Univerzita Karlova v Praze

Matematicko-fyzikální fakulta

RIGORÓZNÍ PRÁCE

Mgr. Marian Rybář

„Regresní modely a jejich výuka“

Katedra didaktiky matematiky

Studijní program: Matematika

Studijní obor: Obecné otázky matematiky a informatiky

Praha 2014

Prohlašuji, že jsem tuto rigorózní práci vypracoval samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle § 60 odst. 1 autorského zákona.

V Praze dne 30.6.2014 podpis

Název práce: „Regresní modely a jejich výuka“

Autor: Mgr. Marian Rybář

Katedra / Ústav: Katedra didaktiky matematiky

Abstrakt:

Regresní modely a jejich výstupy mají obrovské využití nejen v oblasti medicíny, vědy či manažerského rozhodování, ale i mnoha dalších oborech. Otázka srozumitelného vysvětlení regresních modelů absolventům nematematických oborů je potom velmi aktuálním a využitelným tématem z pohledu matematické didaktiky. Hlavním problémem je vysvětlení tématu pochopitelně a pouze s minimálním matematickým pozadím absolventů statistických školení, kteří často nejsou absolventy matematických či technických oborů, ale při své práci regresní modely pravidelně používají.

Rigorózní práce se snaží na konkrétních příkladech navrhnout, jak přehledně laicky vysvětlit problematiku regresních modelů. Na praktických příkladech jsou názorně demonstrovány dopady nejčastějších chyb, kterých se laici při jejich použití v praxi dopouštějí. Hlavním výstupem práce je portfolio příkladů využitelných při školeních regresních modelů.

Klíčová slova:

regresní modely, regresní analýza, analýza závislostí, didaktika, výuka

Title: "Regression models and their teaching"

Author: Mgr. Marian Rybář

Department: Department of Mathematics Education

Abstract:

Regression models and their outputs have a huge utilization not only in the field of medicine, science and managerial decision-making but also in many other fields. From the math didactics point of view is a question of intelligible explanation of regression models to non-math students very topical and exploitable theme. The main problem is an understandable explanation of this topic to statistics schooling members with minimal math application. These people aren´t usually math or technical branch alumni, but they use regression models in their work.

This work tries to suggest simply and clearly explanation of regression models on particular examples. Practical examples show the effects of the most common mistakes that are made by laymen in application of regression models. The main output of this work is a set of examples that could be used in regression models schooling.

Keywords:

regression models, regression analysis, didactics, teaching

Obsah

Úvod / Předmluva

Problematika srozumitelného vysvětlení regresních modelů absolventům nematematických oborů je velmi aktuálním a využitelným tématem z pohledu matematické didaktiky. Na jedné straně leží velké praktické využití regresních modelů v oblasti medicíny, manažerského rozhodování či vědy a velký zájem o pochopení této problematiky. Na straně druhé však leží dosti velká náročnost vysvětlení tématu pochopitelně a pouze s minimálním matematickým pozadím absolventů statistických školení.

S tímto problémem jsem se opakovaně setkával nejen při statistických školeních lékařů, manažerů a vědeckých pracovníků, ale také při řešení matematických projektů, kdy je nezbytné najít společnou řeč mezi matematikem a budoucím uživatelem vytvořeného regresního modelu. Často jsem také dostával při statistických školeních (zvláště od lékařů) dotaz, kterou jednoduchou a přehlednou publikaci s praktickými příklady pro pochopení regresních modelů bych mohl doporučit či poskytnout.

Přes poměrně velké úsilí a studium velkého počtu publikací zabývajících se danou problematikou jsem však nenarazil na publikaci, kterou bych mohl tazatelům doporučit s vědomím, že po jejím přečtení problematiku regresního modelování plně zvládnou. To také ovlivnilo výběr tématu mé rigorózní práce.

Propast mezi matematickými znalostmi zvláště v případě lékařů a matematickou úrovní většiny odborných publikací zabývajících se regresními modely je dle mých zkušeností opravdu velká. Orientace ve vzorcích, rovnicích a matematických termínech, na kterých bývá podstata problému často vysvětlována, činí zájemcům o pochopení problém.

Při vysvětlování regresních modelů nematematikům je potom nezbytné tyto složitější vzorce a teoretická odvození obcházet a spolehnout se především na pochopení látky na ukázkách konkrétních příkladů nejlépe blízkých dané skupině účastníků školení.

Ve své rigorózní práci jsem se snažil právě na konkrétních příkladech, se kterými jsem se setkal ve své praxi, navrhnout, jak přehledně vysvětlit problematiku regresních modelů. Na praktických příkladech jsem se také pokusil ukázat dopady nejčastějších chyb, kterých se laici při jejich použití dopouštějí. Jedním z cílů práce bylo tedy vytvořit sobě a případně dalším kolegům portfolio zajímavých příkladů využitelných při školeních regresních modelů.

Bohužel jsem si vědom, že v některých částech textu musela matematická a statistická korektnost a úplnost částečně ustoupit didaktickým účelům. Dovoluji si však předpokládat, že čtenář, který pochopí základy problematiky a je motivován k dalšímu studiu, sáhne následně po některé z pokročilejších publikací, zabývajících se problematikou regresního modelování podrobněji a více i v teoretické rovině.

Z českojazyčných publikací bych zájemcům o další vzdělávání v oblasti regresního modelování doporučil ze spíše teoreticky zaměřených pramenů například: Zvára K. – Regrese (1), Anděl J. – Statistické metody (2), Hebák P. – Regrese – I. a II. část (3). Z prakticky zaměřených publikací bych potom doporučil: Komárek A., Komárková L. - Statistická analýza závislostí s příklady v R (4).

Z cizojazyčných publikací bych doporučil zejména: Clarke B. – Linear Models (5) ze spíše teoreticky zaměřených publikací a Draper N. – Applied Regression analysis (6), Harrell F. – Regression Modeling Strategies (7), Faraway J. – Linear Models with R (8) z více prakticky zaměřených publikací.

Pro demonstraci statistických výstupů jsem v této práci použil až na několik výjimek software STATISTICA 10 od společnosti StatSoft, Inc. (Tulsa, USA). Tento software bych na základě svých zkušeností začínajícím uživatelům statistiky a regresních modelů doporučil i v konkurenci dalších známých statistických software (SPSS, SAS, R, Statgraphics, SigmaPlot, GraphPad apod.), se kterými jsem měl možnost se v rámci svého zaměstnání podrobně seznámit. Kromě jiných výhod je jedním z důvodů pro doporučení uvedeného software také to, že jako jediný je nabízen v poměrně kvalitní českojazyčné verzi.

Členění práce

Po počátečním Úvodu a krátkém popisu historie regresních modelů je v Kapitole 1 „Návrh optimálního didaktického postupu při vysvětlení regresní analýzy“ na praktickém příkladu závislosti tržeb na investicích do reklamy postupně od jednoduššího po složitější stavěna posloupnost informací potřebných pro pochopení regresních modelů. Od jednoduchého regresního modelu je penzum potřebných znalostí postupně rozšiřováno o použití nelineárních regresních funkcí až po vícenásobný regresní model. V této kapitole jsem se pokusil vysvětlit danou problematiku a teorii jen s použitím minimálního množství matematických odvození, termínů a symboliky za účelem umožnění pochopení i s minimálním matematickým a statistickým pozadím. Vysvětlení teorie s použitím matematických odvození jsem se snažil nahradit ukázkami na konkrétních výstupech a výsledcích daného příkladu.

V Kapitole 2 „Praktické použití vytvořeného regresního modelu pro analýzu nákladů a tržeb a optimalizaci investic ve firmě“ jsem se s využitím jednoduchých derivací pokusil ukázat, jak by bylo možno výsledný regresní model z Kapitoly 1 použít pro praktickou optimalizaci a plánování investic ve firmě.

V Kapitolách 3-17 jsou na jednoduchých příkladech demonstrovány dopady nejčastějších chyb, kterých se laici při použití regresních modelů dopouštějí a které mohou znehodnotit práci s regresním modelem.

V Diskusi a Závěru jsou na konci práce shrnuta autorova doporučení, jak efektivně a didakticky pojmout problematiku vysvětlení regresních modelů. Současně jsou v této části rozděleny problémy při aplikaci regresních modelů na závažné, středně závažné a méně závažné z hlediska dopadu na přesnost regresních modelů.

Cílové skupiny

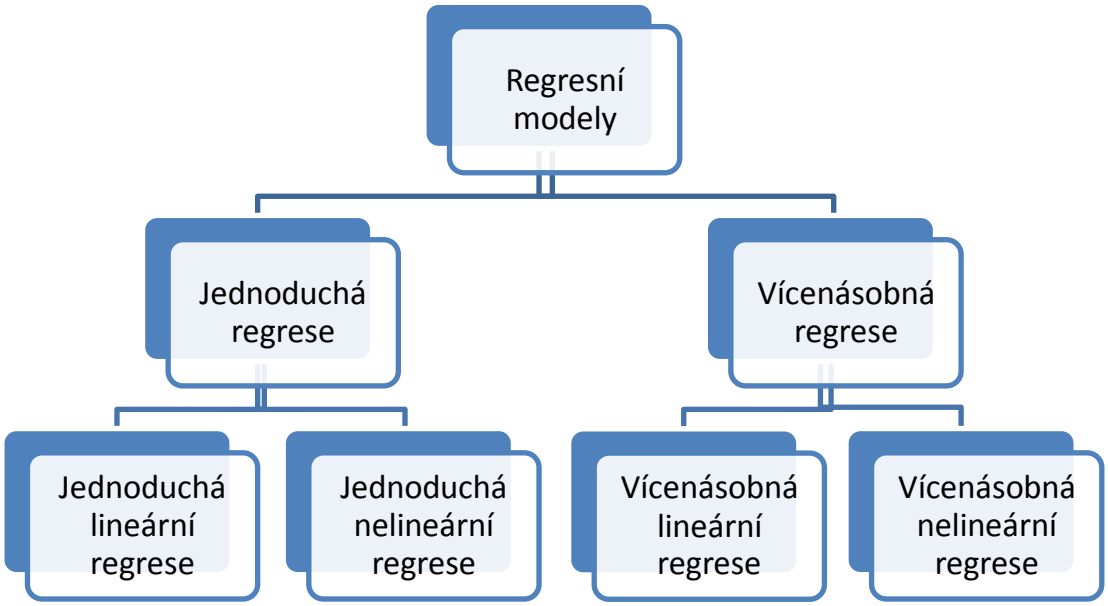
Cílovými skupinami pro vysvětlení regresních modelů jsou v této práci uvažováni zejména lékaři a vědečtí pracovníci, pracovníci technických a IT oborů, manažeři a ekonomové, případně zaměstnanci státní sféry. V Tabulce 1 jsou tyto cílové skupiny charakterizovány blíže z hlediska vstupních znalostí matematiky a statistiky a možného využití regresních modelů v jejich pracovním zaměření.

	Vstupní znalosti matematiky a statistiky	Příklady využití regresních modelů v pracovním zaměření
Lékaři a vědečtí pracovníci	Většinou absolventi lékařských a přírodovědeckých fakult. Znalost matematiky a statistiky spíše na úrovni SŠ. V rámci vysokoškolského studijního plánu matematika a statistika pouze v podobě volitelných přednášek případně jednosemestrálně. Častou motivací pro výběr školy je právě nepřítomnost matematiky a statistiky při studiu. Znalost názvů základních statistických metod z lékařských a vědeckých publikací. Z oblasti regresních modelů povědomí většinou o pojmu korelace - využíváno však v kontextu závislosti jakýchkoliv (i kategoriálních) veličin.	Odhad koncentrace či množství látky v těle na základě vstupního množství aplikovaného léčiva. Odhad potřebného okruhu ozáření na základě referenčního měření. Predikce pravděpodobnosti výskytu onemocnění použitím logistické regrese. Identifikace významných faktorů pro vznik daného onemocnění. Substituce časově či finančně náročného měření jiným vzájemně korelovaným dostupnějším měřením.
Vědečtí pracovníci technických a IT oborů	Většinou absolventi technických a IT fakult. Matematika i statistika většinou tří a vícesemestrální. Velmi dobrá orientace v matematické i statistické symbolice a textu.	Pochopení regrese jako základu pro programování některých aplikací v případě IT oborů. Substituce časově či finančně náročného měření jiným vzájemně korelovaným dostupnějším měřením u vědeckých oborů.
Manažeři a ekonomové	Většinou absolventi ekonomických fakult. Matematika i statistika nejčastěji dvousemestrální. Znalost statistiky pasivní bez schopnosti ovládání statistického software. Z oblasti regresních modelů většinou znalost základů metody a možnosti jejího použití.	Kredit skóring klientů bank před poskytnutím úvěru s využitím logistické regrese. Oceňování obchodních společností a nemovitostí. Optimalizace investic ve společnosti řešením extrémů a inflexních bodů produkčních regresních funkcí. Odhad pravděpodobnosti odchodu zákazníka s využitím logistické regrese. Předpověď ceny komodity na základě externě odhadnutých budoucích cen jiných vysvětlujících komodit.
Zaměstnanci státní sféry	Většinou absolventi ekonomických či obchodních oborů.	Oceňování obchodních společností a nemovitostí. Odhad pravděpodobnosti odchodu zákazníka s využitím logistické regrese. Předpověď ceny komodity na základě externě odhadnutých budoucích cen jiných vysvětlujících komodit.

Tabulka 1 – přehled cílových skupin pro vysvětlení regresních modelů

Rozdělení regresních modelů

Základní regresní modely můžeme rozdělit na jednoduché a vícenásobné. V případě jednoduchých regresních modelů je k dispozici pouze jedna vysvětlující proměnná x, v případě vícenásobných regresních modelů je k dispozici sada vysvětlujících proměnných $x_1,\ldots, x_p$. Jednoduché regresní modely můžeme následně ještě rozdělit na jednoduché lineární modely, kde grafem závislosti je přímka a jednoduché nelineární modely, kde grafem závislosti je nějaká nelineární funkce. Vícenásobné regresní modely jsou v této práci probírány opět na úrovni lineárních modelů a nelineárních modelů. Schématické rozdělení typů regresních modelů probíraných v této práci je možno vidět v Obr. 1.



Obr. 1 – přehled regresních modelů uvažovaných v práci

Pokročilejší regresní modely typu vícenásobných logistických a multinomických regresních modelů či Poissonových regresních modelů nebyly v této práci z důvodu zjednodušení uvažovány. Tyto modely by už navíc vyžadovaly při aplikaci v praxi úzkou spolupráci se statistikem a jejich jednoduché vysvětlení bez matematického a statistického pozadí by bylo zavádějící.

Krátká historie regresních modelů

První zmínky

Metoda nejmenších čtverců - základ regresního modelování - byla v roce 1805 prvně publikována francouzským matematikem Adrienem Mariem Legendrem v díle: *Nouvelles méthodes pour la détermination des orbites des comètes* a později německým matematikem Carlem Friedrichem Gaussem v roce 1809 v díle: *Theoria Motus Corporum Coelestium v Sectionibus Conicis Solem Ambientum*. Legendre a Gauss používali metodu nejmenších čtverců pro určení dráhy nebeských těles kolem Slunce. Přesný popis chování nebeských těles byl klíčem k určení přesných pozic lodí na otevřených oceánech. Následně v roce 1821 Gauss zveřejnil další rozvoj teorie metody nejmenších čtverců v díle: *Theoria combinationis observationum erroribus minimis obnoxiae* včetně tzv. Gauss-Markova teorému. (9)

Původ pojmu slova regrese

Pojem regrese pochází z prací antropologa a meteorologa Francise Galtona, které předložil veřejnosti v letech 1877 až 1885. Galton se zabýval obecnými otázkami dědičnosti a konkrétně se zajímal o vztah mezi výškou otců a jejich prvorozených synů. A to v publikacích: *Typical laws of heredity*. Nature 15, 1877, *Regression towards mediocrity in hereditary stature*. Journ. Anthrop. Inst. XV, 1886 a *Family likeness in stature*. Proc. Roy. Soc. XL, 1886. Pozorováním a analýzou údajů došel k rovnici, ze které vyplývá, že vysocí otcové sice mají i vysoké syny, ale v průměru menší než jsou sami, a podobně i malí otcové mají i malé syny, ale v průměru větší než jsou sami. Tuto tendenci návratu následující generace směrem k průměru nazval Galton regresí (původně tomuto jevu říkal Galton *reversion*, což později změnil na *regression* = krok zpět). Současné pojetí regresní analýzy má sice jen málo společného s původním záměrem Galtona, nicméně myšlenka přístupu k empirickým údajům zůstala zachována a pojem regrese se natolik vžil, že se používá dodnes. (3)

1 Návrh optimálního didaktického postupu při vysvětlení regresní analýzy

Ve většině publikací, vysvětlujících regresní analýzu, je kapitola o regresních modelech tvořena zejména výčtem řady všech vzorců, které se při aplikaci regresních modelů používají. Vzorce jsou uvedeny často včetně jejich odvození. Na závěr kapitoly bývá obvykle uveden jeden zkrácený ukázkový příklad s již vypočtenými výsledky, které si však často čtenář neumí propojit s teoretickou částí.

V této kapitole budeme demonstrovat alternativní postup vysvětlení práce s regresními modely na konkrétním příkladu použití v běžné praxi firem, které se pomocí investic do reklamy a dalších marketingových pobídek snaží maximalizovat své tržby.

Teorie potřebná pro vysvětlení problematiky prostupuje postupně celým příkladem a příslušné vzorce se objevují vždy v té části příkladu, kde je jejich použití nezbytné. Moderní statistika probíhá již několik desetiletí se zpřístupněním statistického software pouze v tomto prostředí a přímý výpočet pomocí vzorců pouze s použitím kalkulátoru již není v případě větších datových souborů ani možný. Přestože je z tohoto důvodu v práci kladen důraz zejména na podrobnou interpretaci výstupů ze statistického software, pro lepší pochopení jsou v některých částech příkladu dnes již spíše historické výpočty přímo ze vzorců demonstrovány.

Níže uvedený příklad si bere za cíl demonstrovat všechny základní potřebné znalosti z oblasti regresních modelů najednou v rámci jednoho zadání. Jednotlivé kapitoly jsou uvedeny vždy konkrétní otázkou, která by mohla uživatele regresního modelu v praxi zajímat.

Příklad 1.1: Představme si, že jsme marketingovým ředitelem střední firmy. Máme zkušenosti o návratnosti investic do reklamy z posledních 12 reklamních akcí. Příslušná data jsou uvedena v Tabulce 1.1.

x (mil.Kč) Investice	y (mil.Kč) Tržby
2	24,9
10	27,9
15	36,5
19	41,1
23	74,5
27	81,2
30	120,6
40	152,7
42	161,1
50	168,2
54	157,5
59	149,2

Tabulka 1.1 – data závislosti tržeb (y) na investicích do reklamy (x)

Úkolem je nyní odpovědět na následující otázky, které by mohly zajímat vedení společnosti.

1.1 Jednoduchá lineární regrese

a) Jakým způsobem závisí tržby společnosti na investicích do reklamy?

Závislost tržby společnosti (y) na investicích do reklamy (x) můžeme popsat lineární regresní rovnicí

$$y = b_0 + b_1.x \tag{1.1}$$

kde regresní parametry b_1 a b_0 je možno bez použití statistického software vypočítat dle vzorců

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \tag{1.2}$$

$$b_0 = \frac{\sum y_i}{n} - b_1 \frac{\sum x_i}{n} \tag{1.3}$$

Příslušné úpravy dat nezbytné pro dosazení do vzorců pro výpočet regresních parametrů b_1 a b_0 jsou v Tabulce 1.2.

x	y	x.y	x^2	y^2
2	24,9	49,8	4	620,01
10	27,9	279	100	778,41
15	36,5	547,5	225	1332,25
19	41,1	780,9	361	1689,21
23	74,5	1713,5	529	5550,25
27	81,2	2192,4	729	6593,44
30	120,6	3618	900	14544,36
40	152,7	6108	1600	23317,29
42	161,1	6766,2	1764	25953,21
50	168,2	8410	2500	28291,24
54	157,5	8505	2916	24806,25
59	149,2	8802,8	3481	22260,64
$\sum$ x	$\sum$ y	$\sum$ x.y	$\sum x^2$	$\sum y^2$
371	**1195,4**	**47773,1**	**15109**	**155736,6**

Tabulka 1.2 – úpravy dat nezbytné pro dosazení do vzorců pro výpočet regresních parametrů b_1 a b_0

Vypočtené hodnoty nyní dosadíme do vzorců regresních parametrů b_1 a b_0

$$b_1 = \frac{12.47773,1 - 371.1195,4}{12.15109 - 371^2} = 2,9721$$

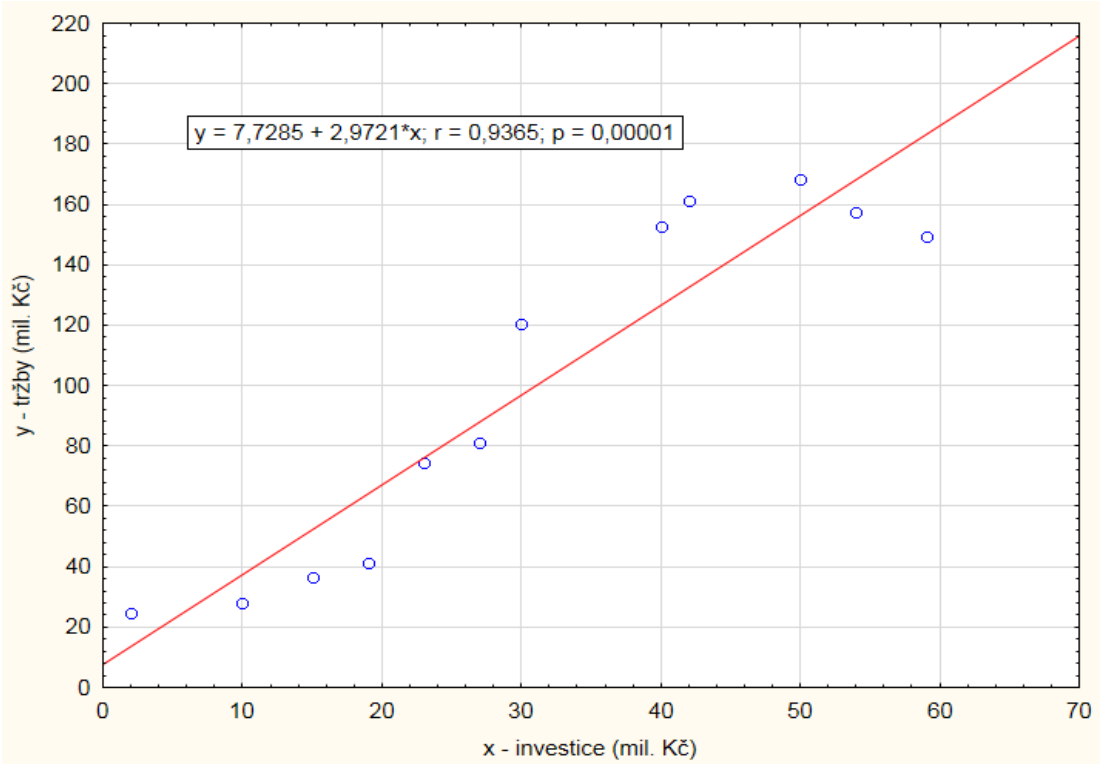
$$b_0 = \frac{1195,4}{12} - 2,9721.\frac{371}{12} = 7,7285$$

Vypočtené regresní parametry b_0 a b_1 dosadíme zpět do regresní rovnice $y = b_0+b_1.x$ a výsledná podoba modelu závislosti tržeb společnosti na investicích do reklamy je tedy $y = 7,7285 + 2,9721.x$.

Stejné výsledky je možno alternativně efektivněji získat použitím statistického software. Srovnáním výsledných výstupů ze statistického software s předchozími ručně získanými výsledky si můžeme potvrdit, že výsledná regresní rovnice se neliší. Hledané odhady regresních koeficientů můžeme najít ve sloupci b v Obr. 1.1 a grafické znázornění v Obr. 1.2. Výsledná rovnice má tedy stejný tvar $y = 7,7285 + 2,9721.x$, jako v případě výpočtu ze vzorce.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,93645686 R2= ,87695146 Upravené R2= ,86464660 F(1,10)=71,269 p<,00001 Směrod. chyba odhadu : 21,238					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			7,728456	12,49237	0,618654	0,549976
x - investice (mil. Kč)	0,936457	0,110927	2,972125	0,35206	8,442082	0,000007

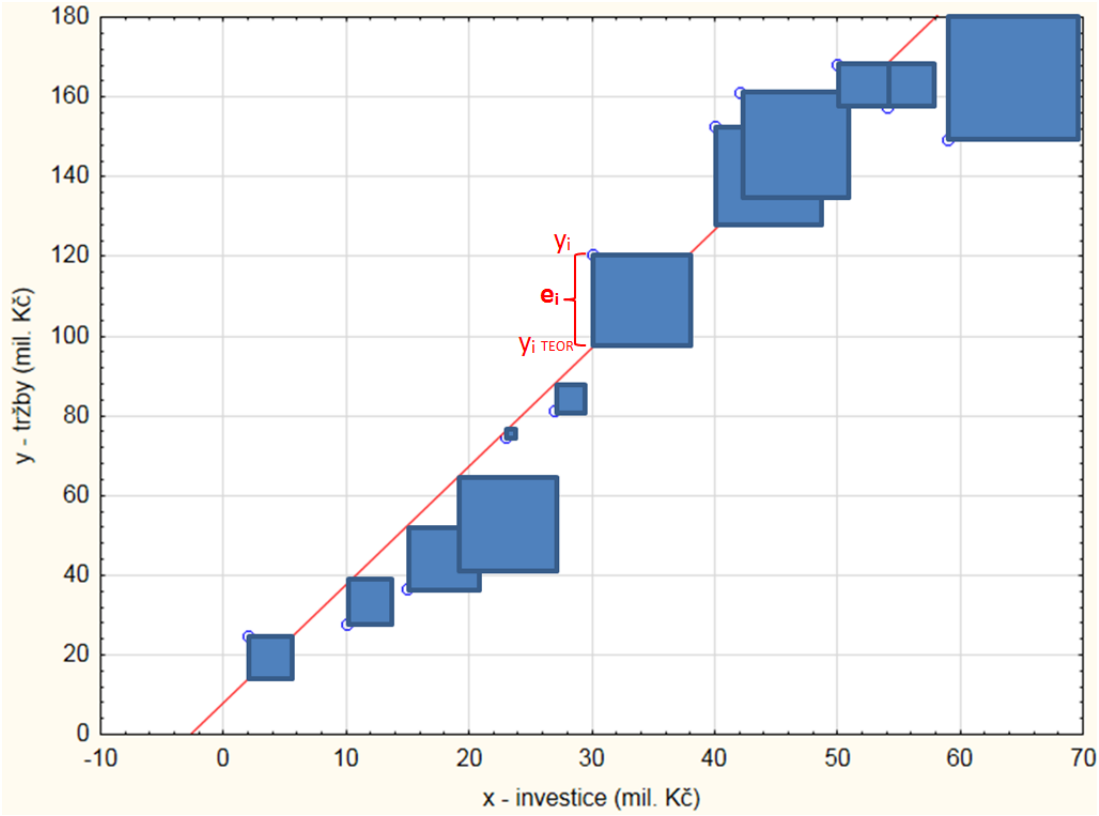
Obr. 1.1 – odhady regresních koeficientů b_0 a b_1 (Výstup ze sw Statistica 10)



Obr. 1.2 – grafické znázornění regresní přímky včetně odhadnuté rovnice závislosti (Výstup ze sw Statistica 10)

Odhady regresních parametrů využitím Metody nejmenších čtverců

Odhady regresních parametrů b_1 a b_0 jsou odvozeny použitím Metody nejmenších čtverců. Podstatu metody je možno dobře demonstrovat na Obrázku 1.3. Optimální regresní parametry b_1 a b_0 přímky $y = b_0+b_1.x$ jsou určeny tak, aby součet čtverců na Obrázku 1.3 dle vzorce $S_R = \sum_{i=1}^{n}(y_i - y_{iTeor})^2$ byl minimální. Jak je možno dobře sledovat na straně popsaného čtverce v obrázku, y_i ve vzorci reprezentuje reálnou naměřenou hodnotu tržeb y, zatímco y_{iTeor} reprezentuje teoretickou hodnotu tržeb y, která vznikne dosazením příslušné hodnoty investic x do regresní rovnice. Rozdíly těchto dvou hodnot, definované vztahem $e_i = y_i - y_{iTeor}$ jsou označovány jako **rezidua** regresního modelu a reprezentují v podstatě strany čtverců v obrázku, které se snažíme minimalizovat.



Obr. 1.3 – grafické znázornění principu Metody nejmenších čtverců (upravený výstup ze sw Statistica 10)

Poznámka: Odvození regresních parametrů b_1 a b_0 je prakticky provedeno minimalizací funkce čtverců

$$S_R = \sum_{i=1}^{n}(y_i - (b_0 + b_1.x))^2$$

která vznikne dosazením výrazu y_{iTeor} = b_0 + b_1.x do dříve uvedeného vzorce pro součet čtverců S_R. Získaná funkce je nejdříve parciálně derivována podle obou proměnných b_1 a b_0 a obě parciální derivace jsou následně položeny rovno 0.

Řešením vzniklé soustavy jsou odvozeny vzorce (1.2) a (1.3) pro odhad regresních parametrů b_1 a b_0. Podrobněji je možno se s metodou nejmenších čtverců seznámit například v publikaci: Hebák P. - Regrese – I. část (3).

b) Jak je možné prakticky interpretovat regresní koeficienty b_0 a b_1 na příkladu tržeb a investic do reklamy?

Regresní koeficient b_0 můžeme graficky interpretovat jako průsečík regresní přímky s osou y (viz Obr. 1.2). Je to hodnota tržeb při nulové hodnotě investice do reklamy (x = 0). V našem ukázkovém příkladu by byla hodnota tržeb při nulové investici do reklamy tedy rovna

y = 7,7285 + 2,9721.0 = 7,7285 mil. Kč.

Regresní koeficient b_1 je možno si představit jako sklon (směrnici) přímky. Pokud je regresní koeficient b_1 kladný, přímka je rostoucí. Pokud je regresní koeficient b_1 záporný, přímka je klesající. Čím je hodnota regresního koeficientu b_1 v absolutní hodnotě vyšší, tím má přímka větší sklon. Pro praktické použití koeficient b_1 říká, o kolik jednotek se zvýší hodnota tržeb y, pokud se hodnota investic do reklamy x zvýší o jednotku.

