovat s grafem – viz výše). Na základě těchto hodnot tedy můžeme napsat rovnici regresní funkce, v našem případě přímky $y = 1{,}061x - 110{,}739$. Hodnota směrnice (tj. 1,061) nám říká, že o 1 cm větší osoba v průměru váží o 1,061 kg více.

Hodnota 1,061 je bodovým odhadem regresního koeficientu. Přesnější je však intervalový odhad, který nám určují červená pole. V našem případě vidíme, že směrnice má s 95% spolehlivostí hodnotu v rozmezí mezi 0,877 a 1,246.

4.2.3 Jednoduchá regresní analýza pomocí doplňku prostředku Excel – obecný model

Poněkud složitější bude tvorba obecnějšího modelu jednoduché regrese. K tomu účelu si nejprve musíme speciálním způsobem upravit data, speciálně oblast obsahující údaje o veličině x. Tato úprava vychází z tvaru matice F, o které jsme mluvili v teoretickém úvodu v části o maticovém vzorci pro výpočet vektoru regresních koeficientů.

Oblast dat obsahující údaje o veličině x bude obsahovat tolik sloupců, kolikrát se x objevuje v rovnici požadovaného regresního modelu, každý z těchto výskytů je ve tvaru nějaké funkce, označme ji $f_i(x)$. Jednotlivé sloupce tedy budou obsahovat $f_i(x)$.

Uveďme si několik příkladů:

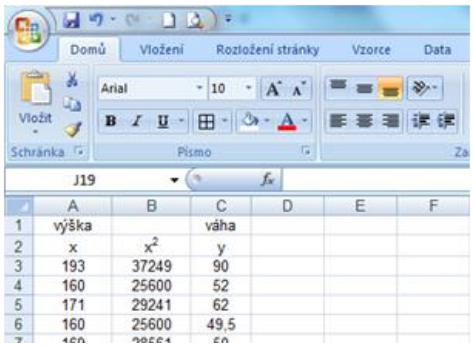
Lineární model má tvar $y = b_1 x + b_0$, proto bude oblast obsahovat jeden sloupec a v něm hodnoty x.

Kvadratický model má tvar $y = b_2 x^2 + b_1 x + b_0$, proto bude oblast obsahovat dva sloupce, v prvním z nich hodnoty x, v druhém x^2.

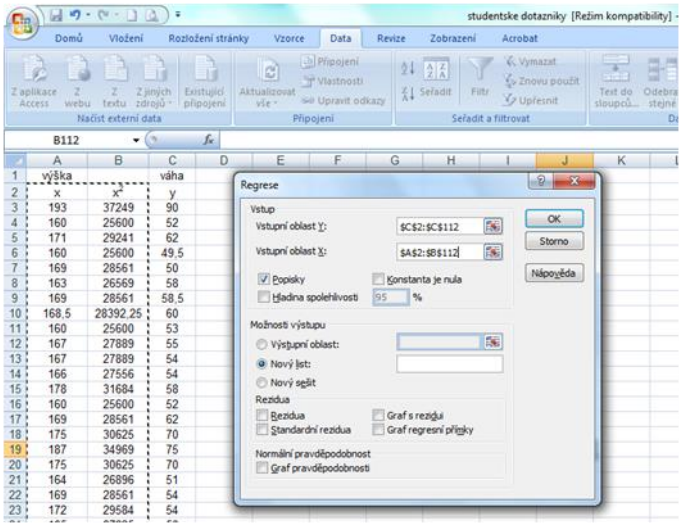
Polynomický model 3. stupně má tvar $y = b_3 x^3 + b_2 x^2 + b_1 x + b_0$, proto bude oblast obsahovat tři sloupce, v prvním z nich hodnoty x, v druhém x^2, ve třetím x^3.

Logaritmický model má tvar $y = b_1 \ln x + b_0$, proto bude oblast obsahovat jeden sloupec a v něm hodnoty ln x.

My nyní chceme vytvořit kvadratický model, proto do listu vložíme nový sloupec obsahující hodnoty x^2 (sloupec obsahující *x* již v souboru máme).



Nyní v nabídce *Data* vybereme doplněk *Analýza dat* a v následném dialogovém okně vybereme možnost *Regrese*. Jako vstupní oblast Y vybereme sloupec obsahující údaje o váze respondentů, jako vstupní oblast X vyznačíme oba sloupce obsahující informace o položce výška. Pozor – jako popisky můžeme vyznačit pouze jeden řádek. Ostatní části dialogového okna vyplníme obdobně, jako jsme již vyplňovali v případě tvorby lineárního modelu.



Excel nám do nového listu umístí následující výstup.

VÝSLEDEK

Regresní statistika	
Násobné R	0,753741
Hodnota spolehlivosti R	0,568126
Nastavená hodnota spolehlivosti R	0,560053
Chyba stř. hodnoty	11,43301
Pozorování	110

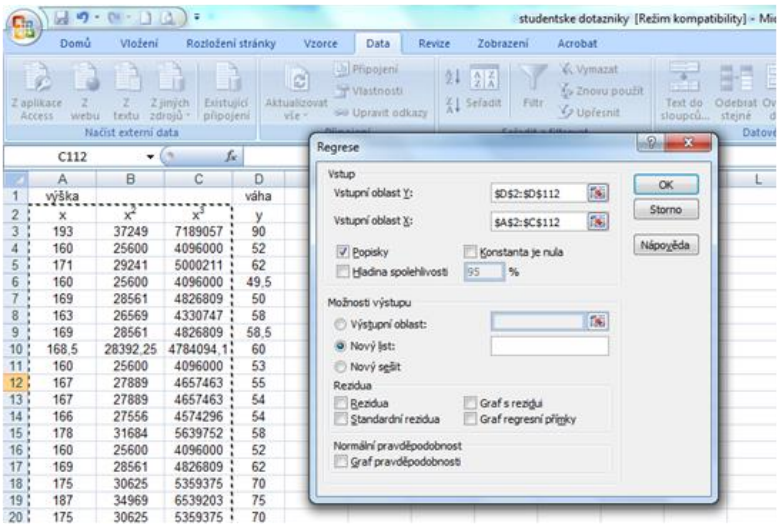
ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	18398,91	9199,453	70,37868	3,1E-20
Rezidua	107	13986,36	130,7136		
Celkem	109	32385,26			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	415,5727	225,7234	1,84107	0,068382	-31,8976	863,043
x	-4,97572	2,584211	-1,92543	0,05683	-10,0986	0,147181
x2	0,017229	0,007371	2,337475	0,021275	0,002617	0,03184

Ve výstupu jsme barevně vyznačili nejdůležitější pole pro vyhodnocení modelu a následný výběr modelu nejlepšího. Tato pole později zaneseme do přehledné tabulky obsahující údaje o všech vytvářených modelech. Jen si nyní uvědomme, že model je významný (p-hodnota jeho významnosti je rovna $3{,}1*10^{-20}$) a taktéž je významný kvadratický koeficient (p-hodnota jeho významnosti je rovna 0,021275), tedy se opravdu jedná o kvadratický model.

Nyní obdobně vytvoříme polynomický model 3. stupně. Nejprve upravíme vstupní oblast X, a to tak, že přidáme další sloupec. Tento sloupec bude obsahovat třetí mocniny *x*. Uvědomme si, že celá vstupní oblast X musí být kompaktní, tedy všechny její sloupce musí být bezprostředně vedle sebe.



	A	B	C	D
1	výška			váha
2	x	x^2	x^3	y
3	193	37249	7189057	90
4	160	25600	4096000	52
5	171	29241	5000211	62
6	160	25600	4096000	49,5
7	169	28561	4826809	50
8	163	26569	4330747	58
9	169	28561	4826809	58,5
10	168,5	28392,25	4784094,1	60
11	160	25600	4096000	53
12	167	27889	4657463	55
13	167	27889	4657463	54
14	166	27556	4574296	54
15	178	31684	5639752	58
16	160	25600	4096000	52
17	169	28561	4826809	62
18	175	30625	5359375	70
19	187	34969	6539203	75
20	175	30625	5359375	70

Excel nám do nového listu umístí následující výstup.

VÝSLEDEK

Regresní statistika	
Násobné R	0,77559
Hodnota spolehlivosti R	0,60154
Nastavená hodnota spolehlivosti R	0,590263
Chyba stř. hodnoty	11,03349
Pozorování	110

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	3	19481,04	6493,68	53,34145	4,27E-21
Rezidua	106	12904,22	121,738		
Celkem	109	32385,26			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	9326,493	2996,716	3,112238	0,002387	3385,212	15267,77
x	-158,902	51,68809	-3,07424	0,002684	-261,378	-56,425
x2	0,900658	0,296394	3,038719	0,002992	0,313028	1,488289
x3	-0,00168	0,000565	-2,98145	0,003559	-0,0028	-0,00056

Ve výstupu jsme opět vyznačili žlutě nejdůležitější pole pro vyhodnocení modelu a následný výběr modelu nejlepšího. Tato pole opět zaneseme do závěrečné tabulky obsahující údaje o všech vytvářených modelech. Vidíme, že tento model je významný (p-hodnota jeho významnosti je rovna $4{,}27*10^{-21}$) a taktéž je významný vedoucí koeficient, tedy koeficient u x^3 (p-hodnota jeho významnosti je rovna 0,003559), tedy se opravdu jedná o polynomický model stupně 3.

A nyní si již můžeme vytvořit již zmíněnou závěrečnou tabulku.

	Významnost modelu	Významnost vůdčího koef.	Index det./ upr.index det.	Rezid. rozptyl	Rovnice
lineární	$3{,}09*10^{-20}$	$3{,}09*10^{-20}$	0,546/ 0,542	136,12	$y = 1{,}061\ x - 110{,}739$
kvadratický	$3{,}1*10^{-20}$	0,021275	0,568/ 0,560	130,71	$y = 0{,}017229\ x^2 - 4{,}97572\ x + 415{,}5727$
polynom 3.st.	$4{,}27*10^{-21}$	0,003559	0,602/ 0,590	121,74	$y = -0{,}00168\ x^3 + 0{,}900658\ x^2 - 158{,}902\ x + 9326{,}493$

Jednotlivé indexy determinace a rovnice jednotlivých regresních modelů můžeme porovnat s výsledky v rámci grafického zpracování. Je však přirozené, že index determinace je vyšší, čím vyšší je mocnina v polynomické funkci. Zvýhodňování složitějších modelů (tj. modelů s vyšším počtem regresních koeficientů) je základní vlastnost a nevýhoda indexu determinace. Proto jsme nemohli udělat hodnověrný závěr už v první fázi zpracování.

Nyní však již máme všechny potřebné údaje. Vidíme, že všechny vytvořené modely jsou významné. Významné jsou i regresní koeficienty u nejvyšší mocniny v polynomu. Jedná se tedy o využitelné modely. Rozhodujícím faktorem vyhodnocení tedy bude porovnání hodnot upravených indexů determinace a reziduálních rozptylů jednotlivých modelů. Z teoretického úvodu připomeňme tři kritéria pro porovnání modelů. Víme, že

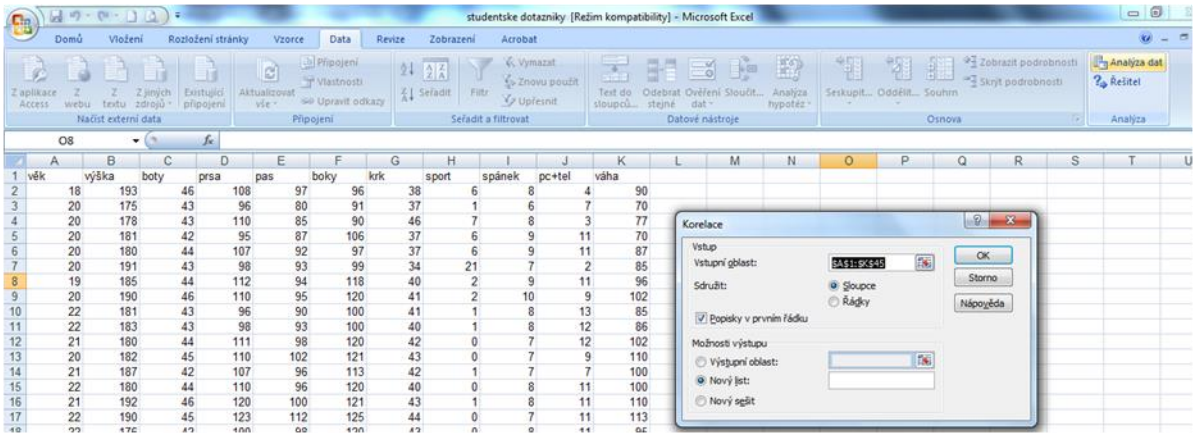
- čím nižší reziduální rozptyl, tím lepší model;
- čím vyšší upravený index determinace, tím kvalitnější model.
- čím vyšší upravený index determinace, tím lepší model.

Ze závěrečného vyhodnocení tedy vychází vítězně polynom 3. stupně. Vidíme, že v rámci tohoto modelu je 60,2 % změn veličiny váha vysvětleno změnami veličiny výška a 39,8 % zůstalo nevysvětleno.

4.2.4 Vícenásobná regresní analýza pomocí doplňku prostředku Excel

Z výše uvedeného je zřejmé, že ještě musíme hledat další faktory mající vliv na váhu člověka. V našem průzkumu byly sledovány ještě veličiny Pohlaví, Věk, Velikost bot, Obvod pasu, Obvod prsou, Obvod boků, Obvod krku, Týdenní počet hodin strávený sportem, Týdenní počet hodin strávený u TV a PC, Denní počet hodin strávený spánkem.

Budeme sledovat vliv všech těchto faktorů. K tomu využijeme vícenásobné regrese, speciálně Stepwise regrese. Než však začneme dělat samotnou regresi, měli bychom zkontrolovat vzájemnou korelaci mezi jednotlivými regresory. K tomu využijeme další z možností doplňku *Analýza dat*, a to *Korelace*.



V dialogovém okně jednak označíme celou souvislou oblast s daty, tedy všechny sloupce obsahující údaje o jednotlivých zmíněných veličinách. Zde je víceméně nutno vybrat i hlavičky sloupců a tedy zatrhnout pole *Popisky v prvním řádku*. Opět doporučujeme ponechat implicitní možnost výstupu do nového listu.

Výsledkem je následující korelační matice (tj. tabulka) obsahující hodnoty vzájemných korelačních koeficientů. Matice je pouze trojúhelníková, a to z důvodu, že korelace je vztah vzájemný, symetrický, nemusíme tedy vyplňovat zbylé údaje. Jedničky na diagonále znamenají, že každá z veličin je sama se sebou dokonale korelovaná, což je naprosto přirozené a nemůže tomu být jinak.

	pohlaví	*věk*	*výška*	*boty*	*prsa*	*pas*	*boky*	*krk*	*sport*	*spánek*	*pc+tel*
pohlaví	1,00										
věk	0,23	1,00									
výška	0,80	0,16	1,00								
boty	0,87	0,16	0,93	1,00							
prsa	0,72	0,19	0,70	0,74	1,00						
pas	0,84	0,19	0,71	0,79	0,82	1,00					
boky	0,64	0,22	0,62	0,66	0,84	0,80	1,00				
krk	0,92	0,18	0,78	0,85	0,80	0,87	0,74	1,00			
sport	0,07	0,02	0,20	0,08	-0,07	-0,12	-0,20	-0,01	1,00		
spánek	0,17	0,07	0,07	0,12	0,37	0,33	0,32	0,25	-0,03	1,00	
pc+tel	0,31	0,11	0,10	0,24	0,36	0,49	0,47	0,34	-0,57	0,33	1,00

Žlutě jsou vyznačeny hodnoty korelačních koeficientů, které jsou větší než 0,5 – tedy značí určitou míru korelace. Hodnoty nad 0,7 znamenají již významnou korelaci. Znaménko mínus u určitých hodnot značí, že daný vztah je nepřímý. Například hodnota -0,57 korelačního koeficientu mezi veličinami Sport a PC+TV značí středně silnou závislost a dále nám dává informaci, že čím více osoby sportují, tím méně tráví času u PC+TV a naopak. Vidíme, že veličina Věk není korelována s žádnou z ostatních veličin. Naopak veličiny Pohlaví a Krk

jsou korelovány s většinou ostatních veličin. Není vhodné, aby model vícenásobné regrese obsahoval vzájemně korelované faktory, protože toto dává zkreslenou informaci o síle působení těchto faktorů. Pokud data obsahují vzájemně korelované faktory, je toto nutno nějak řešit. Možností je mnoho. Nejjednodušší, ne vždy však dobře použitelné je některé z těchto faktorů z analýzy vypustit. Někdy ale nelze dobře určit, který z faktorů vypustit. Pokud nepoužijeme jiný způsob řešení této situace, musíme si aspoň dát pozor, aby se ve výsledném modelu tyto faktory nevyskytovaly společně. Složitějšími možnostmi jsou některé vícerozměrné metody, např. Metoda hlavních komponent, které nám na základě našich vstupních faktorů vytvoří umělé faktory, které jsou již nekorelované.

