14 JEDNODUCHÁ LINEÁRNÍ REGRESE

Často chceme prozkoumat vztah mezi dvěma veličinami, kde jedna z nich, tzv. **nezávisle proměnná x**, má ovlivňovat druhou, tzv. **závisle proměnnou Y**. Předpokládá se, že obě veličiny jsou spojité. Prvním krokem ve zkoumání by mělo být zakreslení dat do bodového grafu, tzv. **korelačního pole** a ověření toho, zda mezi veličinami skutečně existuje předpokládaná závislost, tzv. **regrese**.

Výsledky této části regresní analýzy jsou často na výstupu z počítače prezentovány ve formě **tabulky analýzy rozptylu**.

Nejjednodušší formou regrese je **jednoduchá lineární regrese**, která předpokládá lineární závislost mezi dvěmi veličinami.

Rovnici regresní přímky zapisujeme ve tvaru: $Y_i = \beta_0 + \beta_1 \cdot x_i + e_i$

Odhad regresní přímky nazýváme **vyrovnávací přímka** a zapisujeme jej v jednom z těchto tvarů:

$\hat{Y}_i = b_0 + b_1 \cdot x_i$

$\hat{Y}_i = b_0^* + b_1 \cdot (x_i - \bar{x})$ (tzv. **odchylková forma** zápisu)

$\hat{Y}_i = b_0 + b_1 \cdot x_i + e_i$

(kde e_i označujeme jako chyby predikce (odhadu), resp. rezidua)

Pokud jsou splněny podmínky lineárního regresního modelu, můžeme koeficienty regresní přímky odhadovat **metodou nejmenších čtverců**.

Podmínky lineárního regresního modelu jsou tyto:

$$Y_i = \beta_0 + \beta_1 \cdot x_i + e_i,$$

kde

1. $E(e_i) = 0$ pro každé $i=1,2,\ldots,n$
   Střední hodnota náhodné složky je nulová.
2. $D(e_i) = \sigma^2$ pro každé $i=1,2,\ldots,n$
   Rozptyl náhodné složky je konstantní.
3. $Cov(e_i, e_j) = 0$ pro každé $i \neq j$, kde $i, j = 1,2,\ldots,n$
   Kovariance náhodné složky je nulová.
4. **Normalita:** Náhodné složky e_i mají pro $i = 1,2,\ldots,n$ normální rozdělení.
5. **Regresní parametry β_i mohou nabývat libovolných hodnot**.
6. **Regresní model je lineární v parametrech**.

Podmínky lineárníhu regresního modelu je nutno v rámci regresní analýzy ověřit.

Existenci lineárního vztahu mezi dvěma veličinami zjišťujeme tak, že se formálně ptáme, zda je směrnice β_1 rovna nule. Pokud je odpověď na tuto otázku kladná, znamená to, že směrnice vyrovnávací přímky se liší od nuly pouze náhodně, tzn., že vztah mezi sledovanými veličinami není lineární. (Jde o obdobu testu, který je vyhodnocen v tabulce ANOVA.)

Obdobně můžeme testovat významnost absolutního členu vyrovnávací přímky (b_0). Testům významnosti koeficientů vyrovnávací přímky říkáme **dílčí t-testy**.

Intervalový odhad můžeme při regresi hledat jednak pro střední hodnotu Y při dané úrovni x ($E(Y_0|X=x_0)$), jednak pro jednotlivé pozorování (Y_0). Intervalu spolehlivosti pro jednotlivé pozorování říkáme **interval predikce**. Tyto intervalové odhady pro spojitě se měnící hodnoty x tvoří tzv. **pás spolehlivosti kolem regresní přímky,** resp. **pás predikce kolem regresní přímky**.

Kvalitu regresního modelu udává **index determinace R^2**. Přesněji řečeno udává kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem a kolik zůstalo nevysvětleno.

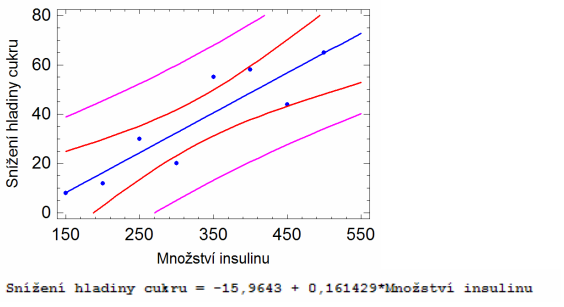
Regresní model nám umožňuje provádět rovněž **extrapolaci**, tj. odhad závisle proměnné pro hodnoty nezávisle proměnné ležící mimo interval naměřených hodnot. Extrapolace je vždy spojena s rizikem, že regresní model mimo interval naměřených hodnot pozbývá platnosti.

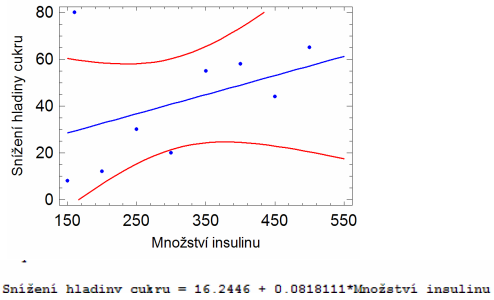
Lépe je znát několik užitečných pravidel, než nastudovat mnoho neužitečných věcí. (Seneca, volně dle Ing Pavla Blažíčka, IV. zjazd Slovenskej spoľočnosti klinickej biochémie, Stará Ľubovňa, květen 2000)

- Závěry plynoucí z našich výsledků platí pouze pro rozsah hodnot, pro které byl model navržen. Jakákoliv extrapolace je přinejmenším ošidná.