V našem případě by tedy každá další investovaná jednotka (1 mil. Kč) do reklamy měla zvýšit tržby o 2,9721 mil. Kč. Vedení společnosti tedy může očekávat, že investice do reklamy bude dobře návratná.

c) **Na příští období plánujeme investiční akci do reklamy ve výši 60 mil. Kč. Jakou hodnotu tržeb můžeme na základě předchozích dat očekávat při této investiční akci?**

Z předchozích dat společnosti již máme odvozen regresní model závislosti tržeb na investicích do reklamy y = 7,7285 + 2,9721.x. Pokud nyní dosadíme hodnotu x = 60 mil. Kč do získaného regresního modelu, vypočteme příslušnou očekávanou hodnotu tržeb

y = 7,7285 + 2,9721.60 = 186,0545 mil. Kč

Při investicích do reklamy ve výši 60 mil. Kč můžeme v příštím období očekávat tržby ve výši 186,0545 mil. Kč.

d) **V jakém intervalu může management společnosti očekávat odhadované tržby pro investici do reklamy ve výši 60 mil. Kč.**

Pro intervalový odhad predikované hodnoty vysvětlované proměnné se běžně používá 95 % konfidenční nebo 95 % predikční interval. Jak bude podrobně popsáno v Kapitole 9, každý z intervalů má jinou interpretaci. Pokud bude management společnosti zajímat **průměrná** hodnota předpovídaných tržeb, použije pro odhad intervalu tržeb konfidenční interval. Pokud bude management společnosti zajímat **jedna konkrétní hodnota** předpovídané tržby pro konkrétní investiční reklamní akci, použije pro odhad intervalu tržeb predikční interval.

Při stejné hodnotě spolehlivosti je predikční interval vždy širší než konfidenční interval.

Pokud tedy bude management společnosti zajímat **průměrná** hodnota předpovídaných tržeb, použije pro odhad intervalu tržeb konfidenční interval.

Proměnná	Předpovězené hodnoty proměnné: y - tržby (mil. Kč)		
	b-váha	Hodnota	b-váha * Hodnot
x - investice (mil. Kč)	2,972125	60,00000	178,3275
Abs. člen			7,7285
Předpověď			186,0560
-95,0%LS			159,4649
+95,0%LS			212,6470

Obr. 1.4 – 95 % konfidenční interval pro průměrnou hodnotu předpovídané tržby (Výstup ze sw Statistica 10)

Management společnosti tedy může očekávat, že průměrná hodnota předpovídaných tržeb pro investici do reklamy ve výši 60 mil. Kč by se měla s 95 % spolehlivostí pohybovat v intervalu (159,4649; 212,6470).

Pokud bude management společnosti naopak zajímat **jedna konkrétní hodnota** předpovídané tržby pro konkrétní investiční reklamní akci, použije pro odhad intervalu tržeb predikční interval.

Proměnná	Předpovězené hodnoty proměnné: y - tržby (mil. Kč)		
	b-váha	Hodnota	b-váha * Hodnot
x - investice (mil. Kč)	2,972125	60,00000	178,3275
Abs. člen			7,7285
Předpověď			186,0560
-95,0%PL			131,7763
+95,0%PL			240,3356

Obr. 1.5 – 95 % predikční interval pro jednu konkrétní hodnotu předpovídané tržby (Výstup ze sw Statistica 10)

Management společnosti tedy může očekávat, že jedna konkrétní hodnota předpovídané tržby pro investici do reklamy ve výši 60 mil. Kč by se měla s 95 % spolehlivostí pohybovat v intervalu (131,7763; 240,3356).

Poznámka: Při podrobné analýze výstupů pro konfidenční a predikční interval v Obr. 1.4 a 1.5 je možno si všimnout, že středy obou intervalů v řádku „Předpověď“ odpovídají předpovězené hodnotě z odstavce c) na předchozí straně.

e) O kolik se zvýší tržby, pokud zvýšíme investice do reklamy o 2 mil. Kč?

Jak již bylo uvedeno, koeficient b_1 říká, o kolik jednotek se zvýší hodnota tržeb y, pokud se investice do reklamy x zvýší o jednotku. Každá další jednotka (1 mil. Kč)

investovaná do reklamy by tedy měla zvýšit tržby o 2,9721 mil. Kč. Z uvedeného vyplývá vztah

$$\Delta y = b_1 . \Delta x \qquad (1.4)$$

Po dosazení tedy získáváme

$\Delta y = 2{,}9721 . 2 = 5{,}9442$ mil. Kč

Pokud tedy zvýšíme investice do reklamy o 2 mil. Kč, tržby se zvýší o 5,9442 mil. Kč.

f) Z důvodu udržení cash flow potřebujeme v příštím období dosáhnout tržeb 120 mil. Kč. Na základě předchozích dat potřebujeme odhadnout, jakou částku bychom měli v následujícím období investovat do reklamy, abychom dosáhli plánovaných tržeb.

Při první úvaze by se nabízela možnost jednoduše dosadit hodnotu tržeb 120 mil. Kč do dříve vypočítané regresní rovnice **y = 7,7285 + 2,9721.x** za y a z rovnice následně vyjádřit hledanou hodnotu potřebných investic x (nesprávný výsledek 37,7797 mil. Kč). Jak bude dále podrobně rozvedeno v Kapitole 6, tento přístup není správný. Nevede sice k úplně nesmyslnému výsledku, ale k odhadu musí být pro přesný odhad použita nově vypočtená sdružená regresní přímka závislosti proměnné x na proměnné y. Pro výpočet stačí ve vzorci zaměnit hodnoty y s hodnotami x, případně ve statistickém software zaměnit vysvětlovanou a vysvětlující proměnnou.

Výsledná rovnice sdružené regresní přímky je

$x = 1{,}5239 + 0{,}2951.y$

Do sdružené regresní rovnice nyní dosadíme za y hodnotu plánovaných tržeb y = 120 mil. Kč.

$x = 1{,}5239 + 0{,}2951.120$ mil. Kč = 36,9359 mil. Kč

Abychom dosáhli v následujícím období plánovaných tržeb 120 mil. Kč, měli bychom investovat do reklamy 36,9359 mil. Kč.

g) Jak velká je síla závislosti mezi tržbami y a investicemi do reklamy x? Neboli: Do jaké míry je možno danému modelu věřit?

Sílu lineární závislosti mezi dvěma číselnými veličinami je možno vyjádřit pomocí Pearsonova korelačního koeficientu r. Korelační koeficient nabývá vždy hodnot v rozmezí -1 až 1. Pokud vyjde r blízko -1, znamená to nepřímou lineární závislost mezi veličinami (čím je větší jedna veličina, tím je menší druhá). Pokud vyjde r blízko +1, znamená to přímou lineární závislost (čím je větší jedna veličina, tím je větší i druhá). Pokud vyjde r blízko 0, znamená to, že veličiny jsou lineárně nezávislé.

Pearsonův korelační koeficient r je možno vypočítat dle vzorce

$$r = \frac{n.\sum x_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} \tag{1.5}$$

Dosazením hodnot do vzorce získáváme

$$r = \frac{12.47773{,}1 - 371.1195{,}4}{\sqrt{[12.15109 - 371^2][12.155737 - 1195{,}4^2]}} = 0{,}9365$$

Stejnou hodnotu korelačního koeficientu získáme také použitím statistického software.

	Korelace	
Proměnná	x - investice (mil. Kč)	y - tržby (mil. Kč)
x - investice (mil. Kč)	1,0000	,9365
	p= ---	p=,000
y - tržby (mil. Kč)	,9365	1,0000
	p=,000	p= ---

Obr. 1.6 – odhad Pearsonova korelačního koeficientu r (Výstup ze sw Statistica 10)

Dle výsledné hodnoty korelačního koeficientu potom rozlišujeme následující typy závislosti:

$0 < |r| \leq 0{,}3$ Slabá závislost

$0{,}3 < |r| \leq 0{,}8$ Středně silná závislost

$0{,}8 < |r| \leq 1$ Silná závislost

Korelační koeficient příslušný k našemu regresnímu modelu vyšel r = 0,9365. Jedná se tedy o velmi silnou přímou závislost mezi tržbami y a investicemi do reklamy x v našem modelu.

h) Je závislost mezi tržbami y a investicemi do reklamy x statisticky významná?

Míru závislosti není možno posuzovat pouze podle hodnoty korelačního koeficientu, ale také podle statistické významnosti příslušného korelačního koeficientu. Ta je kromě samotné hodnoty korelačního koeficientu závislá také na počtu měření n v souboru. Alternativní hypotézou, kterou chceme v tomto případě potvrdit je, že korelační koeficient r se statisticky významně liší od 0.

Testové kritérium, které měří, jak významně je korelační koeficient odlišný od nuly má tvar

$$t = \frac{r}{\sqrt{1-r^2}} . \sqrt{n-2} \tag{1.6}$$

Po dosazení do vzorce získáváme hodnotu testového kritéria

$$t = \frac{0{,}9365}{\sqrt{1-0{,}9365^2}} . \sqrt{12-2} = 8{,}4452$$

Aby byla závislost považována za statisticky významnou, musí být překročena tabulková 5% kritická hodnota na pravé straně kritického oboru

$$t > t_{1-\frac{\propto}{2}}(n-2) \tag{1.7}$$

V tabulkách dohledáme příslušnou tabulkovou 5% kritickou hodnotu z pravé strany kritického oboru.

$$t_{1-\frac{\propto}{2}}(n-2) = 2{,}56$$

Vypočtenou hodnotu testového kritéria 8,4452 nyní porovnáme s tabulkovou kritickou hodnotou 2,56, která stanovuje hranici, od které je již daný korelační koeficient statisticky významný.

$$8{,}4452 \geq 2{,}56$$

Vidíme, že vypočtené testové kritérium překročilo tabulkovou kritickou hodnotu pro α = 5 % a korelační koeficient je na 5 % hladině tedy statisticky významný.

V praxi je však rychlejší statistickou významnost korelačního koeficientu posoudit pomocí výstupu ze statistického software, kde už je hodnota testového kritéria přepočítaná pomocí inverzní funkce na tzv. statistickou významnost p. Pokud je tato hodnota menší než nastavená hladina spolehlivosti α = 5 %, považujeme korelační koeficient za statisticky významný.

	Korelace	
Proměnná	x - investice (mil. Kč)	y - tržby (mil. Kč)
x - investice (mil. Kč)	1,0000	,9365
	p= ---	p=,000
y - tržby (mil. Kč)	,9365	1,0000
	p=,000	p= ---

Obr. 1.7 – posouzení statistické významnosti korelačního koeficientu r (Výstup ze sw Statistica 10)

Korelační koeficient pro závislost tržeb y na investicích do reklamy x je tedy statisticky významný.

i) Z jaké části se v naší společnosti podařilo pomocí regresního modelu vysvětlit závislost tržeb na investicích do reklamy a jaká část zůstala nevysvětlena?

Míra lineární závislosti vysvětlované proměnné y na vysvětlujících proměnných x v daném modelu je vyjádřena koeficientem determinace R^2. Koeficient determinace

nabývá hodnot v rozmezí 0 až 1 (respektive 0 až 100 %) a čím je jeho hodnota vyšší, tím je model vhodnější.

V případě lineární závislosti je možno počítat koeficient determinace R^2 jednoduše jako druhou mocninu klasického korelačního koeficientu r.

$$R^2 = r^2 \tag{1.8}$$

$$0{,}9365^2 = 0{,}877 = 87{,}7\%$$

Hodnota koeficientu determinace je také součástí základních výstupů ze statistického software při odhadu koeficientů regresní rovnice.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,93645686 R2= ,87695146 Upravené R2= ,86464660 F(1,10)=71,269 p<,00001 Směrod. chyba odhadu : 21,238					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			7,728456	12,49237	0,618654	0,549976
x - investice (mil. Kč)	0,936457	0,110927	2,972125	0,35206	8,442082	0,000007

Obr. 1.8 – odhad koeficientu determinace R^2 (Výstup ze sw Statistica 10)

V našem lineárním modelu závislosti tržeb na investicích do reklamy vyšla hodnota koeficientu determinace R^2 = 87,7 %. Tržby v naší společnosti jsou tedy pomocí lineárního regresního modelu závislosti na investicích do reklamy vysvětleny cca z 87,7 %. Zbylých 12,3 % variability tržeb není zatím možno daným modelem vysvětlit. V další části této práce si však ukážeme, že procento vysvětlené variability budeme schopni ještě výrazně zvýšit a model tedy pro naši potřebu zlepšit.

j) Je regresní koeficient b_1 statisticky významný?

Podstatu testování významnosti s použitím statistických vzorců a tabulkové kritické hodnoty jsme si ukázali již v případě korelačního koeficientu. Protože v praxi je mnohem snazší použít pro testování významnosti statistický software, omezíme se již pouze na interpretaci příslušného výstupu ze statistického software.

Důležité je si uvědomit, že test významnosti regresního koeficientu b_1 nám poskytuje jinou informaci než test významnosti korelačního koeficientu r. Podstatou testu

významnosti regresního koeficientu b_1 je zjistit, zda sklon regresní přímky je nenulový, neboli že sledovaná závislost není konstantní. Jak bude popsáno dále v této práci, korelační koeficient a směrnice přímky (regresní koeficient b_1) jsou navzájem do určité míry nezávislé.

	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,93645686 R2= ,87695146 Upravené R2= ,86464660 F(1,10)=71,269 p<,00001 Směrod. chyba odhadu : 21,238					
N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			7,728456	12,49237	0,618654	0,549976
x - investice (mil. Kč)	0,936457	0,110927	2,972125	0,35206	8,442082	0,000007

Obr. 1.9 – posouzení statistické významnosti regresního koeficientu b_1 (Výstup ze sw Statistica 10)

Z výstupu ze statistického software vidíme, že statistická významnost p, je menší než nastavená hladina spolehlivosti $\alpha = 5$ %. Regresní koeficient b_1 můžeme tedy považovat za statisticky významně nenulový a daná závislost tedy není konstantní.

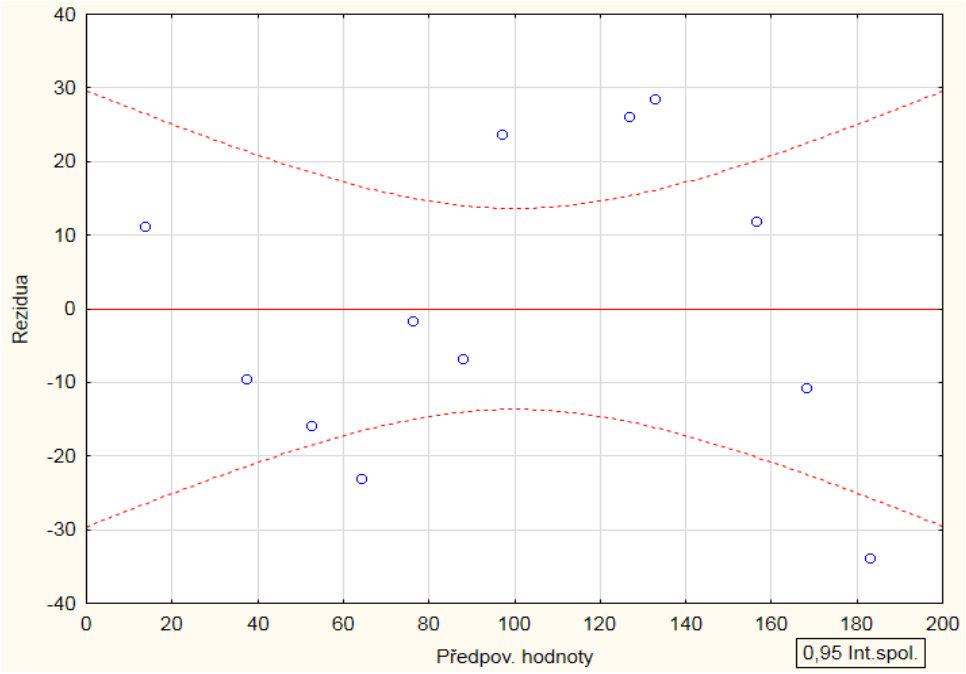
k) Jsou splněny všechny potřebné předpoklady pro použití navrženého regresního modelu a je tedy jeho použití pro predikci opodstatněné?

Jak bude podrobně probráno v dalších kapitolách, při konstrukci regresního modelu je potřeba ošetřit zejména tyto předpoklady:

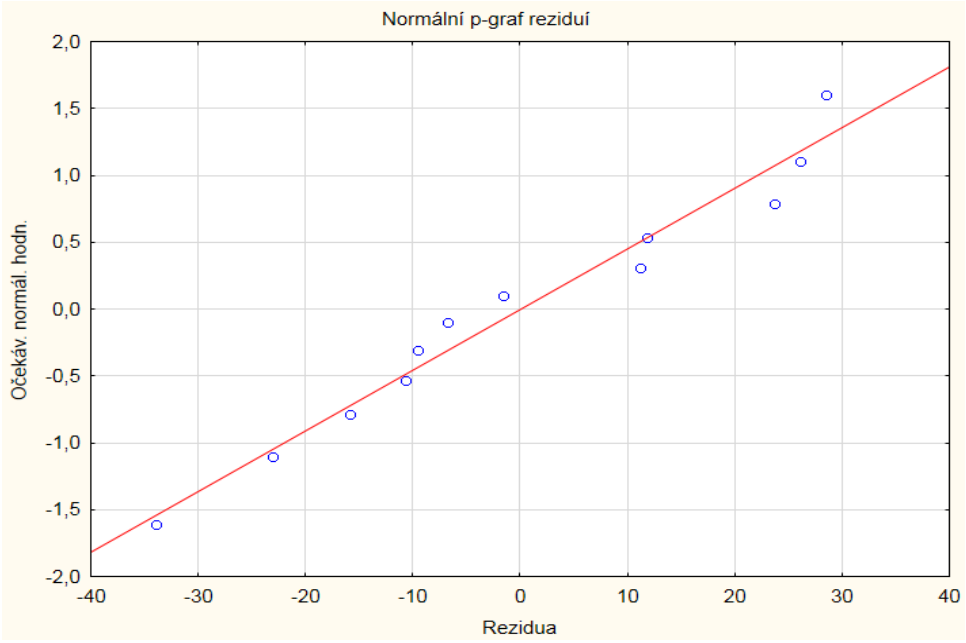
- Správný funkční tvar dané závislosti
- Homoskedasticitu reziduí
- Nepřítomnost autokorelace reziduí
- Nepřítomnost odlehlých hodnot
- Normalitu reziduí

Pro kontrolu výše uvedených předpokladů uvedeného lineárního regresního modelu si vykreslíme graf reziduí proti předpovězeným hodnotám a normální p-graf reziduí. V grafu reziduí proti předpovězeným hodnotám (Obr. 1.10) by měla být rezidua v případě splnění všech předpokladů optimálně rozložena zcela náhodně kolem horizontální osy procházející hodnotou 0 a rezidua by se neměla nacházet mimo 95 % interval spolehlivosti, který je v grafu znázorněn přerušovanou čárou.

V normálním p-grafu (Obr. 1.11) by potom měla rezidua přibližně kopírovat přímku znázorněnou v grafu.



Obr. 1.10 – graf reziduí lineárního regresního modelu proti předpovězeným hodnotám (Výstup ze sw Statistica 10)



Obr. 1.11 – normální p-graf reziduí lineárního regresního modelu (Výstup ze sw Statistica 10)

Z tvaru grafu reziduí proti předpovězeným hodnotám (Obr. 1.10) vidíme, že rezidua nejsou v grafu rozložena náhodně a vykazují trend, který by mohl ukazovat na možné porušení některého z předpokladů regresního modelu, nejspíše předpokladu správného funkčního tvaru dané závislosti a předpokladu nepřítomnosti autokorelace reziduí (bude podrobně probráno v Kapitole 14). Kromě toho se rezidua často vyskytují mimo hranice vyznačeného 95 % intervalu spolehlivosti.

Z normálního p-grafu reziduí dále vidíme, že rezidua příliš nekopírují přímku, což opět ukazuje na možné porušení některého z předpokladů regresního modelu.

Pozn.: Podrobně budou předpoklady regresního modelu probrány v Kapitole 13 -16

Pro modelování závislosti tržeb y na investicích do reklamy x zkusíme tedy použít jinou než lineární funkční závislost.

1.2 Jednoduchá regrese s použitím nelineárních funkcí

l) Je možné najít nějaký vhodnější model pro závislost tržeb v naší společnosti na investicích do reklamy?

Poznámka: V této kapitole budeme nelineárními funkcemi pro jednoduchost označovat pouze lineární funkce s transformovaným nelineárním tvarem vysvětlované proměnné x, přestože v některých publikacích jsou za nelineární označovány zejména složitější funkce nelineární v samotných parametrech.

Kromě modelu lineární regrese existují také další regresní modely, které mohou v některých případech předpovídat vysvětlovanou proměnnou lépe než lineární model. Mezi nejčastěji používané nelineární regresní funkce patří například

Kvadratická funkce

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{1.10}$$

Kubická funkce

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \tag{1.11}$$

Logaritmická funkce

$$y = \beta_0 + \beta_1 \ln x \tag{1.12}$$

Hyperbolická funkce

$$y = \beta_0 + \beta_1 \frac{1}{x} \tag{1.13}$$

Rozhodnutí, který model vybereme jako optimální, provádíme podle několika kritérií:

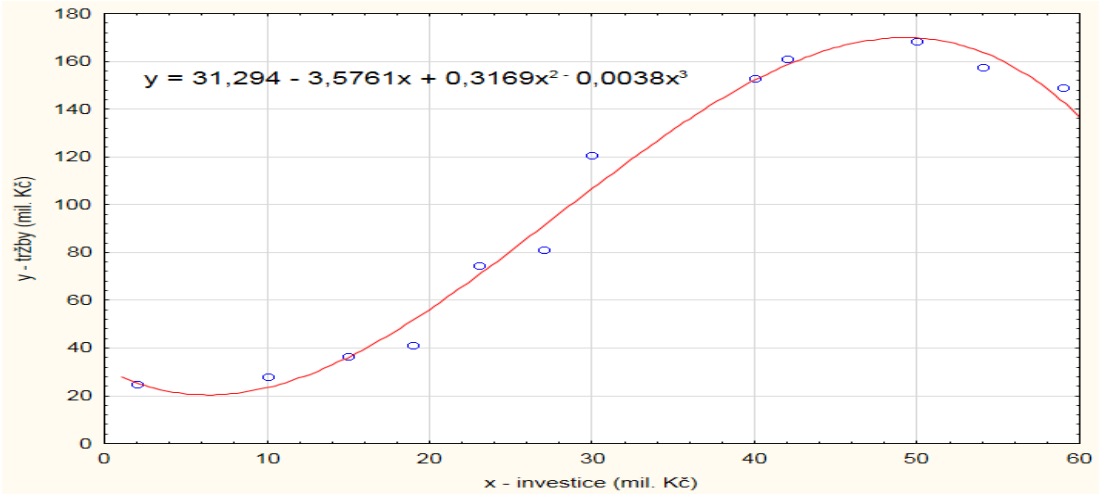
- Grafická analýza tvaru závislosti v bodovém xy grafu
- Hodnota koeficientu determinace R^2 příslušného k danému modelu
- t - testy významnosti jednotlivých parametrů
- Celkový F test vhodnosti modelu
- Splnění ostatních výše uvedených předpokladů regresního modelu

V níže uvedené tabulce je porovnání uvažovaných modelů podle koeficientu determinace R^2 a p hodnoty F-testu. Jako nejvhodnější se jeví kubický model.

Model	Regresní rovnice	Koeficient determinace R^2	p hodnota F-testu
Lineární	y = 7,7285 + 2,9721.x	87,70%	,000
Logaritmický	y = - 60,146 + 50,603.lnx	68,86%	,001
Hyperbolický	y = 117,85 -233,41.(1/x)	29,81%	,066
Kvadratický	y = - 10,458 + 4,6405x - 0,0265x^2	89,48%	,000
Kubický	**y = 31,294 - 3,5761x + 0,3169x^2 - 0,0038x^3**	**98,57%**	**,000**

Tabulka 1.3 – porovnání uvažovaných modelů podle koeficientu determinace R^2 a p hodnoty F-testu

Vhodnost kubického regresního modelu potvrzuje také proložení uvažovaného modelu bodovým grafem závislosti tržeb y na investicích do reklamy x.



Obr. 1.12 - grafické znázornění závislosti tržeb na investicích při proložení kubickým regresním modelem (Výstup ze sw Statistica 10)

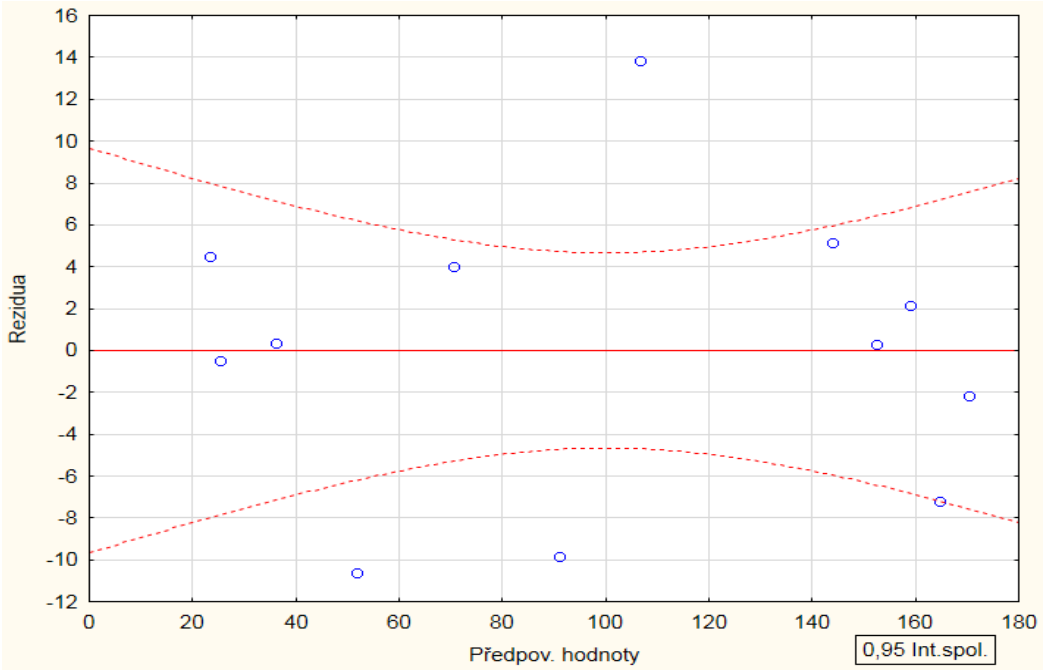
Všechny t-testy významnosti jednotlivých regresních koeficientů kubického modelu jsou statisticky významné, což potvrzuje opodstatnění jeho použití.

Model je: v2=b0+b1*v1+b2*v1^2+b3*v1^3)
Záv.prom.:y - tržby (mil. Kč)
Hladina spolehlivosti:95.0% (alfa =0.050)

	Odhad	Standard chyba	t-hodn. sv = 8	p-hodn.	Dol. sp. Mez	Hor. sp. Mez
b0	31,29368	9,504537	3,29250	0,010981	9,37618	53,21118
b1	-3,57607	1,276860	-2,80067	0,023174	-6,52051	-0,63162
b2	0,31688	0,048982	6,46930	0,000194	0,20393	0,42984
b3	-0,00379	0,000533	-7,11617	0,000100	-0,00502	-0,00256

Obr. 1.13 – t-testy významnosti jednotlivých regresních koeficientů pro kubický model (Výstup ze sw Statistica 10)

Z tvaru grafu reziduí proti předpovězeným hodnotám v případě kubického modelu vidíme, že rezidua jsou již v grafu rozložena lépe a nenasvědčují zásadnímu porušení předpokladů regresního modelu.



Obr. 1.14 – graf reziduí kubického regresního modelu proti předpovězeným hodnotám (Výstup ze sw Statistica 10)

Pro modelování závislosti tržeb na investicích do reklamy navrhneme tedy místo lineárního modelu kubický model s regresní rovnicí $y = 31{,}294 - 3{,}5761x + 0{,}3169x^2 - 0{,}0038x^3$, který splňuje potřebné předpoklady regresního modelu a vystihuje danou závislost lépe než model lineární.

Veškeré praktické odhady pro závislost tržeb na investicích s výjimkou výpočtu koeficientu determinace R^2 je nyní možno provést výše popsanými postupy pro lineární model naprosto stejně i pro kubický model.

Poznámka: Při praktickém použití výsledného kubického modelu pro odhad tržeb je třeba mít na paměti (stejně jako u původního lineárního modelu), že provádění odhadů pro x mimo interval $\langle 2; 59\ \text{mil. Kč} \rangle$, na kterém model vznikal, může být velmi zavádějící a může přinášet překvapivé a nesprávné výsledky. Tento problém, který označujeme pojmem extrapolace, bude podrobně probrán v Kapitole 10.

V případě regresních modelů vyššího stupně je navíc riziko zkreslených predikcí pro x mimo uvedený interval podstatně větší než v případě lineárních modelů, jak bude také demonstrováno v Kapitole 10.

m) Z jaké části se podařilo vysvětlit závislost tržeb na investicích do reklamy pomocí kubického regresního modelu?

Jak bude podrobně popsáno v Kapitole 3, výše definovaný koeficient korelace R a koeficient determinace R^2 je možno použít pouze v případě lineární závislosti. Není možno jej v žádném případě zaměňovat s indexem korelace I a indexem determinace I^2, které se používají v případě nelineární závislosti a většinou bývají označovány stejně jako koeficient korelace R a koeficient determinace R^2.