V případě silné korelovanosti naší veličiny Pohlaví můžeme využít možnosti rozdělení souboru na dvě části – muže a ženy. Další šetření pak budeme provádět zvlášť pro muže a zvlášť pro ženy. Vytvoříme tak dva regresní modely, které pak můžeme vzájemně porovnávat.

Následující dvě tabulky obsahují korelační matice v případě žen a mužů odděleně.

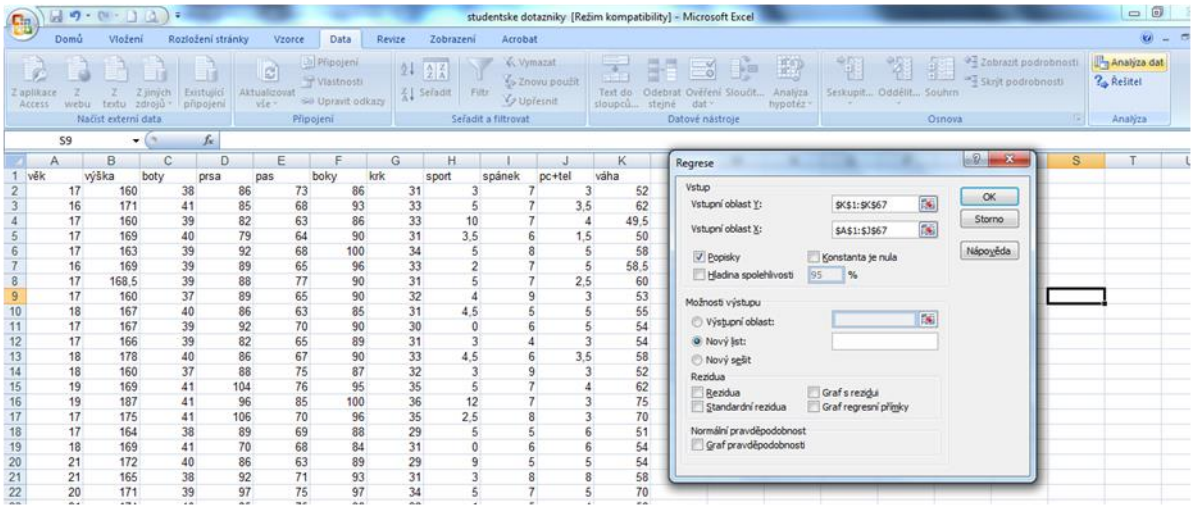
ženy	*věk*	*výška*	*boty*	*prsa*	*pas*	*boky*	*krk*	*sport*	*spánek*	*pc+tel*
věk	1,00									
výška	0,05	1,00								
boty	-0,04	0,76	1,00							
prsa	0,10	0,05	0,00	1,00						
pas	0,09	-0,23	-0,09	0,44	1,00					
boky	0,15	-0,02	0,04	0,60	0,58	1,00				
krk	-0,12	0,07	0,15	0,45	0,41	0,50	1,00			
sport	-0,04	0,26	0,07	0,03	-0,34	-0,11	0,10	1,00		
spánek	0,07	-0,21	-0,16	0,51	0,48	0,48	0,36	-0,07	1,00	
pc+tel	0,21	-0,32	-0,09	0,23	0,62	0,45	0,05	-0,48	0,37	1,00

Vidíme, že v souboru obsahujícím údaje o ženách jsou již faktory korelovány minimálně. Nejzávažnější je korelace mezi veličinami Boty a Výška. V tomto případě můžeme situaci řešit vypuštěním veličiny Boty z další analýzy.

muži	*věk*	*výška*	*boty*	*prsa*	*pas*	*boky*	*krk*	*sport*	*spánek*	*pc+tel*
věk	1,00									
výška	-0,15	1,00								
boty	-0,17	0,82	1,00							
prsa	-0,03	0,54	0,59	1,00						
pas	-0,15	0,63	0,61	0,73	1,00					
boky	0,10	0,42	0,42	0,78	0,78	1,00				
krk	0,01	0,32	0,37	0,56	0,56	0,53	1,00			
sport	0,06	0,19	0,00	-0,34	-0,32	-0,43	-0,44	1,00		
spánek	-0,09	0,10	0,11	0,22	0,06	0,24	0,12	-0,01	1,00	
pc+tel	-0,17	-0,17	-0,06	0,20	0,26	0,36	0,23	-0,75	0,21	1,00

V souboru obsahujícím údaje o mužích se vyskytuje více významných hodnot korelačních koeficientů, musíme si tedy v dalším zpracování dávat větší pozor.

Nyní budeme provádět vícenásobnou regresi v souboru žen. Nejprve použijeme tzv. metodu Enter, tedy do regrese zahrneme všechny sledované faktory. Opět využijeme prostředek *Regrese* v rámci doplňku *Analýza dat*. Abychom mohli dobře označit *Vstupní oblast dat X*, musíme si dát pozor, aby sloupce obsahující tyto faktory byly všechny vedle sebe. Do *Vstupní oblast X* pak tažením myši vyznačíme všechny tyto sloupce naráz.



Výstup jsme si opět nechali umístit do nového listu. Vzniklá tabulka je nyní poněkud složitější než v případě jednoduché regrese.

VÝSLEDEK

Regresní statistika	
Násobné R	0,886917
Hodnota spolehlivosti R	0,786623
Nastavená hodnota spolehlivosti R	0,747827
Chyba stř. hodnoty	5,123442
Pozorování	66

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	10	5322,36	532,236	20,27592	4,77E-15
Rezidua	55	1443,731	26,24966		
Celkem	65	6766,091			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	-159,507	24,33852	-6,55366	2,03E-08	-208,282	-110,731
věk	0,523515	0,340898	1,535697	0,130346	-0,15966	1,206689
výška	0,238201	0,15998	1,488941	0,142215	-0,08241	0,558808
boty	1,404614	0,79341	1,770351	0,082211	-0,18542	2,994644
prsa	0,094408	0,171444	0,550662	0,584095	-0,24917	0,437988
pas	0,711686	0,135465	5,253662	2,49E-06	0,440208	0,983163
boky	0,680456	0,225227	3,021206	0,003816	0,229092	1,131821
krk	-0,56238	0,579918	-0,96975	0,336415	-1,72456	0,599804
sport	-0,11926	0,213659	-0,55817	0,578995	-0,54744	0,308925
spánek	1,838993	0,634988	2,896105	0,005412	0,566448	3,111538
pc+tel	-0,02808	0,391058	-0,07182	0,943008	-0,81178	0,755613

Tmavě modré pole značí hodnotu vícenásobného korelačního koeficientu, hodnotícího působení všech faktorů dohromady. Zelené pole značí hodnotu indexu determinace. V našem případě lze tedy říci, že variabilita veličiny Váha je modelem vystižena z 78,66 %. Růžové pole určuje p-hodnotu významnosti celého modelu. Její hodnota ($4{,}77*10^{-15}$) je velmi nízká, model je tedy významný.

Žlutě jsou vyznačeny p-hodnoty, které jsou vyšší než 0,05, tedy ukazují, že příslušný koeficient není významný. P-hodnoty menší než 0,05 jsme nevyznačili, tyto příslušné koeficienty významné jsou. Povšimněme si modře a červeně vyznačených polí. Červeně vyznačený je koeficient s hodnotou 1,404614, modře pak koeficient s hodnotou 0,711686. Vidíme, že červený koeficient má vyšší hodnotu, přesto dle p-hodnoty není významný. Oproti tomu modrý koeficient má nižší hodnotu a přesto dle p-hodnoty významný je. Toto je

příkladem, že významnost koeficientů nelze posuzovat „od oka“. Opravdu se může stát, že hodnota 1000 bude nevýznamná, naopak hodnota 0,001 významná bude.

Z výstupu vidíme, že se v modelu vyskytují nevýznamné koeficienty, tedy model není dobrý. Použijeme tedy metodu stepwise regrese k postupnému budování výsledného modelu. V prvním kroku vytvoříme tolik jednoduchých lineárních regresí, kolik máme faktorů. Každá z těchto regresí vytvoří výstup do samostatného listu. Do přehledné tabulky si z těchto výstupů přeneseme p-hodnoty významnosti jednotlivých modelů. V následující tabulce jsme si takto vytvořili první sloupec. Řádek Boty je prázdný, protože jsme výše zjistili, že tato veličina je silně korelovaná s veličinou výška a z další analýzy jsme ji vyjmuli. Žlutě jsou v tomto sloupci vyznačeny p-hodnoty větší než 0,05, které ukazují na nevýznamnost příslušného modelu. Z ostatních p-hodnot (tedy hodnot, které ukazovaly na významné modely) jsme vybrali tu nejmenší a označili ji modře. Tato hodnota ($1{,}8295*10^{-12}$) náleží k modelu popisujícímu vztah váhy a obvodu pasu a ukazuje, že tento model je nejvýznamnější. Proto veličinu Obvod pasu vybereme jako nejsilnější faktor do dalšího zpracování. Výsledkem 1. kroku je tedy model s jedním faktorem.

V druhém kroku budeme vytvářet několik modelů vícenásobné regrese, každý z nich bude obsahovat dva regresory. V každém z těchto modelů bude jedním ze dvou regresorů faktor Obvod pasu. Druhým regresorem budou postupně všechny faktory, jejichž modely byly v 1. kroku identifikované jako významné. Do části tabulky obsahující informace o tomto 2. kroku jsme si zaznamenali jednak p-hodnoty regresních koeficientů příslušných přidávaných regresorů, jednak p-hodnoty významnosti celého modelu.

Výsledky stepwise regrese - ženy

	1.krok	2.krok = pas+…		3.krok = pas+boky+…	
	jednoduch.	*p-hodnota*	*významnost*	*p-hodnota*	*významnost*
věk	0,0718902				
výška	0,29303342				
boty					
prsa	3,1714E-06	0,003808	2,985E-13	0,207638	1,2229E-14
pas	1,8295E-12				
boky	8,0392E-11	2,94E-05	3,071E-15		
krk	0,00098189	0,213433	9,3E-12		
sport	0,08071724				
spánek	1,7736E-06	0,006516	4,869E-13	0,063926	4,8609E-15
pc+tel	1,6632E-05	0,512528	1,64E-11		

Žlutě jsme vyznačili případy, kdy daný model nepřináší vylepšení modelu z předchozího kroku. Poznáme to jak na p-hodnotě daného koeficientu, a to tak, že tato hodnota je větší než 0,05. Také to poznáme z p-hodnoty významnosti daného modelu (např. u faktoru Krk je tato

hodnota $9{,}3*10^{-12}$) – ta je větší než p-hodnota významnosti modelu z předchozího kroku ($1{,}8295*10^{-12}$).

Ze zbylých modelů, tedy modelů, které přinášejí určité vylepšení, vybereme ten, který má p-hodnotu významnosti ze všech nejmenší, daný model je tedy nejvýznamnější. V našem případě je to modře vyznačená hodnota $3{,}071*10^{-15}$, která určuje, že dalším faktorem vstupujícím do modelu je veličina Boky.

Obdobně postupujeme i v 3. kroku. Zde vytváříme modely vícenásobné regrese obsahující tři regresory, dvěma z nich jsou vybrané faktory Obvod pasu a Obvod boků. Třetím faktorem jsou postupně všechny faktory, jejichž modely v 2. kroku byly významné (v našem případě to budou postupně faktory Obvod prsou a Množství spánku). Do tabulky si opět zaznamenáme dvě vybrané hodnoty pro každý z těchto modelů.

Jak vidíme, pro oba tyto modely platí, že nejsou významnější než model z 2. kroku (jejich p-hodnoty významnosti nejsou menší, respektive p-hodnoty regresních koeficientů jsou větší než 0,05).

Jako příklad ze všech výstupů jednotlivých regresí si uvedeme výstup závěrečného kroku stepwise regrese.

VÝSLEDEK

Regresní statistika	
Násobné R	0,808604
Hodnota spolehlivosti R	0,65384
Nastavená hodnota spolehlivosti R	0,642851
Chyba stř. hodnoty	6,097295
Pozorování	66

ANOVA

	Rozdíl	*SS*	*MS*	*F*	*Významnost F*
Regrese	2	4423,94	2211,97	59,49834	3,07E-15
Rezidua	63	2342,151	37,177		
Celkem	65	6766,091			

	Koeficienty	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*
Hranice	-78,7155	16,73245	-4,70436	1,44E-05	-112,153	-45,2783
boky	0,979094	0,217337	4,504949	2,94E-05	0,54478	1,413408
pas	0,698306	0,126248	5,531227	6,54E-07	0,44602	0,950592

Zeleně jsou vyznačeny důležité hodnoty. Jednak jsme vyznačili hodnotu indexu determinace, která nám říká, že výsledný model vystihuje variabilitu veličiny Váha z 65,38 %, tedy zbývajících 34,62 % popisují ještě jiné faktory.

Výsledný model má tedy tvar:

V = -78,7155 + 0,979094*B + 0,698306*P,

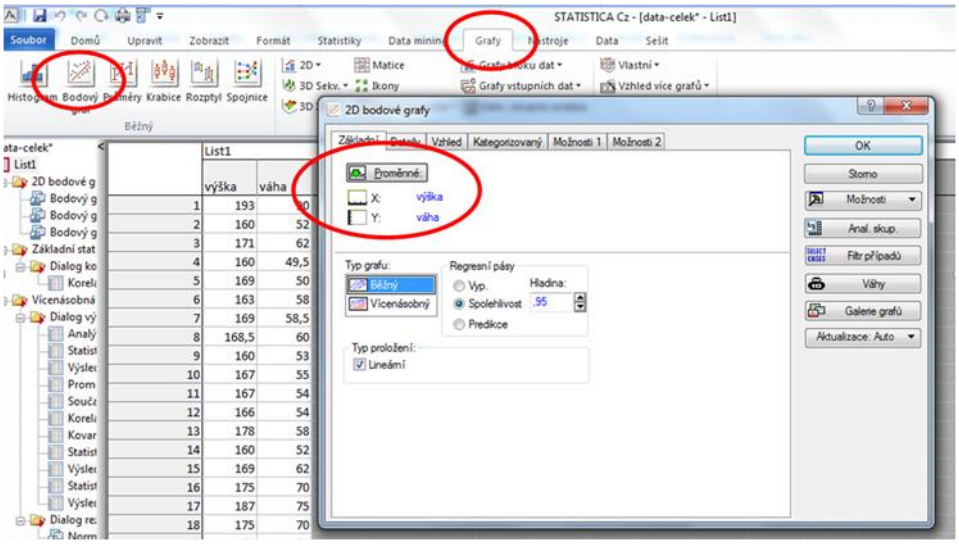
kde V je váha v kilogramech, B je obvod boků v centimetrech a P je obvod pasu v centimetrech.

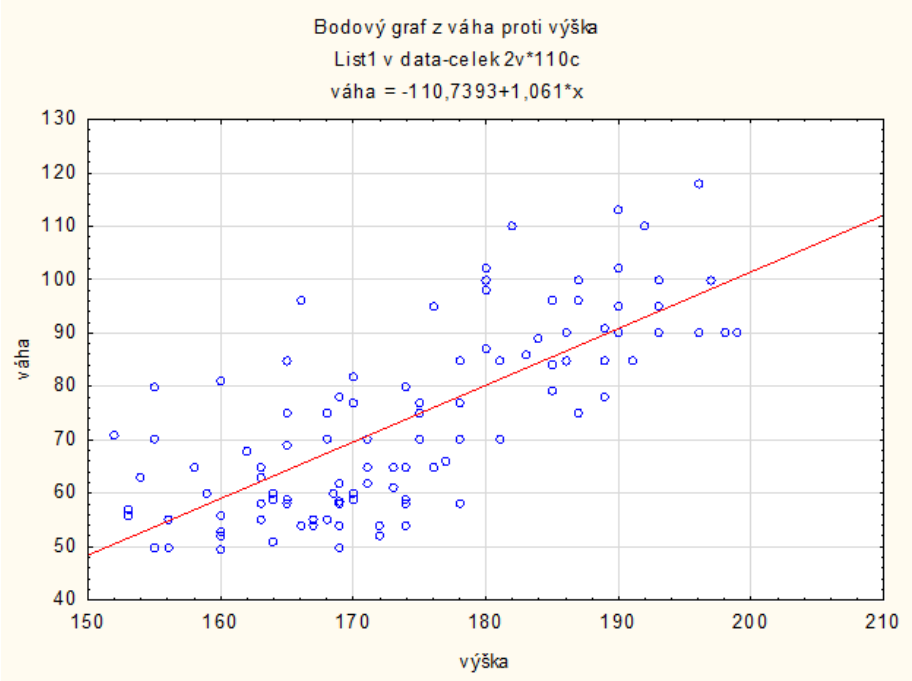
Vidíme, že pokud má žena v pase o 1 cm více, bude těžší průměrně o 0,698 kg a pokud bude mít žena přes boky o 1 cm více, bude těžší průměrně o 0,979 kg.