- Na data se vždy nejprve "podíváme" pomocí korelačního pole. Z korelačního pole usuzujeme, zda nejsou přítomny tzv. **vlivné** resp. vychýlené **body**. Bod, který je silně vychýlený ve směru pouze jedné ze souřadnic, často nazýváme odlehlý (outlier). Bod, který je vychýlený ve směru obou souřadnic, označujeme často jako extrém. Terminologie není ustálená. Vlivné body mohou mít silný vliv na odhadovanou regresní funkci.

- Problém odlehlých bodů bývá často řešen tím, že jsou z výběrového souboru vyloučeny a to na základě odhadu (jsou patrné už na výše zmíněném korelačním poli). Jiný vhodný způsob jejich odhalení je zkonstruování a posouzení tzv. diagnostických grafů (např. z-souřadnice, $x_{0,5}$-souřadnice) nebo provedení numerických testů (Dixonův, Grubbsův). Pokud je dostatečné množství dat, je někdy účelné odlehlý bod (body) vyloučit z dalšího zpracování. Nikdy bychom však neměli vlivný bod vyloučit, aniž bychom vysvětlili příčinu jeho vzniku nebo se přesvědčili, že se jedná o artefakt (např. hrubá chyba).





- Pokud používáme korelační koeficient, je třeba mít na paměti, že tento koeficient je pouze mírou lineární závislosti výsledků. "Pěkný" korelační koeficient (hodnota blízká jedné nebo minus jedné) ještě vůbec neznamená, že srovnávané metody dávají "pěkně" shodné

výsledky. Znamená to pouze silnou lineární závislost mezi výsledky oběma metodami. "Špatný" (malý v absolutní hodnotě) korelační koeficient vůbec neznamená, že závislost je málo silná. Může (ale nemusí!) jít např. o silnou nelineární závislost, např. kvadratickou.

- Použití lineární regrese je vhodné pouze v některých případech. Řekněme, že chceme provést lineární regresi vysvětlované proměnné Y na vysvětlující proměnné x. Tato regrese má svoje oprávnění pouze tehdy, jestliže:
  - rozptyl (neurčitost) při získávání (měření) hodnot vysvětlující proměnné je alespoň o řád menší než rozptyl (neurčitost) při měření hodnot vysvětlované proměnné. Důvod je docela prozaický. Uvědomme si, že při výpočtu koeficientů optimální vyrovnávací křivky metodou nejmenších čtverců se vlastně hledá taková vyrovnávací křivka, aby součet čtverců odchylek jednotlivých (naměřených) bodů od této křivky byl nejmenší možný. Matematicky řečeno hledáme globální minimum. Drtivá většina algoritmů (počítačových programů) provádí měření vzdálenosti bodů od vyrovnávací křivky ve směru vysvětlované proměnné. Jinak řečeno, postup výpočtu předpokládá, že ve směru vysvětlující proměnné jsou neurčitosti jednotlivých bodů zanedbatelné oproti směru vysvětlované proměnné.
  - Dále je třeba, aby každá proměnná měla v ideálním případě normální (Gaussovo) anebo v praxi alespoň symetrické rozdělení dat. Při troše zkušenosti to poznáme už z korelačního pole eventuelně z empirické hustoty (histogramu) příslušné proměnné.

- Jestliže jsou některé hodnoty při testování statisticky významné, nemusí to znamenat, že jsou významné i prakticky. Obdobně, jestliže jsou některé hodnoty při testování statisticky nevýznamné, nemusí to znamenat, že jsou nevýznamné i prakticky.

Podle L. Dohnala (posbíráno na Internetu)

14.1. Byl vyvinut nový druh insulinu a zkoumá se závislost snížení hladiny cukru v krvi pacienta na množství podaného insulinu určitou dobu před měřením. Náhodně vybraným 8 pacientům byla naočkována různá množství insulinu a po určité době bylo těmto pacientům změřeno snížení cukru v krvi. Výsledky měření:

Množství insulinu $[\mu l]$	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
Snížení hladiny cukru $[\%]$	**8**	**12**	**30**	**20**	**55**	**58**	**44**	**65**

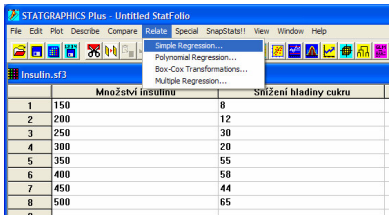
a) Znázorněte korelační pole a zvolte vhodný typ lineárního regresního modelu pro popis závislosti snížení hladiny cukru na množství podaného inzulínu.
b) Ověřte oprávněnost použití vybraného modelu.
c) Proveďte dílčí t-testy.
d) Ověřte kvalitu modelu – resp. vyberte nejvhodnější lin. regresní model pro popis dané závislosti (zvolili-li jste regr. model jiný než původní, vraťte se k bodu a)).
e) Ověřte, zda byly splněny předpoklady pro použití vybraného lin. regr. modelu.
f) Zapište rovnici vyrovnávací funkce.
g) Určete střední hodnotu $E(Y_O|X=325)$ **snížení hladiny cukru při množství podaného insulinu 325 µl, včetně 95%-ního intervalu spolehlivosti. Vyjádřete slovně, co znamená 95%-ní interval spolehlivosti** $E(Y_O|X=x_0)$ **pro** $x_0 = 325\,\mu l$ **.**
h) Odhadněte, o kolik se sníží hladina cukru pacienta, jemuž se podá 325 µl insulinu (včetně 95%-ního intervalu predikce).