Výpočetní tvary pro určení hodnoty indexu korelace I a indexu determinace I^2 jsou

$$I = \sqrt{1 - \frac{S_R}{S_y}} \tag{1.14}$$

$$I^2 = 1 - \frac{S_R}{S_y} \tag{1.15}$$

kde

$S_y = \sum_{i=1}^{n}(y_i - y_{Prům})^2$ je celkový součet čtverců modelu (1.16)

$S_R = \sum_{i=1}^{n}(y_i - y_{iTeor})^2$ je reziduální součet čtverců modelu (1.17)

Podstatu výpočtu koeficientu determinace budeme demonstrovat v MS Excel. Sloupce x_i a y_i reprezentují původní data. Hodnoty ve sloupci y_{iTeor} získáme dosazením hodnot ze sloupce x_i do regresní rovnice výsledného kubického modelu $y = 31{,}294 - 3{,}5761x + 0{,}3169x^2 - 0{,}0038x^3$.

x_i	y_i	y_{iTeor}	$y_i - y_{iTeor}$	$(y_i - y_{iTeor})2$	$y_i - y_{Prům}$	$(y_i - y_{Prům})2$
2	24,9	25,38	-0,48	0,23	-80,75	6520,56
10	27,9	23,42	4,48	20,04	-77,75	6045,06
15	36,5	36,13	0,37	0,14	-69,15	4781,72
19	41,1	51,68	-10,58	112,04	-64,55	4166,70
23	74,5	70,45	4,05	16,41	-31,15	970,32
27	81,2	90,96	-9,76	95,34	-24,45	597,80
30	120,6	106,62	13,98	195,41	14,95	223,50
40	152,7	152,09	0,61	0,37	47,05	2213,70
42	161,1	158,58	2,52	6,38	55,45	3074,70
50	168,2	169,74	-1,54	2,37	62,55	3912,50
54	157,5	163,90	-6,40	40,98	51,85	2688,42
59	149,2	142,99	6,21	38,53	43,55	1896,60
$\sum x_i$	$\sum y_i$			$S_R = \sum(y_i - y_{iTeor})2$		$S_Y = \sum(y_i - y_{Prům})2$
371	**1195,4**			**528,23**		**37091,61**

Tabulka 1.4 – demonstrace výpočtu indexu determinace pomocí MS Excel

$$I^2 = 1 - \frac{528{,}23}{37091{,}61} = 98{,}57\%$$

Index determinace $I^2 = 98{,}57$ % je tedy významně vyšší než koeficient determinace v případě lineárního regresního modelu $R^2 = 87{,}7$ %. V případě, že by byl index determinace pro kubický model jen o málo vyšší než koeficient determinace v případě lineárního modelu, nedoporučuje se dobře interpretovatelný lineární model nahrazovat složitější podobou modelu.

Poznámka: Protože pojem index determinace I^2 je často v literatuře i statistických software značen R^2 a v případě lineárního modelu jsou tyto pojmy dokonce totožné, budeme dále v této práci značit koeficient determinace R^2 i index determinace I^2 shodně již pouze jedním značením R^2.

1.3 Vícenásobná lineární regrese

V případě, že vysvětlovaná proměnná y (tržby) závisí na více vysvětlujících proměnných např. x_1 (investice do reklamy), x_2 (průměrná výše slevy) a x_3 (počet direct mailů), je nutno použít model vícenásobné regrese. V případě, že pro popis dané závislosti je model lineární vícenásobné regrese jen o málo horší než případné jiné složitější nelineární vícenásobné regresní modely, bývá pro praktické použití upřednostňován. Důvodem je smysluplná interpretace regresních koeficientů a s tím související možnost kvantifikace očištěného vlivu jednotlivých vysvětlujících proměnných x_i na vysvětlovanou proměnnou y. Toto u nelineárních modelů není možné. Kvantifikace vlivu jednotlivých vysvětlujících proměnných x_i na vysvětlovanou proměnnou y je přitom jedním z hlavních praktických výstupů regresních modelů v komerční i vědecké sféře.

Vícenásobný lineární regresní model má tvar

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_px_p \tag{1.17}$$

V případě pouze dvourozměrného regresního modelu

$$y = b_0 + b_1x_1 + b_2x_2 \tag{1.18}$$

je možno bez použití statistického software provést výpočet jednotlivých regresních koeficientů a koeficientu determinace pomocí vzorců

$$b_1 = \frac{s_y}{s_{x_1}} . \frac{r_{yx_1} - r_{x_1x_2} . r_{yx_2}}{1 - r_{x_1x_2}^2} \tag{1.19}$$

$$b_2 = \frac{s_y}{s_{x_2}} . \frac{r_{y2} - r_{x_1x_2} . r_{yx_1}}{1 - r_{x_1x_2}^2} \tag{1.20}$$

$$b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2} \tag{1.21}$$

$$R^2 = \frac{r_{yx_1}^2 + r_{yx_2}^2 - 2r_{x_1x_2} . r_{yx_1} r_{yx_2}}{1 - r_{x_1x_2}^2} \tag{1.22}$$

kde

b_1 = odhadovaný regresní koeficient b_1 modelu

b_2 = odhadovaný regresní koeficient b_2 modelu

b_0 = odhadovaný konstantní člen b_0 modelu

$r_{x_1x_2}, r_{yx_1}, r_{yx_2}$= párové korelační koeficienty

$s_{x_1}, s_{x_1}, s_{x_1}$ = výběrové směrodatné odchylky

Výpočet ručně pomocí vzorců je časově velmi náročný. V případě tří a více rozměrných regresních modelů se již bez použití software při výpočtu regresních koeficientů a dalších charakteristik modelu neobejdeme.

Pro praktické vysvětlení podstaty vícerozměrné regrese rozšíříme zadání dříve demonstrovaného příkladu závislosti tržeb na investicích do reklamy o další dvě vysvětlující proměnné sledované marketingovým oddělením společnosti.

Příklad 1.1 - pokračování: V předchozí části příkladu jsme zjistili, že tržby společnosti významně závisí na investicích do reklamy (x_1). Kromě této proměnné jsou ve společnosti také sledovány informace o průměrné výši slevy v % (x_2) a počtu rozeslaných direct mailů (x_3). Chceme nyní zjistit, zda tržby nějak závisí i na těchto dvou proměnných a případně sestavit model pro predikci tržeb na základě všech těchto proměnných. Máme opět k dispozici data z posledních 12 reklamních akcí.

y tržby	x_1 reklama	x_2 průměrná výše	x_3 počet direct
24,9	2	5,1	1100
27,9	10	3,8	1320
36,5	15	5,6	987
41,1	19	2,3	1570
74,5	23	6,1	1921
81,2	27	4,2	1260
120,6	30	8,2	1350
152,7	40	7,2	1430
161,1	42	10,5	1230
168,2	50	8,1	1540
157,5	54	10,1	1650
149,2	59	6,9	1360

Tabulka 1.5 – data závislosti tržeb (y) na investicích do reklamy (x_1), průměrné výši slevy v % (x_2) a počtu rozeslaných direct mailů (x_3)

Odstranění multikolinearity

V případě vícerozměrných regresních modelů je nutné před začátkem tvorby modelu odstranit případný výskyt vysoké multikolinearity. Multikolinearita je chápána jako nežádoucí závislost (korelace) mezi vysvětlujícími proměnnými na pravé straně regresního modelu. Za vysokou multikolinearitu je považována korelace mezi vysvětlujícími proměnnými přesahující v absolutní hodnotě 0,8. Problematika multikolinearity bude podrobně probrána v Kapitole 12 této práce.

Případné nesplnění předpokladu multikolinearity v modelu je možno posoudit na základě matice párových korelačních koeficientů.

Proměnná	Korelace (Data reklama a tržby - vícerozměrná regrese) Označ. korelace jsou významné na hlad. p < ,05000 N=12 (Celé případy vynechány u ChD) y - tržby (mil. Kč)	x1 - investice do reklamy (mil. Kč)	x2 - průměrná výše slevy (%)	x3 - počet direct mailů (ks)
y - tržby (mil. Kč)	1,0000	,9365	,8284	,2863
	p= ---	p=,000	p=,001	p=,367
x1 - investice do reklamy (mil. Kč)	,9365	1,0000	,6975	,3377
	p=,000	p= ---	p=,012	p=,283
x2 - průměrná výše slevy (%)	,8284	,6975	1,0000	,1115
	p=,001	p=,012	p= ---	p=,730
x3 - počet direct mailů (ks)	,2863	,3377	,1115	1,0000
	p=,367	p=,283	p=,730	p= ---

Obr. 1.15 – matice párových korelačních koeficientů mezi vysvětlujícími proměnnými x_1, x_2 a x_3 pro posouzení výskytu vysoké multikolinearity v modelu (Výstup ze sw Statistica 10)

Z matice párových korelačních koeficientů vidíme, že žádný z korelačních koeficientů mezi vysvětlujícími proměnnými x_1, x_2 a x_3 nepřekročil v absolutní hodnotě hodnotu 0,8. Mezi vysvětlujícími proměnnými není tedy v modelu vysoká multikolinearita a mohou být bez dalších úprav použity v modelu.

S použitím statistického software vypočítáme regresní koeficienty a t-testy pro statistickou významnost jednotlivých parametrů.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,96795640 R2= ,93693960 Upravené R2= ,91329195 F(3,8)=39,621 p<,00004 Směrod. chyba odhadu : 16,998					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(8)	p-hodn.
Abs.člen			-24,8306	32,57537	-0,762249	0,467787
x1 - investice do reklamy (mil. Kč)	0,691054	0,133084	2,1933	0,42238	5,192607	0,000830
x2 - průměrná výše slevy (%)	0,344796	0,126052	7,9929	2,92209	2,735351	0,025633
x3 - počet direct mailů (ks)	0,014510	0,095962	0,0033	0,02192	0,151207	0,883556

Obr. 1.16 – koeficient determinace, regresní koeficienty a t-testy pro statistickou významnost všech proměnných x_1, x_2 a x_3 (Výstup ze sw Statistica 10)

Koeficient determinace výsledného vícenásobného lineárního modelu s třemi vysvětlujícími proměnnými x_1, x_2 a x_3 je R^2 = 93,69 %, upravený koeficient determinace R^2 = 91,33 %.

Poznámka: V případě vícerozměrného regresního modelu by měl být důležitějším ukazatelem kvality modelu právě upravený koeficient determinace, který penalizuje nadměrný počet proměnných v modelu. Rozdíl mezi koeficientem determinace a upraveným koeficientem determinace bude podrobně vysvětlen v Kapitole 8 této práce.

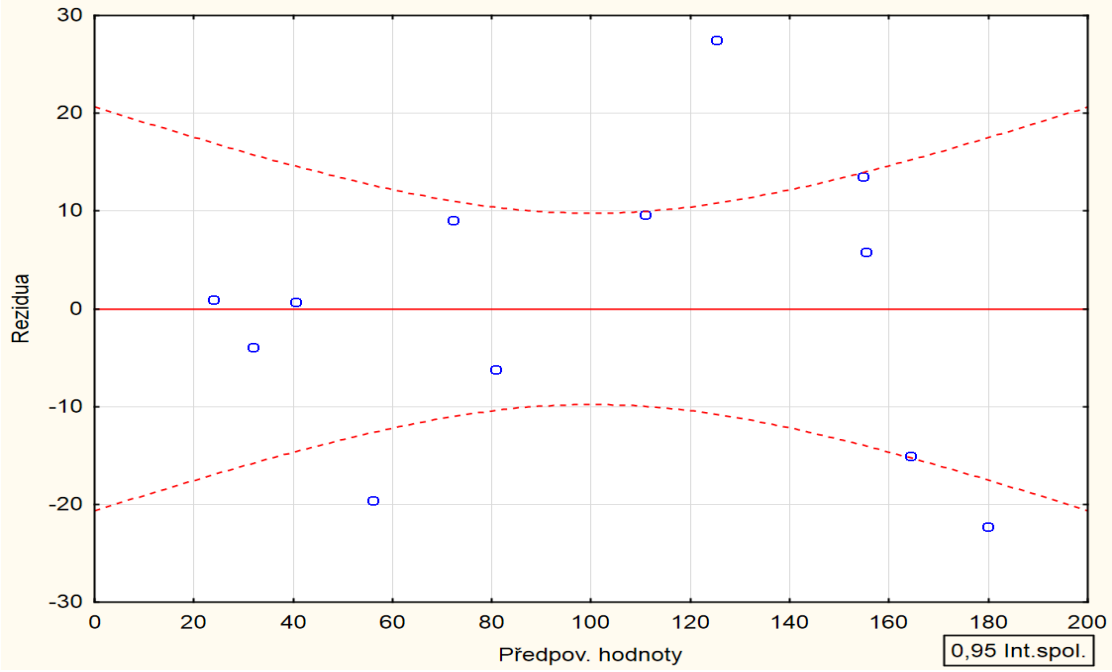
Z výsledků t-testů vidíme, že statisticky významné jsou pouze proměnné investice do reklamy (x_1) a průměrná výše slevy v % (x_2). Význam proměnné x_3 - počet rozeslaných direct mailů vyšel statisticky neprůkazně. Počet rozeslaných direct mailů tedy neovlivňuje významně tržby y a tuto proměnnou můžeme z modelu vypustit. Ponechání nevýznamné proměnné v modelu zbytečně model zesložiťuje bez dalšího přínosu pro kvalitu predikce. Nyní znovu přepočítáme model pouze pro významné proměnné x_1 a x_2.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,96786331 R2= ,93675938 Upravené R2= ,92270591 F(2,9)=66,657 p<,00000 Směrod. chyba odhadu : 16,049					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			-20,4044	13,49464	-1,51204	0,164814
x1 - investice do reklamy (mil. Kč)	0,698399	0,116983	2,2166	0,37128	5,97010	0,000210
x2 - průměrná výše slevy (%)	0,341291	0,116983	7,9117	2,71185	2,91744	0,017103

Obr. 1.17 – koeficient determinace, regresní koeficienty a t-testy pro statistickou významnost vybraných statisticky významných proměnných x_1 a x_2 (Výstup ze sw Statistica 10)

V modelu nyní zůstaly pouze statisticky významné proměnné investice do reklamy (x_1) a průměrná výše slevy v % (x_2). Koeficient determinace výsledného modelu pouze s dvěma významnými proměnnými x_1 a x_2 je R^2 = 93,68 %, tedy nepatrně menší než v případě plného modelu. Upravený koeficient determinace vyšel R^2 = 92,27 %, tedy vyšší než v případě plného modelu s třemi vysvětlujícími proměnnými x_1, x_2 a x_3.

Pro kontrolu splnění předpokladů lineárního regresního modelu si vykreslíme opět graf reziduí proti předpovězeným hodnotám.



Obr. 1.18 – graf reziduí proti předpovězeným hodnotám vícerozměrného lineárního modelu (Výstup ze sw Statistica 10)

Z tvaru grafu reziduí proti předpovězeným hodnotám v případě výsledného vícerozměrného lineárního modelu vidíme, že rezidua nenasvědčují zásadnímu porušení některého z nutných předpokladů lineárního regresního modelu.

Výsledná podoba vícerozměrného regresního modelu je tedy

$y = -20{,}4044 + 2{,}2166.x_1 + 7{,}9117.x_2$

kde

y = tržby (mil. Kč)

x_1 = investice do reklamy (mil. Kč)

x_2 = průměrná výše slevy (%)

Na základě výsledků vícerozměrného regresního modelu chceme nyní odpovědět na další otázky, které by mohly zajímat vedení společnosti.

a) **Jak je možné prakticky interpretovat vícenásobné regresní koeficienty b_0, b_1 a b_2 na uvedeném příkladu závislosti tržeb na investicích do reklamy (x_1) a průměrné výši slevy v % (x_2)?**

Regresní koeficient b_0 je možno charakterizovat jako hypotetickou hodnotu tržeb y při nulové hodnotě obou vysvětlujících proměnných x_1 a x_2. V našem ukázkovém příkladu by tedy byla hodnota tržeb při nulové investici do reklamy x_1 a nulové výši slevy x_2 rovna -20,4044 mil. Kč. V tomto místě je však nutno opět upozornit na problém extrapolace, kdy dosazujeme do modelu nulové hodnoty x_1 i x_2, které jsou mimo oba intervaly, na kterých model vznikal. Výsledek je tedy spíše hypotetický.

Regresní koeficient b_1 říká, o kolik jednotek se zvýší hodnota tržeb y, pokud se proměnná x_1 (investice do reklamy) zvýší o jednotku, přičemž druhá proměnná x_2 (průměrná výše slevy v %) zůstane neměnná. V našem případě by tedy každý 1 mil. Kč investovaný do reklamy při neměnné výši slevy v % měl zvýšit tržby o 2,2166 mil. Kč. Tedy se jedná o návratnou investici.

Regresní koeficient b_2 říká, o kolik jednotek se zvýší hodnota tržeb y, pokud se proměnná x_2 (průměrná výše slevy v %) zvýší o jednotku, přičemž druhá proměnná x_1 (investice do reklamy) zůstane neměnná. V našem případě by tedy každý 1 procentní bod slevy při neměnné výši investic měl zvýšit tržby o 7,9117 mil. Kč.

b) Na příští období plánujeme investiční akci do reklamy ve výši 60 mil. Kč a nastavení výše slevy na 10 %. Jakou hodnotu tržeb můžeme na základě předchozích dat očekávat při této investiční akci?

Z předchozích dat společnosti již máme odvozen vícerozměrný lineární regresní model závislosti tržeb na investicích do reklamy (x_1) a průměrné výši slevy v % (x_2). Pokud nyní dosadíme hodnotu x_1 = 60 mil. Kč a hodnotu x_2 = 10 % do získaného regresního modelu, vypočteme příslušnou očekávanou hodnotu tržeb.

y = -20,4044 + 2,2166 . 60 + 7,9117 . 10 = 191,4086 mil. Kč

Při investicích do reklamy ve výši 60 mil. Kč a nastavené výši slevy 10 % můžeme v příštím období očekávat tržby ve výši 191,4086 mil. Kč.

c) V jakém intervalu může management společnosti očekávat odhadované tržby pro investici do reklamy ve výši 60 mil. Kč a výši slevy 10 %.

Stejně jako u jednorozměrného regresního modelu se pro intervalový odhad predikované hodnoty vysvětlované proměnné používá 95 % konfidenční nebo 95 % predikční interval. Pokud bude management společnosti zajímat průměrná hodnota předpovídaných tržeb, použije pro odhad intervalu tržeb konfidenční interval. Pokud bude management společnosti zajímat jedna konkrétní hodnota předpovídané tržby pro konkrétní investiční reklamní akci a konkrétní nastavenou výši slevy, použije pro odhad intervalu tržeb predikční interval. Při stejné hodnotě spolehlivosti je predikční interval opět vždy širší než konfidenční interval.

Proměnná	Předpovězené hodnoty proměnné: y - tržby (mil. Kč)		
	b-váha	Hodnota	b-váha * Hodnot
x1 - investice do reklamy (mil. Kč)	2,216578	60,00000	132,9947
x2 - průměrná výše slevy (%)	7,911679	10,00000	79,1168
Abs. člen			-20,4044
Předpověď			191,7071
-95,0%LS			170,8407
+95,0%LS			212,5735

Obr. 1.19 – 95 % konfidenční interval pro průměrnou hodnotu předpovídaných tržeb (Výstup ze sw Statistica 10)

Management společnosti tedy může očekávat, že průměrná hodnota předpovídaných tržeb by se měla s 95 % spolehlivostí pohybovat v intervalu (170,8407; 212,5735).

Proměnná	Předpovězené hodnoty proměnné: y - tržby (mil. Kč)		
	b-váha	Hodnota	b-váha * Hodnot
x1 - investice do reklamy (mil. Kč)	2,216578	60,00000	132,9947
x2 - průměrná výše slevy (%)	7,911679	10,00000	79,1168
Abs. člen			-20,4044
Předpověď			191,7071
-95,0%PL			149,8329
+95,0%PL			233,5813

Obr. 1.20 – 95 % predikční interval pro jednu konkrétní hodnotu předpovídané tržby (Výstup ze sw Statistica 10)

Management společnosti tedy může očekávat, že jedna konkrétní hodnota předpovídané tržby by se měla s 95 % spolehlivostí pohybovat v intervalu (149,8329; 233,5813).

d) O kolik se zvýší tržby, pokud zvýšíme investice do reklamy o 2 mil. Kč a výše slevy v % zůstane konstantní?

Koeficient b_1 říká, o kolik jednotek se zvýší hodnota y (tržby) pokud se x_1 (investice do reklamy) zvýší o jednotku a výše slevy v % x_2 zůstane konstantní. Každá další jednotka (1 mil. Kč) investovaná do reklamy by tedy měla zvýšit tržby o 2,2166 mil. Kč. Z uvedeného vyplývá vztah

$\Delta y = b1 \,.\, \Delta x$

Po dosazení

$\Delta y = 2{,}2166 . 2 = 4{,}4332$ mil. Kč

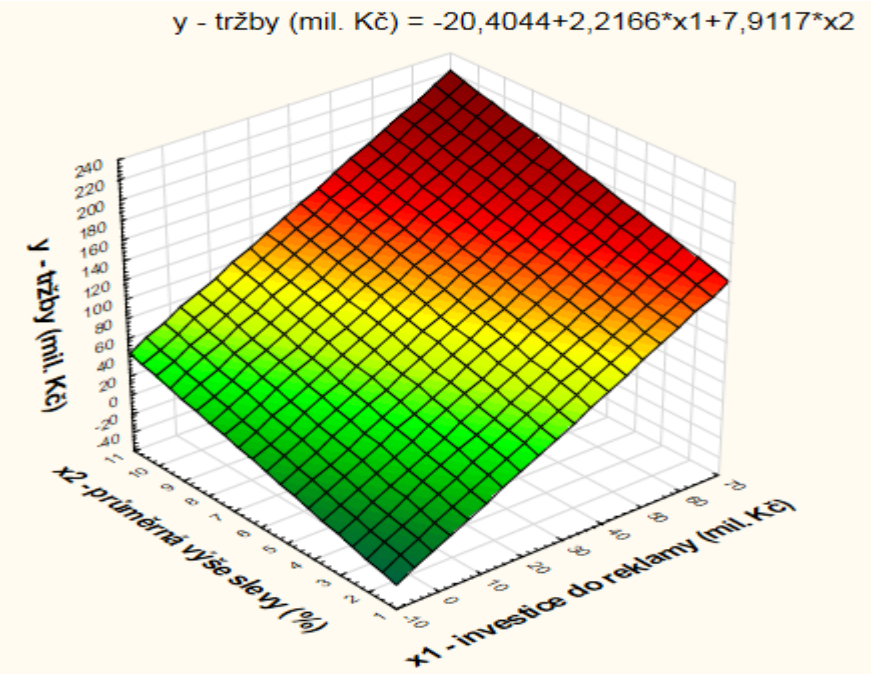
Pokud tedy zvýšíme investice do reklamy o 2 mil. Kč, přičemž výše slevy v % x_2 zůstane konstantní, tržby se zvýší o 4,4332 mil. Kč.

1.4 Vícenásobná regrese s použitím nelineárních funkcí

Jak již bylo uvedeno výše, v případě, že je model lineární vícenásobné regrese jen o málo horší než jiné nelineární modely a jeho použití odpovídá navíc reálné představě, bývá pro praktické použití upřednostňován. Důvodem je smysluplná interpretace regresních koeficientů, což u nelineárních modelů není možné.

V předchozím lineárním modelu závislosti tržeb y na investicích do reklamy x_1 a výši slevy v % x_2 vyšla hodnota koeficientu determinace R^2 = 93,67 % a hodnota upraveného koeficientu determinace R^2 = 92,27 %. Jak si ukážeme dále v tomto příkladu, použitím nelineárních funkcí se nám podaří zvýšit hodnotu koeficientu determinace R^2 na 99,40 % a hodnotu upraveného koeficientu determinace R^2 na 98,90 %. Toto vylepšení modelu určitě není zanedbatelné.

Dalším racionálním argumentem pro nahrazení lineárního vícerozměrného modelu modelem nelineárním je tvar funkční závislosti neodpovídající realitě. V 3D povrchovém grafu pro lineární model v Obr. 1.21 je možno demonstrovat neklesající hodnotu tržeb při vysokých hodnotách investic do reklamy a průměrné výše slevy.



Obr. 1.21 – 3D povrchový graf vícerozměrného lineárního modelu (Výstup ze sw Statistica 10)

Tento tvar funkční závislosti však neodpovídá častému progresivně-degresivnímu tvaru produkční funkce známému z ekonomie. Pro extrémně vysoké hodnoty vysvětlujících proměnných x_1 a x_2 se stávají hodnoty tržeb y často klesajícími. Pro lepší pochopení si můžeme představit situaci, kdy tržby y vyjadřují tržby za prodej určitého software dosti konzervativnímu a opatrnému bankovnímu sektoru. V případě extrémně masivní kampaně prostřednictvím investic do reklamy x_1 a extrémní výši slevy například 50 % a výše by se stával produkt v tomto segmentu zákazníků podezřelým a tržby by od určité hranice začaly klesat.

V další části této kapitoly se tedy pokusíme vícerozměrný lineární regresní model ještě vylepšit použitím nelineárních funkcí.

Při dřívější jednorozměrné analýze závislosti tržeb y samostatně na investicích do reklamy (x_1) a samostatně na výši slevy v % (x_2) bylo zjištěno, že obě proměnné vysvětlují tržby y dobře v kubickém (v našem případě progresivně-degresivním) tvaru.

Pokud bychom postavili plný regresní model včetně všech lineárních, kvadratických i kubických členů (Obrázek 1.22), vyšla by hodnota koeficientu determinace R^2 = 99,42 % a hodnota upraveného koeficientu determinace R^2 = 98,73 %. Oproti lineárnímu modelu se tedy hodnota koeficientu determinace zvýšila téměř o 6%.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,99711618 R2= ,99424067 Upravené R2= ,98732947 F(6,5)=143,86 p<,00002 Směrod. chyba odhadu : 6,4978					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(5)	p-hodn.
Abs.člen			26,90185	36,86867	0,72967	0,498328
x1 - investice do reklamy (mil. Kč)	-0,79152	0,369387	-2,51213	1,17236	-2,14280	0,085016
x2 - průměrná výše slevy (%)	-0,37268	0,832508	-8,63927	19,29889	-0,44766	0,673130
x1**2 - investice do reklamy kvadraticky	5,25560	1,015230	0,25698	0,04964	5,17676	0,003535
x1**3 - investice do reklamy kubicky	-3,76167	0,660986	-0,00312	0,00055	-5,69100	0,002336
x2**2 - průměrná výše slevy kvadraticky	1,52671	1,915970	2,64239	3,31611	0,79684	0,461690
x2**3 - průměrná výše slevy kubicky	-1,04989	1,098124	-0,16182	0,16926	-0,95608	0,382941

Obr. 1.22 – koeficient determinace, regresní koeficienty a t-testy pro statistickou významnost proměnných plného regresního modelu včetně všech lineárních, kvadratických i kubických členů (Výstup ze sw Statistica 10)

Jak vidíme z výše uvedených výstupů pro plný regresní model včetně všech lineárních, kvadratických i kubických členů, některé koeficienty v modelu nejsou

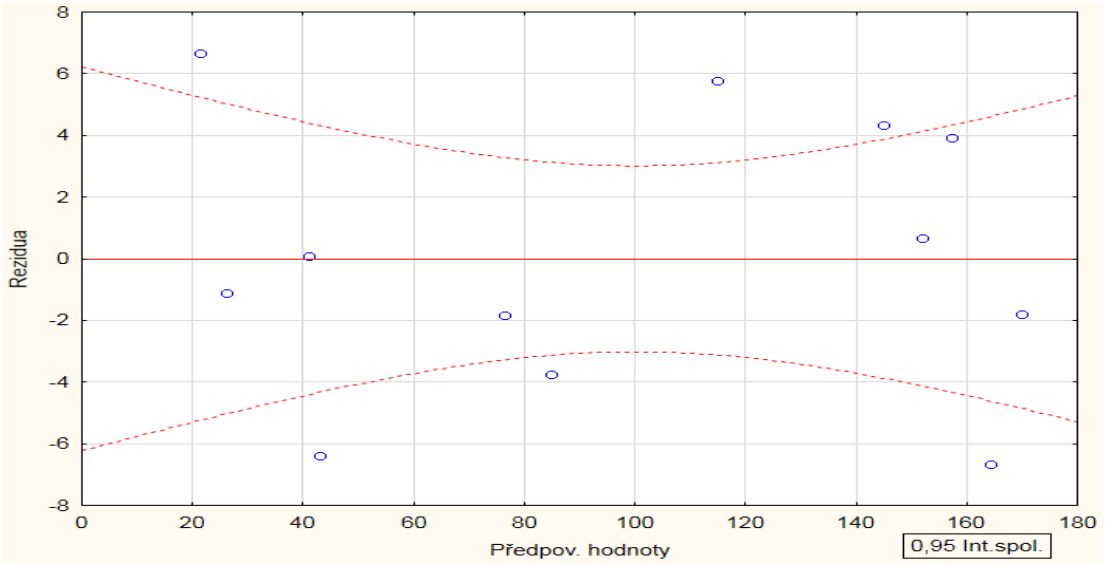
statisticky významné. Zkusíme tedy pomocí zpětné krokové regrese z modelu vyjmout proměnnou x_2 v lineárním tvaru, která má nejvyšší p hodnotu a je tedy v modelu nejméně významná.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,99700042 R2= ,99400984 Upravené R2= ,98901804 F(5,6)=199,13 p<,00000 Směrod. chyba odhadu : 6,0494 b*	Sm.chyba z b*	b	Sm.chyba z b	t(6)	p-hodn.
Abs.člen			11,10709	9,957264	1,11548	0,307320
x1 - investice do reklamy (mil. Kč)	-0,77672	0,342513	-2,46516	1,087068	-2,26771	0,063868
x1**2 - investice do reklamy kvadraticky	5,27875	0,943937	0,25812	0,046156	5,59228	0,001391
x1**3 - investice do reklamy kubicky	-3,78844	0,612843	-0,00315	0,000509	-6,18175	0,000824
x2**2 - průměrná výše slevy kvadraticky	0,67703	0,243295	1,17179	0,421088	2,78276	0,031881
x2**3 - průměrná výše slevy kubicky	-0,57040	0,225370	-0,08792	0,034737	-2,53096	0,044622

Obr. 1.23 – koeficient determinace, regresní koeficienty a t-testy pro regresní model po odebrání nejméně významné proměnné x_2 v lineárním tvaru (Výstup ze sw Statistica 10)

Vidíme, že po zjednodušení modelu vynecháním nejméně významné proměnné x_2 v lineárním tvaru se hodnota koeficientu determinace R^2 snížila z 99,42 % na hodnotu 99,40 % avšak hodnota upraveného koeficientu determinace R^2 se zvýšila z 98,73 % na 98,90 %.