4.3 Regresní analýza v SW STATISTICA

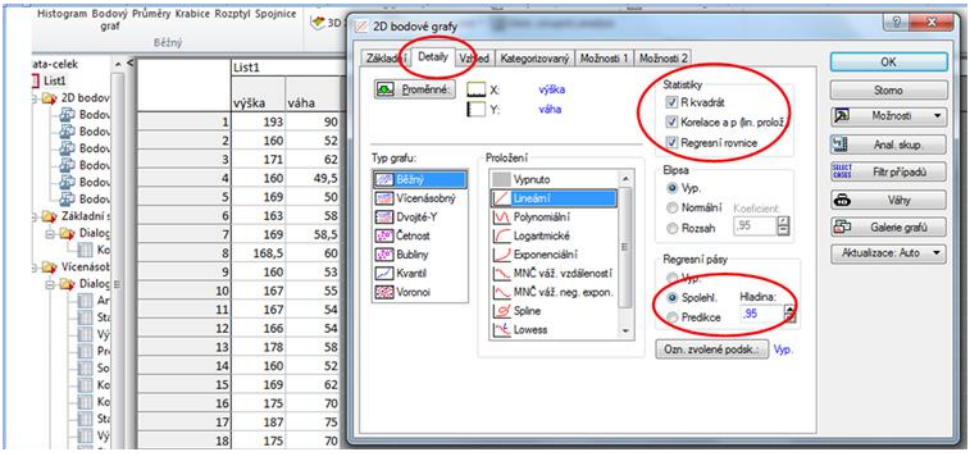
4.3.1 Grafické znázornění

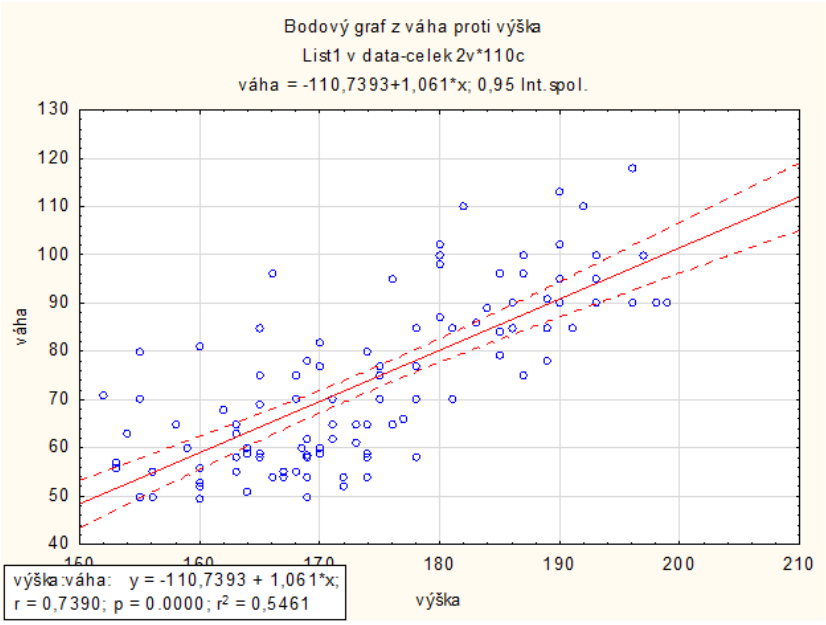
Pokud chceme provést regresní analýzu, vytvoříme si nejprve graf. V SW STATISTICA je pro tyto účely nejvhodnější bodový graf. V nabídce zvolíme *Grafy - Bodové grafy*. V dialogovém okně vybereme jako nezávislou proměnnou výšku a jako závislou proměnnou váhu. Zkontrolujeme, zda je zaškrtnuto pole *Typ proložení: Lineární*.



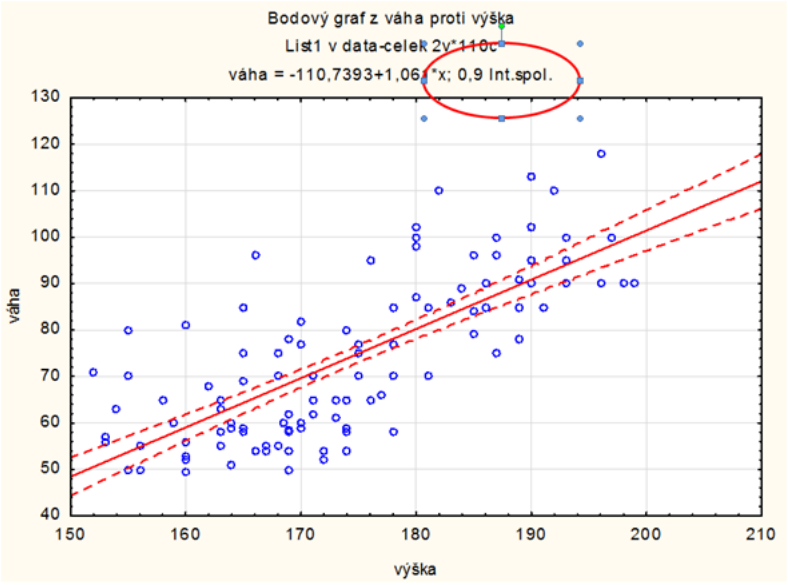


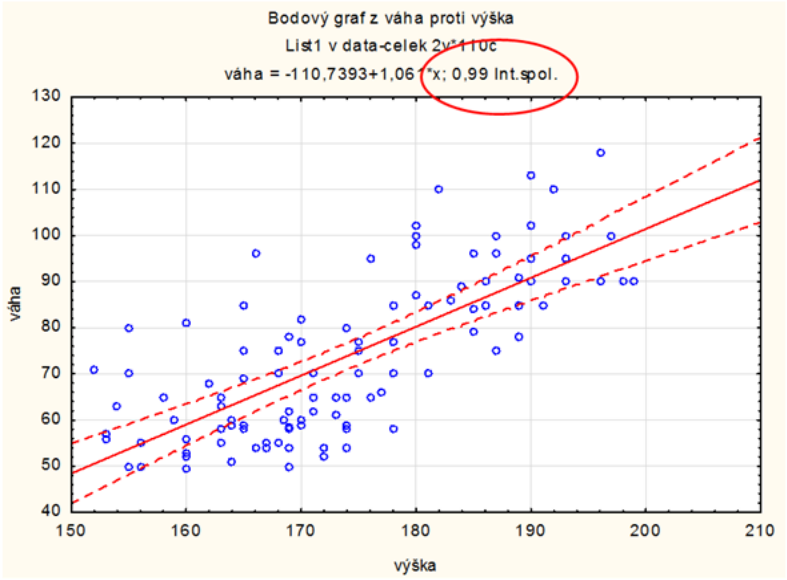
Pokud bychom chtěli mít graf s podrobnějšími informacemi, zvolíme záložku *Detaily* a nastavíme své požadavky. Mezi možnostmi je např. zobrazení regresních pásů. V tomto případě můžeme volit spolehlivost pro intervalové odhady. Volit můžeme mezi regresními pásy pro střední hodnotu (v dialogovém okně označeno *Spolehlivost* – naše volba), nebo regresní pásy pro individuální hodnotu (v dialogovém okně označeno *Predikce*). Zároveň můžeme zvolit možnost zobrazení regresní rovnice, indexu determinace, korelačního koeficientu a p-hodnoty významnosti modelu.



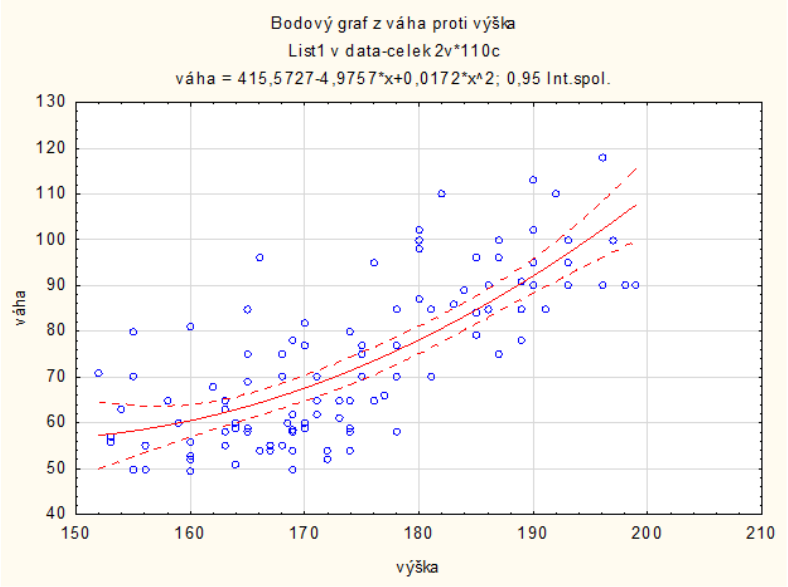


Z následujících výstupů je vidět, jak s rostoucí spolehlivostí vzrůstá i šíře těchto odhadů, tedy i zobrazených pásů.



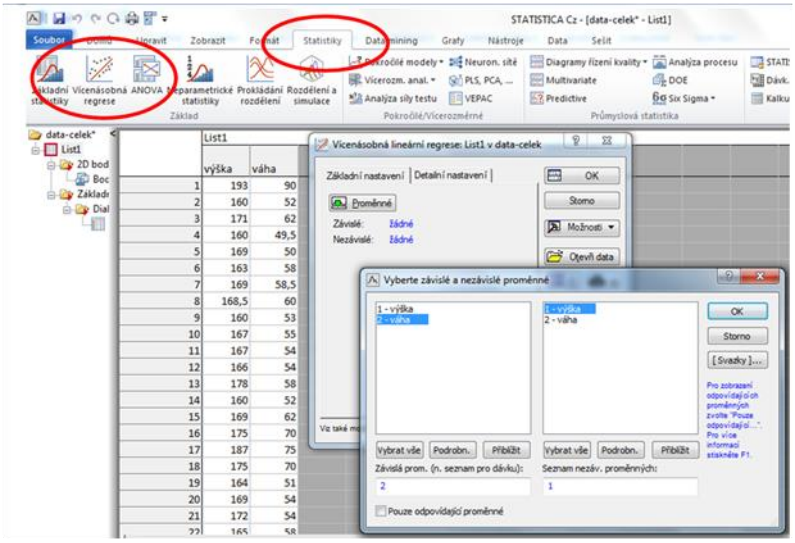


Případně můžeme změnit typ regresního modelu, např. na kvadratický.

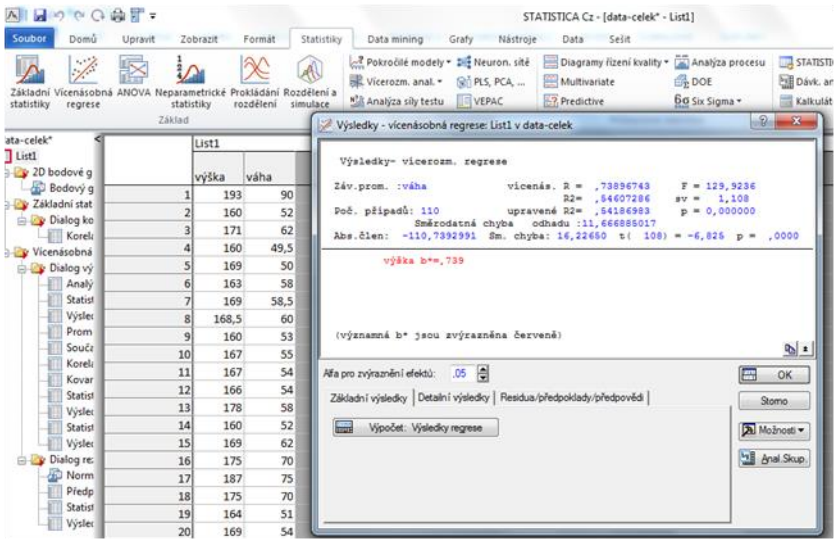


4.3.2 Jednoduchá regresní analýza – lineární model

Grafický výstup sám o sobě není dostačující pro určení modelu. V následujícím textu si tedy ukážeme, jak provést podrobnou regresní analýzu. K tomu vybereme v záložce *Statistiky* možnost *Vícenásobná regrese*. Opět zpracováváme závislost váhy na výšce u sledovaných osob. Jako závislou proměnnou vybereme veličinu Váha, nezávislou pak Výška. Pak v obou dialogových oknech zvolíme možnost *OK*.



Poté nám SW STATISTICA zobrazí následující výstup.



Výsledné okno obsahuje např. následující důležité informace.

Záv. prom. - obsahuje jméno závislé proměnné – v našem případě Výška.

Vícenás. R - hodnota koeficientu vícerozměrné korelace, což je odmocnina hodnoty R^2 (resp. I^2, neboli koeficientu determinace).

R2 – hodnota indexu determinace. Jak už jsme v teoretické části psali, tato veličina nám udává, jaký podíl celkové variability závisle proměnné je vysvětleno naším modelem.

Upravené R2 – hodnota upraveného indexu determinace, který bere do úvahy také počet regresorů zahrnutých v modelu.

F – hodnota testového kritéria týkajícího se testu významnosti celého modelu

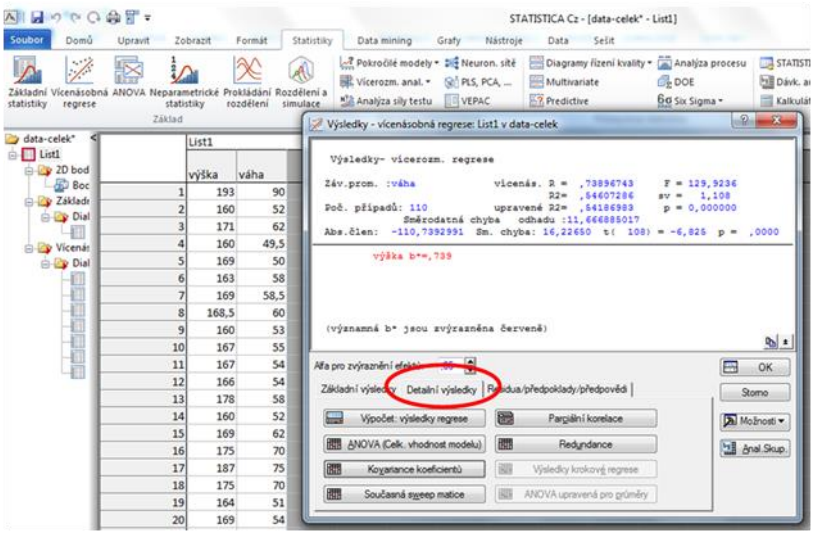
P - odpovídající p-hodnota, tedy opět slouží k vyhodnocení významnosti celého modelu

Směrodatná chyba odhadu - představuje míru rozptýlení pozorovaných hodnot okolo regresní přímky.

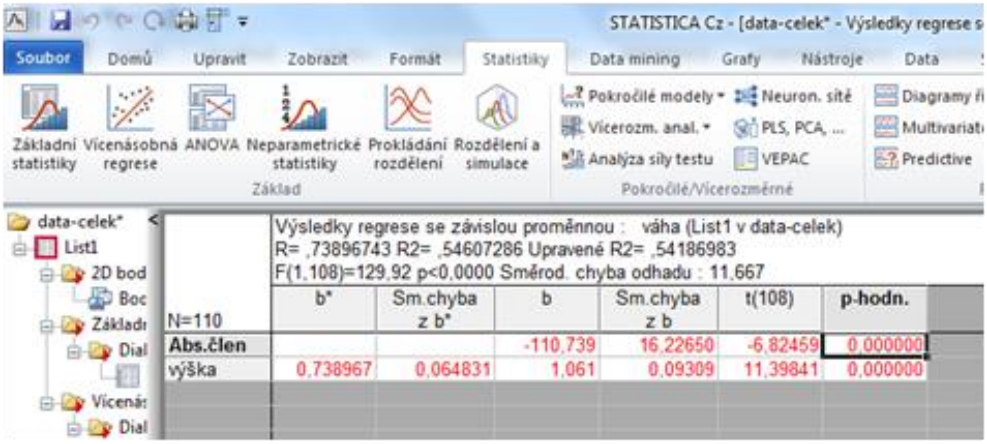
Abs. člen. - obsahuje odhad b_0 (tj. absolutního členu) regresní rovnice.