i) Odhadněte na základě zvoleného regresního modelu o kolik se sníží hladina cukru pacienta, jemuž se podá 700 µl insulinu (včetně 95%-ního intervalu predikce). Pojednejte o oprávněnosti této predikce.

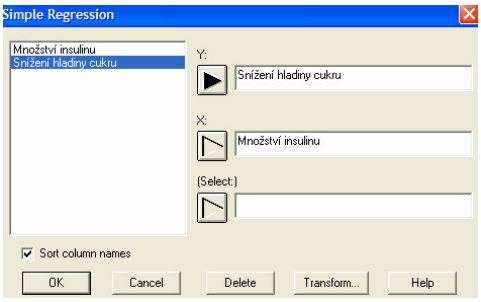
Řešení ve Statgraphicsu:

Nejdříve data zadáme do Statgraphicsu, popř. použijeme soubor **Insulin.sf3**.

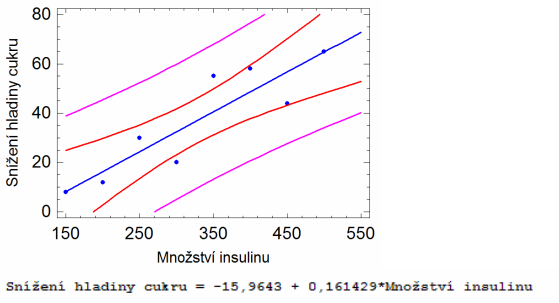
Pro jednoduchou regresi volíme menu **Relate/Simple Regression...**

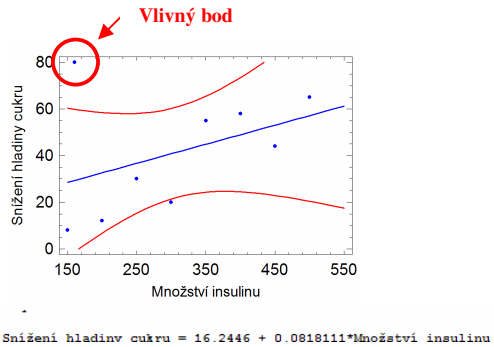


Vysvětlovanou proměnnou (Snížení hladiny cukru) zadáme jako Y, vysvětlující proměnnou (Množství insulinu) zadáme jako X.

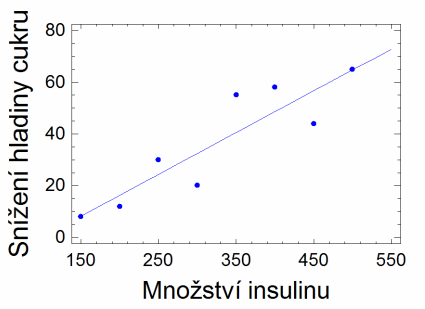


ada) Následující obrázek je ilustrací toho, co mohou způsobit **vlivné body** obsažené v datech (při použití metody nejmenších čtverců). Z obrázku je zřejmé, že jediný vlivný bod dokáže odhad regresní funkce znehodnotit. Nikdy bychom však neměli vlivný bod vyloučit, aniž bychom vysvětlili příčinu jeho vzniku nebo se přesvědčili, že jde o hrubou chybu. (Tyto body mohou například signalizovat, zvláště při malém počtu pozorovaných bodů, datovou oblast, kterou jsme měřením nepokryli.)





Proto si nejdříve prohlédneme **korelační pole** (scatter plot, bodový graf) a zjistíme zda data vlivné body neobsahují.



Je zřejmé, že tato data vlivné body neobsahují.

Zároveň se pokusíme na základě této vizualizace **odhadnout vhodný typ lineárního regresního modelu**. Bývá zvykem volit regresní funkci s co nejmenším počtem regresních koeficientů, avšak dostatečně flexibilní a s požadovanými vlastnostmi (monotónie, asymptoty, …). Většinou se vychází ze zkušenosti, popř. v dnešní době, kdy je běžné pro regresní analýzu využívat statistický software, využíváme vhodnou databázi regresních funkcí.