Pro kontrolu splnění předpokladů lineárního regresního modelu si vykreslíme opět graf reziduí proti předpovězeným hodnotám.



Obr. 1.24 – graf reziduí proti předpovězeným hodnotám - finální model (Výstup ze sw Statistica 10)

Při snaze o další zjednodušování modelu se již snižuje výrazně hodnota koeficientu determinace R^2 i adjustovaného koeficientu determinace R^2.

1.5 Závěr

Na základě všech dosavadních analýz v této kapitole tedy můžeme vedení společnosti navrhnout jako finální model posledně uvedený vícerozměrný nelineární kubický model

$$y = 11{,}1071 - 2{,}4652.x_1 + 0{,}2581.x_1^2 - 0{,}00315.x_1^3 + 1{,}17179.x_2^2 - 0{,}0879.x_2^3$$

Tento model má ze všech výše uvedených modelů v této kapitole nejlepší vlastnosti a management společnosti by měl tedy při jeho správném použití získat nejrelevantnější výsledky.

2 Praktické použití vytvořeného regresního modelu pro optimalizaci investic ve firmě

Kromě předpovědi neznámé hodnoty tržeb pro danou úroveň investic a kvantifikace vlivu jednotlivých vysvětlujících proměnných x_i na vysvětlovanou proměnnou y je možno regresní model využít také pro optimalizaci investic ve firmě, která může velmi zvýšit úspěšnost dané společnosti. Tato aplikace regresních modelů není ve většině publikací, zabývajících se problematikou regresního modelování zmíněna.

Dle zkušeností ze školení vedoucích pracovníků a manažerů firem tomu také odpovídají téměř nulové znalosti z této oblasti následné optimalizace, se kterou se na vysoké škole nebylo možno seznámit.

Při výpočtech budeme používat jednorozměrný kubický regresní model závislosti tržeb y na investicích x, který jsme odvodili v předchozí kapitole. Koeficient determinace modelu je $R^2 = 98,57$ %. Je tedy možno říct, že investice do reklamy x vysvětlují tržby y velmi dobře.

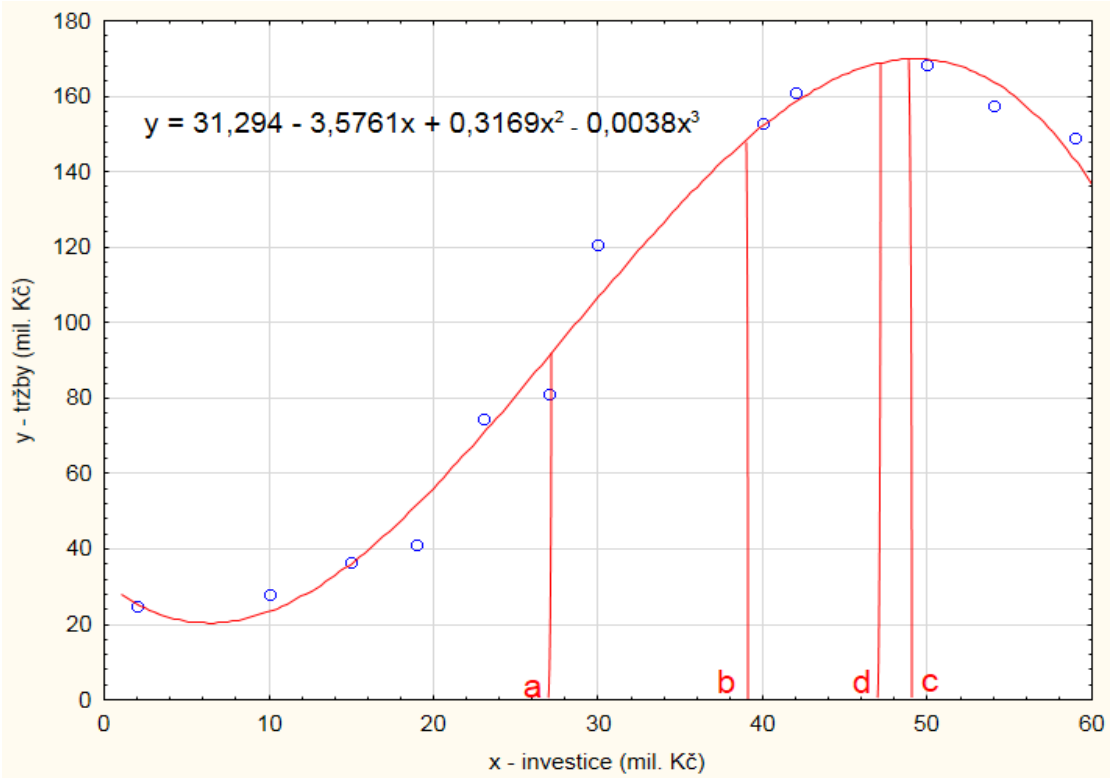
Vypočtená regresní rovnice modelu má tvar

$$y = 31,294 - 3,5761x + 0,3169x^2 - 0,0038x^3$$

Pro optimalizaci investic ve firmě je velmi účelné umět stanovit následující body znázorněné v grafu progresivně-degresivní funkce tržeb: (10)

- úroveň nákladů, při které bude dosaženo maximálních celkových tržeb (maximalizace hrubých tržeb) – bod c – bod maxima grafu
- úroveň nákladů, při které bude růst funkce tržeb nejrychlejší (maximalizace rychlosti růstu tržeb) – bod a – inflexní bod grafu
- úroveň nákladů, při které bude dosaženo maximálních průměrných tržeb v přepočtu na jednotku nákladů (maximalizace průměrných tržeb) – bod b – bod tečny vedené z počátku

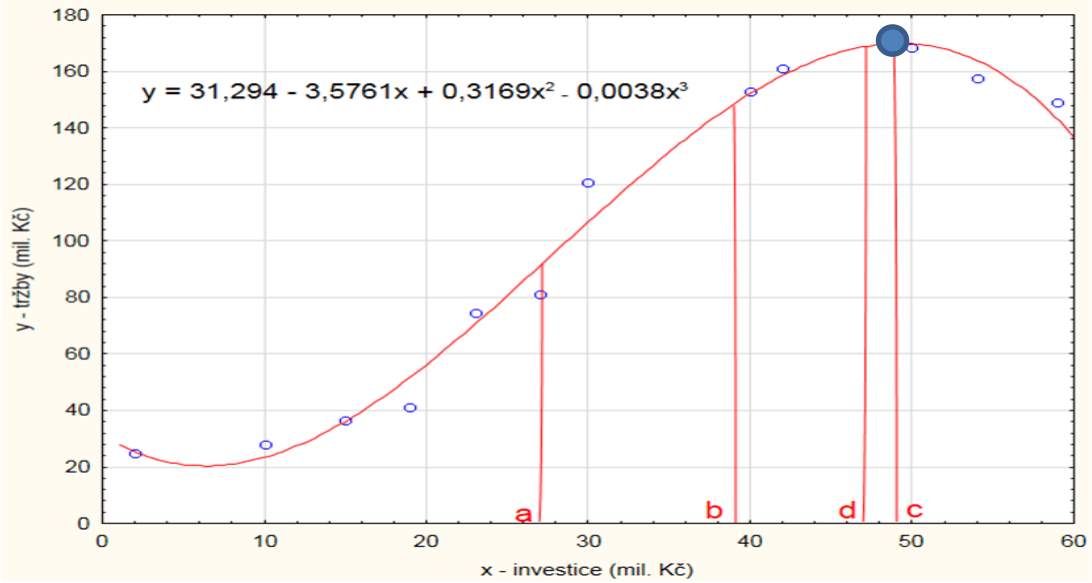
- úroveň nákladů, při které bude dosaženo maximálního zisku (maximalizace zisku) – bod d



Obr. 2.1 – grafické znázornění významných bodů grafu pro závislost tržeb na investicích při proložení kubickým regresním modelem (upravený výstup ze sw Statistica 10)

2.1 Maximalizace hrubých tržeb

Podstatou problému je nalezení úrovně nákladů, při které bude dosaženo maximálních celkových tržeb – bod c – bod maxima grafu. Za tímto bodem již budou při dalším zvyšování nákladů tržby společnosti klesat.



Obr. 2.2 – grafické znázornění úrovně nákladů, při které bude dosaženo maximálních celkových tržeb (maximalizace hrubých tržeb) – bod c – bod maxima grafu (upravený výstup ze sw Statistica 10)

Úroveň nákladů, při které bude dosaženo maximálních tržeb, vypočteme podle vzorce

$$y' = 0 \qquad (2.1)$$

Po zderivování kubické funkce závislosti tržeb na investicích získáváme kvadratickou rovnici

$$y' = -3,5761 + 0,6338x - 0,0114x^2 = 0$$

$$-0,0114x^2 + 0,6338x - 3,5761 = 0$$

Dosazením koeficientů a = - 0,0114, b = + 0,6338 a c = - 3,5761 do vzorce pro výpočet kořenů kvadratické rovnice získáváme 2 kořeny:

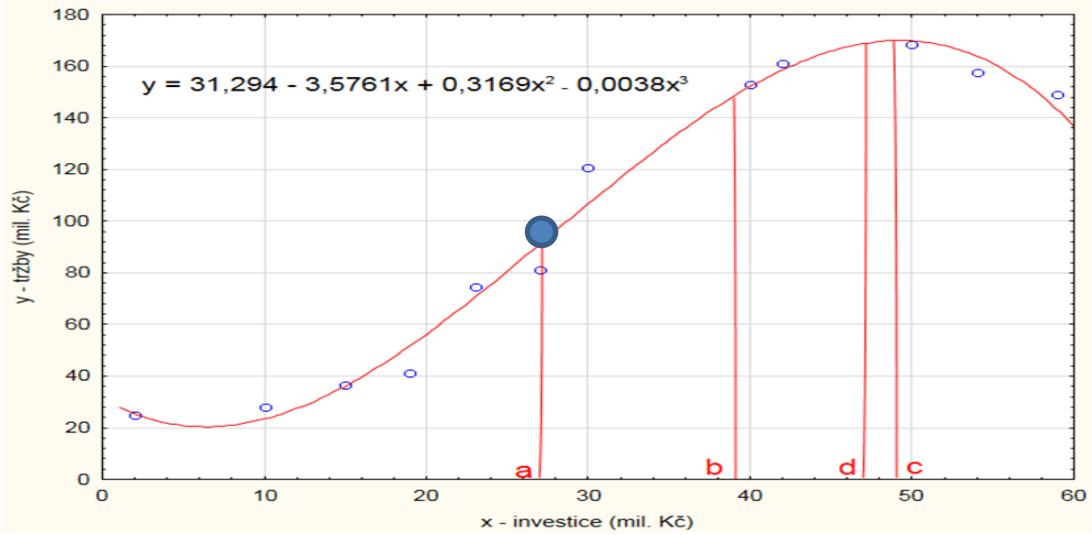
$x_1 = 49,2237$ a $x_2 = 6,3728$

Z grafu uvedené progresivně-degresivní funkce tržeb plyne, že v bodě $x_2 = 6,3728$ je lokální minimum tržeb, zatímco v bodě $x_1 = 49,2237$ je lokální maximum tržeb.

Při úrovni nákladů 49,2237 mil. Kč bude dosaženo maximálních tržeb společnosti.

2.2 Maximalizace rychlosti růstu tržeb

Úkolem je nyní nalezení úrovně nákladů, při které bude růst funkce tržeb nejrychlejší – bod a – inflexní bod grafu. Představme si situaci krizového manažera, který má za úkol přijít do firmy a provést investice, které během velmi krátkého časového úseku zásadně zvýší tržby společnosti. V tom případě je nejvhodnějším bodem právě inflexní bod a, ve kterém je tempo přírůstků tržeb nejrychlejší. Pokud v tomto bodě zvýšíme náklady na ose x o jednotku, přírůstek tržeb na ose y bude maximální.



Obr. 2.3 – grafické znázornění úrovně nákladů, při které bude růst funkce tržeb nejrychlejší – bod a – inflexní bod grafu (upravený výstup ze sw Statistica 10)

Úroveň nákladů, při které bude dosaženo maximální rychlosti růstu tržeb, vypočteme podle vzorce

$$y'' = 0 \qquad (2.2)$$

Po dvojnásobném zderivování kubické funkce závislosti tržeb na investicích získáváme lineární rovnici

$$y'' = 0{,}6338 - 0{,}0228.\ x = 0$$

Výpočtem vzniklé lineární rovnice získáváme jeden kořen:

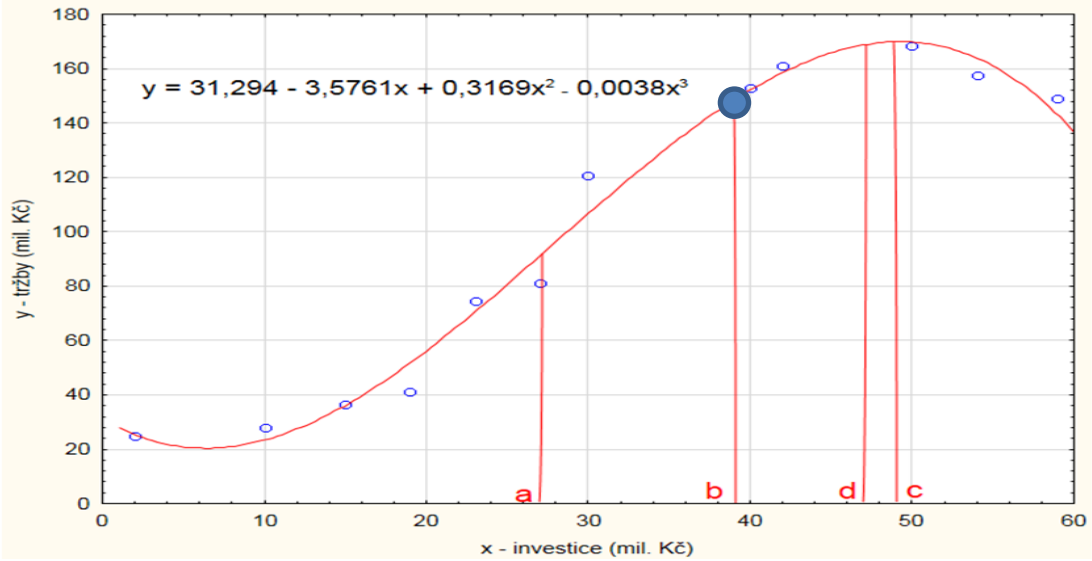
$$0{,}6338 = 0{,}0228.\ x$$

x = 27,7982

Úroveň nákladů, při které bude dosaženo maximální rychlosti růstu tržeb, je tedy 27,7982 mil. Kč.

2.3 Maximalizace průměrných tržeb

Dalším praktickým úkolem může v realitě být nalezení úrovně nákladů, při které bude dosaženo maximálních průměrných tržeb z jednotky nákladů – bod b – bod tečny vedené z počátku.



Obr. 2.4 – grafické znázornění úrovně nákladů, při které bude dosaženo maximálních průměrných tržeb z jednotky nákladů – bod b – bod tečny vedené z počátku (upravený výstup ze sw Statistica 10)

Úroveň nákladů, při které bude dosaženo maximálních průměrných tržeb, vypočteme podle vzorce

$$y' = \frac{y}{x} \tag{2.3}$$

Levou stranu rovnice získáme zderivováním původní kubické funkce y = 31,294 - 3,5761x + 0,3169x^2 - 0,0038x^3, pravou stranu potom jejím vydělením proměnnou x. Získáváme kvadratickou rovnici

-3,5761 + 0,6338x - 0,0114x^2 = 31,294/x - 3,5761 + 0,3169x - 0,0038x^2 / . x

$-3,5761.x + 0,6338x^2 - 0,0114x^3 = 31,294 - 3,5761.x + 0,3169x^2 - 0,0038x^3$

$- 0,0076.x^3 + 0,3169.\ x^2 - 31,294 = 0$

Výpočtem kubické rovnice získáváme 3 kořeny:

$x_1 = 38,9886$

$x_2 = 11,7199$

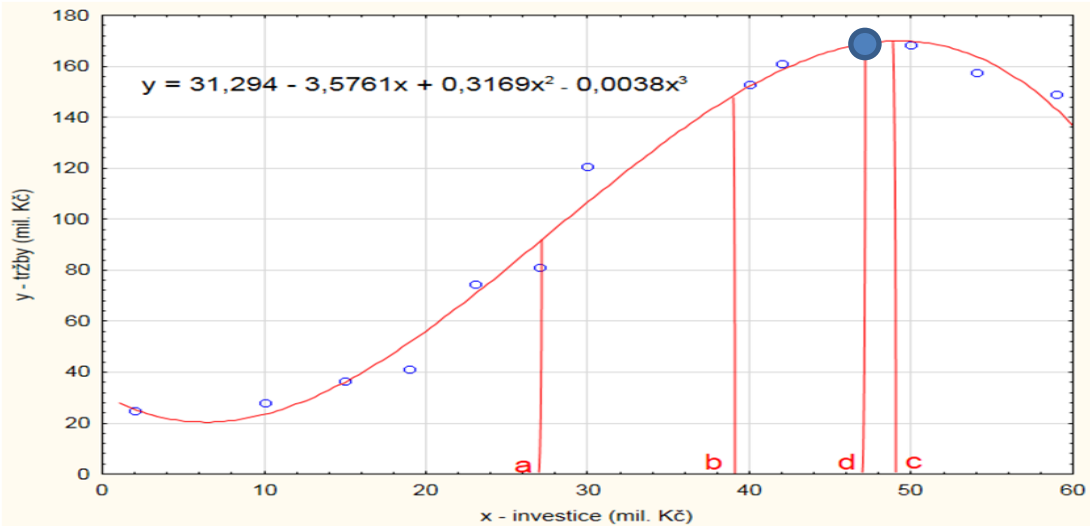
$x_3 = -9,0112$

Z grafu uvedené progresivně-degresivní funkce tržeb plyne, že hledaný bod tečny je v bodě x = 38,9886 mil. Kč.

Úroveň nákladů, při které bude společnost schopna dosáhnout maximálních průměrných tržeb je tedy 38,9886 mil. Kč.

2.4 Maximalizace zisku

Z hlediska firemních optimalizací je nejdůležitějším problémem řešení tzv. kritéria optimality, tedy nalezení úrovně nákladů, při které bude dosaženo maximálního zisku – bod d



Obr. 2.5 – grafické znázornění úrovně nákladů, při které bude dosaženo maximálního zisku – bod d (upravený výstup ze sw Statistica 10)

Úroveň nákladů, při které bude dosaženo maximálního zisku, vypočteme podle vzorce

$$y' = \frac{Cy}{Cx} \tag{2.4}$$

kde

C_y = cena za jednotku y

C_x = cena za jednotku x

V našem případě platí $C_y = C_x = 1$

Levou stranu rovnice získáme zderivováním původní kubické funkce $y = 31,294 - 3,5761x + 0,3169x^2 - 0,0038x^3$, pravou stranu potom jako podíl cen C_y a C_x. Získáváme kvadratickou rovnici

$-3,5761 + 0,6338x - 0,0114x^2 = 1$

$- 0,0114x^2 + 0,6338x - 4,5761 = 0$

Dosazením koeficientů a = - 0,0114, b = + 0,6338 a c = - 4,5761 do vzorce pro výpočet kořenů kvadratické rovnice získáváme 2 kořeny:

$x_1 = 47,0681$

$x_2 = 8,5283$

Úroveň nákladů, při které bude dosaženo maximálního zisku, je tedy 47,0681 mil. Kč.

3 Důsledky použití korelačního koeficientu pro nelineární závislost

3.1 Teoretická část

Mezi začínajícími uživateli regresních modelů je často bráno jako samozřejmost, že míra závislosti mezi dvěma kvantitativními veličinami je měřena korelačním koeficientem r dle (1.5) případně z něho odvozeným koeficientem determinace R^2 dle (1.8).

$$r = \frac{n.\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

$$R^2 = r^2$$

Často však bývá opomenuto, že korelační koeficient r se používá pouze jako míra **lineární** závislosti mezi dvěma kvantitativními veličinami. Pokud vyjde korelační koeficient r = 0, nemusí to ještě zdaleka znamenat nezávislost dvou měřených kvantitativních veličin, ale může se jednat o některou z nelineárních závislostí (např. kvadratickou, logaritmickou apod.) I při nulovém korelačním koeficientu na sobě tedy veličiny mohou záviset, pouze tento vztah nelze vyjádřit lineární funkcí.

V případě, že máme na základě grafického posouzení či věcné znalosti problému podezření na nelineární formu závislosti, měli bychom pro měření míry závislosti použít index korelace I a příslušný index determinace I^2 dle vztahů (1.14) a (1.15).

$$I = \sqrt{1 - \frac{S_R}{S_y}}$$

$$I^2 = 1 - \frac{S_R}{S_y}$$

kde

$S_y = \sum_{i=1}^{n}(y_i - y_{Prům})^2$ je celkový součet čtverců modelu

$S_R = \sum_{i=1}^{n}(y_i - y_{iTeor})^2$ je reziduální součet čtverců modelu

3.2 Praktická část

Jako ukázkový příklad špatné interpretace výsledné hodnoty korelačního koeficientu r si můžeme uvést regresní model závislosti výše tržeb na počtu zákazníků

Příklad 3.1: Mějme informace o počtech zákazníků a tržbách v jednotlivých dnech. Management společnosti chce posoudit, zda spolu tyto dvě proměnné vzájemně souvisí. Pro posouzení však použije nesprávně korelační koeficient r namísto indexu korelace I.

	1 x - počet zákazníků	2 y - tržby (tis. Kč)
1	34	13
2	35	13
3	37	14
4	38	14
5	39	14
6	42	15
7	44	15
8	48	14
9	49	13
10	51	13

Obr. 3.1 – data pro závislost výše tržeb na počtu zákazníků

Nejdříve vypočítáme běžný korelační koeficient pro závislost výše tržeb na počtu zákazníků. Pro výpočet je možno použít dříve uvedeného vzorce (1.5)

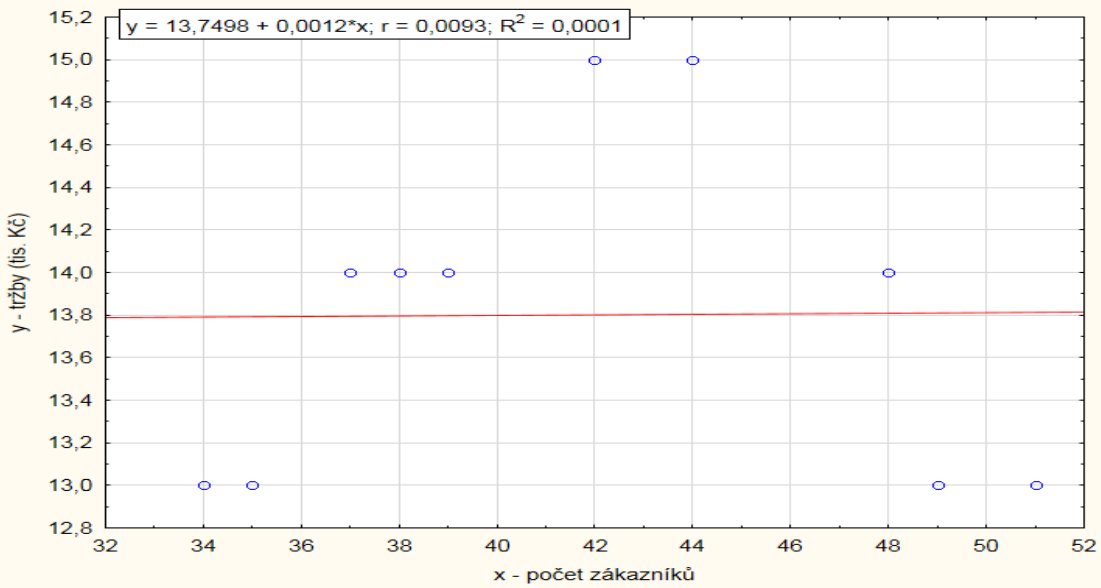
$$r = \frac{n.\sum xy - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Případně je možno korelační koeficient r vypočítat rychleji použitím statistického software

Proměnná	Korelace N=10 x - počet zákazníků	y - tržby (tis. Kč)
x - počet zákazníků	1,0000	,0093
	p= ---	p=,980
y - tržby (tis. Kč)	,0093	1,0000
	p=,980	p= ---

Obr. 3.2 – korelační matice pro závislost výše tržeb na počtu zákazníků (Výstup ze sw Statistica 10)

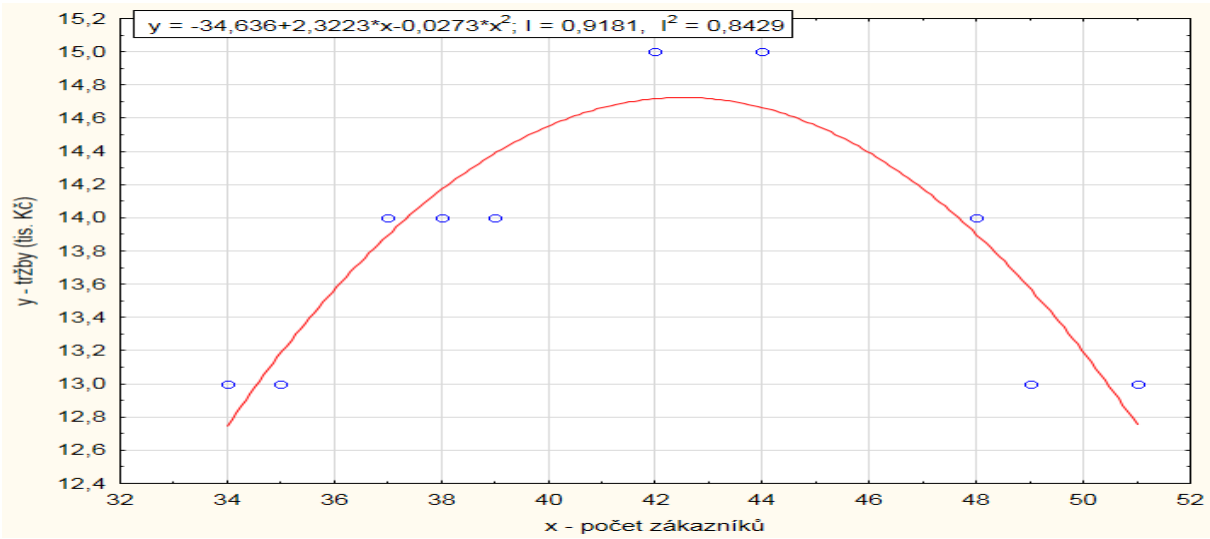
Na základě výsledné hodnoty korelačního koeficientu r = 0,0093 bychom mohli usuzovat na naprostou nezávislost mezi výší tržeb a počtem zákazníků. Jak již bylo uvedeno výše, korelační koeficient r je mírou lineární závislosti. Grafické znázornění proloženého lineárního regresního modelu by tedy vypadalo následovně.



Obr. 3.3 - grafické znázornění proloženého lineárního regresního modelu (Výstup ze sw Statistica 10)

Jak je možno vidět z grafického znázornění proloženého lineárního regresního modelu, tento typ závislosti je pro analyzovaná data závislosti tržeb na počtu zákazníků naprosto nevhodný a vhodnější by bylo použít kvadratický model závislosti. To potvrzuje i praktická znalost problematiky, kdy při velkém počtu zákazníků se dá očekávat pokles tržeb z důvodu snížení pohodlí při nákupu. Kvadratickou závislost však klasický korelační koeficient vypočítaný ze vzorce (1.5) vůbec nereflektuje.

V následujícím obrázku je pro proložení bodů použit vhodnější kvadratický trend. Místo korelačního koeficientu r je použit v tomto případě index korelace I a příslušný index determinace I^2 dle vztahů (1.14) a (1.15). Jejich výpočet je podrobně popsán v Kapitole 1.



Obr. 3.4 - grafické znázornění proloženého kvadratického regresního modelu (Výstup ze sw Statistica 10)

Výsledný index korelace I má nyní hodnotu 0,9181 (ve srovnání s korelačním koeficientem r = 0,0093). Výsledný index determinace I^2 má nyní hodnotu 84,29 % (ve srovnání s koeficientem determinace R^2 = 0,01 %).

3.3 Závěr

Vidíme, že pokud pro posouzení závislosti použijeme v našem příkladu nesprávný klasický korelační koeficient r (případně příslušný koeficient determinace R^2), dostáváme informaci o naprosté nezávislosti obou proměnných.

Pokud ovšem použijeme pro posouzení závislosti správný index korelace I (případně příslušný index determinace I^2), zjistíme, že proměnné spolu korelují velmi dobře. Hledaná závislost však byla jiná než lineární.

4 Problém nezávislosti korelačního koeficientu a směrnice přímky

4.1 Teoretická část

Jak vyplývá ze studia publikací z oboru lékařství ale i jiných oborů, významná část uživatelů statistických metod je přesvědčena, že pokud má regresní přímka velký sklon (vysokou hodnotu směrnice b_1), musí mít nutně i vysokou hodnotu korelačního koeficientu (obě proměnné spolu lépe korelují). Naopak, u téměř vodorovné přímky je očekáván korelační koeficient blízko 0 (obě proměnné spolu téměř nekorelují). Tato úvaha však v některých případech neplatí.