*Výška b** - koeficient(y) nezávisle proměnné (resp. proměnných). Toto však neodpovídá odhadům parametrů z uvažovaného regresního modelu. Jedná se o speciálně upravené odhady parametrů z jiného modelu, které nám umožňují porovnat relativní vliv jednotlivých regresorů na závisle proměnnou. Statisticky významné regresní koeficienty jsou zvýrazněny červenou barvou.

Pro podrobnější informace zvolíme záložku *Detailní výsledky.*

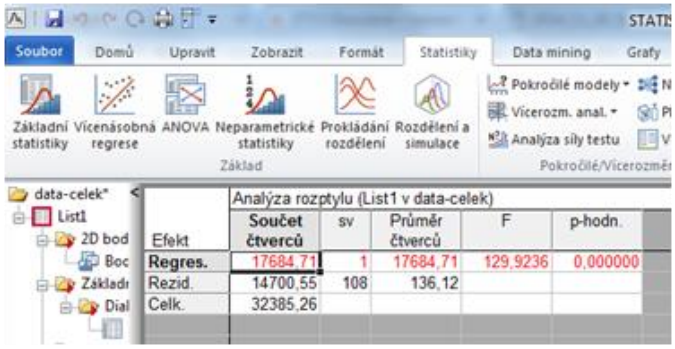


Souhrnné výsledky regresní analýzy obdržíme zvolením možnosti *Výpočet, výsledky regrese*



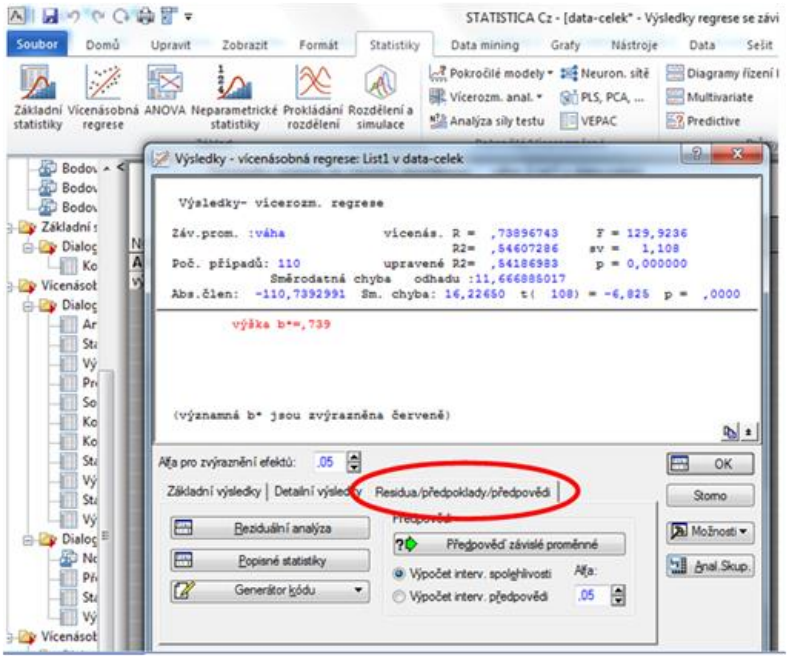
V této tabulce již obdržíme všechny koeficienty tak, jak očekáváme, jak známe např. z Excelu. Můžeme tedy určit výsledný tvar regresní rovnice: $y = -110{,}739 + 1{,}061x$

Tabulku ANOVA obsahující informace o významnosti modelu obdržíme volbou *ANOVA (celk. vhodnost modelu)*.

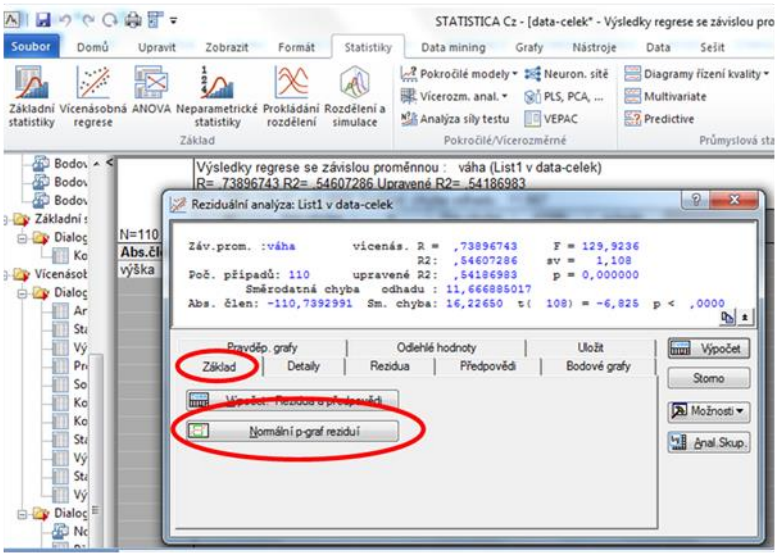


Zde opět vidíme jednak hodnotu testového kritéria a p-hodnotu, na jejichž základě můžeme vyhodnotit významnost modelu. V našem případě vidíme, že p-hodnota je po zaokrouhlení na šest desetinných míst rovna nule (uvědomme si však, že nule rovna ve skutečnosti není), tedy je menší než hladina významnosti (ať už 5% nebo 10%). Znamená to, že námi vytvořený model je významný.

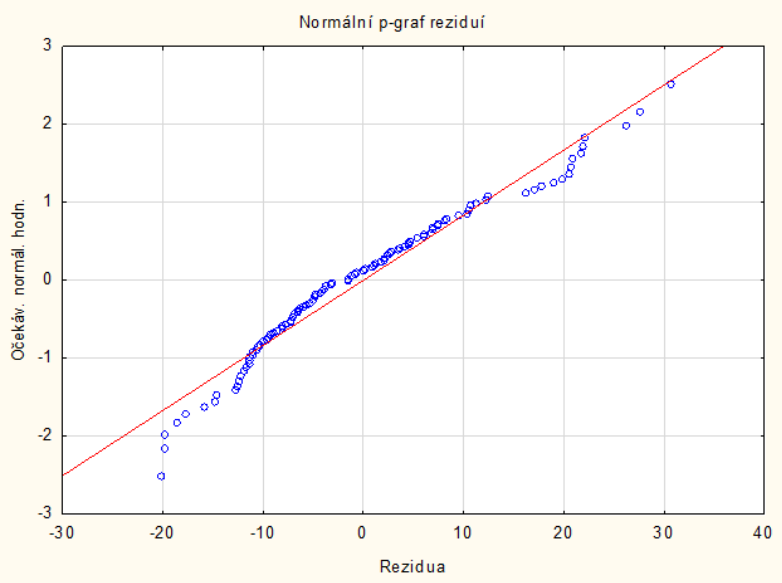
Pro podrobnější vyhodnocení vhodnosti modelu provedeme ještě verifikaci chování reziduí. K tomuto účelu zvolíme záložku *Rezidua/předpoklady/předpovědi*.



Asi nejjednodušším způsobem zjištění, zda se rezidua chovají „rozumně“, tedy zda mají normální rozdělení, je vytvoření Normálního p-grafu reziduí. Toto provedeme výběrem možnosti *Reziduální analýza* v zobrazeném dialogovém okně a následně výběrem možnosti *Normální p-graf reziduí* na záložce *Základní*.



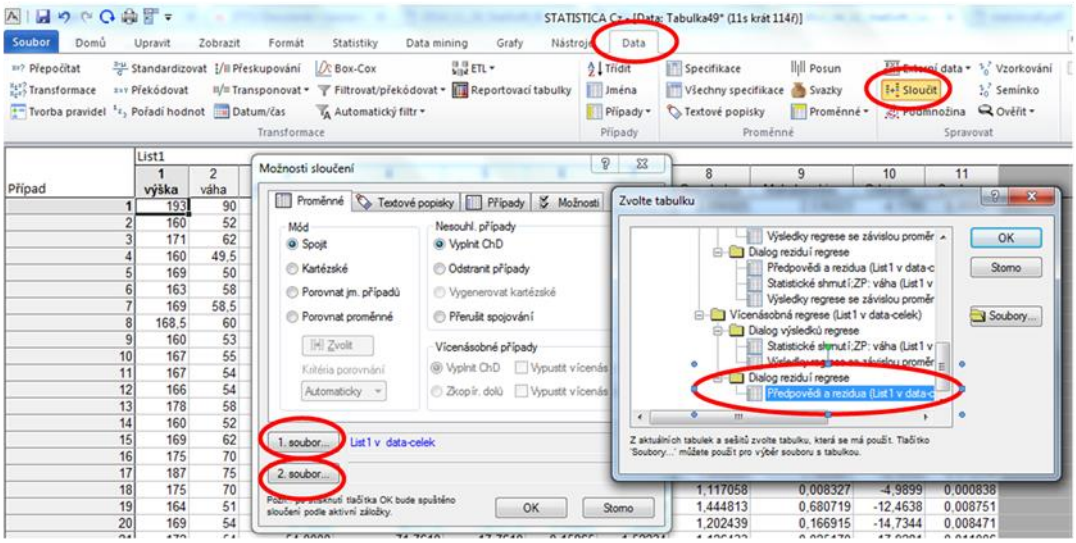
Rezidua by měla mít normální rozdělení, což znamená, že v zobrazeném grafu by měla ležet co nejblíže přímky.



V našem případě vidíme, že se body ve spodní části grafu od přímky relativně významně odchylují, tedy normalita není zcela splněna, stav je minimálně hraniční.

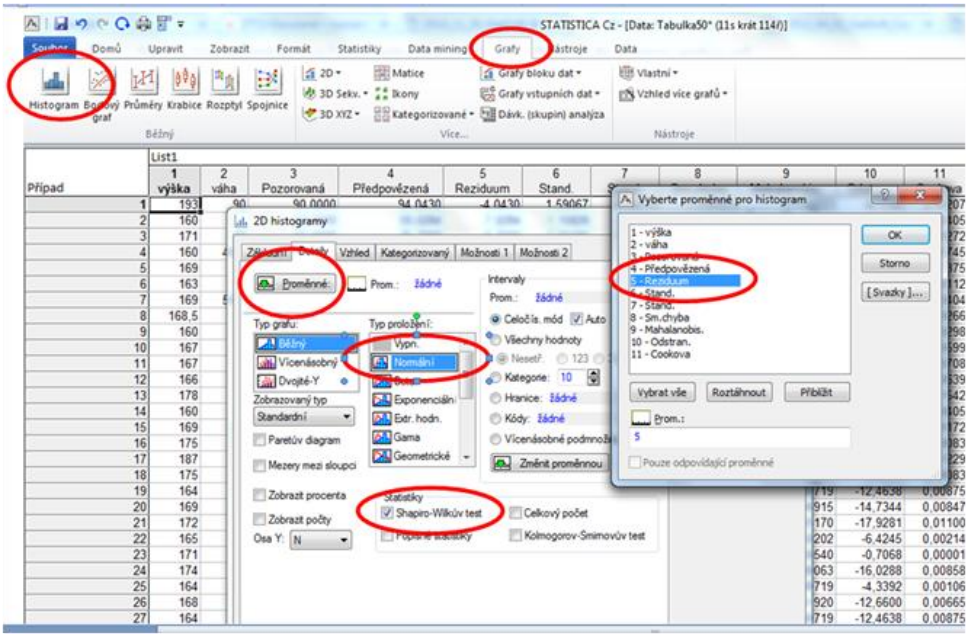
Pro podrobnější vyhodnocení ještě můžeme vytvořit histogram reziduí. K tomuto účelu si musíme sloučit soubory, a to jednak původní soubor a dále soubor reziduí. Ten získáme pomocí možnosti *Předpověď závislé proměnné* v záložce *Rezidua/předpoklady/předpovědi*.

Vlastní sloučení pak provedeme přes záložku *Data*, kde v nabídce vybereme možnost *Sloučit*. V následujícím dialogovém okně pak pomocí tlačítek *1. soubor* a *2. soubor* postupně vybereme oba slučované soubory. Na následujícím obrázku je znázorněn výběr souboru s rezidui.

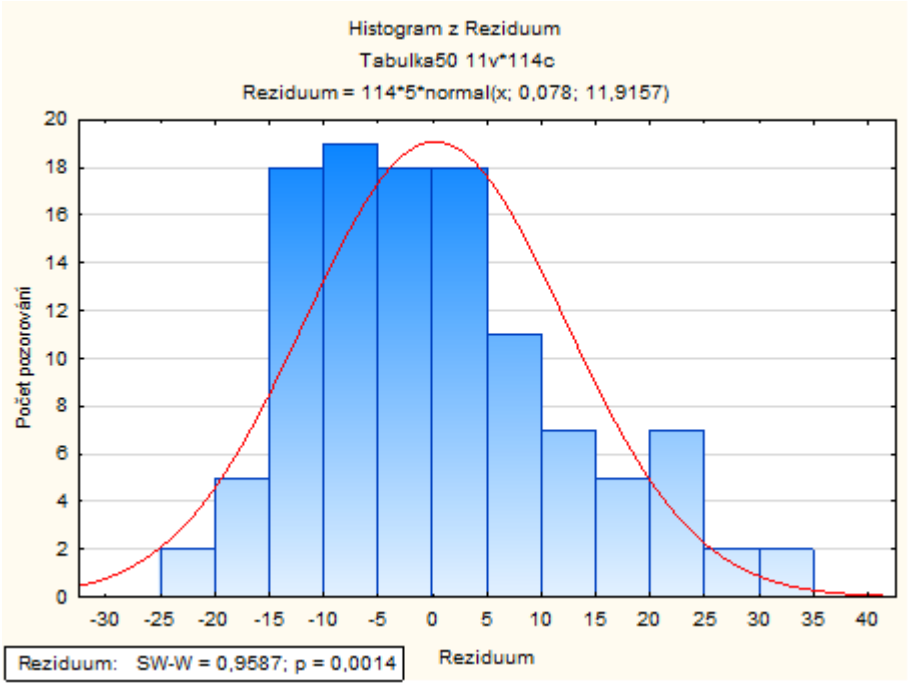


Následně nám vznikne a zobrazí se soubor, který jednak obsahuje sloupce našeho zpracovávaného souboru a jednak sloupce ze souboru obsahujícího informace výsledcích předpovědí a o reziduích.

Na základě tohoto souboru vytvoříme histogram reziduí. K tomuto účelu v záložce *Grafy* vybereme v nabídce možnost *Histogram*. V následném dialogovém okně pak vybereme proměnnou pro histogram. Touto proměnnou bude veličina Reziduum. Dále si vybereme *Typ proložení Normální* a můžeme též zaškrtnout možnost *Shapiro-Wilkův test* pro vyhodnocení normality reziduí.



Jak z grafického zobrazení, tak z p-hodnoty Shapiro-Wilkova testu je vidět, že rezidua v našem případě normální rozdělení nemají, což není na závadu při odhadování regresních koeficientů a můžeme tedy říci, že regresní rovnice je v pořádku (významnost modelu byla potvrzena). Nemůžeme se však zcela spolehnout na významnost regresních koeficientů a především na správnost intervalů spolehlivosti. Toto bychom se měli snažit odstranit, např. vhodnou transformací dat, nebo zjištěním a odstraněním odlehlých pozorování. Ale těchto možných příčin a tím i řešení je mnoho.