Statgraphics má jako výchozí lin. regresní model přednastavenou lineární regresní funkci, která by (na základě vizuální kontroly) mohla být v tomto případě použita. Na základě konzultace se zadavatelem úlohy bychom mohli rovněž zvolit funkci kvadratickou, resp. funkci logaritmickou.

adb) – adf)

Nyní si všimneme textového výstupu.

```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: Snížení hladiny cukru
Independent variable: Množství insulinu
-----------------------------------------------------------------------------
                                Standard          T
Parameter          Estimate        Error     Statistic       P-Value
-----------------------------------------------------------------------------
Intercept         -15,9643      11,1856      -1,42721        0,2034
Slope             0,161429    0,0324596       4,97321        0,0025
-----------------------------------------------------------------------------

                          Analysis of Variance
-----------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square     F-Ratio       P-Value
-----------------------------------------------------------------------------
Model                     2736,21      1      2736,21       24,73        0,0025
Residual                  663,786      6      110,631
-----------------------------------------------------------------------------
Total (Corr.)             3400,0       7

Correlation Coefficient = 0,897089
R-squared = 80,4769 percent
R-squared (adjusted for d.f.) = 77,223 percent
Standard Error of Est. = 10,5181
Mean absolute error = 7,42857
Durbin-Watson statistic = 2,78236 (P=0,0340)
Lag 1 residual autocorrelation = -0,391276

  Snížení hladiny cukru = -15,9643 + 0,161429*Množství insulinu
```

Typ modelu, rovnice vyrovnávací funkce

Závisle a nezávisle proměnná

Bodové odhady koeficientů regresní přímky

Bodové odhady směrodatných odchylek koeficientů regresní přímky

Výsledky dílčích t-testů

Součty čtverců pro model, reziduální a celkový

Reziduální výběrový rozptyl

Výsledek F-testu pro regresi

Korelační koeficient

Koeficient determinace

Výběrová reziduální směrodatná odchylka

Rovnice vyrovnávací přímky

Jak jsme si již uvedli, Statgraphics zahajuje regresní analýzu použitím lineární regresní funkce (je to nejjednodušší lineární regresní model). Hned vedle názvu modelu je obecná rovnice vyrovnávací křivky (my značíme koeficienty b_0, b_1, Statgraphics a, b). Odhady regresních koeficientů nalezneme pod zápisem o vysvětlované a vysvětlující proměnné. V této tabulce jsou uvedeny jak bodové odhady regresních koeficientů (intercept ... absolutní člen, b_0; slope

... směrnice, b_1), odhady jejich směrodatných odchylek, tak i vyhodnocení dílčích t-testů o významnosti regresních koeficientů.

Následuje tabulka ANOVA (výstup pro F-test v regresi), která vypovídá o vhodnosti vybraného regresního modelu. V tabulce ANOVA najdeme, mimo příslušného p-value, součty čtverců pro model, reziduální a celkový součet čtverců (jde o hodnoty pomocí nichž se určuje koeficient determinace) a výběrový reziduální rozptyl.

Pod tabulkou ANOVA nacházíme hodnoty korelačního koeficientu (míra lineární závislosti mezi proměnnými), koeficientu determinace R^2 (vypovídá o vhodnosti použitého modelu) a výběrové reziduální směrodatné odchylky (odmocnina z výběrového reziduálního rozptylu uvedeného v tabulce ANOVA).

Ve spodní části textového výstupu pak nalezneme odhadnutou rovnici vyrovnávací křivky.

adb) Vhodnost použití zvoleného lineárního regresního modelu ověříme pomocí **analýzy rozptylu (F-test) v regresi**. Tato analýza vychází ze vztahu:

$$SS_Y = SS_{\hat{Y}} + SS_R,$$

kde $SS_Y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$ je celkový součet čtverců odchylek od průměru,

$SS_{\hat{Y}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ je součet čtverců modelu (tzv. regresní (vysvětlený) součet čtverců)

a

$SS_R = \sum_{i=1}^{n}(\hat{e}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ je reziduální (nevysvětlený) součet čtverců.

Vhodný regresní model musí mít vysvětlený součet čtverců větší než reziduální součet čtverců. Pro testování tohoto předpokladu se ukazuje jako vhodný F-test známý z ANOVY (H_0: Zvolená funkční závislost mezi závisle a nezávisle proměnnou neexistuje.). Výstupem tohoto testu je tabulka ANOVA.

Zdroj proměnlivosti	**Součet čtverců**	**Stupně volnosti**	**Průměrný čtverec**	**Testová stat. F-poměr**	**P-value**
Model	$SS_{\hat{Y}} = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	1	$MS_{\hat{Y}} = SS_{\hat{Y}}$		
Rezidua	$SS_R = \sum_{i=1}^{n}(\hat{e}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$n-2$	$MS_R = \frac{SS_R}{n-2}$	$F-ratio = \frac{MS_{\hat{Y}}}{MS_R}$	$1 - F(F-ratio)$
Celkový	$SS_Y = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$	$n-1$			

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2736,21	1	2736,21	24,73	0,0025
Residual	663,786	6	110,631		
Total (Corr.)	3400,0	7			

V našem případě lze říci, že lineární závislost mezi snížením hladiny cukru a množstvím podaného insulinu existuje.

adc.) Nyní se zaměříme na zjištění toho, zda nalezený model nelze zjednodušit – zda některé regresní koeficienty nelze z modelu vypustit (otestujeme, zda není možné některé regresní koeficienty považovat za nulové). Tento proces nazýváme **dílčími t-testy** (jejich konstrukce je popsána ve skriptech).