V Kapitole 1 byl vztahem (1.2) definován pojem směrnice regresní přímky b_1

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

a vztahem (1.5) pojem korelačního koeficientu r

$$r = \frac{n.\sum x_i - \sum x_i \sum y_i}{\sqrt{[n\sum {x_i}^2 - (\sum x_i)^2][n\sum {y_i}^2 - (\sum y_i)^2]}}$$

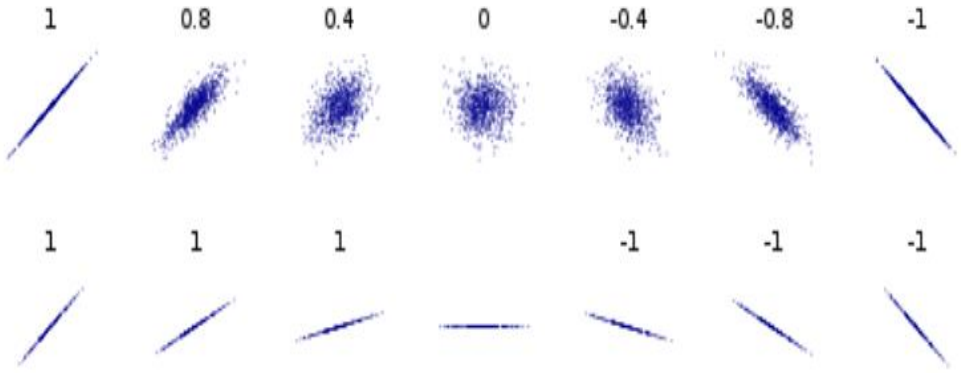
Srovnáním výše uvedených vzorců pro směrnici b_1 a korelační koeficient r vidíme, že čitatele jsou v obou případech stejné. Jmenovatele se však liší. Vzájemný vztah mezi regresním a korelačním koeficientem můžeme potom charakterizovat pomocí vztahů (4.1) a (4.2).

$$r = b_{yx}.\frac{s_x}{s_y} \tag{4.1}$$

$$b_{yx} = r.\frac{s_y}{s_x} \tag{4.2}$$

Na Obr. 4.1 jsou znázorněny hodnoty korelačních koeficientů pro různé tvary korelačních polí. Korelačním polem rozumíme grafické znázornění datových bodů popisujících danou závislost. Jak vidíme z obrázků jednotlivých korelačních polí, předpoklad o vzájemné jednoznačné závislosti sklonu (směrnice) přímky a hodnoty

korelačního koeficientu v praxi často neplatí. Vidíme, že některá korelační pole mají velmi prudký sklon (směrnici) a zároveň nízký korelační koeficient. Jiná korelační pole mají naopak malý sklon (směrnici) a zároveň vysoký korelační koeficient.



Obr. 4.1 – Hodnoty korelačních koeficientů pro různé tvary korelačních polí (11)

4.2 Praktická část

Nejlepším způsobem vyvrácení mylné domněnky posluchačů statistických kurzů je uvedení protipříkladu.

Příklad 4.1: V datové matici máme uvedeny dvě vysvětlované proměnné y_1 a y_2 v závislosti na jedné vysvětlující proměnné x (Obr. 4.2).

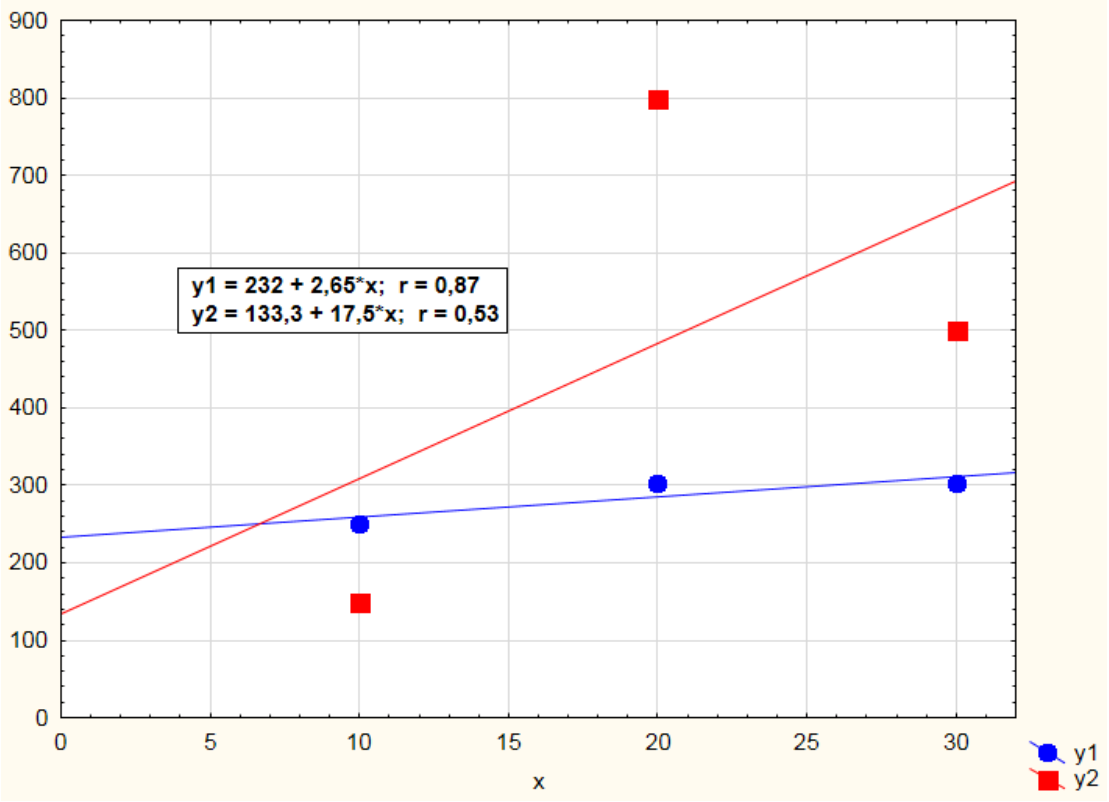
	1 x	2 y1	3 y2
1	10	250	150
2	20	302	800
3	30	303	500

Obr. 4.2 - ukázka datové matice pro demonstraci nezávislosti směrnice přímky a korelačního koeficientu

O neplatnosti vzájemné závislosti sklonu (směrnice) přímky a korelačního koeficientu se přesvědčíme porovnáním hodnot směrnic a korelačních koeficientů pro závislosti y_2 na x a y_1 na x.

V Obr. 4.3 můžeme sledovat dva protichůdné příklady. Závislost vysvětlované proměnné y_1 na vysvětlující proměnné x (znázorněno modře) je popsána lineárním

regresním modelem y_1=232+2,65.x. Směrnice přímky b_1 je poměrně malá (b_1=2,65), zatímco korelační koeficient r nabývá vysoké hodnoty 0,87. Závislost vysvětlované proměnné y_2 na vysvětlující proměnné x (znázorněno červeně) je popsána lineárním regresním modelem y_2=133,3+17,5.x. Směrnice přímky b_1 je poměrně vysoká (b_1=17,5), zatímco korelační koeficient r nabývá hodnoty pouze 0,53. Tento příklad vyvrací představu o vzájemné jednoznačné závislosti směrnice a korelačního koeficientu.



Obr. 4.3 – grafická demonstrace nezávislosti směrnice přímky a korelačního koeficientu (Výstup ze sw Statistica 10)

4.3 Závěr

Představa o vzájemné jednoznačné závislosti směrnice a korelačního koeficientu neplatí. Pokud má regresní přímka velký sklon (vysokou hodnotu směrnice b_1), nemusí mít nutně i vysokou hodnotu korelačního koeficientu (obě proměnné spolu lépe korelují). Zatímco hodnota směrnice přímky ukazuje na sklon (intenzitu) závislosti, korelační koeficient ukazuje spíše na „přiléhavost“ bodů ke grafu daného regresního modelu.

5 Problém zdánlivé korelace

5.1 Teoretická část

V základních kurzech statistiky je často probírán pouze klasický Pearsonův korelační koeficient, případně neparametrický Spearmanův korelační koeficient pořadové korelace pro nenormálně rozdělená data. Tyto základní míry měření závislosti dvou kvantitativních veličin však neumožňují při výpočtu odstranit zavádějící vliv nějaké další kvantitativní proměnné, která často velmi zásadně výsledek korelace dvou sledovaných proměnných ovlivňuje.

Důsledkem použití jednoduchého párového korelačního koeficientu potom může být velmi zavádějící interpretace povahy dané závislosti z důvodu výskytu tzv. „zdánlivé“ (nesmyslné) korelace.

Zdánlivá korelace je situace, kdy charakteristika závislosti indikuje závislost statistických znaků, které však jsou prakticky nezávislé. Dochází k ní v případě, že souvislost hodnot statistických znaků je dána jejich závislostí na nějakém třetím znaku.

Řešením problému je použití parciálních korelačních koeficientů, jejichž hlavním přínosem je schopnost odfiltrovat vliv nějaké další proměnné na obě sledované proměnné.

Parciální korelační koeficient závislosti proměnné y na proměnné x_1 při odstranění vlivu proměnné x_2 je dán vztahem

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2}.r_{x_1x_2}}{\sqrt{\left(1-r_{yx_2}^2\right)\left(1-r_{x_1x_2}^2\right)}} \tag{5.1}$$

kde r_{yx_1}, r_{yx_2} a $r_{x_1x_2}$ jsou klasické párové korelační koeficienty.

Parciální korelační koeficient je někdy také interpretován jako závislost proměnné y na proměnné x_1 při neměnné (konstantní) úrovni proměnné x_2.

5.2 Praktická část

O tom, jak velká může být míra dezinterpretace v případě nepoužití parciálních koeficientů pro odstranění vlivu další proměnné, se můžeme přesvědčit v následujících dvou vybraných příkladech.

Příklad 5.1: (Vymyšlený demonstrační příklad) V rámci výzkumu osteoporózy byly měřeny mimo jiné hodnoty hustoty kostí BMD (Bone mineral density), počet šedých vlasů pacienta a věk pacienta.

	1 počet šedých vlasů	2 BMD - Bone mineral density (g/cm3)	3 věk
1	2	1,57	14
2	158	0,48	68
3	69	0,84	41
4	5	1,12	20
5	120	0,56	58
6	75	0,78	52
7	110	0,51	71
8	198	0,42	85
9	45	0,9	38
10	56	0,79	29

Obr. 5.1 - ukázka datového souboru pro kvantifikaci závislosti proměnných BMD, počet šedých vlasů a věk pacienta

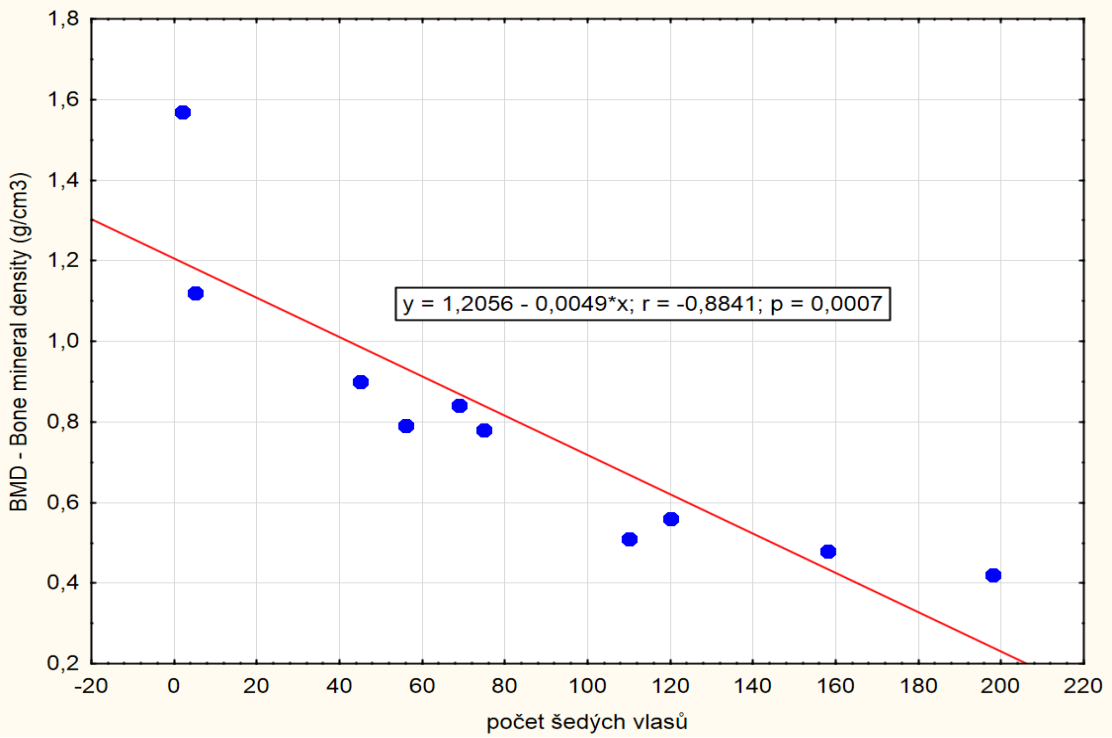
Již při zevrubné analýze datového souboru vidíme, že se zvyšujícím se počtem šedých vlasů pacientů se snižuje hustota kostní hmoty (BMD).

Z korelační matice párových Pearsonových korelačních koeficientů můžeme vyčíst hodnotu korelace r = - 0,88, tedy by se mělo jednat o silnou nepřímou korelaci mezi počtem šedých vlasů a BMD pacienta. Tedy, čím má pacient více šedých vlasů, tím má nižší hustotu kostní hmoty BMD. Dosažená hladina testu p = 0,001ukazuje, že na hladině spolehlivosti 5 % je daná závislost statisticky vysoce průkazná.

Proměnná	Korelace N=10 počet šedých vlasů	Korelace N=10 BMD - Bone mineral density (g/cm3)
počet šedých vlasů	1,0000	-,8841
	p= ---	p=,001
BMD - Bone mineral density (g/cm3)	-,8841	1,0000
	p=,001	p= ---

Obr. 5.2 - korelační matice pro párový Pearsonův korelační koeficient závislosti BMD na počtu šedých vlasů pacienta (Výstup ze sw Statistica 10)

Grafické znázornění závislosti mezi počtem šedých vlasů a BMD pacienta pomocí bodového grafu potvrzuje na první pohled domněnku o silné nepřímé závislosti.



Obr. 5.3 - grafické znázornění závislosti proměnných „počet šedých vlasů“ a „BMD“ pomocí bodového grafu (Výstup ze sw Statistica 10)

Při předpokládané znalosti problematiky vývoje hodnoty BMD u pacientů by měl být výsledek značně odporující reálné představě o možné závislosti. Je však nutno si uvědomit, že mnohdy při předběžné korelační analýze sledovaných parametrů takto dobrou znalost problematiky nemáme a naopak si povědomí o možných závislostech budujeme teprve na základě hodnot výsledných korelačních koeficientů.

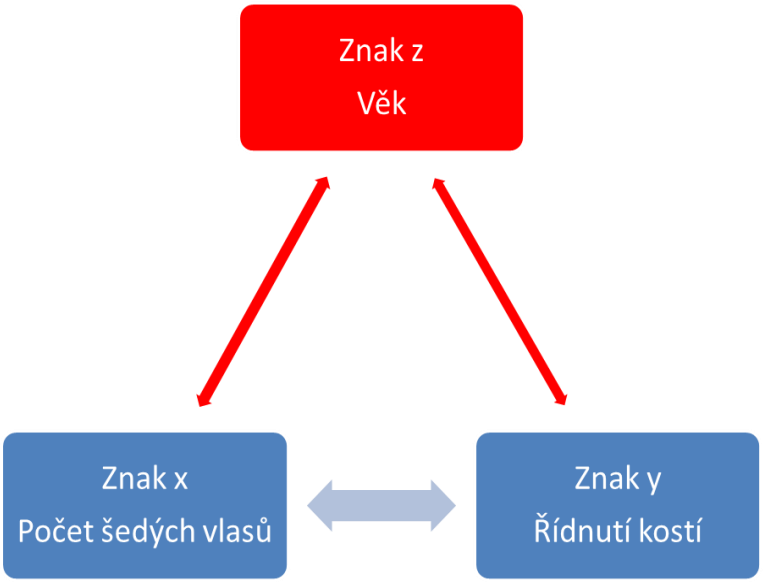
Na základě znalosti dané problematiky by mělo být očekávatelné, že obě sledované proměnné jsou zároveň ovlivňovány věkem. Pro hodnocení závislosti obou sledovaných proměnných tedy očistíme závislost od vlivu proměnné „věk“.

Proměnná	Parciální korelace (Data - BMD) s vyloučením vlivu:věk N=10	
	počet šedých vlasů	BMD - Bone mineral density (g/cm3)
počet šedých vlasů	1,0000	-,1733
	p= ---	p=,656
BMD - Bone mineral density (g/cm3)	-,1733	1,0000
	p=,656	p= ---

Obr. 5.4 - korelační matice pro parciální korelační koeficient mezi počtem šedých vlasů a BMD při odstranění vlivu proměnné věk (Výstup ze sw Statistica 10)

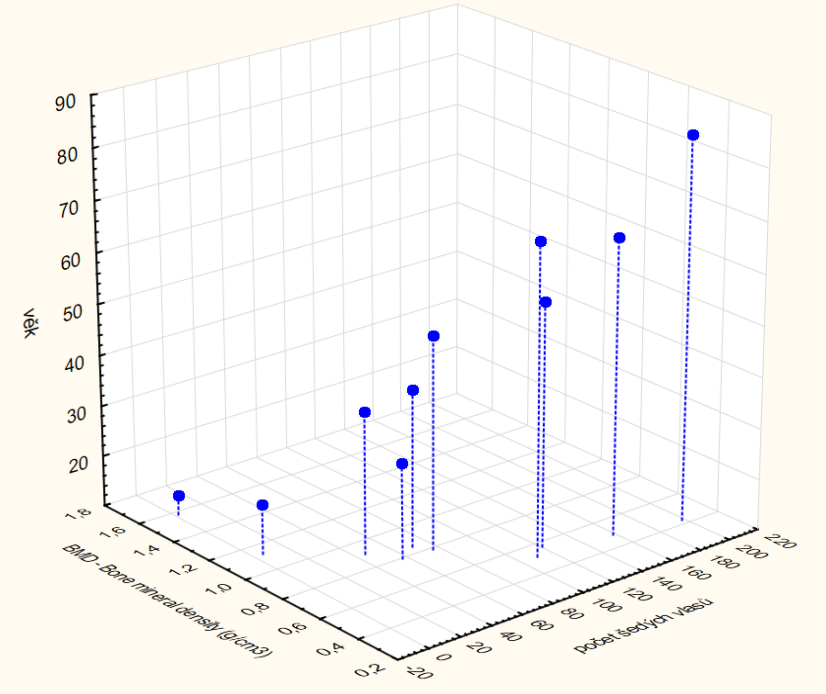
Z korelační matice parciálních korelačních koeficientů vidíme hodnotu parciální korelace mezi počtem šedých vlasů a BMD při odstranění vlivu věku r = - 0,17. Hodnota parciální korelace je na hladině spolehlivosti 5 % neprůkazná. Tedy, při odstranění vlivu třetí proměnné „věk“ již spolu obě sledované proměnné nekorelují.

Důvodem je právě v této kapitole demonstrovaná „zdánlivá korelace“, kterou je v tomto případě možno schematicky znázornit dle Obr. 5.5.



Obr. 5.5 - schematické znázornění zdánlivé korelace mezi proměnnými z Příkladu 5.1

Ze schématu „zdánlivé“ korelace vidíme, že obě sledované proměnné „počet šedých vlasů“ i „řídnutí kostí BMD“ jsou výrazně ovlivňovány třetí proměnnou „věk“. To potvrzuje i 3D bodový graf v Obr. 5.6.



Obr. 5.6 – grafické znázornění závislosti proměnných „počet šedých vlasů“, „BMD“ a „věk“ pomocí 3D XYZ bodového grafu (Výstup ze sw Statistica 10)

Použití jednoduchých párových korelací vede v uvedeném případě k nesprávnému závěru. Na základě prostého Pearsonova korelačního koeficientu se zdá, že počet šedých vlasů jasně souvisí s řídnutím kostí. Realita je ale úplně jiná. Díky parciálnímu korelačnímu koeficientu vidíme, že při odstranění vlivu věku spolu obě proměnné téměř nekorelují. Nesmyslný by byl tedy závěr o příčinném působení těchto dvou proměnných. Korelační závislost je zdůvodněna proměnnou „věk“, jež je společnou příčinnou obou proměnných.

Příklad 5.2: V rámci psychologického výzkumu mezi žáky základních škol byly provedeny znalostní testy. Zkoumány byly zejména jazykové a matematické schopnosti. Zároveň byl u žáků proveden klasický inteligenční test. Cílem bylo zjistit, zda mezi výsledky jazykových a matematických testů existuje nějaká závislost a v případě, že ano, jaká. (12)

	1 Jazyk	2 Mat	3 IQ
1	67	71	97
2	85	70	109
3	100	86	118
4	68	48	97
5	95	99	121
6	89	82	106
7	53	62	96
8	56	36	84
9	81	56	96
10	96	76	121
11	63	69	109
12	91	46	102
13	38	30	85
14	100	77	114
15	82	100	113
16	98	77	99
17	78	58	96
18	88	92	112
19	51	53	91
20	75	89	105
21	86	88	117
22	96	93	115
23	65	73	96
24	59	69	94
25	67	72	97
26	83	64	105
27	55	45	88
28	81	86	104
29	73	88	100
30	74	61	101
31	70	56	90
32	34	19	65
33	97	56	105
34	60	70	93

Obr. 5.7 - ukázka datového souboru pro kvantifikaci závislosti proměnných „Jazyk“, „Mat“ a „IQ“ u žáků základních škol

Při primární analýze datového souboru zjistíme, že žáci s lepšími výsledky z matematiky dosahují i lepších výsledků z jazyka. Z korelační matice párových Pearsonových korelačních koeficientů můžeme mezi proměnnými „Jazyk“ a „Mat“ vyčíst hodnotu korelace r = + 0,65, mělo by se proto jednat o silnou přímou korelaci mezi výsledky z jazyka a matematiky. Tedy, čím je žák nadanější na jazyky, tím je nadanější i na matematiku.

Proměnná	Korelace N=34 Jazyk	Mat	IQ
Jazyk	1,0000	,6534	,8341
	p= ---	p=,000	p=,000
Mat	,6534	1,0000	,8131
	p=,000	p= ---	p=,000
IQ	,8341	,8131	1,0000
	p=,000	p=,000	p= ---

Obr. 5.7 - korelační matice pro párový Pearsonův korelační koeficient závislosti „Mat“, „Jazyk“ a „IQ“ (Výstup ze sw Statistica 10)

Mezi všemi proměnnými v tabulce existují poměrně silné, statisticky významné korelace. Opět je zde významná třetí proměnná IQ, která ovlivňuje současně obě sledované proměnné. Žáci s vysokým IQ budou tedy podávat pravděpodobně nadprůměrné výkony i při zbylých dvou testech. To nás opět vede k úvaze, že pro seriózní posouzení vzájemného vztahu mezi výsledky matematických a jazykových testů bude nezbytné parciálně očistit danou závislost od vlivu celkové inteligence IQ.

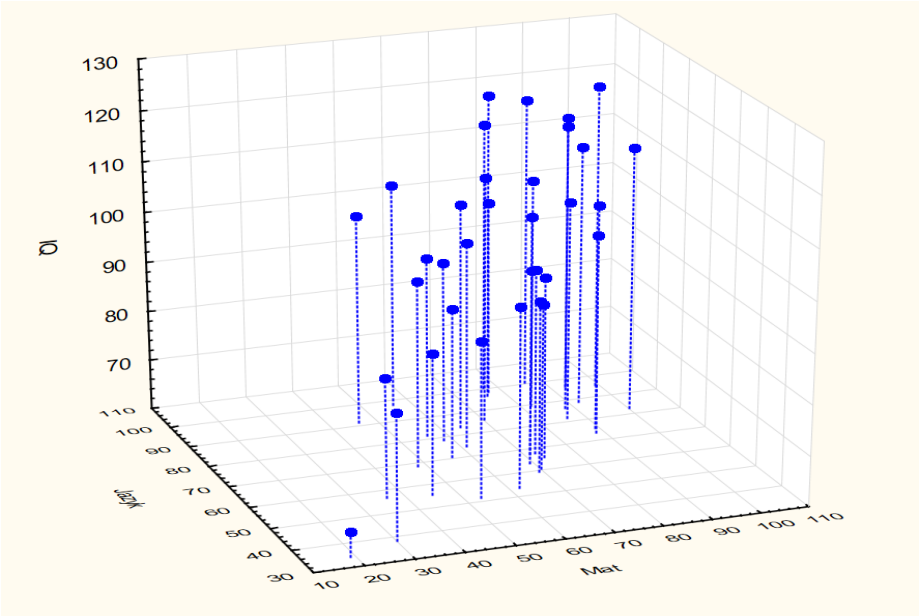
Proměnná	Parciální korelace s vyloučením vlivu:IQ N=34 Jazyk	Mat
Jazyk	1,0000	-,0772
	p= ---	p=,669
Mat	-,0772	1,0000
	p=,669	p= ---

Obr. 5.8 - korelační matice pro parciální korelační koeficient mezi výsledky matematických a jazykových testů při odstranění vlivu proměnné „IQ“ (Výstup ze sw Statistica 10)

Z korelační matice parciálních korelačních koeficientů vidíme hodnotu parciální korelace mezi proměnnými „Jazyk“ a „Mat“ při odstranění vlivu IQ r = - 0,0772. Tedy, při odstranění vlivu celkové inteligence IQ již spolu obě sledované proměnné významně nekorelují. Jinými slovy: mezi žáky, kteří jsou zhruba stejně inteligentní, není významná korelace mezi výsledky v jazykovém a matematickém testu.

Navíc záporná (i když nevýznamná) hodnota korelačního koeficientu ukazuje spíše na často tradovanou nepřímou závislost mezi nadáním na jazyky a nadáním na

matematiku. Závislost mezi proměnnými „Jazyk“, „Mat“ a „IQ“ je znázorněna v 3D bodovém grafu v Obr. 5.9.



Obr. 5.9 - grafické znázornění závislosti proměnných „Jazyk“, „Mat“ a „IQ“ pomocí 3D XYZ bodového grafu (Výstup ze sw Statistica 10)

5.3 Závěr

Použití jednoduchých párových korelací může v některých případech vést k velmi zavádějícím interpretacím povahy dané závislosti z důvodu výskytu tzv. „zdánlivé“ korelace. Na první pohled se může zdát, že dvě proměnné spolu významně korelují. Při odstranění vlivu významné proměnné, která obě sledované proměnné ovlivňuje, se však může sledovaná závislost ukázat již jako nevýznamná či dokonce obrácená než se zdálo při použití jednoduchých párových korelací.

Na základě znalosti problematiky a komunikace mezi statistikem a zadavatelem analýzy (lékařem, vědeckým pracovníkem apod.) by mělo vyplynout, od kterých proměnných je nutné výsledky korelací adjustovat. Například v lékařství jsou velmi často adjustovanými proměnnými věk pacienta, BMI či délka nemoci.

Řešením problému je použití parciálních korelačních koeficientů, jejichž hlavním přínosem je schopnost odfiltrovat vliv nějaké další proměnné na obě sledované proměnné.

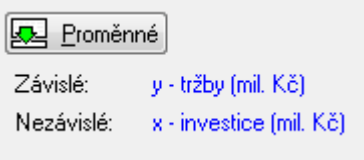
6 Původní a sdružená regresní přímka

6.1 Teoretická část

Jednou z velmi častých chyb při finálním použití modelu je odhad očekávané hodnoty x při známé hodnotě y přímo z inverzního vyjádření proměnné x z původní regresní rovnice. Jak bude dále rozvedeno, tento přístup není správný. Nevede sice k úplně nesmyslnému výsledku, ale k odhadu musí být pro přesný odhad použita na základě MNČ znovu vypočtená sdružená regresní přímka závislosti proměnné x na proměnné y, jak již bylo zmíněno v Kapitole 1.

6.2 Praktická část

Příklad 6.1: Při předchozí tvorbě regresního modelu závislosti tržeb y na investicích do reklamy x jsme získali na základě výpočtu regresních koeficientů ve statistickém software (Obr. 6.1) výslednou regresní rovnici y = 7,7285 + 2,9721.x (Kapitola 1, str.14).



Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč)
R= ,93645686 R2= ,87695146 Upravené R2= ,86464660
F(1,10)=71,269 p<,00001 Směrod. chyba odhadu : 21,238

N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			7,7285	12,49237	0,618654	0,549976
x - investice (mil. Kč)	0,936457	0,110927	2,9721	0,35206	8,442082	0,000007

Obr. 6.1 – regresní koeficienty pro původní regresní přímku závislosti tržeb y na investicích x (výstup ze sw Statistica 10)

Z důvodu udržení cash flow potřebujeme v příštím období dosáhnout tržeb 120 mil. Kč. Na základě předchozích dat potřebujeme odhadnout, jakou částku bychom měli v následujícím období investovat do reklamy, abychom dosáhli plánovaných tržeb.

Postup 1 (nesprávný): Při první úvaze by se nabízela možnost jednoduše dosadit hodnotu tržeb 120 mil. Kč do dříve vypočítané regresní rovnice **y = 7,7285 + 2,9721.x** za y a z rovnice následně vyjádřit hledanou hodnotu potřebných investic x.

Budeme tedy úmyslně demonstrovat častý, ale nesprávný postup při odhadu hodnoty x z „inverzní" regresní rovnice namísto použití nově vypočtené sdružené regresní rovnice.

Pokud vyjádříme x z rovnice y = 7,7285 + 2,9721.x, získáváme tvar modelu

$y - 7{,}7285 = 2{,}9721.x$ *//:2,9721*

$$\frac{1.y}{2{,}9721} - \frac{7{,}7285}{2{,}9721} = x$$

$$x = -2{,}6003 + 0{,}3365.y$$

Do vypočtené „nesprávné" regresní rovnice nyní dosadíme za y hodnotu plánovaných tržeb y = 120 mil. Kč.

$$x = -2{,}6003 + 0{,}3365.120 = \mathbf{37{,}7797\ mil.Kč}$$

Jak již bylo uvedeno výše, tento přístup není správný. Nevede sice k úplně nesmyslnému výsledku, ale k odhadu musí být pro přesný odhad použita znovu vypočtená sdružená regresní přímka závislosti proměnné x na proměnné y.

Postup 2 (správný): Pro výpočet nové sdružené regresní přímky závislosti proměnné x na proměnné y stačí ve výpočtu vzorce nebo při zadání proměnných ve statistickém software zaměnit hodnoty y s hodnotami x.