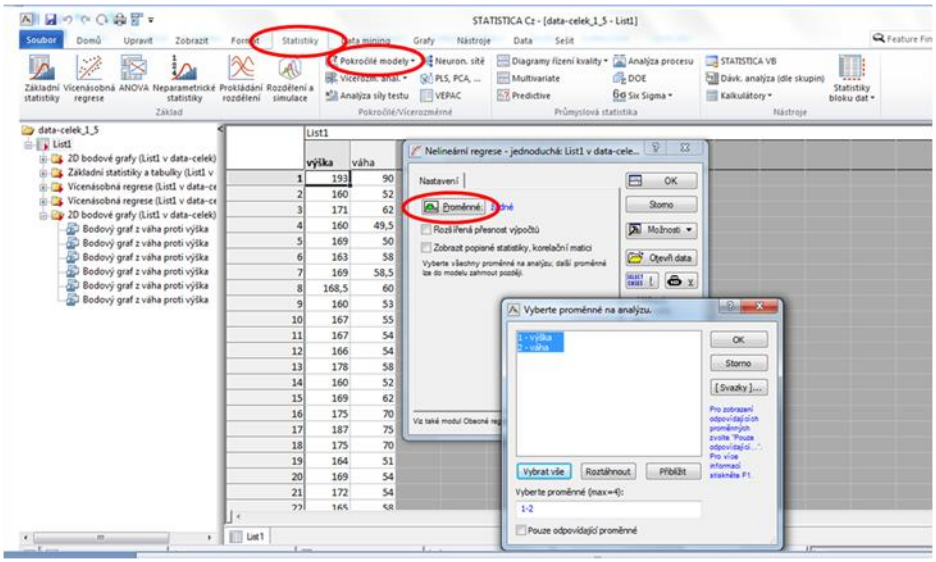


4.3.3 Jednoduchá regresní analýza – obecný model

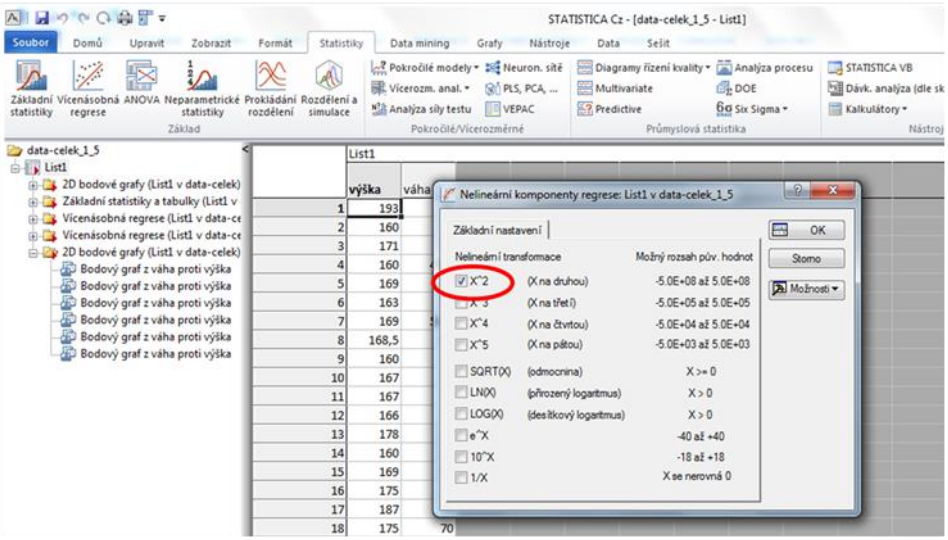
Ukázali jsme si jeden ze způsobů, jak je v SW STATISTICA vytvořit a vyhodnotit lineární regresní model vyjadřující vztah mezi dvěma veličinami. Jak už jsme v úvodní teoretické části tak i v části o zpracování pomocí Excelu, ne vždy je však lineární model jediný možný a nejlepší.

Ukážeme si nyní, jak bychom v SW STATISTICA vytvoříme takovýto model. Postup si ukážeme například na kvadratickém modelu. Grafický způsob jsme si již ukázali, nyní provedeme detailní regresní analýzu.

Opět se vrátíme k listu obsahujícímu data o všech studentech bez rozdílu pohlaví. V záložce *Statistiky* vybereme možnost *Pokročilé modely* a v otevřeném seznamu pak vybereme *Nelineární regrese – jednoduchá*. V zobrazeném dialogovém nastavíme proměnné regresního modelu.



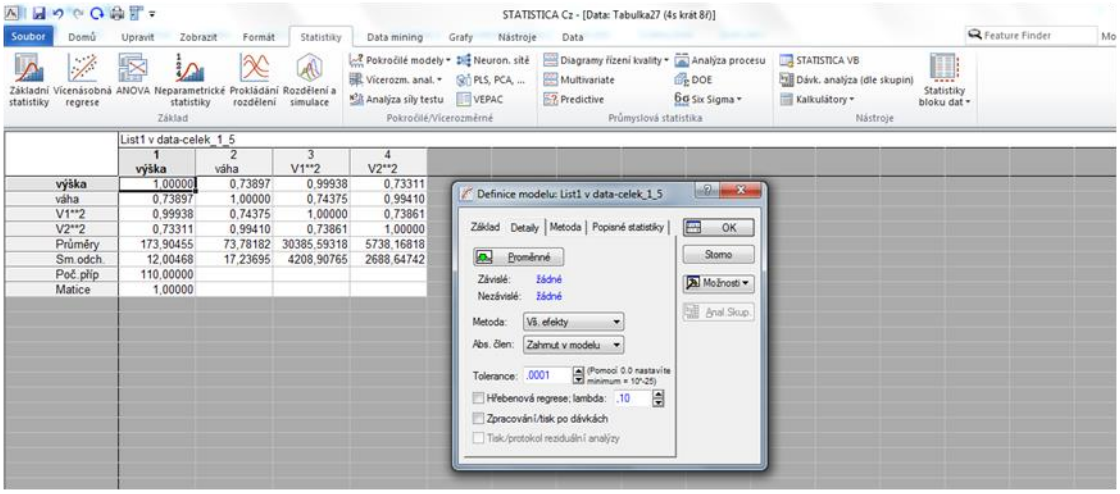
V obou oknech pak zadávání potvrdíme stisknutím klávesy *OK*, následně se otevře okno *Nelineární komponenty regrese*.



V tomto okně volíme z nabídky nelineárních transformací. V našem případě je cílem vytvořit kvadratický model, proto zvolíme transformaci X^2.

Kdybychom chtěli vytvořit polynomický model 3. stupně, zvolili bychom X^3, v případě, že bychom tímto způsobem tvořili lineární model, ignorujeme transformace a stiskneme OK ihned.

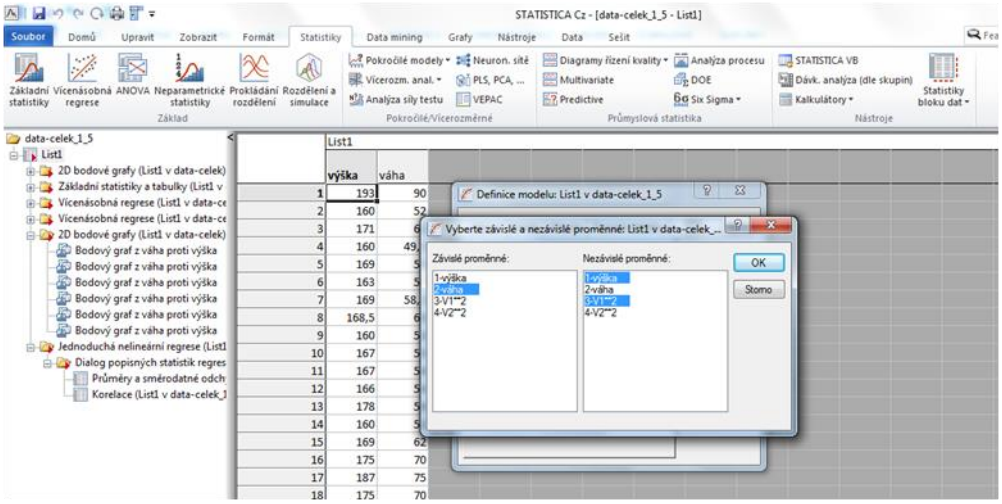
Následně se nám zobrazí dialogové okno *Definice modelu*. Nastává jedna z nejobtížnějších a nejméně přehledných, přitom velmi důležitých etap naší práce, a to výběr proměnných.



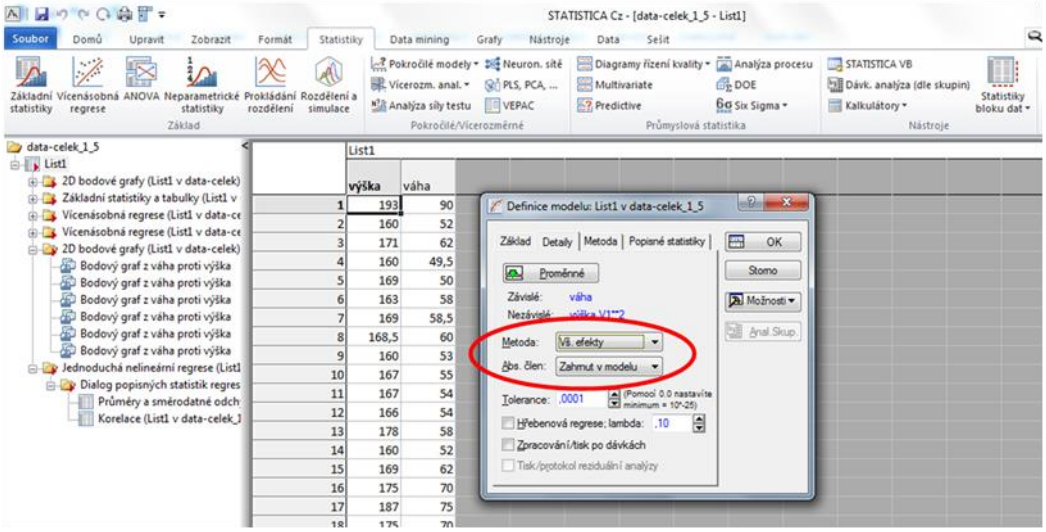
Jakmile stiskneme tlačítko *Proměnné*, objeví se nabídka všech variant výběru závisle i nezávisle proměnných.

Nevýhodou je, že transformované proměnné jsou označeny pořadím (např. V1**2 či V2**2), nikoli jménem (např. výška**2 nebo váha**2). Musíme být tedy velmi opatrní. V našem případě požadujeme, aby závislou proměnnou v modelu byla proměnná Váha a do pozice nezávislých proměnných se dostanou veličiny Výška a druhá mocnina výšky, tedy proměnná V1**2.

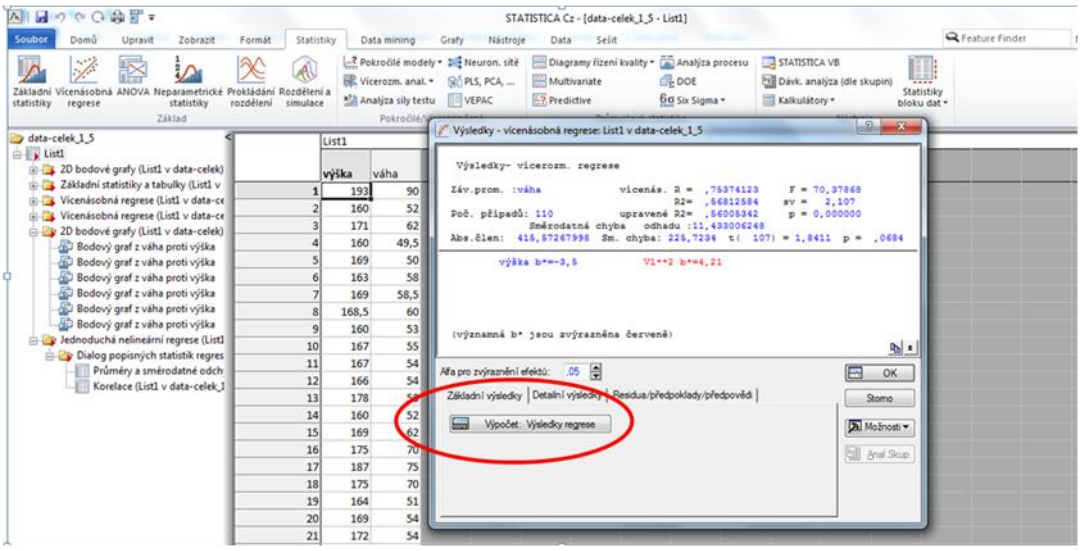
Můžeme si uvědomit paralelu s maticí F z maticového vzorce pro výpočet regresních koeficientů (viz teoretický úvod) či s přidávanými sloupci při tvorbě kvadratického modelu pomocí doplňku Analýza dat v Excelu (viz kapitola o zpracování regrese v Excelu).



Výběr proměnných potvrdíme stisknutím *OK*. Vrátíme se do okna *Definice modelu*.



Zde ještě zkontrolujeme, zda je nastavena metoda *Vš. efekty* a zda je absolutní člen zahrnut v modelu. Poté provedeme výpočet stiskem *OK*. Otevře se okno *Výsledky – vícenásobná regrese*. V záložce *Základní výsledky* zvolíme tlačítko *Výpočet: Výsledky regrese*.



Zobrazí se následující tabulka. V její horní části je řada užitečných informací. Pro nás jsou důležité hodnoty R (korelační koeficient) a R2 (index determinace), podle toho, zda pracujeme s lineárním či nelineárním modelem.

N=110	Výsledky regrese se závislou proměnnou : váha (List1 v data-celek_1_5) R= ,75374123 R2= ,56812584 Upravené R2= ,56005342 F(2,107)=70,379 p					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(107)	p-hodn.
Abs.člen			415,5727	225,7234	1,84107	0,068382
výška	-3,46534	1,799775	-4,9757	2,5842	-1,92543	0,056830
V1**2	4,20693	1,799775	0,0172	0,0074	2,33747	0,021275

Dále si v této tabulce budeme především všímat sloupce b, který obsahuje jednotlivé regresní koeficienty. V prvním sloupci poznáme příslušnost jednotlivých koeficientů. Náš model má tedy tvar:

$y = 415{,}5727 - 4{,}9757x + 0{,}0172x^2$ s indexem determinace $I^2 = 0{,}5681$.

V posledním sloupci jsou uvedeny p-hodnoty významnosti jednotlivých koeficientů. Nejdůležitější je poslední uvedená hodnota náležející ke kvadratickému členu. Kdyby tento člen nebyl významný, nejednalo by se totiž o kvadratický model. V našem případě vidíme, že p-hodnota je 0,02, tedy koeficient významný je. Zbývající dvě p-hodnoty se rovnají 0,068, resp. 0,057, tedy příslušné koeficienty významné nejsou. Pro nás z toho plyne, že model lze zjednodušit tím, že vypustíme absolutní, resp. lineární člen.

Vsuvka:

To bychom provedli návratem do lišty *Výsledky-vícerozměrné*. Ta je zobrazena v levé spodní části obrazovky. Objeví se opět okno *Výsledky vícerozměrné regrese: Tabulka*, zde stiskneme tlačítko *Storno*. Vrátíme se tak do okna *Definice modelu*.

Vyloučení lineárního členu bychom provedli v tomto okně změnou výběru proměnných vstupujících do modelu, a to tak, že bychom odebrali nezávislou proměnnou Výška a ponechali pouze nezávislou proměnnou V1**2.

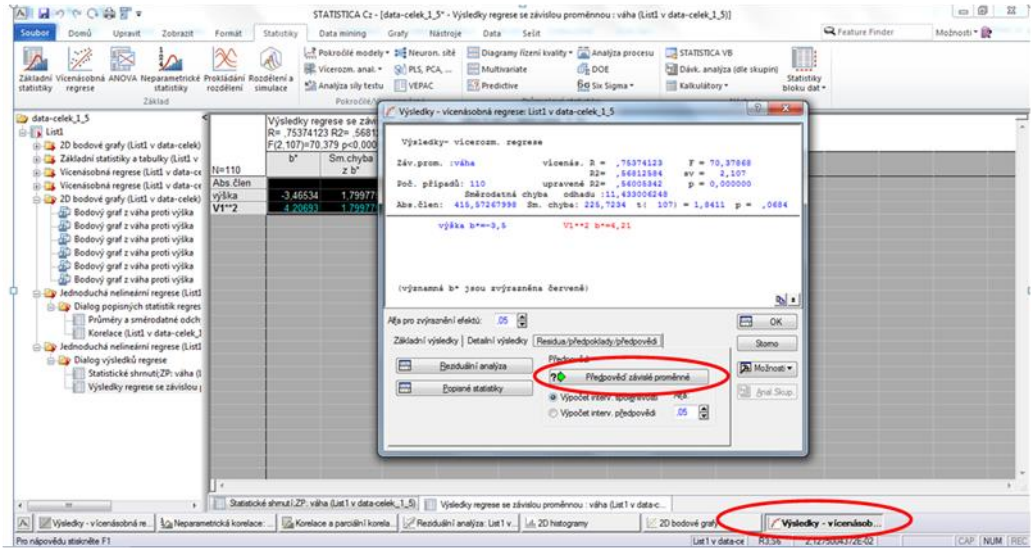
Vyloučení absolutního členu bychom provedli v okně *Definice modelu*, kde v záložce *Detaily* rozbalíme položku *Abs. člen* a vybereme nabídku *Nastaven na 0*. Klikneme na *OK*.