Výsledky dílčích t-testů jsou v našem případě tyto:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Snížení hladiny cukru
Independent variable: Množství insulinu
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept       -15,9643       11,1856       -1,42721         0,2034
Slope           0,161429     0,0324596        4,97321         0,0025
------------------------------------------------------------------------------
```

H_0: $\beta_0 = 0$

H_A: $\beta_0 \neq 0$

p-value = 0,2034 $\Rightarrow$ nezamítáme H_0, tzn. koeficient β_0 bychom mohli z modelu vypustit.

H_0: $\beta_1 = 0$

H_A: $\beta_1 \neq 0$

p-value = 0,0025 $\Rightarrow$ zamítáme H_0, tzn. koeficient β_1 z modelu vypustit nemůžeme.

Vyrovnávací přímku bychom tedy mohli zapisovat ve tvaru:

Snížení hladiny cukru = 0,16. Množství insulinu

add.) Kvalitu regresního modelu můžeme hodnotit pomoci **indexu determinace R^2**. Index determinace udává, kolik procent rozptylu vysvětlované proměnné je vysvětleno modelem.

Hodnotu indexu determinace najdeme v textovém výstupu procedury Simple Regression.

```
R-squared = 80,4769 percent
```

V našem případě model vysvětluje cca 80% celkového rozptylu, což svědčí o poměrně vhodné volbě modelu.

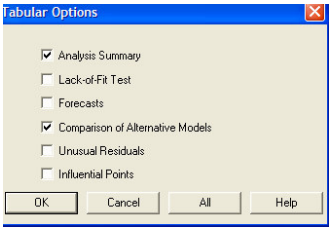
Nyní si ještě ukážeme, **jak najít nejvhodnější model lineární regrese** pro daná data.

Pozor!!! *„lineární“ znamená lineární vzhledem ke koeficientům regresní funkce, nikoliv regrese lineární funkcí (přímkou).*

Mezi další modely lineární regrese patří například model kvadratický, exponenciální, reciproční, apod.

Chceme-li zjistit, zda pro naše data není vhodnější jiná funkce než lineární, provedeme porovnání jednotlivých funkcí pomocí indexu determinace.

Nacházíme-li se ve výstupním okně procedury jednoduchá regrese (simple regression), klikneme na ikonu **Tabular Options** a zvolíme položku **Comparisson of Alternative Models** (porovnání dalších modelů).

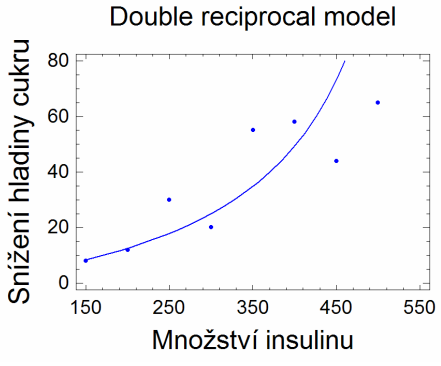


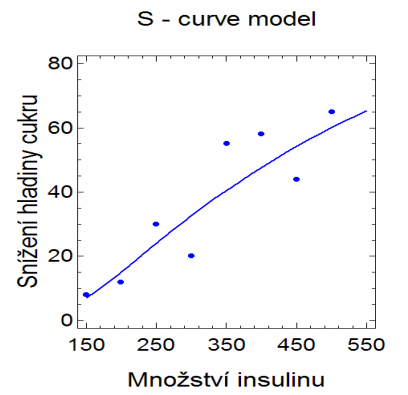
Comparison of Alternative Models

Model	Correlation	R-Squared
Double reciprocal	0,9575	91,68%
S-curve	-0,9350	87,43%
Multiplicative	0,9320	86,87%
Square root-Y	0,9041	81,73%
Square root-X	0,9007	81,13%
Exponential	0,8994	80,90%
Linear	0,8971	80,48%
Logarithmic-X	0,8969	80,45%
Reciprocal-X	-0,8665	75,08%
Reciprocal-Y	<no fit>	
Logistic	<no fit>	
Log probit	<no fit>	

Z modelů s nejvyššími indexy determinace vybereme ten, který nejlépe odpovídá předpokládanému vztahu (v praxi je při výběru nutné spolupracovat s odborníkem na studovanou problematiku).