N=12	Výsledky regrese se závislou proměnnou : x - investice (mil. Kč) R= ,93645686 R2= ,87695146 Upravené R2= ,86464660 F(1,10)=71,269 p<,00001 Směrod. chyba odhadu : 6,6915					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			1,5239	3,981654	0,382731	0,709928
y - tržby (mil. Kč)	0,936457	0,110927	0,2951	0,034951	8,442082	0,000007

Obr. 6.2 – regresní koeficienty pro novou sdruženou regresní přímku závislosti investic x na tržbách y (Výstup ze sw Statistica 10)

Výsledná rovnice nové sdružené regresní přímky má tedy nyní tvar

$$x = 1{,}5239 + 0{,}2951.y$$

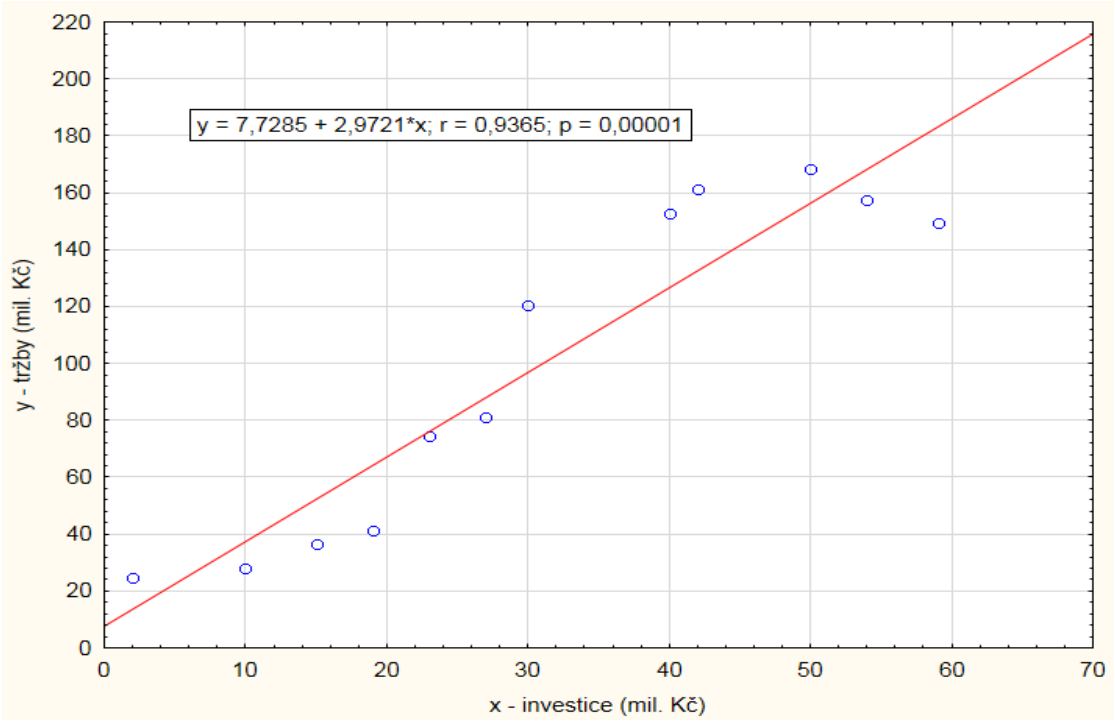
Do nové sdružené regresní rovnice nyní dosadíme za y hodnotu plánovaných tržeb y = 120 mil. Kč.

$$x = 1{,}5239 + 0{,}2951.120 = \mathbf{36{,}9359\ mil.Kč}$$

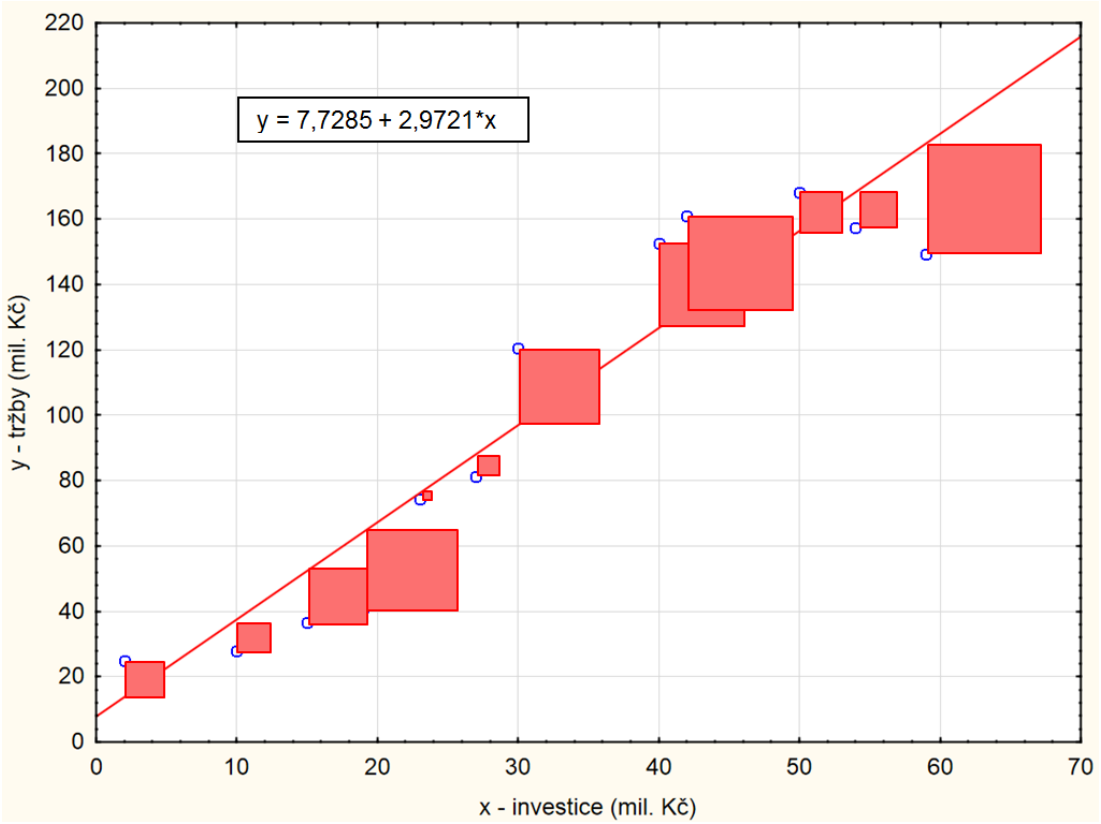
Abychom dosáhli v následujícím období plánovaných tržeb 120 mil. Kč, měli bychom investovat do reklamy 36,9359 mil. Kč.

Pokud nyní porovnáme „nesprávnou“ vypočtenou hodnotu 37,7797 mil. Kč se „správnou“ hodnotou 36,9359 mil. Kč vypočtenou při použití sdružené regresní přímky, vidíme, že v případě takto vysokých částek může být i malý rozdíl dosti zásadní.

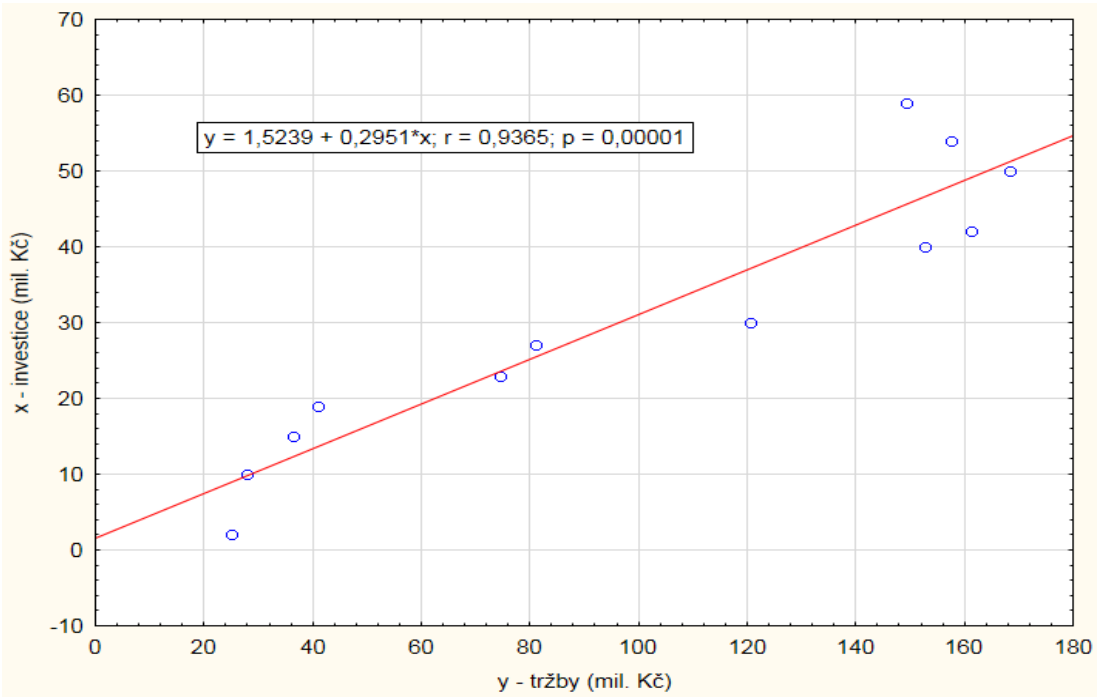
Důvodem neshody obou přístupů je jiné postavení reziduálních čtverců v Metodě nejmenších čtverců při záměně proměnných x a y. V případě závislosti y na x (Obr. 6.3 a 6.4) se kolmice z bodů vedou vertikálně kolmo na osu x, zatímco v případě závislosti x na y (Obr. 6.5 a 6.6) se kolmice z bodů vedou vertikálně kolmo na osu y.



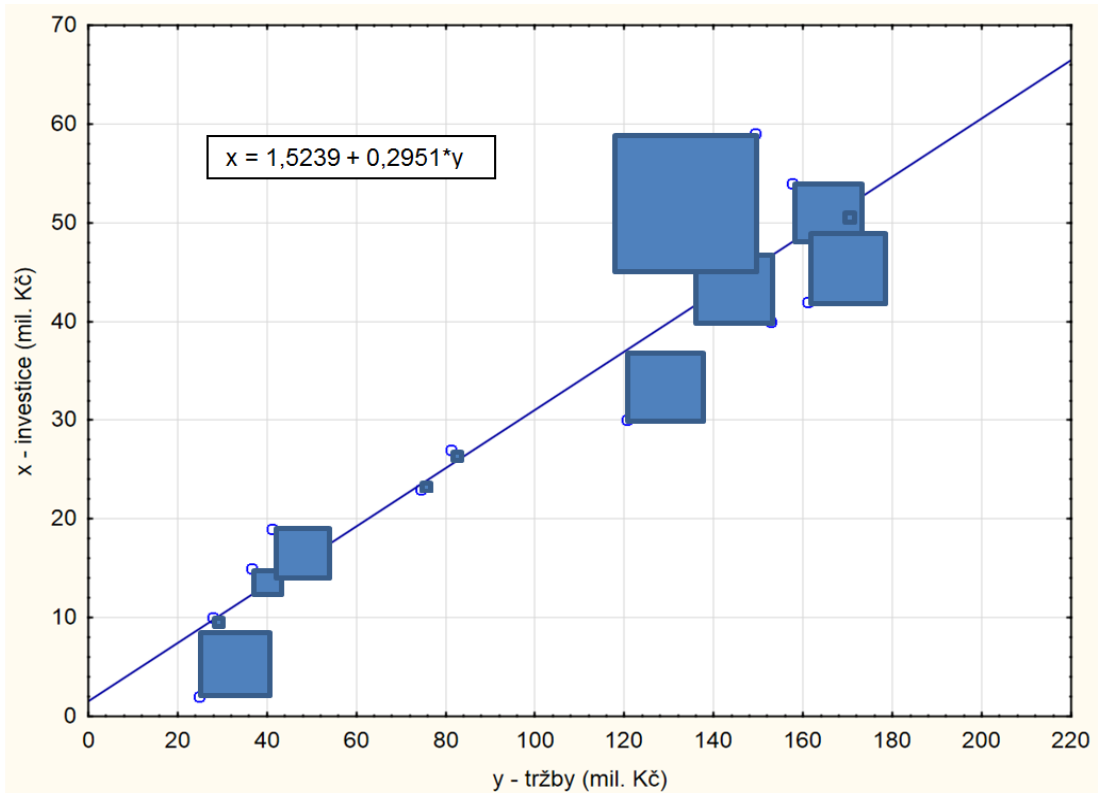
Obr. 6.3 – postavení bodů vůči regresní přímce v MNČ pro závislost y na x (Výstup ze sw Statistica 10)



Obr. 6.4 – postavení bodů vůči regresní přímce v MNČ pro závislost y na x s vykreslením nejmenších čtverců - kolmice z bodů se vedou vertikálně kolmo na osu x (upravený výstup ze sw Statistica 10)

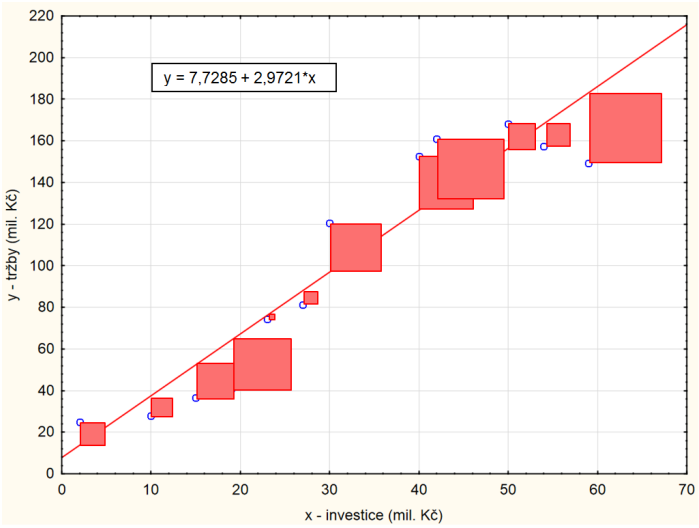


Obr. 6.5 – postavení bodů vůči regresní přímce v MNČ pro závislost x na y (Výstup ze sw Statistica 10)

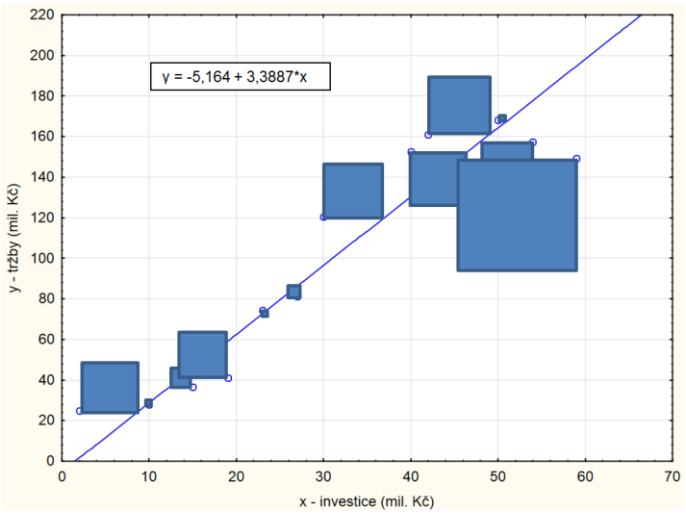


Obr. 6.6 – postavení bodů vůči regresní přímce v MNČ pro závislost x na y s vykreslením nejmenších čtverců - kolmice z bodů se vedou vertikálně kolmo na osu y (upravený výstup ze sw Statistica 10)

Pro ještě lepší pochopení rozdílnosti postavení čtverců při závislosti y na x a při závislosti x na y je možno inverzně otočit osy v obrázku 6.6 znázorňujícím závislost x na y (získáme Obr. 6.7 v dolní části této stránky – modré čtverce) a následně srovnat postavení čtverců se čtverci v obrázku 6.4 znázorňujícím standardní závislost y na x (červené čtverce). Zatímco v případě standardní závislosti y na x v obrázku 6.4 se kolmice z bodů vedou vertikálně kolmo na osu x, při inverzně otočené závislosti x na y v Obr. 6.7 se kolmice z bodů vedou kolmo na osu y, ale z důvodu inverzního otočení nyní horizontálně.

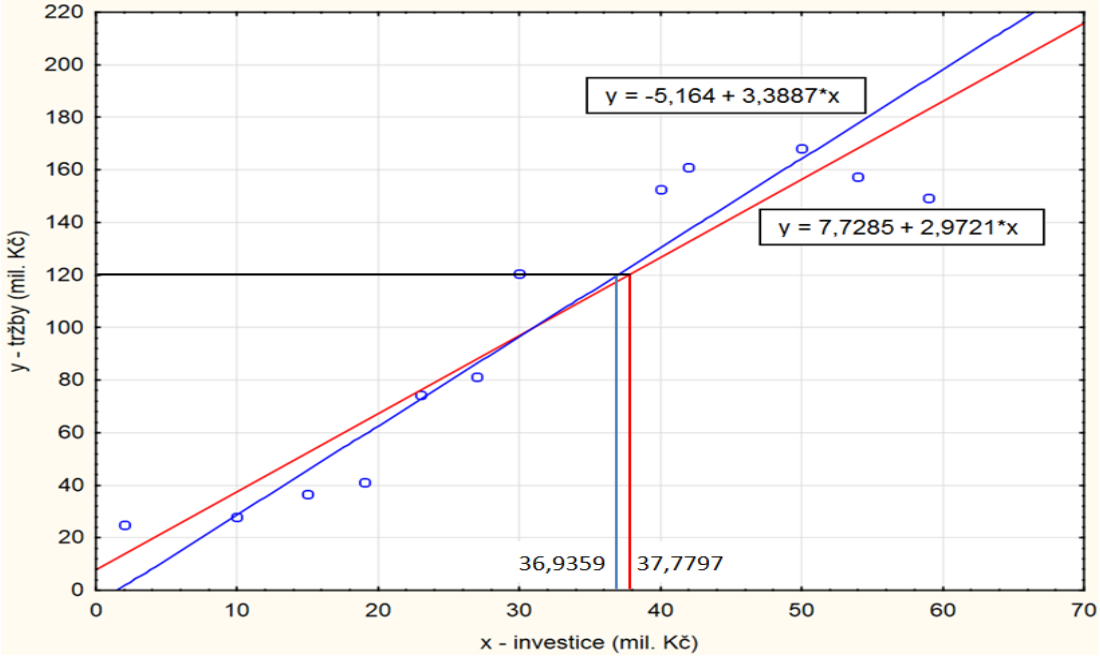


Obr. 6.4 – postavení bodů vůči regresní přímce v MNČ pro závislost y na x s vykreslením nejmenších čtverců - kolmice z bodů se vedou vertikálně kolmo na osu x (upravený výstup ze sw Statistica 10)



Obr. 6.7 – Inverzně otočené osy z obrázku 6.6 původně znázorňujícího závislost x na y - kolmice z bodů vedou kolmo na osu y, ale z důvodu inverzního otočení nyní horizontálně (výstup ze sw Statistica 10)

Jak bylo možno vidět ve srovnání obrázků 6.4 a 6.7, postavení čtverců se při závislosti y na x a závislosti x na y liší, což má za následek i jiné postavení výsledných regresních přímek a tedy i rozdílné odhady potřebných investic x při známé hodnotě tržeb y = 120 mil. Kč (Obr 6.8).



Obr. 6.8 – srovnání postavení regresních přímek při závislosti y na x a závislosti x na y (upravený výstup ze sw Statistica 10)

Pokud nyní porovnáme „nesprávnou“ vypočtenou hodnotu 37,7797 mil. Kč pomocí vyjádření x z původní regresní přímky (červená barva) se „správnou“ hodnotou 36,9359 mil. Kč vypočtenou při správném použití sdružené regresní přímky (modrá barva), vidíme, že se vypočtené hodnoty potřebných investic x s použitím obou přístupů opravdu liší. Z úhlu, který svírají obě regresní přímky, navíc vyplývá, že při odhadu potřebných investic x například pro tržby y = 160 mil. Kč by byl rozdíl mezi odhadnutými hodnotami potřebných investic x ještě markantnější.

6.3 Závěr

Odhad očekávané hodnoty x při známé hodnotě y přímo z inverzního vyjádření proměnné x z původní regresní rovnice vede k nepřesným výsledkům a tento přístup není správný. K odhadu musí být pro přesný odhad použita na základě MNČ znovu vypočtená sdružená regresní přímka závislosti proměnné x na proměnné y.

7 Korelační koeficient a významnost p

7.1 Teoretická část

Při hodnocení míry závislosti mezi číselnými veličinami se někdy setkáváme se dvěma nesprávnými přístupy k interpretaci výsledků na základě korelačního koeficientu r a hodnoty statistické významnosti p:

Problém 1: Uživatel interpretuje míru závislosti pouze na základě statistické významnosti p bez současného přihlédnutí k hodnotě korelačního koeficientu r

Problém 2: Uživatel interpretuje míru závislosti pouze na základě korelačního koeficientu r bez současného přihlédnutí k hodnotě statistické významnosti p

S oběma problémy se často setkáváme i v jiných částech statistiky, než jen při analýze závislosti číselných veličin. Uživatelé statistiky si často neuvědomují, že hodnota statistické významnosti p je na rozdíl od korelačního koeficientu r (či jiné popisné statistiky) úměrná velikosti analyzovaného souboru. V případě, že je počet dvojic měření ve vzorku velký, může nastat na první pohled paradoxní situace, kdy korelační koeficient r ukazuje na slabou závislost, zatímco hodnota p ukazuje na statisticky významnou závislost. Naopak, v případě, že je počet dvojic měření ve vzorku malý, může nastat situace, kdy korelační koeficient r ukazuje na silnou závislost, zatímco hodnota p ukazuje na statisticky nevýznamnou závislost.

7.2 Praktická část

Ad Problém 1: Hodnocení míry závislosti pouze na základě statistické významnosti p bez současného přihlédnutí k hodnotě korelačního koeficientu r

Příklad 7.1: V rámci výzkumu výskytu kyseliny močové v těle byla testována závislost mezi množstvím kyseliny močové, množstvím GMT (gamaglutamyl-transferázy) a cholesterolu. Z důvodu problematického měření GMT byly u pacientů kompletní hodnoty u kyseliny močové (N=75) a cholesterolu (N=75), ale pouze N=14 hodnot u GMT.

	1 kyselina močová	2 GMT	3 cholesterol
1	262	1,14	2,1
2	242		2,2
3	319		7,1
4	292		5,5
5	321	0,36	7,1
6	277		5,9
7	571		5,1
8	147		5,3
9	168		5
10	314		5,4
11	401	1,5	7,7
12	436	0,9	4,6
13	245		5,7
14	266	0,54	5,6
15	336		4,7
16	318		4,9
17	314		5

Obr. 7.1 - data pro závislost mezi množstvím kyseliny močové, množstvím GMT a cholesterolu. V tabulce ukázka pouze části dat, celkový počet dat pro uvedené parametry je postupně 75,14,75.

Z výstupu ze software (Obr. 7.2) se zdá, že množství kyseliny močové v těle koreluje statisticky významně s množstvím cholesterolu. S množstvím GMT naopak množství kyseliny močové v těle statisticky významně nekoreluje.

	Korelace		
Proměnná	kyselina močová	GMT	cholesterol
kyselina močová	1,0000	,5154	,2291
	N=75	N=14	N=75
	p= ---	p=,059	p=,048
GMT	,5154	1,0000	,0692
	N=14	N=14	N=14
	p=,059	p= ---	p=,814
cholesterol	,2291	,0692	1,0000
	N=75	N=14	N=75
	p=,048	p=,814	p= ---

Obr. 7.2 - závislost mezi množstvím kyseliny močové, množstvím GMT a cholesterolu (Výstup ze software Statistica 10)

Naprosto chybnou interpretací v tomto případě je však uvést, že množství kyseliny močové koreluje lépe s množstvím cholesterolu než s množstvím GMT.

Nižší (průkaznější) hodnota p mezi kyselinou močovou a cholesterolem než mezi kyselinou močovou a GMT (p=0,048 vs p=0,059) je způsobena vyšším počtem dvojic měření (N=75 vs N=14), nikoliv však vyšší mírou závislosti (r = 0,2291 vs r = 0,5154).

Z výše uvedeného příkladu tedy můžeme vidět, že při nestejném množství dvojic měření může nastat situace, kdy pro jednu dvojici proměnných může být při statisticky významné hodnotě p menší korelační koeficient r a pro druhou dvojici proměnných může být při statisticky nevýznamné hodnotě p vyšší korelační koeficient r.

Ad Problém 2: Hodnocení závislosti pouze na základě korelačního koeficientu r bez současného přihlédnutí k hodnotě statistické významnosti p

Příklad 7.2: (Vymyšlený demonstrační příklad) Představme si situaci, kdy pro odhad síly závislosti hmotnosti mužů na jejich výšce bylo provedeno měření pouze u třech mužů. Dílem náhody se navíc v měřeném souboru objevili právě dva vysocí a hubení muži a jeden malý a silný muž.

	1 výška (cm)	2 váha (kg)
1	192	89
2	198	82
3	162	102

Obr. 7.3 - data pro hodnocení závislosti hmotnosti mužů na jejich výšce

Pro posouzení míry závislosti mezi výškou a hmotností byl spočítán korelační koeficient r a statistická významnost p.

	Korelace N=3	
Proměnná	výška (cm)	váha (kg)
výška (cm)	1,0000	-,9809
	p= ---	p=,125
váha (kg)	-,9809	1,0000
	p=,125	p= ---

Obr. 7.3 - korelační matice pro hodnocení závislosti mezi výškou a hmotností mužů (Výstup ze software Statistica 10)

Výsledná hodnota korelačního koeficientu vyšla r = -0,9808, jedná se tedy na první pohled o velmi silnou nepřímou závislost mezi výškou a váhou mužů. Nezkušený uživatel statistiky by tedy mohl bez přihlédnutí k hodnotě statistické významnosti p a znalosti reality výsledek interpretovat tak, že čím je muž vyšší, tím má menší hmotnost.

Pokud však při interpretaci výsledků zahrneme do úvah i hodnotu statistické významnosti p, zjistíme, že síla závislosti není významná z důvodu velmi malého počtu dvojic měření. Hodnotě korelačního koeficientu r = -0,9808 tedy není možno věřit a vyvozovat z ní obecně platné úsudky.

7.3 Závěr

Pro posouzení míry závislosti dvou sledovaných proměnných není možno v žádném případě použít pouze hodnotu statistické významnosti p bez současného přihlédnutí k hodnotě korelačního koeficientu r a naopak není možno posuzovat hodnotu korelačního koeficientu r bez současného posouzení statistické významnosti p.

8 Záměna koeficientu determinace a adjustovaného koeficientu determinace

8.1 Teoretická část

Jak již bylo uvedeno v Kapitole 1, jednou z nejčastěji používaných měr kvality (přiléhavosti) regresního modelu je koeficient determinace R^2 definovaný vzorcem

$$R^2 = 1 - \frac{S_R}{S_y}$$

kde

$S_y = \sum_{i=1}^{n}(y_i - y_{Prům})^2$ je celkový součet čtverců modelu

$S_R = \sum_{i=1}^{n}(y_i - y_{iTeor})^2$ je reziduální součet čtverců modelu

Koeficient determinace R^2 říká, jaká část rozptylu vysvětlované proměnné y je v daném modelu vysvětlena pomocí vysvětlujících proměnných x_i a nabývá vždy hodnot v rozmezí 0 až 1 (respektive 0 až 100 %). Čím je tento koeficient determinace vyšší, tím by měl být daný model vhodnější.

Klasický koeficient determinace R^2 však nereflektuje skutečnost, že snahou by mělo být vytvoření pokud možno co nejjednoduššího modelu, ve kterém nebudou nadbytečné proměnné. Jednou z vlastností koeficientu determinace R^2 tedy je, že nezohledňuje počet vysvětlujících proměnných v modelu.

Pro hodnocení vícerozměrných regresních modelů je proto lepší použít adjustovaný (upravený, korigovaný) koeficient determinace, který odstraňuje problém s rostoucí hodnotou R^2 na základě většího počtu proměnných. Jinými slovy je možno říct, že adjustovaný koeficient determinace penalizuje nadměrný počet proměnných v modelu a je pro posouzení vícerozměrného regresního modelu tedy vhodnější.

Adjustovaný koeficient determinace je možno vypočítat ze vzorce

$$R^2_{adj} = 1 - (1 - R^2).\frac{n-1}{n-p} \qquad (8.1)$$

kde p označuje počet parametrů v regresním modelu a n počet měření.

8.2 Praktická část

V příkladu 8.1 si ukážeme podrobně význam Adjustovaného koeficientu determinace při tvorbě vícerozměrného regresního modelu závislosti výše tržeb y na vysvětlujících proměnných x_1 – investice do reklamy a x_2 – průměrná výše slevy.

Příklad 8.1: Při hledání optimálního modelu v Kapitole 1 jsme jako základní model postavili plný regresní model včetně všech lineárních, kvadratických i kubických členů (Obr. 8.1). Pro tento plný model vyšla hodnota koeficientu determinace R^2 = 99,42 % a hodnota adjustovaného koeficientu determinace vyšla R^2_{adj} = 98,73 %.

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,99711618 R2= ,99424067 Upravené R2= ,98732947 F(6,5)=143,86 p<,00002 Směrod. chyba odhadu : 6,4978					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(5)	p-hodn.
Abs.člen			26,90185	36,86867	0,72967	0,498328
x1 - investice do reklamy (mil. Kč)	-0,79152	0,369387	-2,51213	1,17236	-2,14280	0,085016
x2 - průměrná výše slevy (%)	-0,37268	0,832508	-8,63927	19,29889	-0,44766	0,673130
x1**2 - investice do reklamy kvadraticky	5,25560	1,015230	0,25698	0,04964	5,17676	0,003535
x1**3 - investice do reklamy kubicky	-3,76167	0,660986	-0,00312	0,00055	-5,69100	0,002336
x2**2 - průměrná výše slevy kvadraticky	1,52671	1,915970	2,64239	3,31611	0,79684	0,461690
x2**3 - průměrná výše slevy kubicky	-1,04989	1,098124	-0,16182	0,16926	-0,95608	0,382941

Obr. 8.1 - regresní koeficienty a t-testy pro plný regresní model včetně všech lineárních, kvadratických i kubických členů (Výstup ze sw Statistica 10)

Jak vidíme z výše uvedených výstupů pro plný regresní model, některé koeficienty v modelu nejsou statisticky významné. Zkusíme tedy pomocí zpětné krokové regrese z modelu vyjmout proměnnou x_2 v lineárním tvaru, která má nejvyšší p hodnotu a je tedy v modelu nejméně významná (Obr 8.2).