Provede se nový výpočet. Před tím však budeme upozorněni, že nelze srovnávat R^2 původní výstupní sestavy s hodnotou R^2 v sestavě zjednodušené. Je totiž počítán podle jiného vzorce. Nová výstupní sestava již nebude mít absolutní člen.

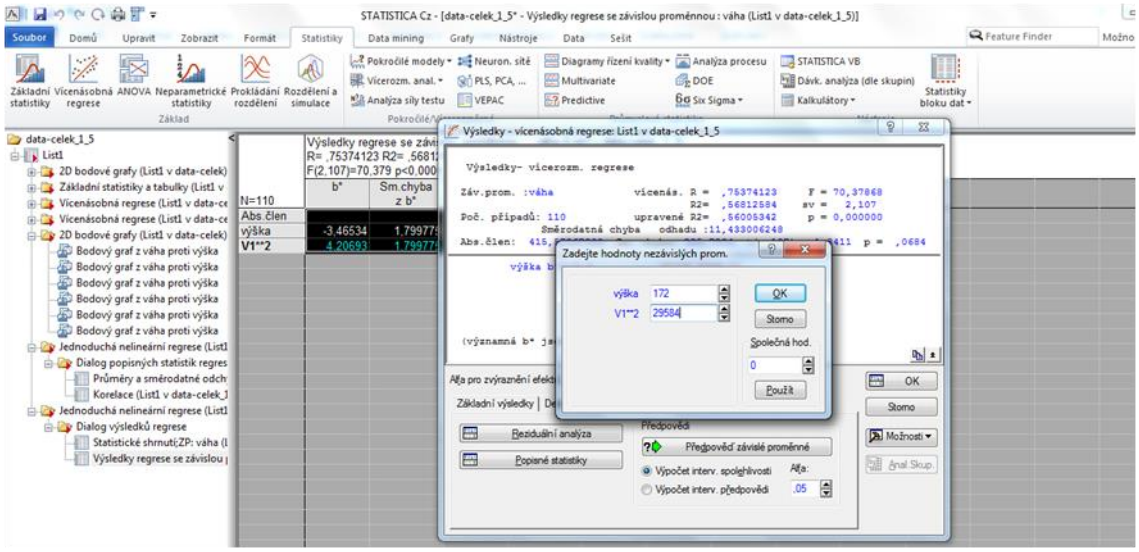
Tento krok však není nezbytně nutný.

Predikce

Predikci umožní provést nastavení záložky Residua/předpoklady/předpovědi v okně *Výsledky –vícerozměrná regrese.*



V zobrazeném okně vyplníme hodnotu Výšky, pro kterou nás zajímá předpověď Váhy. Nevýhodou je, že musíme vyplnit i druhou mocninu této hodnoty.



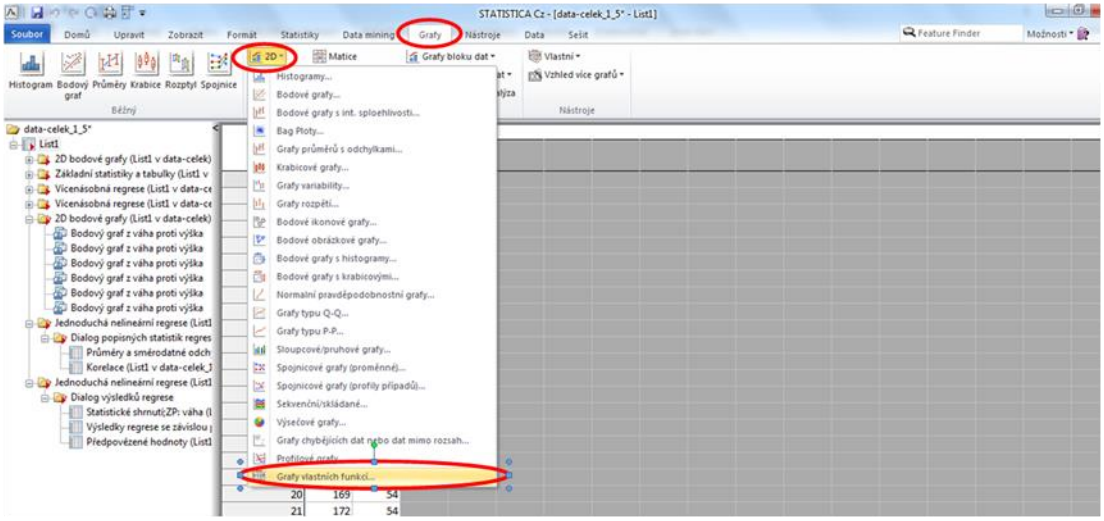
Výsledkem je následující tabulka:

Proměnná	Předpovězené hodnoty (List1 v data-celek_1_5) proměnné: váha		
	b-váha	Hodnota	b-váha * Hodnot
výška	-4,97572	172,00	-855,823
V1**2	0,01723	29584,00	509,698
Abs. člen			415,573
Předpověď			69,448
-95,0%LS			66,509
+95,0%LS			72,387

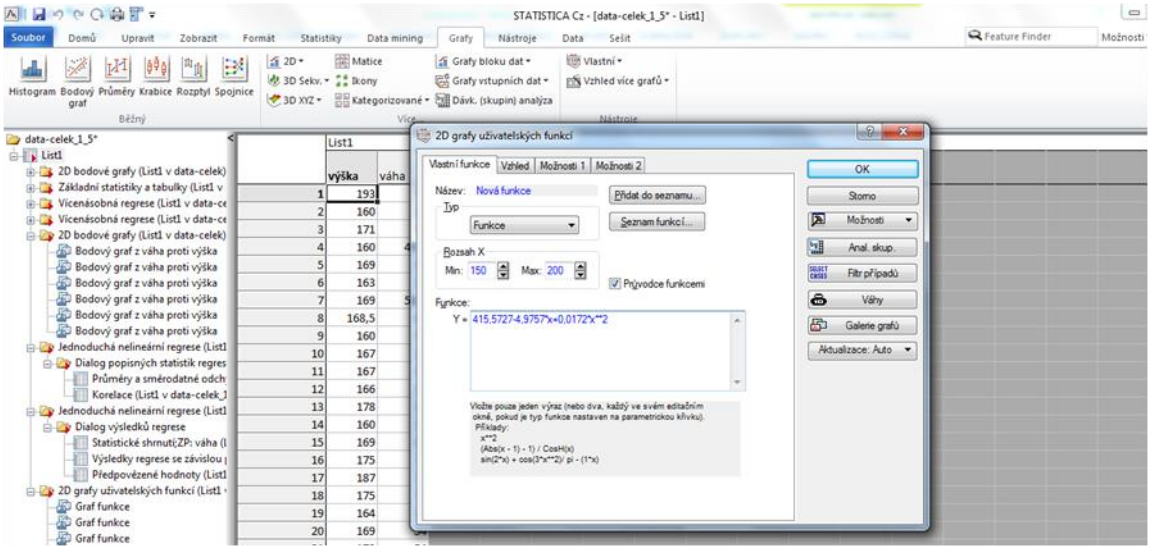
V této tabulce jsou pro nás nejdůležitější tři žlutá pole. Vidíme, že student/ka měřící 172 cm by průměrně měl/a vážit 69,45 kg. 95% intervalovým odhadem je rozmezí od 66,51 do 72,39 kg.

Grafické znázornění nalezeného regresního modelu

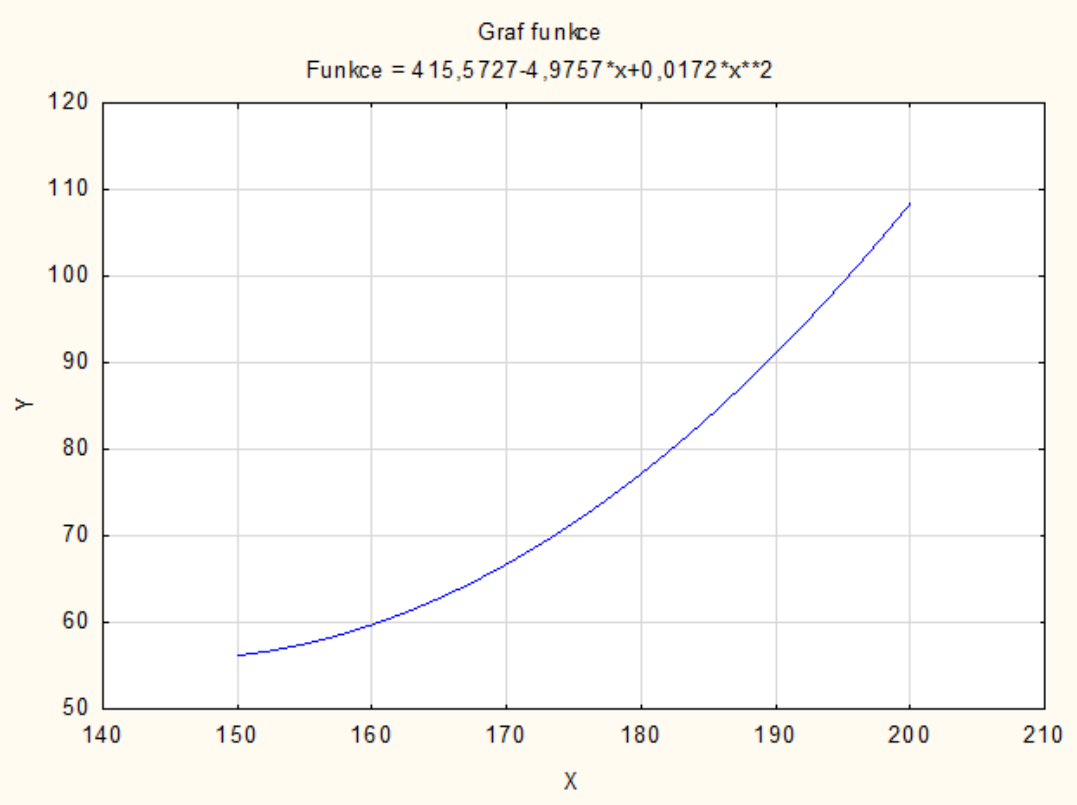
Známe-li předpis pro regresní model, můžeme si model graficky zobrazit. V záložce *Grafy*, zvolíme nabídku *2D grafy* a v něm nabídku *Grafy vlastních funkcí.*



V okně, které se otevře, zvolíme *Rozsah X* (tedy rozsah naší závisle proměnné), tj. nastavíme políčka *Min.* , *Max.* a předepíšeme tvar funkce. Zadání potvrdíme stisknutím *OK*.



Následně se nám zobrazí graf, který můžeme upravovat dle vlastních potřeb (viz kapitola o grafech)

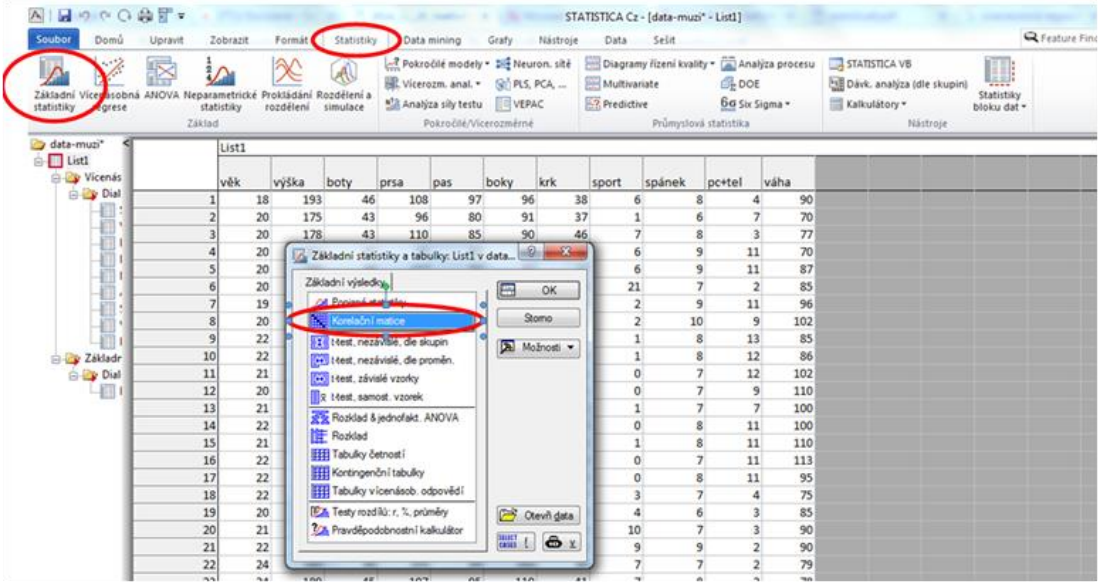


4.3.4 Vícenásobná regresní analýza pomocí SW STATISTICA

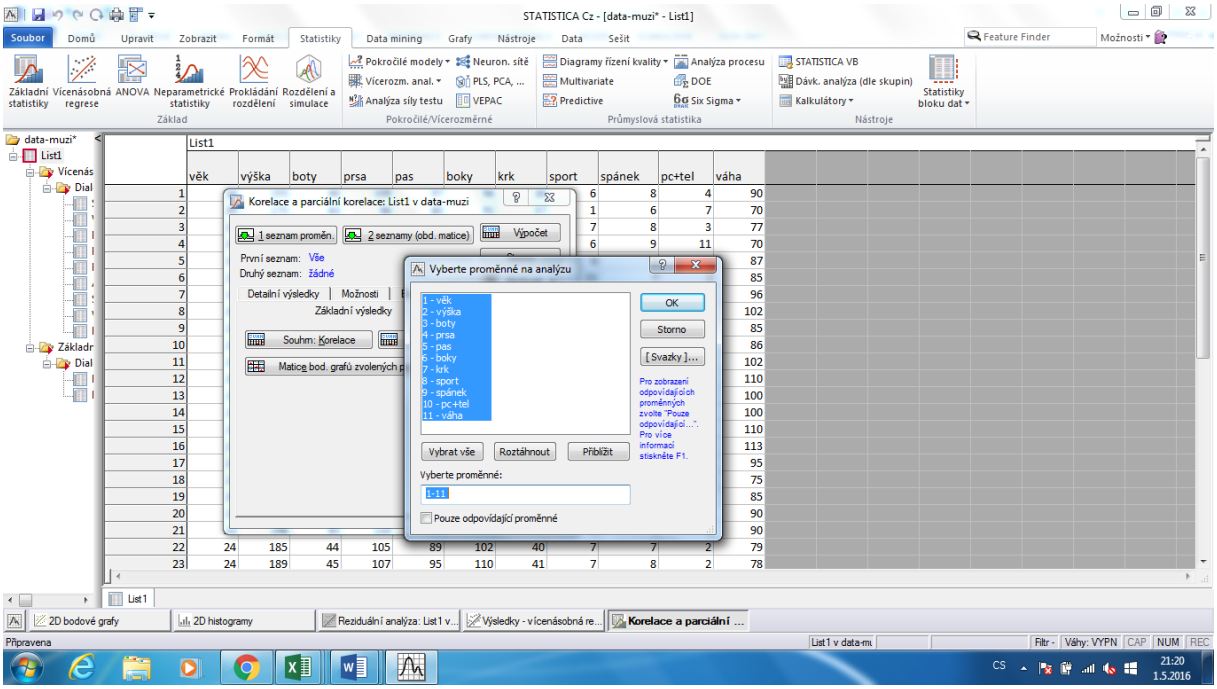
Nyní si ukážeme zpracování vícenásobné regresní analýzy v SW STATISTICA. V této části budeme zkoumat vliv různých faktorů (věk, výška, množství spánku, množství času u TV a PC, …) na váhu u mladých mužů, přesněji řečeno u studentů středních škol.

Jak jsme si již v teoretické části i části zpracování pomocí Excelu řekli, nejdříve bychom měli zkontrolovat, zda jednotlivé faktory nejsou mezi sebou silně korelované. Tento fakt by mohl narušit kvalitu výsledného regresního modelu a analýzu síly vlivu jednotlivých faktorů.