Vzhledem k povaze našich dat (nedá se očekávat, že s rostoucím množstvím insulinu bude docházet k prudkému snížení hladiny cukru (model Double reciprocal) nevolíme v tomto případě model s nejvyšším indexem determinace, raději se přikloníme k modelu S-curve. V tuto chvíli by však pro výběr modelu byla opravdu nejvhodnější konzultace se zadavatelem úlohy. Volbu modelu provedeme RC na textový výstup a v menu **Analysis Options** zvolíme vybraný model.





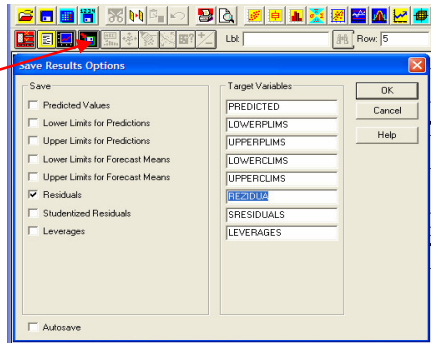
Pokud bychom se skutečně rozhodli pro užití jiného než původně vybraného modelu, museli bychom **znovu posoudit** korelační pole**, vyhodnotit** anovu pro regresi a dílčí t-testy.

ade.) Vyhodnocení předpokladů pro použití lineárního regresního modelu provádíme pomocí posouzení reziduí. Ověříme:

1. normalitu reziduí
2. nulovou střední hodnotu reziduí
3. nulovou kovariancí reziduí

Nejdříve si rezidua zapíšeme do datové tabulky. Nemusíme používat zadání proměnné pomocí vzorce, můžeme použít předdefinované vztahy Statgraphicsu. Nacházíme-li se ve výstupním okně procedury jednoduchá regrese (simple regression), klikneme na ikonu **Save Results** a zvolíme, kterou z předdefinovaných hodnot chceme (a pod jakým názvem) zapsat do tabulky.

Ikona Save Results



Insulin.sf3

	Množství insulinu	Snížení hladiny cukru	REZIDUA
1	150	8	-0,25
2	200	12	-4,32143
3	250	30	5,60714
4	300	20	-12,4643
5	350	55	14,4643
6	400	58	9,39286
7	450	44	-12,6786
8	500	65	0,25

Pozn.: *Z nabízených hodnot by nás mohly ještě zajímat očekávané hodnoty ($\hat{Y}_i$, Predicted Values), dolní, resp. horní mez intervalu predikce (Lower, resp. Upper Limits for Predictions), dolní, resp. horní mez intervalu spolehlivosti pro $E(Y_O|X = x_0)$ (Lower, resp. Upper Limits for Forecast Means).*

ad1.) Testování normality (jak Q-Q grafem, tak statistickými testy) provedeme např. známým způsobem v menu **Describe/Distributions/Distributions Fitting (Uncensored Data)...**

ness-of-Fit Tests for RESIDUALS

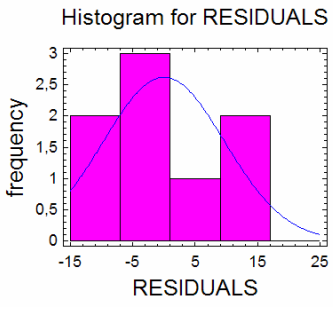
Chi-Square Test

Lower Limit	Upper Limit	Observed Frequency	Expected Frequency	Chi-Square
at or below	-2,46709	3	3,20	0,01
above	-2,46709	5	4,80	0,01

Insufficient data to conduct Chi-Square test.

Estimated Kolmogorov statistic DPLUS = 0,149724
Estimated Kolmogorov statistic DMINUS = 0,114756
Estimated overall statistic DN = 0,149724
Approximate P-Value = 0,993909

EDF Statistic	Value	Modified Form	P-Value
Kolmogorov-Smirnov D	0,149724	0,466981	>=0.10*
Anderson-Darling A^2	0,205946	0,232493	0,7997*



Z výsledků Kolmogorovova – Smirnovova testu je zřejmé, že normalita reziduí nebyla zamítnuta.

ad2.) Rovněž testování nulové střední hodnoty, by pro nás již mělo být jednoduché – menu: **Describe/Numeric Data/One-Variable…**, jako proměnnou zadáme Residuals, ikona **Tabular Options** – Hypothesis Tests….

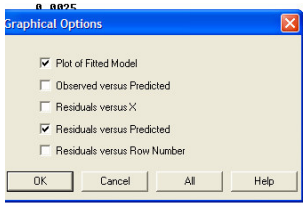
Připomeňme si, že normalitu reziduí jsme již potvrdili v předcházejícím kroku (předpoklad testu tedy byl ověřen).