N=12	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,99700042 R2= ,99400984 Upravené R2= ,98901804 F(5,6)=199,13 p<,00000 Směrod. chyba odhadu : 6,0494					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(6)	p-hodn.
Abs.člen			11,10709	9,957264	1,11548	0,307320
x1 - investice do reklamy (mil. Kč)	-0,77672	0,342513	-2,46516	1,087068	-2,26771	0,063868
x1**2 - investice do reklamy kvadraticky	5,27875	0,943937	0,25812	0,046156	5,59228	0,001391
x1**3 - investice do reklamy kubicky	-3,78844	0,612843	-0,00315	0,000509	-6,18175	0,000824
x2**2 - průměrná výše slevy kvadraticky	0,67703	0,243295	1,17179	0,421088	2,78276	0,031881
x2**3 - průměrná výše slevy kubicky	-0,57040	0,225370	-0,08792	0,034737	-2,53096	0,044622

Obr. 8.2 - regresní koeficienty a t-testy po odebrání nejméně významné proměnné x_2 v lineárním tvaru (Výstup ze sw Statistica 10)

Následně zkontrolujeme, jak se po zjednodušení modelu změnily hodnoty koeficientu determinace R^2 a hodnoty adjustovaného koeficientu determinace R^2_{adj}. Vidíme, že po zjednodušení modelu vynecháním nejméně významné proměnné x_2 v lineárním tvaru se hodnota koeficientu determinace R^2 snížila z 99,42 % na hodnotu 99,40 %. Na základě prostého koeficientu determinace by se tedy zdálo, že se model vynecháním této nevýznamné proměnné zhoršil. Z podstaty koeficientu determinace je však jeho snížení při odebrání proměnné očekávatelné. Toto kritérium tedy nelze použít jako ukazatel kvality modelu při jeho postupném zjednodušování odebíráním nevýznamných proměnných.

Adjustovaný koeficient determinace R^2_{adj} se však ve stejném případě zvýšil z 98,73 % na 98,90 %. Pokud tedy použijeme jako kritérium adjustovaný koeficient determinace, ukazuje se, že vhodnějším modelem je nový model po odstranění nadbytečné nevýznamné proměnné. To také odpovídá reálnému požadavku na co největší jednoduchost regresního modelu při zachování srovnatelné kvality modelu.

8.3 Závěr

Pro hodnocení vícerozměrných regresních modelů je vhodnější použít tzv. adjustovaný (upravený, korigovaný) koeficient determinace, který penalizuje nadměrný počet proměnných v modelu.

9 Záměna predikčního a konfidenčního intervalu

9.1 Teoretická část

Častým problémem při finálním použití již hotového regresního modelu je záměna predikčního a konfidenčního intervalu.

Jedním z cílů regresní analýzy je možnost získat odpověď na otázku typu: Jaké hodnoty nabude odezva y_h, je-li hodnota prediktoru rovna nějaké zadané hodnotě x_h? V praxi se za takto položenou otázkou skrývá obvykle jeden ze dvou následujících problémů. (4)

Intervalový odhad **průměrné** odezvy při (nové, budoucí) hodnotě prediktoru x_h, tj. odhad pro $E(y_h)$ – **Konfidenční interval spolehlivosti**

Intervalový odhad odezvy **jako takové** při (nové, budoucí) hodnotě prediktoru x_h, tj. odhad pro y_h – **Predikční interval spolehlivosti**

Bodový odhad pro $E(y_h)$ se neliší od bodového odhadu pro y_h, neboť $E(y_h) = y_h$. Odlišnost nalezneme až u odhadu intervalového.

Při pevné spolehlivosti $(1-\alpha)$ je interval spolehlivosti pro y_h vždy širší než interval spolehlivosti pro $E(y_h)$. Důvodem je, že při konstrukci intervalu spolehlivosti pro $E(y_h)$ se ve vzorci objevuje pouze výběrová variabilita odhadu $E(y_h)$, zatímco při konstrukci intervalu spolehlivosti pro y_h je nutno k výběrové variabilitě odhadu $E(y_h)$ přidat i variabilitu (neznámého) chybového členu h. (4)

9.2 Praktická část

Rozdíl v šířce konfidenčního a predikčního intervalu spolehlivosti si můžeme jednoduše ukázat na dvou praktických příkladech předpovědi z již hotového regresního modelu v software Statistica 10.

Příklad 9.1: V rámci laboratorního výzkumu účinnosti rodenticid byla sledována závislost délky přežití divokých potkanů na množství zkonzumované hubící látky.

	1 Dávka (mg)	2 Délka přežití (h)
1	5	98
2	5	77
3	10	40
4	10	35
5	15	15
6	20	8
7	20	4

Obr. 9.1 - data závislosti délky přežití divokých potkanů na množství zkonzumované hubící látky

Jako optimální byl navržen kvadratický regresní model závislosti s výslednou rovnicí $y = 156{,}7585-15{,}9810.x+0{,}4245.x^2$. Koeficient determinace daného modelu má hodnotu $R^2 = 96{,}7$ %.

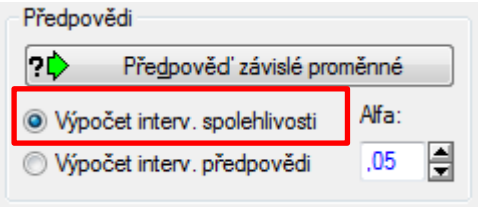
N=7	Výsledky regrese se závislou proměnnou : Délka přežití (h) R= ,98333777 R2= ,96695317 Upravené R2= ,95042976 F(2,4)=58,520 p<,00109 Směrod. chyba odhadu : 7,9768 b*	Sm.chyba z b*	b	Sm.chyba z b	t(4)	p-hodn.
Abs.člen			156,3103	16,75848	9,32724	0,000735
Dávka (mg)	-2,83783	0,556116	-15,9810	3,13172	-5,10295	0,006968
V1**2	1,93260	0,556116	0,4245	0,12215	3,47517	0,025464

Obr. 9.2 - regresní koeficienty a t-testy pro použitý kvadratický model závislosti délky přežití divokých potkanů na množství zkonzumované hubící látky (Výstup ze sw Statistica 10)

Představme si nyní, že potřebujeme odpovědět na dvě následující otázky:

Otázka 1: Jaký je intervalový odhad **průměrné délky přežití** jedinců kteří zkonzumují dávku 8 mg?

V tomto případě musíme stanovit **Konfidenční interval spolehlivosti** (odhad pro $E(y_h)$). Technicky stačí v software Statistica 10 pouze vybrat volbu „Výpočet intervalu spolehlivosti“.



Obr. 9.3 - ukázka zadání výpočtu konfidenčního intervalu spolehlivosti v software Statistica 10

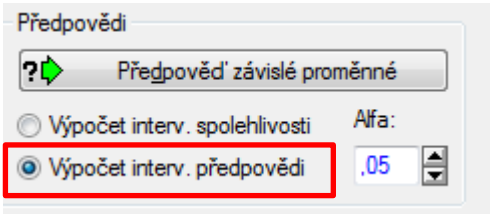
Výsledkem je potom 95 % **konfidenční interval spolehlivosti** (44,438 ; 66,820) s bodovým odhadem 55,629 dní.

Proměnná	Předpovězené hodnoty proměnné: Délka přežití (h)		
	b-váha	Hodnota	b-váha * Hodnot
Dávka (mg)	-15,9810	8,00000	-127,848
V1**2	0,4245	64,00000	27,167
Abs. člen			156,310
Předpověď			55,629
-95,0%LS			44,438
+95,0%LS			66,820

Obr. 9.4 - konfidenční interval spolehlivosti (Výstup ze sw Statistica 10)

Otázka 2: Jaký je intervalový odhad očekávané **délky přežití pro konkrétního jedince**, který zkonzumuje dávku 8 mg?

V tomto případě musíme stanovit **Predikční interval spolehlivosti** (odhad pro y_h). Technicky stačí v software Statistica 10 pouze vybrat volbu „Výpočet intervalu předpovědi".



Obr. 9.5 - ukázka zadání výpočtu predikčního intervalu spolehlivosti v software Statistica 10

Výsledkem je potom 95 % **predikční interval spolehlivosti** (30,815 ; 80,443) s bodovým odhadem 55,629 dní.

Proměnná	Předpovězené hodnoty proměnné: Délka přežití (h)		
	b-váha	Hodnota	b-váha * Hodnot
Dávka (mg)	-15,9810	8,00000	-127,848
V1**2	0,4245	64,00000	27,167
Abs. člen			156,310
Předpověď			55,629
-95,0%PL			30,815
+95,0%PL			80,443

Obr. 9.6 - predikční interval spolehlivosti (Výstup ze sw Statistica 10)

9.3 Závěr

Při finální predikci z již hotového regresního modelu je nutno rozlišovat, za jakým účelem konstruujeme potřebný interval spolehlivosti. Pokud nás zajímá intervalový odhad průměrné odezvy při (nové, budoucí) hodnotě prediktoru x_h, použijeme Konfidenční interval spolehlivosti. Pokud nás zajímá intervalový odhad odezvy jako takové při (nové, budoucí) hodnotě prediktoru x_h, použijeme Predikční interval spolehlivosti.

Z výše uvedeného srovnání výsledků pro konfidenční a predikční interval v praktické části vyplývá, že bodové odhady se v případě obou intervalů spolehlivosti neliší. Odlišnost nalezneme u odhadu intervalového. Při pevné spolehlivosti je Predikční interval spolehlivosti vždy širší než Konfidenční interval spolehlivosti.

10 Problém interpolace a extrapolace

10.1 Teoretická část

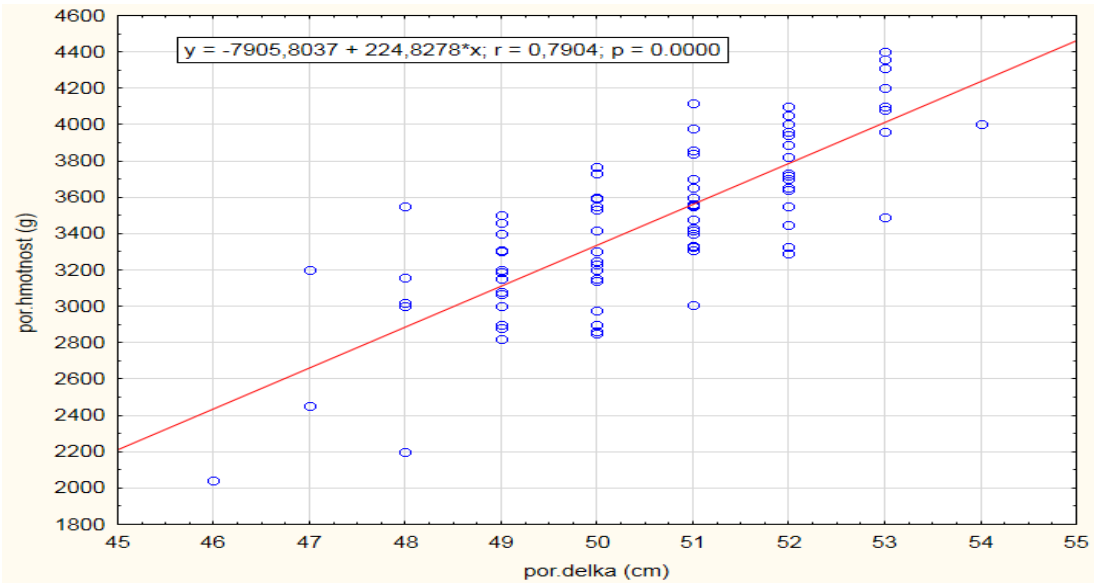
Další častou chybou při finální aplikaci regresních modelů v praxi je následná předpověď pro proměnnou x, která se vyskytuje mimo interval, na kterém byl model vytvořen (tzv. extrapolace). Výsledná hodnota predikce potom často vychází nereálně a kromě špatné předpovědi snižuje také důvěru ve vytvořený regresní model. Důvodem je, že model je tvořen na nějakém oboru x a není možno explicitně předpokládat stejné chování závislosti i pro hodnoty mimo tento obor.

10.2 Praktická část

Příklad 10.1: V rámci sledování závislosti porodní hmotnosti dítěte na jeho porodní délce byl navržen jako optimální lineární regresní model. Výsledná regresní rovnice má tvar: por. hmotnost = - 7905,8 + 224,83. por.delka.

	Výsledky regrese se závislou proměnnou : por.hmotnost R= ,79044422 R2= ,62480207 Upravené R2= ,62093405 F(1,97)=161,53 p<0,0000 Směrod. chyba odhadu : 271,66					
N=99	b*	Sm.chyba z b*	b	Sm.chyba z b	t(97)	p-hodn.
Abs.člen			-7905,80	895,4493	-8,82887	0,000000
por.delka	0,790444	0,062193	224,83	17,6898	12,70945	0,000000

Obr. 10.1 – výsledné regresní koeficienty závislosti porodní hmotnosti dítěte na jeho porodní délce (Výstup ze sw Statistica 10)



Obr. 10.2 - grafické znázornění regresní závislosti porodní hmotnosti dítěte na jeho porodní délce (Výstup ze sw Statistica 10)

Pokud by se uživatel modelu snažil následně odhadnout hypotetickou porodní hmotnost pro novorozence dlouhého např. 30 cm, dosadil by hodnotu 30 cm do vytvořeného regresního modelu.

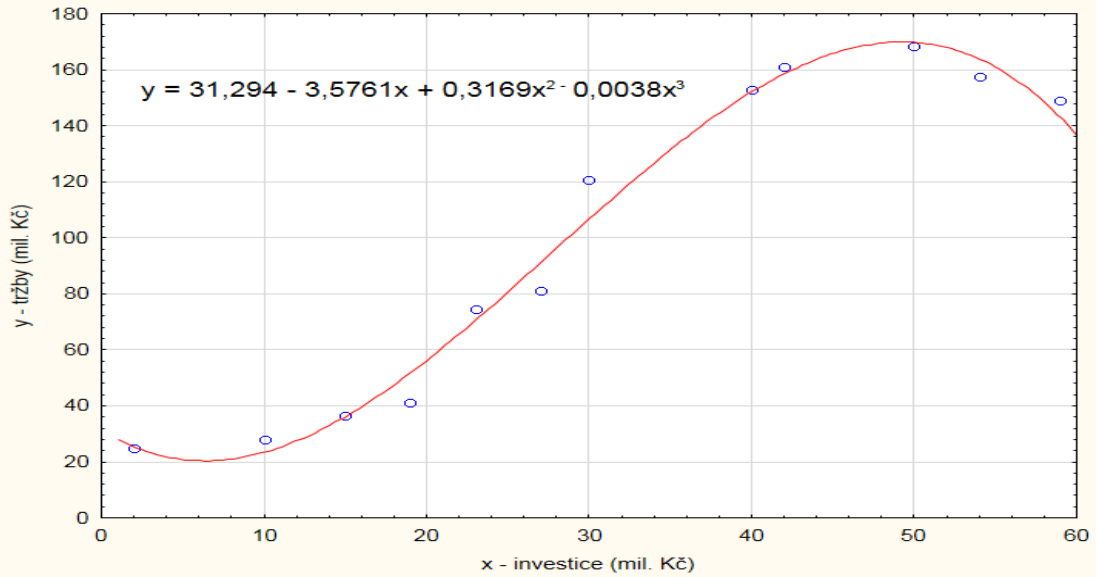
por. hmotnost = - 7905,8 + 224,83. 30 = - 1160,9 g

U dítěte dlouhého 30 cm můžeme tedy očekávat s použitím regresního modelu porodní hmotnost -1160,9 g, což je samozřejmě nereálné. Jak bylo uvedeno výše, důvodem je, že model vznikal na intervalu reálných porodních délek dětí $\langle 46\text{ cm}; 54\text{ cm}\rangle$. Predikovaná porodní hmotnost pro dítě s porodní délkou 30 cm může proto vyjít naprosto nereálně a jakákoliv predikce mimo tento interval může být velmi zavádějící.

Příklad 10.2: V Kapitole 1 byl pro modelování závislosti tržeb na investicích do reklamy navržen místo lineárního modelu kubický model s regresní rovnicí $y = 31{,}294 - 3{,}5761x + 0{,}3169x^2 - 0{,}0038x^3$, který splňoval potřebné předpoklady regresního modelu a vystihoval danou závislost lépe než model lineární. Příslušná data, na kterých model vznikal, jsou uvedena v Tabulce 10.1, grafické znázornění navrženého regresního modelu v Obr. 10.3.

x (mil.Kč) Investice	y (mil.Kč) Tržby
2	24,9
10	27,9
15	36,5
19	41,1
23	74,5
27	81,2
30	120,6
40	152,7
42	161,1
50	168,2
54	157,5
59	149,2

Tabulka 10.1 – data závislosti tržeb (y) na investicích do reklamy (x)



Obr. 10.3 - grafické znázornění závislosti tržeb na investicích při proložení kubickým regresním modelem (Výstup ze sw Statistica 10)

Problém extrapolace si můžeme demonstrovat na příkladu odhadu tržeb pro investice x = 80 mil. Kč, které jsou již mimo interval investic $\langle 2\text{ mil. Kč}\,;59\text{ mil. Kč}\rangle$, na kterém model vznikal. Po dosazení do kubického modelu získáváme:

$$y = 31{,}294 - 3{,}5761 \cdot 80 + 0{,}3169 \cdot 80^2 - 0{,}0038 \cdot 80^3 = -170 \text{ mil. Kč}$$

Při výši investic x = 80 mil. Kč by tedy model po dosazení predikoval tržby -170 mil. Kč, což je v praxi velmi nepravděpodobné. Přestože v některých ekonomických situacích při přesáhnutí extrému „c“ z Kapitoly 2 již tržby klesají, takto prudký pokles by byl velmi výjimečný a je způsoben právě nesmyslnou extrapolací mimo „povolený“ interval, kde by již firma neměla predikci provádět.

V případě regresních modelů vyššího stupně je navíc riziko zkreslených predikcí pro x mimo uvedený interval často podstatně větší než v případě lineárních modelů, což je způsobeno větší rychlostí změny grafu při extrapolaci. Pro srovnání provedeme extrapolaci tržeb pro investice x = 80 mil. Kč pro lineární model z Kapitoly 1:

$$y = 7{,}7285 + 2{,}9721 \cdot 80 = 245{,}5 \text{ mil. Kč}$$

Při výši investic x = 80 mil. Kč by tedy lineární model po dosazení predikoval tržby 245,5 mil. Kč, což je reálnější hodnota než v případě kubického modelu. Lineární

model sice neprokládá data na daném intervalu tak dobře jako model kubický, ale v případě nezbytné extrapolace mimo tento interval nepřináší natolik zavádějící prognózy.

10.3 Závěr

Předpověď pro hodnotu x, která se vyskytuje mimo interval, na kterém byl model vytvořen (tzv. extrapolace), může vést při finálním použití modelu k velmi zavádějícím výsledkům. Pro získání smysluplných výsledků je nutné do výsledného modelu dosazovat při predikci pouze x z oboru, na kterém byl regresní model konstruován.

11 Důsledky použití univariantní analýzy namísto multivariantní regrese

11.1 Teoretická část

Ve většině impaktovaných vědeckých časopisů je při sledování vlivu jedné proměnné na druhou proměnnou při očekávatelném vlivu dalších souvisejících proměnných vyžadována analýza multivariantní regresi. Výhodou je, že pokud vstupují vysvětlující proměnné do multivariantní regresní analýzy současně, je jejich vliv na vysvětlovanou proměnnou y na rozdíl o univariantní analýzy očištěn (adjustován) od vlivu všech ostatních vysvětlujících proměnných.

Z uvedených důvodů je použití univariantní analýzy nedostatečné a může vést k velmi zavádějícím výsledkům, jak bude demonstrováno dále v praktické části.

11.2 Praktická část

Příklad 11.1: V rámci výzkumu léčby osteoporózy byly sledovány tři druhy terapie pacientů a jejich význam na změnu důležitého markeru kostní tvorby PINP.

	1 Jméno	2 PINP(%)	3 terapie	4 CTX(%)	5 CRP(%)
1	BK	105,619	1	-49,367	-94,550
2	HD	25,897	0	24,345	-94,061
3	KA	304,558	1	70,930	-93,688
4	MM	102,943	2	-40,972	357,295
5	NJ	10,399	2	-52,375	178,651
6	PL	164,081	0	108,647	-39,871
7	PN	19,217	0	-11,702	-48,651
8	PA	-27,743	0	4,069	1160,684
9	SP	6,940	1	-37,931	-90,408
10	SJ	133,201	1	157,377	-99,300
11	ZM	-28,270	2	-29,851	-46,079
12	DJ	34,167	0	-17,048	-83,836
13	FP	37,657	0	-38,704	-96,870
14	MJ	0,052	0	70,803	21,701
15	RL	-11,810	2	-3,409	21,151
16	RL	69,933	1	150,000	47,093
17	HM	-20,883	2	14,286	-83,483
18	KH	28,792	0	241,667	-8,708
19	SK	-53,690	2	-77,482	21,937

Obr. 11.1 - ukázka datového souboru pro kvantifikaci vlivu jednotlivých proměnných na marker kostní tvorby PINP

V původní publikaci byly 3 druhy terapie porovnány běžnou univariantní analýzou (v tomto případě neparametrický Kruskal-Wallis test) a zjištěna statisticky významná rozdílnost mezi léčbami.

Kruskal-Wallisova ANOVA založ. na poř.; PINP(%)
Nezávislá (grupovací) proměnná : terapie
Kruskal-Wallisův test: H (2, N= 19) =6,646316 p =,0360

Závislá: PINP(%)	Kód	Počet platných	Součet pořadí	Prům. Pořadí
0	0	8	82,00000	10,25000
1	1	5	73,00000	14,60000
2	2	6	35,00000	5,83333

Souhrnné výsledky
Popisné statistiky

Proměnná	terapie	N platných	Průměr	Medián
PINP(%)	1	5	124,0502	105,6191
PINP(%)	0	8	35,26507	27,34430
PINP(%)	2	6	-0,218431	-16,3462

Obr. 11.2 - výsledky původní univariantní (jednorozměrné) analýzy pomocí neparametrického Kruskal-Wallisova testu (Výstup ze sw Statistica 10)

Z výsledků univariantní analýzy by se zdálo, že změna PINP statisticky průkazně závisí na typu terapie (p = 0,036). Problémem je, že v případě univariantní analýzy nejsou brány v potaz další parametry ovlivňující PINP (například CTX a CRP), které byly mezi skupinami dle terapie značně rozdílné (viz Obr. 11.3).

Rozkladová tabulka popisných statistik
N=19

terapie	PINP(%) N	PINP(%) průměr	PINP(%) medián	CTX(%) N	CTX(%) průměr	CTX(%) medián	CRP(%) N	CRP(%) průměr	CRP(%) medián
0	8	35,2651	27,3443	8	47,7594	14,2065	8	101,2985	-44,2610
1	5	124,0502	105,6191	5	58,2018	70,9302	5	-66,1708	-93,6881
2	6	-0,2184	-16,3462	6	-31,6339	-35,4115	6	74,9119	21,5438
Vš.skup.	19	47,4243	25,8970	19	25,4358	-3,4091	19	48,8950	-46,0794

Obr. 11.3 - Popisné statistiky pro PINP, CTX, CRP (Výstup ze sw Statistica 10)

Na základě univariantní analýzy by se tedy zdálo, že se terapie vzájemně liší a nejúčinnější je terapie č. 1.

V rámci recenzního řízení vznesl recenzent požadavek na provedení multivariantní analýzy závislosti změny PINP na typu terapie při očištění (adjustaci) od dvou

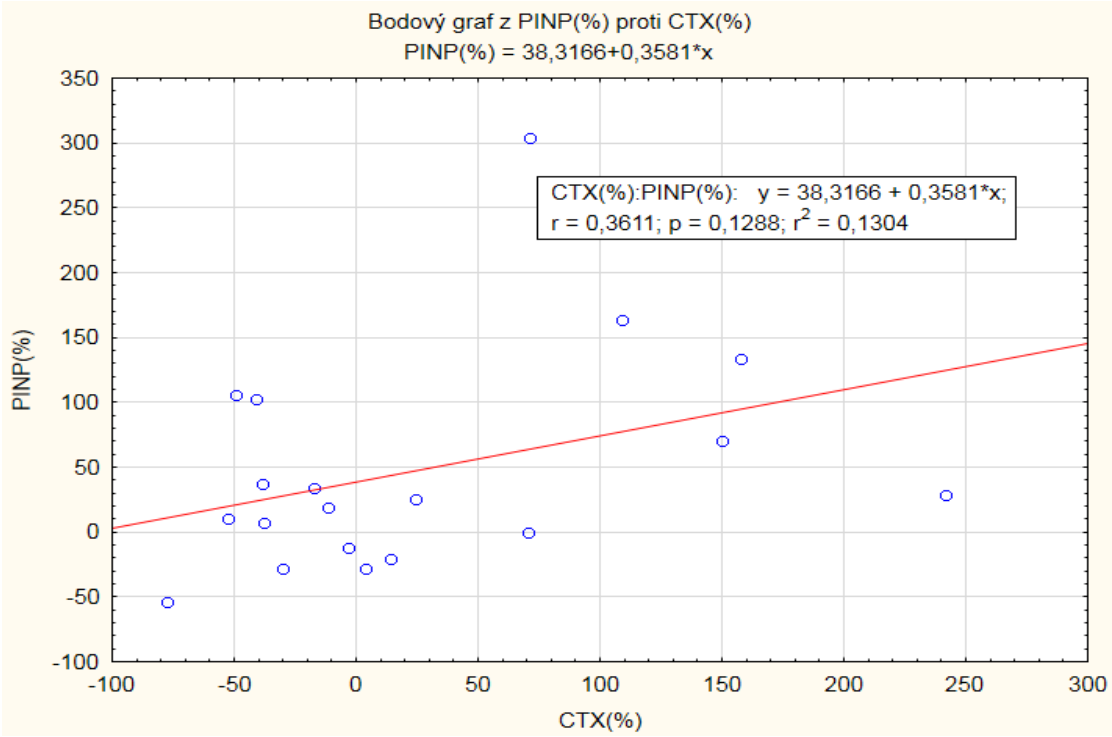
spojitých veličin: změna CTX (marker kostní resorpce) a změna CRP (C-reactive protein), jelikož by tyto proměnné mohly svým skrytým vlivem výsledky léčby ovlivňovat.

Z důvodu splnění požadavku recenzenta byla provedena analýza závislosti PINP na typu terapie multivariantní regresní analýzou, kde do vysvětlujících proměnných byly kromě terapie přidány i další dvě proměnné CTX a CRP. Z výstupů ze statistického software (Obr. 11.4) nyní plyne, že vliv typu terapie na změnu PINP již nevychází statisticky průkazně. Výsledek z multivariantní analýzy je totiž nutno chápat jako „vliv terapie na PINP, pokud by CTX i CRP bylo ve všech skupinách konstantní“ což z dříve uvedené tabulky zjevně není.

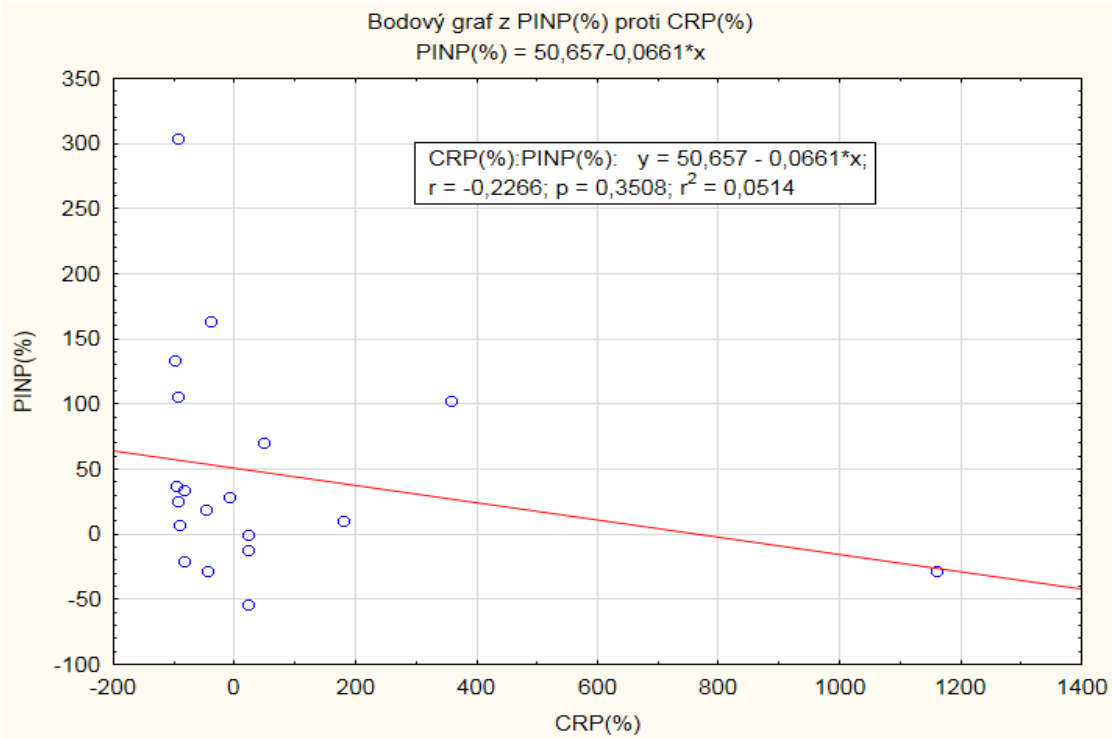
	Jednorozměrné testy významnosti pro PINP(%) Sigma-omezená parametrizace Dekompozice efektivní hypotézy				
Efekt	SČ	Stupně volnosti	PČ	F	p
Abs. člen	39155,02	1	39155,02	6,720139	0,021293
CTX(%)	3830,80	1	3830,80	0,657477	0,431021
CRP(%)	1014,74	1	1014,74	0,174158	0,682771
terapie	27812,21	2	13906,11	2,386692	0,128262
Chyba	81571,27	14	5826,52		

Obr. 11.4 – multivariantní analýza závislosti změny PINP na terapii, CTX a CRP (Výstup ze sw Statistica 10)

Důvodem původního zdání jasného vlivu terapie na změnu PINP byla tedy pravděpodobně situace, kdy ve skupině 1 byly nejvyšší hodnoty změny CTX a nejnižší hodnoty změny CRP. Přičemž právě změna CTX zvyšuje míru změny PINP a změna CRP změnu P1NP snižuje, jak vyplývá z níže uvedených grafů.