K tomuto účelu zvolíme vytvoření korelační matice obsahující korelační koeficienty mezi jednotlivými faktory, a to tak, že v záložce *Statistiky* vybereme nabídku *Základní statistiky* a zde ze seznamu vybereme možnost *Korelační matice*. Pak už jen svůj výběr potvrdíme tlačítkem *OK*.



V následujícím dialogovém okně pak vybereme seznam proměnných, mezi kterými nás zajímá vzájemná korelovanost.



Po potvrzení našeho výběru pak v původním okně vybereme možnost *Výpočet*. Obdržíme korelační matici v následujícím tvaru. Jak vidíme, tato matice je obohacena o dva sloupce, které obsahují průměr a směrodatnou odchylku dané veličiny.

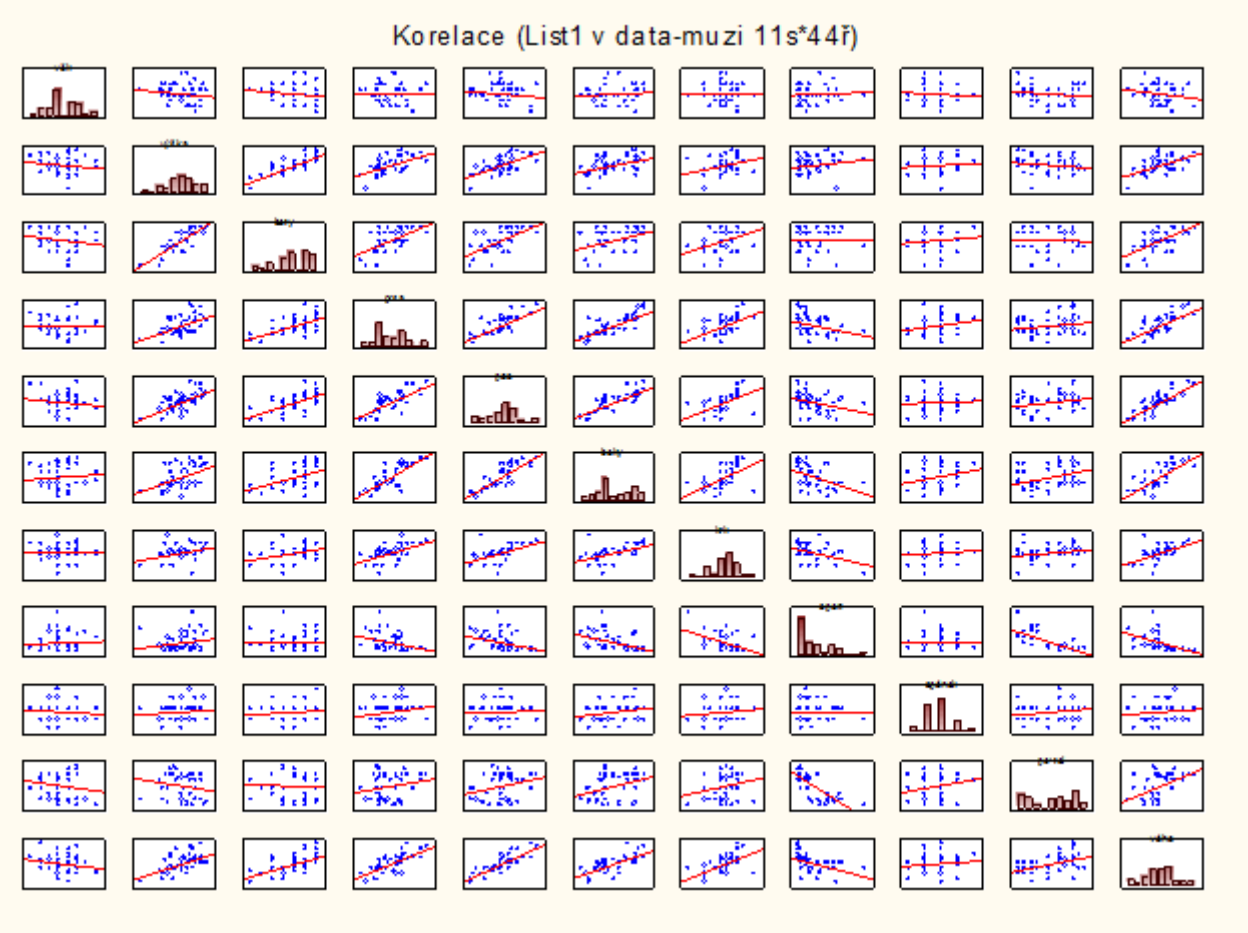
Korelace (List1 v data-muzi) Označ. korelace jsou významné na hlad. p < ,05000 N=44													
	Průměry	Sm. odch.	věk	výška	boty	prsa	pas	boky	krk	sport	spánek	pc+tel	váha
věk	20,34	1,57	1,00	-0,15	-0,17	-0,03	-0,15	0,10	0,01	0,06	-0,09	-0,17	-0,21
výška	185,55	7,57	-0,15	1,00	0,82	0,54	0,63	0,42	0,32	0,19	0,10	-0,17	0,52
boty	44,07	1,61	-0,17	0,82	1,00	0,59	0,61	0,42	0,37	0,00	0,11	-0,06	0,59
prsa	103,52	8,16	-0,03	0,54	0,59	1,00	0,73	0,78	0,56	-0,34	0,22	0,20	0,71
pas	94,93	7,26	-0,15	0,63	0,61	0,73	1,00	0,78	0,56	-0,32	0,06	0,26	0,82
boky	106,05	11,20	0,10	0,42	0,42	0,78	0,78	1,00	0,53	-0,43	0,24	0,36	0,73
krk	40,66	2,28	0,01	0,32	0,37	0,56	0,56	0,53	1,00	-0,44	0,12	0,23	0,60
sport	4,64	4,79	0,06	0,19	0,00	-0,34	-0,32	-0,43	-0,44	1,00	-0,01	-0,75	-0,45
spánek	7,70	0,88	-0,09	0,10	0,11	0,22	0,06	0,24	0,12	-0,01	1,00	0,21	0,12
pc+tel	7,50	3,48	-0,17	-0,17	-0,06	0,20	0,26	0,36	0,23	-0,75	0,21	1,00	0,48
váha	90,30	11,43	-0,21	0,52	0,59	0,71	0,82	0,73	0,60	-0,45	0,12	0,48	1,00

V zobrazené tabulce jsou statisticky významné korelace označeny červeně. Poslední řádek a sloupec obsahuje údaje týkající se veličiny Váha, což je naše vysvětlovaná proměnná, nikoliv regresor, proto nás v tomto kroku příliš nezajímá. Pokud bychom chtěli mít analýzu a vývody „dokonale přesné“, měli bychom si všímat všech červených hodnot. Většina učebnic však uvádí kritérium, že pro násobnou regresi jsou problémové hodnoty v absolutní hodnotě nad 0,8. Vidíme, že takovouto hodnotu má pouze korelační koeficient mezi veličinami Boty a Výška. Ze znalosti problematiky asi většina lidí usoudí, že spíše závisí velikost bot na výšce, než obráceně, proto by asi bylo nejlepší veličinu Boty z další analýzy vypustit.

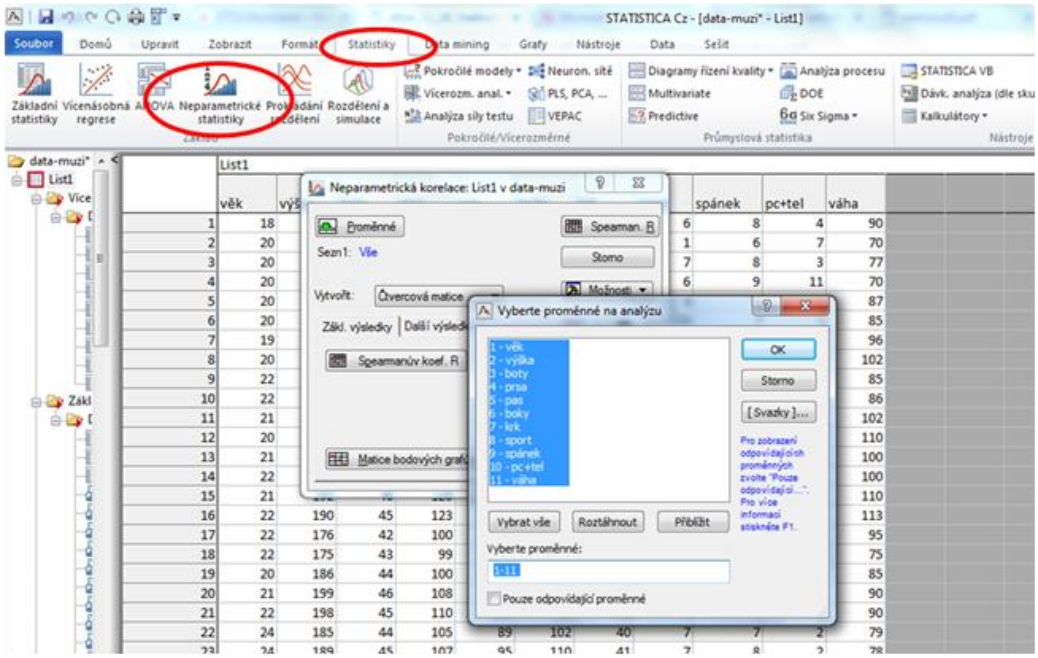
Jednou z nejjednodušších ale i nejméně přesných možností vyhodnocení vhodnosti použití násobné regrese je grafické zobrazení vzájemných vztahů jednotlivých veličin.

Na diagonále grafické matice vidíme histogramy jednotlivých veličin. I nezkušeným okem můžeme vidět, že se většina z nich normálním rozdělením příliš neřídí. V ostatních políčkách matice vidíme bodové grafy popisující vztahy daných veličin s proloženými regresními přímkami. Zde vidíme (tak, jak už jsme přesněji viděli i v korelační matici), že na většině obrázků body příliš přímku nekopírují, že tedy korelace příliš silné nejsou (připomeňme, že si nemáme všímat posledního sloupce a řádku náležícího veličině Váha). Naopak, čeho si v posledním řádku můžeme všimnout, že nejsou zřejmé jiné než lineární závislosti (možná až na jedinou výjimku – viz osmý obrázek v posledním řádku), což by také mohlo být překážkou bezproblémového použití vícenásobné regrese.

Takže vidíme, že nejzásadnějším problémem je chování veličin odlišné od normálního rozdělení. Nejvýraznější je tento fakt v osmém řádku, kde je histogram silně asymetrický. Osmá veličina, což je veličina vyjadřující počet hodin strávených sportem, je tedy problematická ze dvou důvodů. To však neznamená, že násobnou regresi nemůžeme použít vůbec. Jednou z možností je veličinu Sport vyloučit ze zpracování, což by nám ale možná bylo líto, protože v povědomí lidí je, že množstvím sportu svou váhu ovlivňujeme. Proto ji v analýze ponecháme, ale budeme při případném vyhodnocení velmi opatrní.



Fakt, že jednotlivé veličiny nemají příliš normální rozdělení, je někdy považován za problém při vyhodnocování korelace pomocí Pearsonova korelačního koeficientu, což jsme právě prováděli. Proto ještě raději vyhodnotíme korelaci pomocí Spearmanova koeficientu pořadové korelace. K tomu účelu zvolíme v záložce *Statistiky* možnost *Neparametrické statistiky*. V zobrazeném dialogovém okně pak opět vybereme všechny veličiny, mezi nimiž chceme korelovanost vyšetřovat.

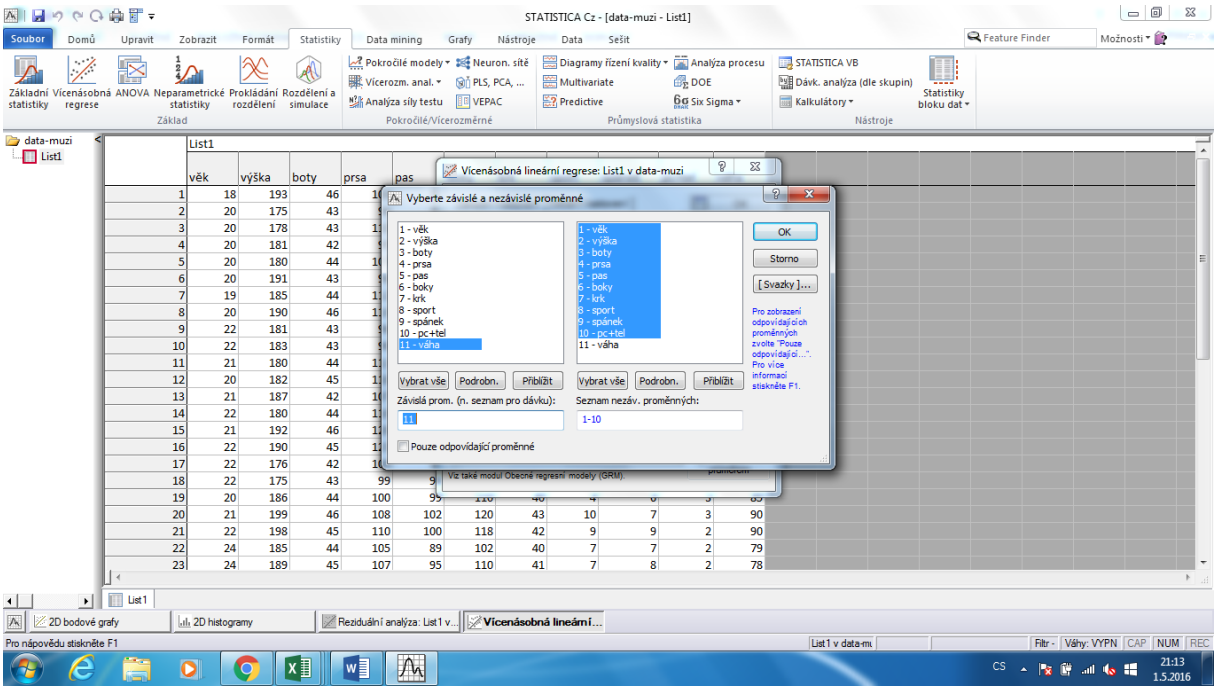


Po potvrzení se nám zobrazí další tabulka.

Spearmanovy korelace (List1 v data-muzi) Označ. korelace jsou významné na hl. p <,05000											
	věk	výška	boty	prsa	pas	boky	krk	sport	spánek	pc+tel	váha
věk	1,00	-0,17	-0,18	-0,06	-0,09	0,19	0,04	-0,02	-0,11	-0,07	-0,16
výška	-0,17	1,00	0,81	0,48	0,59	0,37	0,35	0,21	0,09	-0,25	0,47
boty	-0,18	0,81	1,00	0,56	0,60	0,41	0,42	0,03	0,10	-0,11	0,56
prsa	-0,06	0,48	0,56	1,00	0,72	0,75	0,64	-0,38	0,25	0,15	0,73
pas	-0,09	0,59	0,60	0,72	1,00	0,79	0,73	-0,41	0,05	0,19	0,80
boky	0,19	0,37	0,41	0,75	0,79	1,00	0,64	-0,53	0,17	0,35	0,75
krk	0,04	0,35	0,42	0,64	0,73	0,64	1,00	-0,43	0,14	0,20	0,66
sport	-0,02	0,21	0,03	-0,38	-0,41	-0,53	-0,43	1,00	0,07	-0,77	-0,60
spánek	-0,11	0,09	0,10	0,25	0,05	0,17	0,14	0,07	1,00	0,22	0,14
pc+tel	-0,07	-0,25	-0,11	0,15	0,19	0,35	0,20	-0,77	0,22	1,00	0,48
váha	-0,16	0,47	0,56	0,73	0,80	0,75	0,66	-0,60	0,14	0,48	1,00

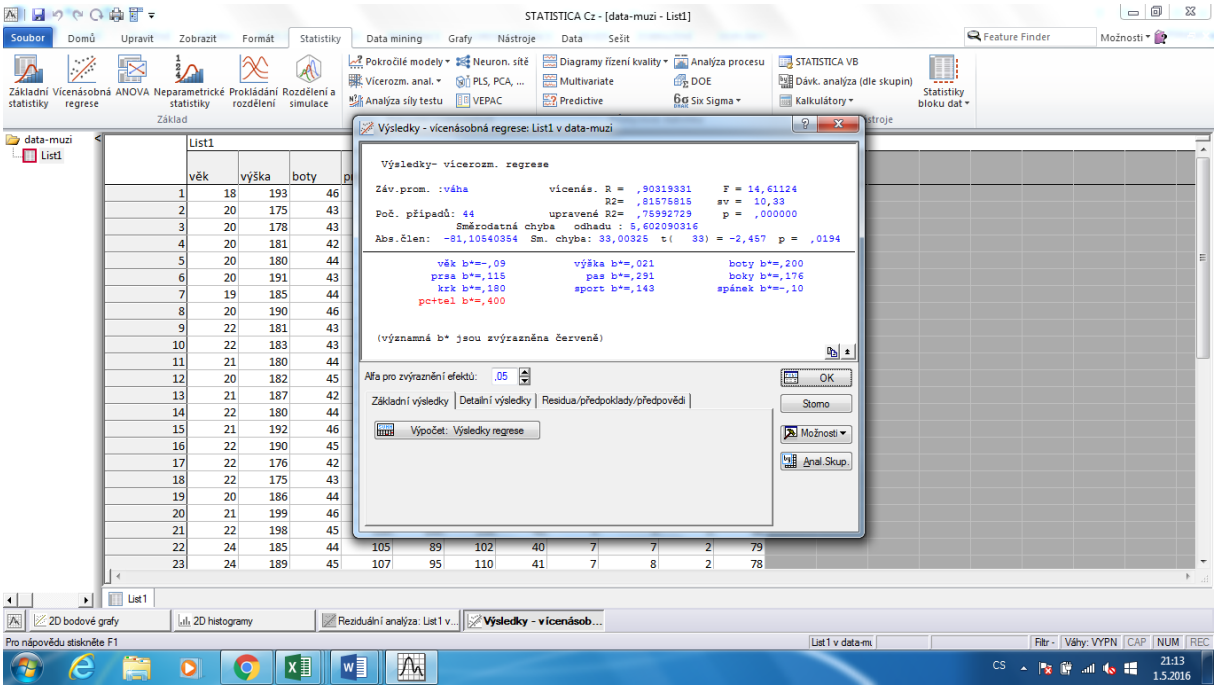
Vidíme, že hodnoty obou druhů korelačních koeficientů vyšly obdobně, hodnoty Spearmanovy korelace vyšly poněkud nižší, ale nijak výrazně.