```
Hypothesis Tests for RESIDUALS

Sample mean = -0,00000375
Sample median = 0,0


t-test
------
Null hypothesis: mean = 0,0
Alternative: not equal

Computed t statistic = -0,00000108921
P-Value = 0,999999

Do not reject the null hypothesis for alpha = 0,05.
```

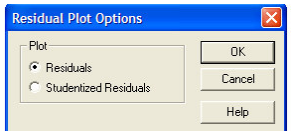
Nulová střední hodnota reziduí nebyla zamítnuta.

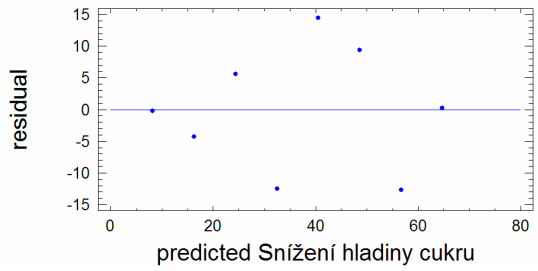
ad3.) Nulovou kovariancí reziduí ověříme pouze pomocí exploračních grafů. Zobrazíme si korelační pole reziduí vůči odhadovaným hodnotám a pokud v něm nebude patrná žádná funkční závislost, odlehlá pozorování ani „střídání znamének“ (střídání kladných a záporných reziduí), budeme považovat kovarianci za nulovou.

Jsme-li ve výstupním okně procedury jednoduchá regrese, korelační pole reziduí vs. očekávané hodnoty získáme kliknutím na ikonu **Graphical Options** a volbou položky **Residuals versus Predicted**.



Ještě musíme na osu y dostat skutečná rezidua a to provedeme RC na příslušný graf a nastavením položky **Residuals** v menu **Pane Options**.





Rezidua jsou náhodně rozmístěna kolem nuly a nemají žádný zřejmý vztah k předpovídaným hodnotám: ani se systematicky nezvyšují ani se systematicky nesnižují spolu s rostoucími předpovídanými hodnotami a není zde ani náznak nelineárního vztahu, nedochází ke „střídání znamének“ ani zde nevidíme odlehlá pozorování, lze tedy předpokládat, že kovariance reziduí je nulová.

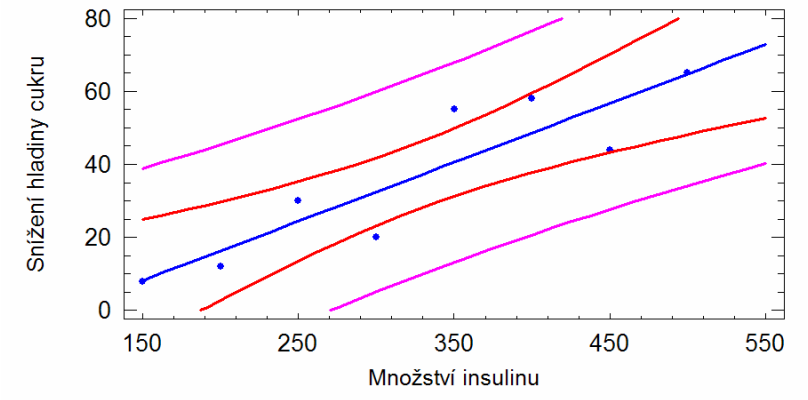
Nyní můžeme konstatovat, že předpoklady lineárního regresního modelu byly splněny.

adf.) Za regresní rovnici tedy budeme považovat:

Snížení hladiny cukru = 0,16. Množství insulinu

Na závěr regresní analýzy si předvedeme využití získaných výsledků. Tím je predikce očekávaných hodnot závislé proměnné při zvolené hodnotě proměnné nezávisle.

Regresní analýza nám umožňuje **odhad podmíněné střední hodnoty** $E(Y_O|X=x_0)$ a **odhad individuální hodnoty** Y_0. V obou případech můžeme získat jak bodový tak i intervalový odhad. Podmíněná střední hodnota $E(Y_O|X=x_0)$ nám v našem případě říká jaká je střední hodnota snížení hladiny cukru pro pacienty, kterým bylo podáno množství insulinu x_0. Oproti tomu individuální hodnota Y_0 udává jaké je snížení hladiny cukru u jediného pacienta, kterému bylo podáno množství insulinu x_0. Bodové odhady podmíněné střední hodnoty a individuální hodnoty jsou totožné. Dále je zřejmé, že intervalový odhad podmíněné střední hodnoty bude „užší" než intervalový odhad individuální hodnoty (při stejně zvolené hladině významnosti). Aby bylo jednoduše rozpoznatelné, který interval spolehlivosti máme na mysli, mluvíme o **intervalu spolehlivosti** (pro podmíněnou střední hodnotu) a **intervalu predikce** (pro individuální hodnotu). Tyto intervalové odhady pro spojitě se měnící hodnoty x tvoří tzv. **pás spolehlivosti** kolem regresní přímky**,** resp. **pás predikce** kolem regresní přímky.



Při odhadech v regresi je nutné ještě sledovat, zda se jedná o **interpolaci** (odhad uvnitř intervalu naměřených dat) nebo o **extrapolaci** (odhad mimo interval naměřených dat). Extrapolaci můžeme považovat za důvěryhodnou pouze v případě, že jsme přesvědčeni o platnosti používaného modelu v oblasti extrapolace.

adg.) Odhad podmíněné střední hodnoty:

Bodový odhad $E(Y_O|X=x_0)$**:** $\hat{Y}(x_0) = (-15{,}96) + 0{,}16 \cdot x_0$ $\hat{Y}(325\,\mu l) = 36{,}04\%$

95%ní interval spolehlivosti $E(Y_O|X=x_0)$**:**

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2736,21	1	2736,21	24,73	0,0025
Residual	663,786	6	110,631		
Total (Corr.)	3400,0	7			

s^2

Summary Statistics for Množství insulinu

Count = 8
Average = 325,0
Variance = 15000,0 s_x^2
Standard deviation = 122,474
Minimum = 150,0
Maximum = 500,0
Range = 350,0
Stnd. skewness = 0,0
Stnd. kurtosis = -0,69282

$$P\left(E(Y_0|X=x_0)\in\left(\hat{Y}(x_0)\mp s\cdot\sqrt{\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)\cdot s_x^2}\right)}\cdot t_{0,975,n-2}\right)\right)=0,95$$

$$P\left(E(Y_0|X=x_0)\in\left((-15,96)+0,16\cdot x_0\mp\sqrt{110,63}\cdot\sqrt{\left(\frac{1}{8}+\frac{(x_0-325)^2}{7\cdot 15000}\right)}\cdot t_{0,975,8-2}\right)\right)=0,95$$

$$P\left(E(Y_0|X=x_0)\in\left((-15,96)+0,16\cdot x_0\mp\sqrt{110,63}\cdot\sqrt{\left(\frac{1}{8}+\frac{(x_0-325)^2}{7\cdot 15000}\right)}\cdot 2,45\right)\right)=0,95$$

Pro $x_0=325\,\mu l$:

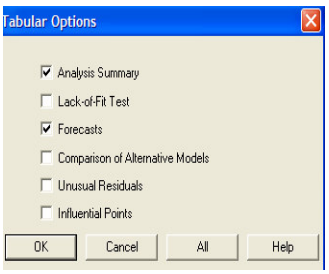
$$P\left(E(Y_0|X=325)\in\left((-15,96)+0,16\cdot 325\mp\sqrt{110,63}\cdot\sqrt{\left(\frac{1}{8}+\frac{(325-325)^2}{7\cdot 15000}\right)}\cdot 2,45\right)\right)=0,95$$

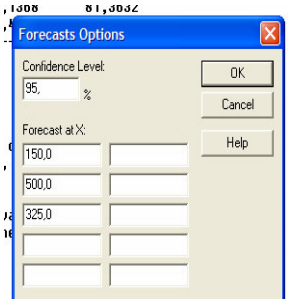
$$P\left(E(Y_0|X=325)\in(36,04\mp 9,11)\right)=0,95$$

$$P\left(E(Y_0|X=325)\in(28,94;45,15)\right)=0,95$$

Statgraphics:

Klikneme na ikonu **Tabular Options** a zvolíme položku **Forecasts**, v okně Forecasts Options zadáme hodnotu x_0, v níž chceme nalézt odhad:





Predicted Values

X	Predicted Y	95,00% Prediction Limits Lower	Upper	95,00% Confidence Limits Lower	Upper
150,0	8,25	-22,3831	38,8831	-8,36316	24,8632
500,0	64,75	34,1169	95,3831	48,1368	81,3632
325,0	36,5	9,2018	63,7982	27,4006	45,5994

Mírné odchylky oproti „ručně“ vypočtenému intervalu jsou způsobeny zaokrouhlováním.

Lze tedy tvrdit, že průměrné snížení hladiny cukru při dávce insulinu 325 µl bude 36,0%. S 95%-ní spolehlivosti bude průměrné snížení hladiny cukru při dávce insulinu 325 µl v rozmezí cca (28,9%; 45,2%).

adh.) Odhad individuální hodnoty:

Bodový odhad $\hat{Y}(x_0)$: $\hat{Y}(x_0) = (-15{,}96) + 0{,}16 \cdot \mathrm{x}_0$

$\hat{Y}(325\,\mu l) = 36{,}04\%$

95%-ní interval predikce:

$$P\left(Y_0 \in \left(\hat{Y}(x_0) \mp s \cdot \sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)\cdot s_x^2} + 1\right)} \cdot t_{0{,}975,n-2}\right)\right) = 0{,}95$$

$$P\left(Y_0 \in \left((-15{,}96) + 0{,}16 \cdot \mathrm{x}_0 \mp \sqrt{110{,}63} \cdot \sqrt{\left(\frac{1}{8} + \frac{(x_0 - 325)^2}{7 \cdot 15000} + 1\right)} \cdot t_{0{,}975,8-2}\right)\right) = 0{,}95$$

$$P\left(Y_0 \in \left((-15{,}96) + 0{,}16 \cdot \mathrm{x}_0 \mp \sqrt{110{,}63} \cdot \sqrt{\left(\frac{1}{8} + \frac{(x_0 - 325)^2}{7 \cdot 15000} + 1\right)} \cdot 2{,}45\right)\right) = 0{,}95$$

Pro $x_0 = 325\,\mu l$:

$$P\left(Y(x=325) \in \left((-15{,}96) + 0{,}16 \cdot 325 \mp \sqrt{110{,}63} \cdot \sqrt{\left(\frac{1}{8} + \frac{(325 - 325)^2}{7 \cdot 15000} + 1\right)} \cdot 2{,}45\right)\right) = 0{,}95$$

$$P\left(Y(x=325) \in (36{,}04 \mp 27{,}33)\right) = 0{,}95$$

$$P\left(Y(x=325) \in (8{,}71;63{,}37)\right) = 0{,}95$$

Statgraphics: Použijeme výstup, který jsme získali při hledání odhadu podmíněné střední hodnoty:

Predicted Values

X	Predicted Y	95,00% Prediction Limits Lower	Upper	95,00% Confidence Limits Lower	Upper
150,0	8,25	-22,3831	38,8831	-8,36316	24,8632
500,0	64,75	34,1169	95,3831	48,1368	81,3632
325,0	36,5	9,2018	63,7982	27,4006	45,5994

Mírné odchylky oproti „ručně“ vypočtenému intervalu jsou opět způsobeny zaokrouhlováním.

Lze říci, že snížení hladiny cukru u pacienta jemuž bylo podáno 325 µl insulinu bude 36,0%. S 95%-ní spolehlivosti se snížení hladiny cukru u tohoto pacienta bude pohybovat v rozmezí cca (8,7%; 63,4%).

adi.) Vzhledem k tomu, že měření byla prováděna pro množství insulinu v rozsahu 150 µl – 500 µl, odhad snížení hladiny cukru pro 700 µl insulinu je extrapolací. V tomto případě nemáme žádné informace o možné platnosti modelu pro x_0 = 700 µl a proto tento odhad určovat nebudeme (nemohli bychom jej považovat za důvěryhodný).