Nyní již tedy můžeme přistoupit k vlastní regresní analýze. Nejdříve provedeme Vícenásobnou regresi, a to metodu Enter. Postupujeme obdobně, jako už jsme postupovali při jednoduché lineární regresi. V záložce *Statistiky* vybereme možnost *Vícenásobná regrese*. Rozdíl však nastane v následujícím kroku, při výběru proměnných vstupujících do analýzy. Za závislou proměnnou vybereme opět veličinu Váha, ale za nezávislé proměnné vybereme všechny uvažované faktory.



S výstupem jsme se opět již setkali v části o jednoduché lineární regresi. Vidíme, že jediným rozdílem je seznam koeficientů b*, což jsou koeficienty nezávisle proměnných. Tyto

však neodpovídají odhadům parametrů z uvažovaného regresního modelu. Jedná se o speciálně upravené odhady parametrů z jiného modelu, které nám umožňují porovnat relativní vliv jednotlivých regresorů na závisle proměnnou. Statisticky významné regresní koeficienty jsou zvýrazněny červenou barvou.



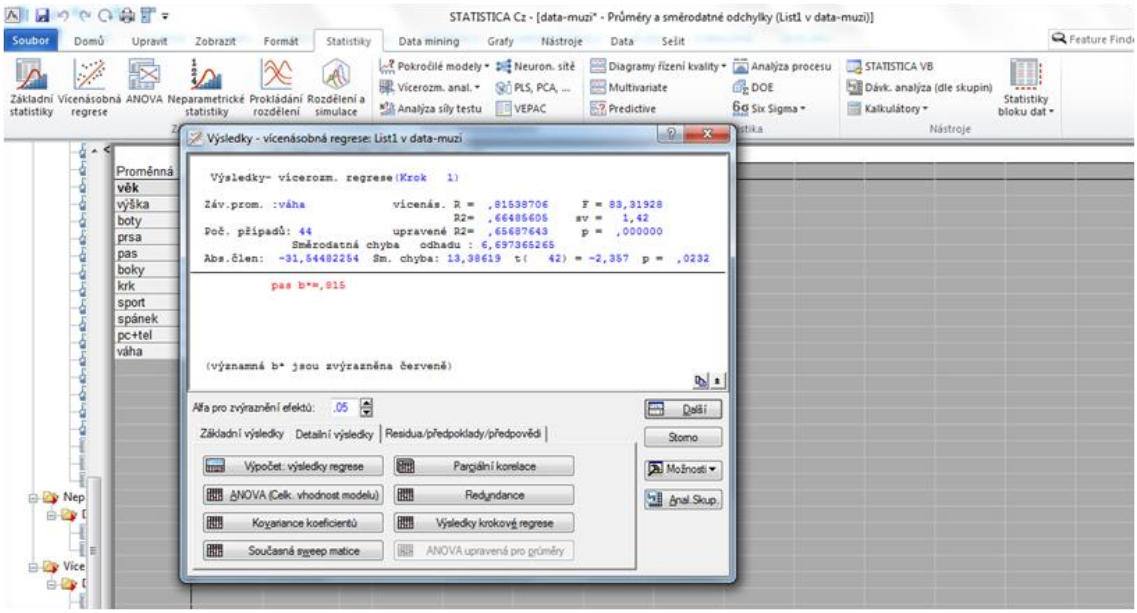
Podrobnější výstup obsahující i hodnoty regresních koeficientů z hledaného modelu získáme výběrem možnosti *Výpočet: Výsledky regrese* v šedé části zobrazeného dialogového okna. Výsledkem je následující tabulka.

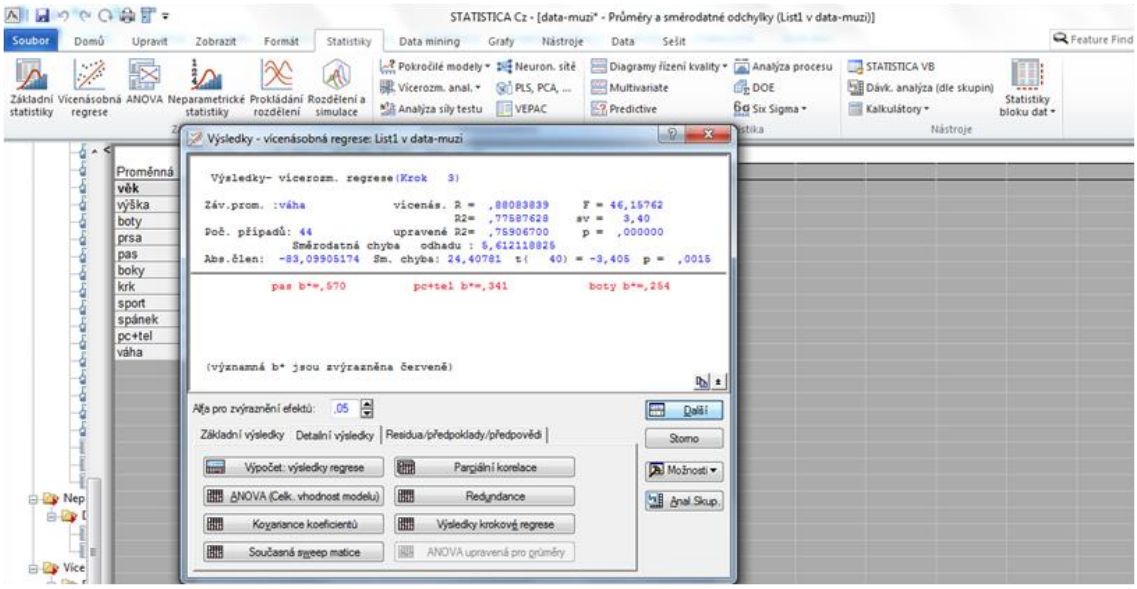
N=44	Výsledky regrese se závislou proměnnou : váha (List1 v data-muzi) R= ,90319331 R2= ,81575815 Upravené R2= ,75992729 F(10,33)=14,611 p					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(33)	p-hodn.
Abs.čl			-81,1054	33,00325	-2,45750	0,019414
věk	-0,094636	0,085961	-0,6895	0,62632	-1,10091	0,278899
výška	0,020790	0,159101	0,0314	0,24016	0,13067	0,896826
boty	0,199542	0,140520	1,4213	1,00090	1,42003	0,164978
prsa	0,115434	0,141274	0,1618	0,19802	0,81710	0,419733
pas	0,290648	0,165021	0,4575	0,25975	1,76128	0,087455
boky	0,176413	0,165383	0,1801	0,16883	1,06669	0,293854
krk	0,179674	0,102106	0,9003	0,51162	1,75969	0,087729
sport	0,142656	0,145810	0,3402	0,34776	0,97837	0,335012
spánek	-0,103072	0,087690	-1,3420	1,14173	-1,17541	0,248240
pc+tv	0,399720	0,126830	1,3130	0,41660	3,15161	0,003445

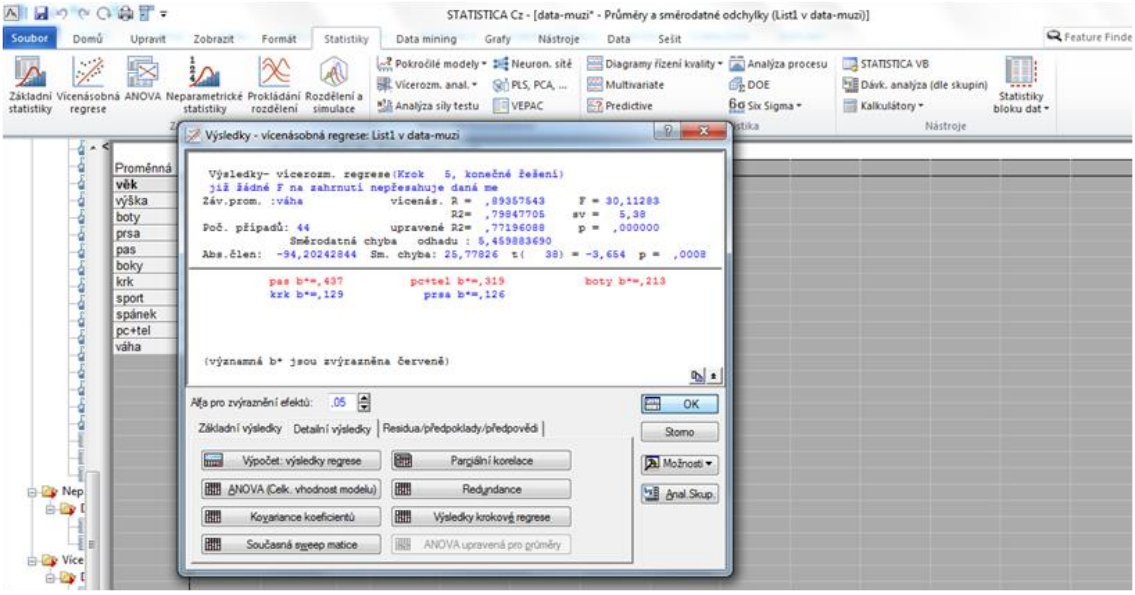
Z výsledných p-hodnot v posledním sloupci je zřejmé, že velká většina regresních koeficientů není významná, proto bychom se v dalším kroku měli pokusit vybrat pouze „důležité“ faktory. K tomu využijeme krokové regrese (neboli stepwise regrese).

K tomu účelu se v dialogovém okně Vícenásobné regrese přepneme na záložku *Detailní výsledky*. Opakovaným stisknutím tlačítka *Další* pak budeme postupně spouštět jednotlivé regresní kroky. Postupně se nám budou zobrazovat další koeficienty b*, tedy další faktory

přidané do modelu. Dokud jsou výsledné b* zbarveny červeně, jsou dané faktory významné, jakmile se přidávané koeficienty b* zbarví namodro, příslušné faktory již významné nejsou, proto je do výsledného modelu nezapracujeme. Poslední krok zpracování poznáme jednoduše, a to tak, že se popis tlačítka *Další* změní na *OK*.



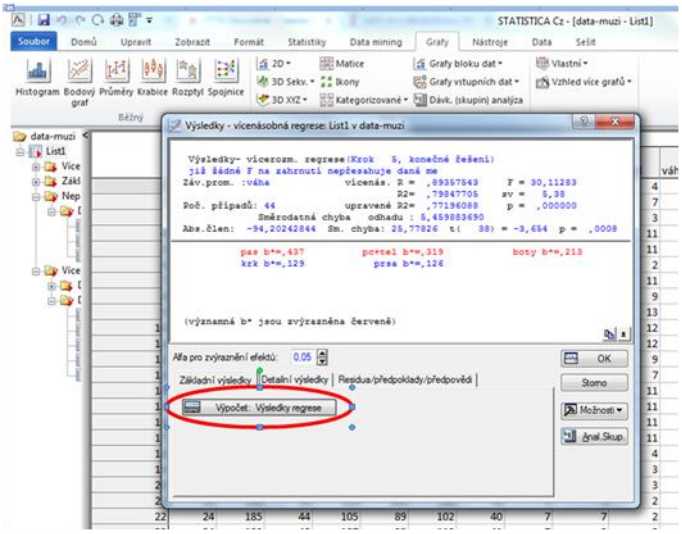




Po provedení posledního kroku zvolíme možnost *Výsledky krokové regrese* a tím obdržíme následující tabulku obsahující závěrečné informace o celé krokové regresi.

Proměnná	Výsledky krokové regrese ; ZP: váha (List1 v data-muzi)						
	Krok +do/-ven	Vícenás. R	Vícenás. R^2	R^2 změna	F zah/vyjm	p-hodn.	Zahrnuté proměnné
pas	1	0,815387	0,664856	0,664856	83,31928	0,000000	1
pc+tv	2	0,859609	0,738928	0,074072	11,63253	0,001468	2
boty	3	0,880838	0,775876	0,036949	6,59434	0,014072	3
krk	4	0,890004	0,792107	0,016231	3,04492	0,088866	4
prsa	5	0,893575	0,798477	0,006370	1,20107	0,280005	5

Dále v záložce *Základní výsledky* vybereme možnost *Výpočet: Výsledky regrese*.



Poté obdržíme tabulku obsahující údaje o jednotlivých regresorech a jejich koeficientech.

N=44	Výsledky regrese se závislou proměnnou : váha (List1 v data-muzi) R= ,89357543 R2= ,79847705 Upravené R2= ,77196088 F(5,38)=30,113 p					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(38)	p-hodn.
Abs. člen			-94,2024	25,77826	-3,65434	0,000776
pas	0,436787	0,121088	0,6875	0,19060	3,60719	0,000888
pc+tv	0,318795	0,079413	1,0471	0,26085	4,01440	0,000271
boty	0,212814	0,099877	1,5158	0,71141	2,13076	0,039640
krk	0,128902	0,091647	0,6459	0,45922	1,40650	0,167701
prsa	0,126218	0,115169	0,1769	0,16143	1,09593	0,280005

Z obou tabulek můžeme vidět, že z devíti uvažovaných faktorů jsou významné pouze tři, a to Obvod pasu, Počet hodin strávených u TV+PC a Velikost bot. Pořadí jejich významnosti je shodné s pořadím uvedení.

Výsledný model má tedy tvar:

V = -94,2024 + 0,6875*P +1,0471*T + 1,5158*O,

kde V je váha v kilogramech, P je obvod pasu v centimetrech, T je čas strávený u PC+TV a O je velikost bot.

Z této rovnice je například zřejmé, že pokud by průměrný mladý muž v pase přibral o 1 cm, pak se to na váze v průměru projeví přírůstkem 0,69 kg. Obdobně i pro ostatní faktory.

Vidíme, že i když jsme problémové faktory nevyřadili ze zpracování, k problémům nedošlo. Vysoká korelovanost mezi veličinami Boty a Výška nevadí, protože do výsledného modelu byla vybrána pouze jedna z nich. Také problematičnost veličiny Sport řešit nemusíme, ani ona nebyla do závěrečného modelu vybrána.

Vidíme, že modely žen (byl vytvářen pomocí Excelu) a mužů (právě jsme vytvořili) se liší. Pouze faktor Obvod pasu se ukázal významným v obou modelech. U žen byl jediným dalším významným faktorem identifikován Obvod boků, kdežto u mužů byly kromě Obvodu pasu identifikovány ještě dva významné faktory, a to Čas strávený u TV+PC a Velikost bot. Vzhledem k tomu, že mezi faktory Čas strávený u TV+PC a Čas strávený sportem byla identifikována silná nepřímá závislost, je vidět, že i tento čas je skrytě v modelu obsažen.