Obr. 11.5 – graf závislosti změny PINP na CTX (Výstup ze sw Statistica 10)



Obr. 11.6 – graf závislosti změny PINP na CRP (Výstup ze sw Statistica 10)

11.3 Závěr

Při analýze vlivu jedné proměnné na druhou proměnnou při očekávatelném vlivu dalších souvisejících proměnných je nutné namísto jednoduché univariantní analýzy použít multivariantní regresní analýzu. Pokud vstupují vysvětlující proměnné do multivariantní regresní analýzy současně, je jejich vliv na vysvětlovanou proměnnou y na rozdíl o univariantní analýzy očištěn (adjustován) od vlivu všech ostatních vysvětlujících proměnných.

12 Důsledky neodstranění multikolinearity z regresního modelu

12.1 Teoretická část

Multikolinearitou rozumíme vzájemnou statistickou závislost, tj. korelaci, mezi vysvětlujícími proměnnými ve vícenásobném lineárním regresním modelu. Informaci o této vzájemné závislosti poskytuje matice výběrových korelačních koeficientů

$$R = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

Jsou-li některé nediagonální prvky matice R nenulové, hovoříme o multikolinearitě. Je-li multikolinearita vysoká, hovoříme o škodlivé multikolinearitě. V tom případě se determinant matice R blíží k nule a metoda nejmenších čtverců dává odhady regresních koeficientů s širokými intervaly spolehlivosti, takže výsledky jsou prakticky neupotřebitelné. Za škodlivou multikolinearitu je většinou považováno, když alespoň jeden nediagonální prvek matice R je v absolutní hodnotě větší než 0,8. Výskyt multikolinearity v modelu se často projevuje vysokým koeficientem determinace (blízkým k 1) a zároveň jsou individuální koeficienty statisticky nevýznamné. Model jako celek je naopak často statisticky významný. (13)

Příčiny vzniku multikolinearity

- Přeurčený regresní model obsahující nadměrný počet vysvětlujících proměnných

- Nevhodná volba kombinací vysvětlujících proměnných. V modelu se vyskytující dvojice či skupiny navzájem korelovaných proměnných

Důsledky výskytu multikolinearity

- Nelze odděleně sledovat a interpretovat skutečný vliv jednotlivých vysvětlujících vstupních proměnných na vysvětlovanou proměnnou. Může dojít k nadhodnocení důležitosti některé vysvětlující proměnné

- Silná náchylnost odhadnutého vektoru parametrů b na malé změny v matici dat X

- Vznik pochybností o modelu v důsledku znamének regresních parametrů b neodpovídajících realitě

- Koeficient vícenásobné determinace R^2 vychází blízko 1 a současně jsou t-testy odhadnutých parametrů statisticky nevýznamné (14)

Možnosti odstranění multikolinearity z modelu

- V případě přeurčeného regresního modelu může být řešením identifikace a vypuštění zbytečných vysvětlujících proměnných, které způsobují multikolinearitu

- Pokud mají data charakter časových řad, je možno odstranit multikolinearitu nahrazením jedné z korelovaných proměnných postupnými diferencemi

- Použití regrese na hlavních komponentách. Na základě Metody hlavních komponent (PCA) jsou z množiny všech vysvětlujících proměnných vybrány pouze nekorelované vysvětlující proměnné tak, aby pokrývaly co největší část informace

Identifikace přítomnosti multikolinearity v datech

- Determinant korelační matice. Při silné vzájemné lineární závislosti vysvětlujících proměnných se determinant korelační matice málo liší od nuly

- Jednoduché korelační koeficienty dvojic vysvětlujících proměnných. Hodnoty v absolutní hodnotě vyšší než 0,8 naznačují multikolinearitu

- Vícenásobné korelační koeficienty j-té vysvětlující proměnné vzhledem k ostatním vysvětlujícím proměnným blízké ±1 indikují silnou multikolinearitu (14)

12.2 Praktická část

Příklad 12.1: Jako ukázkový příklad modelu s výskytem vysoké multikolinearity můžeme uvést v této práci již zmíněný model pro odhad tržeb společnosti, avšak s jinými vysvětlujícími proměnnými než byly uvedeny v původním demonstračním příkladu v Kapitole 1. Budeme nyní chtít sestavit vícerozměrný regresní model na základě vysvětlujících proměnných x_1 - investice do reklamy a x_5 – benefity pro stálé zákazníky.

	1 y - tržby (mil. Kč)	2 x1 - investice do reklamy (mil. Kč)	3 x5 - benefity pro stálé zákazníky (mil. Kč)
1	24,9	2	0,21
2	27,9	10	0,9
3	36,5	15	1,6
4	41,1	19	2
5	74,5	23	2,2
6	81,2	27	2,7
7	120,6	30	3
8	152,7	40	4
9	161,1	42	4,3
10	168,2	50	5
11	157,5	54	5,4
12	149,2	59	5,8

Obr. 12.1 - tabulka vstupních dat závislosti tržeb společnosti y na vysvětlujících proměnných x_1 - investice do reklamy a x_5 – benefity pro stálé zákazníky

Výstupy regresního modelu obsahují výsledky, které jsou překvapivé a odporují teoretickému očekávání. Vysoký koeficient determinace i nízká p hodnota F-testu ukazují na dobrý model. t-testy pro obě vysvětlující proměnné x_1 i x_2 však ukazují na statistickou nevýznamnost obou proměnných.

	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,93960574 R2= ,88285895 Upravené R2= ,85682761 F(2,9)=33,915 p<,00006 Směrod. chyba odhadu : 21,842					
N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(9)	p-hodn.
Abs.člen			6,70656	12,93736	0,518387	0,616691
x1 - investice do reklamy (mil. Kč)	-0,960657	2,818261	-3,04893	8,94459	-0,340869	0,741025
x5 - benefity pro stálé zákazníky (mil. Kč)	1,898670	2,818261	60,52478	89,83901	0,673703	0,517427

Obr. 12.2 - regresní model závislosti tržeb společnosti y na vysvětlujících proměnných x_1 - investice do reklamy a x_5 – benefity pro stálé zákazníky (Výstup ze sw Statistica 10)

Překvapivé výsledky poskytuje také interpretace regresních koeficientů. Výsledná regresní rovnice má tvar

$$y = 6{,}707 - 3{,}049.x_1 + 60{,}525.x_2$$

Pokud budeme interpretovat hodnotu regresního koeficientu $b_1 = -3{,}049$, zjistíme, že při zvýšení investic do reklamy o 1 mil. Kč a nezměněných benefitech pro stálé zákazníky by se tržby společnosti snížily o 3,049 mil.Kč. To samozřejmě neodpovídá ekonometrické představě a snižuje důvěryhodnost modelu pro uživatele.

Důvodem paradoxních výsledků je právě vysoká multikolinearita mezi oběma vysvětlujícími proměnnými v modelu, která plyne z níže uvedené korelační matice.

Proměnná	Korelace N=12		
	y - tržby (mil. Kč)	x1 - investice do reklamy (mil. Kč)	x5 - benefity pro stálé zákazníky (mil. Kč)
y - tržby (mil. Kč)	1,000	0,936	0,939
x1 - investice do reklamy (mil. Kč)	0,936	1,000	0,999
x5 - benefity pro stálé zákazníky (mil. Kč)	0,939	0,999	1,000

Obr. 12.3 - korelační matice pro regresní model závislosti tržeb společnosti y na vysvětlujících proměnných x_1 - investice do reklamy a x_5 – benefity pro stálé zákazníky (Výstup ze sw Statistica 10)

Koeficient korelace mezi vysvětlujícími proměnnými x_1 - investice do reklamy a x_5 – benefity pro stálé zákazníky má hodnotu 0,999, což ukazuje na velmi silnou závislost. Při podrobnějším průzkumu dat zjistíme, že management společnosti investuje pravidelně do benefitů pro stálé zákazníky (x_5) přibližně desetkrát méně prostředků než jsou investice do reklamy (x_1). Obě vysvětlující proměnné jsou tedy prakticky svým násobkem, což způsobuje problémy v regresním modelu.

Odstranění multikolinearity z modelu

Odstranění multikolinearity z modelu provedeme pomocí vypuštění jedné z korelovaných proměnných. Z podstaty regresního modelu by bylo vhodnější vypustit vysvětlující proměnnou, která méně koreluje s vysvětlovanou proměnnou.

Protože obě vysvětlující proměnné mají téměř stejné korelační koeficienty, vypustíme proměnnou x_5 - benefity pro stálé zákazníky, která je pro management společnosti z důvodů menších nákladů méně zajímavá než vysvětlující proměnná x_1 - investice do reklamy. Výstupy nového regresního modelu po vypuštění vysvětlující proměnné x_5 - benefity pro stálé zákazníky vypadají následovně.

	Výsledky regrese se závislou proměnnou : y - tržby (mil. Kč) R= ,93645686 R2= ,87695146 Upravené R2= ,86464660 F(1,10)=71,269 p<,00001 Směrod. chyba odhadu : 21,238					
N=12	b*	Sm.chyba z b*	b	Sm.chyba z b	t(10)	p-hodn.
Abs.člen			7,728456	12,49237	0,618654	0,549976
x1 - investice do reklamy (mil. Kč)	0,936457	0,110927	2,972125	0,35206	8,442082	0,000007

Obr. 12.4 - regresní model závislosti tržeb společnosti y pouze na vysvětlující proměnné x_1 - investice do reklamy po vypuštění proměnné x_5 – benefity pro stálé zákazníky (Výstup ze sw Statistica 10)

Z výše uvedených výstupů plyne, že při vypuštění proměnné x_5 – benefity pro stálé zákazníky se adjustovaný koeficient determinace i hodnota F-testu zlepšily. Vysvětlující proměnná x_1 - investice do reklamy je po vypuštění x_5 – benefity pro stálé zákazníky nyní statisticky významná.

Výsledná regresní rovnice má nyní tvar

$$y = 7{,}728 + 2{,}972.x_1$$

Znaménko u regresního koeficientu b_1 již odpovídá ekonomickému očekávání. Pokud budeme interpretovat hodnotu regresního koeficientu $b_1 = +2{,}972$, zjistíme že při zvýšení investic do reklamy o 1 mil. Kč by se tržby společnosti zvýšily o 2,972 mil.Kč.

12.3 Závěr

Před začátkem stavby regresního modelu je nutné otestovat, zda se mezi vysvětlujícími proměnnými v modelu nevyskytuje silná multikolinearita. V případě jejího výskytu mohou výstupy poskytovat nereálné výsledky zvláště v případě interpretace jednotlivých regresních koeficientů. Multikolinearitu je možné z modelu odstranit vypuštěním některé z vysvětlujících proměnných, zavedením postupných diferencí nebo použitím Metody hlavních komponent.

13 Nesplnění předpokladu homoskedasticity

13.1 Teoretická část

Důležitým předpokladem klasického lineárního regresního modelu je homoskedasticita. Jde o vlastnost, která spočívá v tom, že rozptyl reziuí ε_i v regresním modelu je konstantní. Pokud podmínka není splněna, hovoříme o heteroskedasticitě.

Příčiny vzniku heteroskedasticity

Heteroskedasticita může být způsobena různými příčinami. Častou příčinou heteroskedasticity je fakt, že při sběru dat se technika sběru postupně zlepšuje a chyba se zmenšuje nebo se naopak chyba postupně zvětšuje.

Důsledky heteroskedasticity

Přítomnost heteroskedasticity v regresním modelu je silně nežádoucí, a to zejména z těchto důvodů:

- Prognózy s využitím regresního modelu obsahujícího heteroskedasticitu jsou nespolehlivé

- Přítomnost heteroskedasticity způsobuje neplatnost odhadů rozptylů regresních koeficientů a tudíž také odhadů jejich intervalů spolehlivosti a testů hypotéz o jejich statistické významnosti (t testy, F testy). (4)

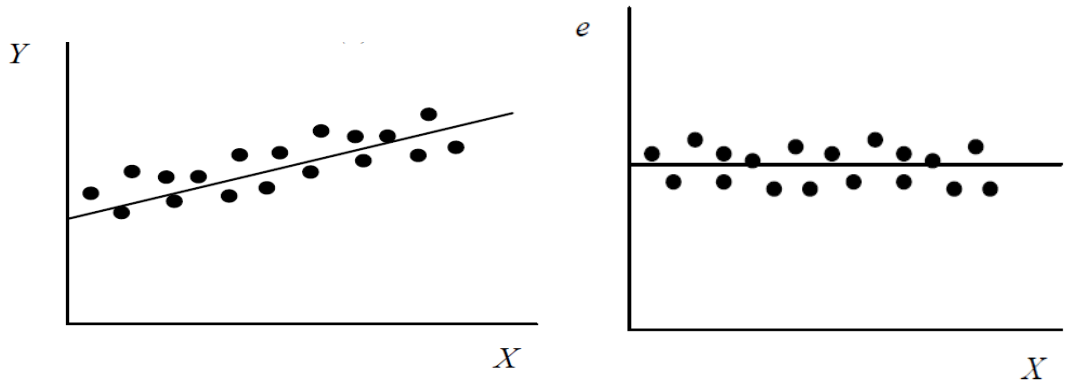
Identifikace přítomnosti heteroskedasticity v datech

Velmi často poznáme přítomnost heteroskedasticity z věcné povahy problému. Například je známo, že s rostoucím věkem zaměstnanců se zvětšuje rozptyl jejich platů.

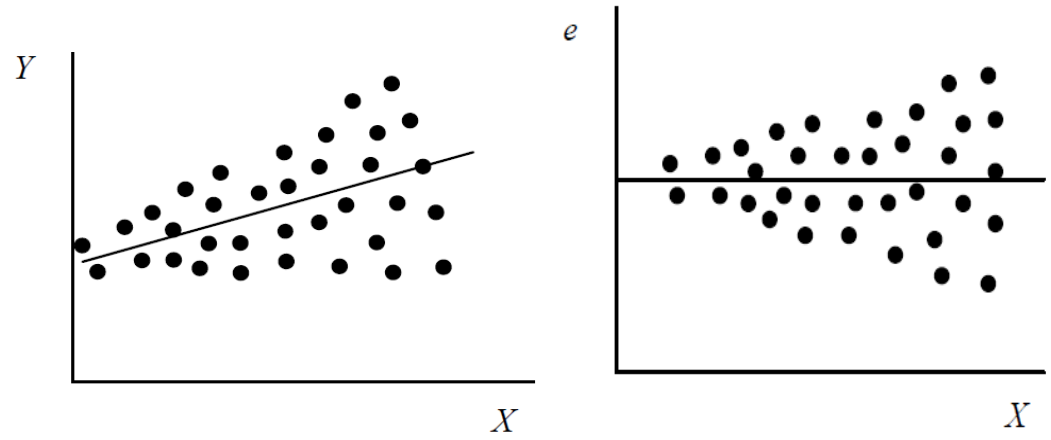
Pokud však nemáme podobné předběžné empirické informace o povaze problému, předpokládáme, že heteroskedasticita není přítomna, že tudíž je rozptyl náhodné složky modelu konstantní. Takové tvrzeni pak můžeme podrobit zkoumání při grafické analýze nebo statistickému testu reziduí e_i.

Grafická analýza

V rámci grafické analýzy přítomnosti heteroskedasticity v modelu můžeme provést hodnocení přímo podle grafu závislosti y na x, případně podle grafu závislosti reziduí e_i na vysvětlující proměnné x.



Obr. 13.1 - ukázka grafického znázornění regresního modelu bez přítomnosti heteroskedasticity (11)



Obr. 13.2 - ukázka grafického znázornění regresního modelu s přítomností heteroskedasticity (11)

Z výše uvedených grafických znázornění regresních modelů bez a s přítomností heteroskedasticity je možno vidět, že v modelech s výskytem heteroskedasticity není splněn předpoklad o konstantním rozptylu reziduí.

Testy heteroskedasticity

Detekce heteroskedasticity s pomocí statistického testu je obvykle založena na nulové hypotéze, že rozptyly náhodné složky εi^2 jsou konstantní, přičemž se analyzují jejich odhady, tj. rezidua ei^2. V literatuře můžeme nalézt podrobné testy heteroskedasticity s názvy jako Breusch – Paganův test, Whiteův test, Glejserův test, Goldfeld-Quandtův test aj. Tyto statistické testy lze provádět pomocí specializovaných statistických software.

Pro pochopení podstaty předpokladu homoskedasticity reziduí je možno zájemcům navíc demonstrovat tzv. Bartletův test heteroskedasticity, který představuje zjednodušený Goldfeld-Quandtův test a lze k jeho provedení využít funkce v Excelu. (13)

Bartleyův test

Test vychází z rozdělení souboru dat podle velikosti některé vysvětlující proměnné do dvou částí: $x_i \leq$ medián a $x_i >$ medián.

Následně se testuje hypotéza o rovnosti rozptylů reziduí v obou částech souboru. Pokud se hypotéza o rovnosti rozptylu reziduí v obou částech zamítá, potom se hypotéza o konstantnosti rozptylu náhodné složky, neboli hypotéza o přítomnosti heteroskedasticity, přijímá. Nevýhodou Bartleyova testu je citlivost na porušení předpokladu o normálním rozdělení. Pokud je předpoklad o normálním rozdělení splněn, je Bartletův test nejsilnější z dostupných testů. (1)

Odstranění heteroskedasticity z modelu

Nejznámější metodou k odstranění heteroskedasticity z modelu je metoda vážených nejmenších čtverců MVNČ. V MVNČ předpokládáme určitý typ nekonstantního chovaní rozptylu náhodné složky.

Budeme předpokládat, že rozptyl náhodné složky je přímo úměrný kvadrátu vysvětlující proměnné x, tj.

$$E\left(\varepsilon_i^2\right) = \sigma^2 . x_i^2, i = 1,2, \dots, n.$$

Transformovaný regresní model získáme tak, že regresní rovnici

$$y_i = \beta_0 + \beta_1 . x_i + \varepsilon_i, i = 1,2, \dots, n.$$

vydělíme hodnotou x_i, čímž obdržíme

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i} = \beta_0 . \frac{1}{x_i} + \beta_1 + \delta_i, i = 1,2, \dots, n.$$

kde pro novou náhodnou chybu δ_i platí

$$E(\delta_i^2) = E\left(\frac{\varepsilon_i^2}{x_i^2}\right) = \sigma^2, i = 1,2, \dots, n.$$

Provedením transformace

$$y_i^{'} = \frac{y_i}{x_i}, x_i^{'} = \frac{1}{x_i}, i = 1,2, \dots, n.$$

obdržíme novy regresní model

$$y_i^{'} = \beta_1 + \beta_0 . x_i^{'} + \delta_i, i = 1,2, \dots, n.$$

což je nový lineární regresní model bez heteroskedasticity. (13)

13.2 Praktická část

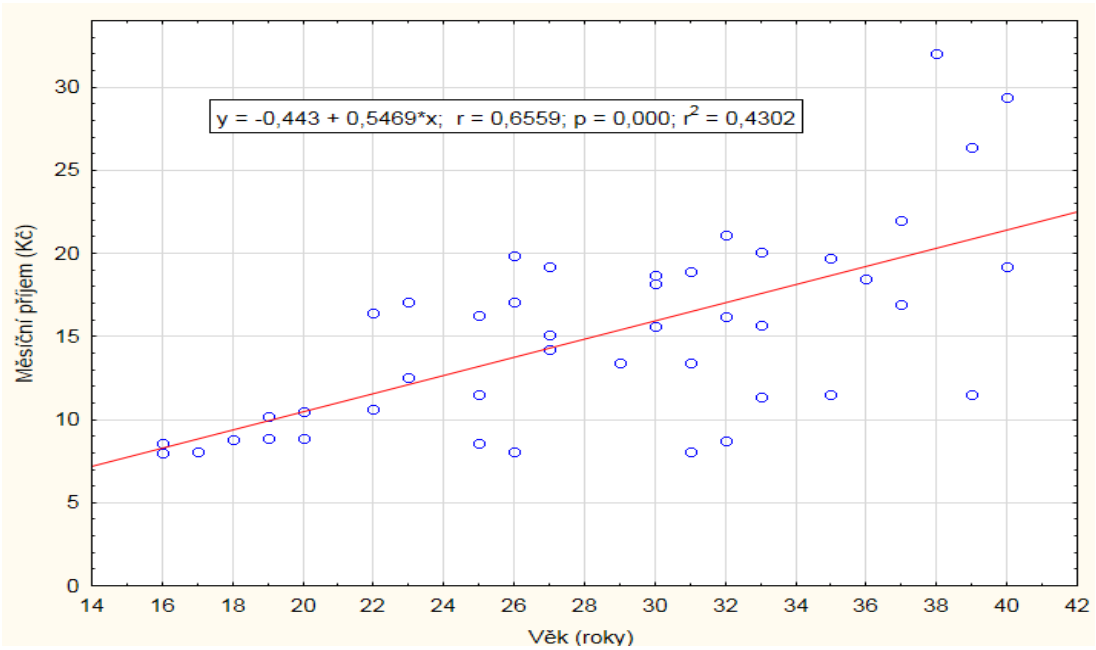
Příklad 13.1: Jako ukázkový příklad modelu s výskytem heteroskedasticity byl vytvořen regresní model závislosti měsíčních příjmů zaměstnanců v tis. Kč na věku daného zaměstnance v letech. Úkolem je osvojení si detekce výskytu heteroskedasticity v modelu a odstranění heteroskedasticity z modelu.

poř.č.	x-věk	y-měsíční příjem (tis Kč)	poř.č.	x-věk	y-měsíční příjem (tis Kč)
1	16	8	23	30	15,6
2	16	8,6	24	30	18,2
3	17	8,1	25	30	18,7
4	18	8,8	26	31	8,1
5	19	8,9	27	31	13,4
6	19	10,2	28	31	18,9
7	20	8,9	29	32	8,7
8	20	10,5	30	32	16,2
9	22	10,6	31	32	21,1
10	22	16,4	32	33	11,4
11	23	17,1	33	33	15,7
12	23	12,5	34	33	20,1
13	25	8,6	35	35	11,5
14	25	16,3	36	35	19,7
15	25	11,5	37	36	18,5
16	26	8,1	38	37	16,9
17	26	17,1	39	37	22
18	26	19,9	40	38	32
19	27	15,1	41	39	11,5
20	27	19,2	42	39	26,4
21	27	14,2	43	40	19,2
22	29	13,4	44	40	29,4

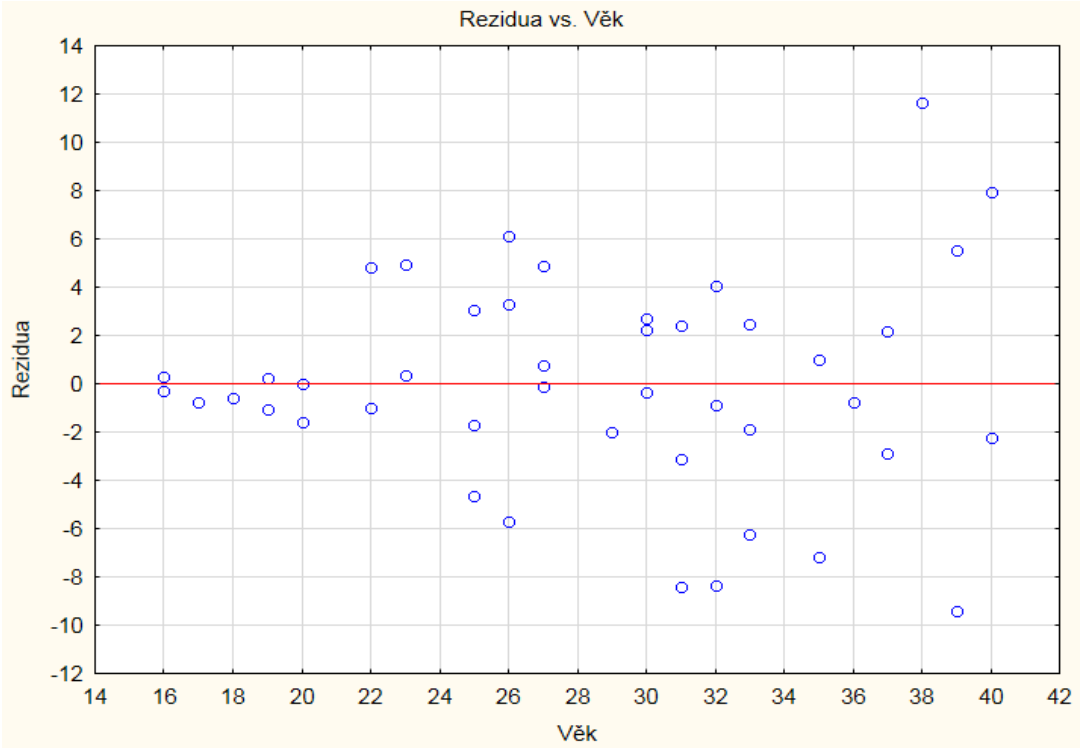
Obr. 13.3 - tabulka vstupních dat závislosti měsíčních příjmů na věku zaměstnance

Grafická analýza výskytu heteroskedasticity v modelu

V rámci grafické analýzy byl vytvořen bodový graf závislosti měsíčních příjmů y na věku zaměstnance x a graf reziduí proti nezávislé proměnné x



Obr. 13.4 - bodový graf závislosti měsíčních příjmů na věku zaměstnance demonstrující graficky výskyt heteroskedasticity v modelu (Výstup ze sw Statistica 10)



Obr. 13.5 - graf reziduí proti nezávislé proměnné x demonstrující graficky výskyt heteroskedasticity v modelu (Výstup ze sw Statistica 10)

V obou uvedených grafech je zřejmé porušení předpokladu konstantního rozptylu reziduí.

Bartletův test heteroskedasticity v sw Excel

Podstatu heteroskedasicity je možno dobře didakticky demonstrovat ukázkou rozdílnosti rozptylu reziduí ve dvou intervalech vysvětlující proměnné x. Nejdříve jsou pro všechny vysvětlující proměnné x vypočteny dosazením do regresní rovnice teoretické hodnoty y_{teor}. Hodnoty reziduálních složek e_i jsou následně vypočítány podle vztahu $e_i = y_i - y_{iteor}$. Podle mediánu hodnot vysvětlující proměnné x jsou následně rezidua rozdělena do dvou podmnožin (e_1 a e_2) a použitím F testu o shodě rozptylů jsou rozptyly reziduí v těchto dvou podmnožinách porovnány.

poř.č.	x	y	y_{teor}	e_1	poř.č.	x	y_{teor}	y odhad	e_2
1	16	8	8,31	-0,31	23	30	15,6	15,96	-0,36
2	16	8,6	8,31	0,29	24	30	18,2	15,96	2,24
3	17	8,1	8,85	-0,75	25	30	18,7	15,96	2,74
4	18	8,8	9,40	-0,60	26	31	8,1	16,51	-8,41
5	19	8,9	9,95	-1,05	27	31	13,4	16,51	-3,11
6	19	10,2	9,95	0,25	28	31	18,9	16,51	2,39
7	20	8,9	10,50	-1,60	29	32	8,7	17,06	-8,36
8	20	10,5	10,50	0,00	30	32	16,2	17,06	-0,86
9	22	10,6	11,59	-0,99	31	32	21,1	17,06	4,04
10	22	16,4	11,59	4,81	32	33	11,4	17,60	-6,20
11	23	17,1	12,14	4,96	33	33	15,7	17,60	-1,90
12	23	12,5	12,14	0,36	34	33	20,1	17,60	2,50
13	25	8,6	13,23	-4,63	35	35	11,5	18,70	-7,20
14	25	16,3	13,23	3,07	36	35	19,7	18,70	1,00
15	25	11,5	13,23	-1,73	37	36	18,5	19,25	-0,75
16	26	8,1	13,78	-5,68	38	37	16,9	19,79	-2,89
17	26	17,1	13,78	3,32	39	37	22	19,79	2,21
18	26	19,9	13,78	6,12	40	38	32	20,34	11,66
19	27	15,1	14,32	0,78	41	39	11,5	20,89	-9,39
20	27	19,2	14,32	4,88	42	39	26,4	20,89	5,51
21	27	14,2	14,32	-0,12	43	40	19,2	21,43	-2,23
22	29	13,4	15,42	-2,02	44	40	29,4	21,43	7,97

Obr. 13.6 - demonstrace výpočtu Bartletova testu heteroskedasticity v sw Excel

Dvouvýběrový F-test pro rozptyl		
	Soubor 1	*Soubor 2*
Stř. hodnota	0,426809091	-0,428027273
Rozptyl	9,180872726	29,50232176
Pozorování	22	22
Rozdíl	21	21
F	0,311191533	
P(F<=f) (1)	0,005024302	
F krit (1)	0,479803022	

Obr. 13.7 - F test o shodě dvou rozptylů (Výstup ze software MS Excel)

Z výsledné p hodnoty F testu o shodě dvou rozptylů v sw Excel (p=0,005) vyplývá, že nulová hypotéza o shodě rozptylů reziduí v obou podmnožinách se zamítá, v modelu je tedy přítomna vysoká heteroskedasticita.

V souvislosti s použitím testů o splnění předpokladu je potřeba opět připomenout častý problém nesprávné interpretace při malém počtu dat v modelu. V tomto případě

test často nezamítne nulovou hypotézu (předpoklad není porušen) i při datech velmi výrazně nesplňujících daný předpoklad. Nesprávnou interpretací potom v tomto případě je, že v modelu není porušen sledovaný předpoklad.

Odstranění heteroskedasticity z modelu

Dle postupu popsaného v teoretické části této kapitoly se nyní pokusíme odstranit vysokou heteroskedasticitu v modelu pomocí vhodné transformace dat.

Nejdříve transformujeme v datové matici vstupní proměnné x a y pomocí transformace

$$y_i^{'} = \frac{y_i}{x_i}, x_i^{'} = \frac{1}{x_i}, i = 1,2, \dots, n.$$

poř.č.	x	y	y´	x´	y´teor	e_1	poř.č.	x	y	y´	x´	y´teor	e_2
1	16	8	0,500	0,063	0,519	-0,019	23	30	15,6	0,520	0,033	0,532	-0,012
2	16	8,6	0,538	0,063	0,519	0,018	24	30	18,2	0,607	0,033	0,532	0,075
3	17	8,1	0,476	0,059	0,521	-0,044	25	30	18,7	0,623	0,033	0,532	0,091
4	18	8,8	0,489	0,056	0,522	-0,033	26	31	8,1	0,261	0,032	0,533	-0,271
5	19	8,9	0,468	0,053	0,524	-0,055	27	31	13,4	0,432	0,032	0,533	-0,100
6	19	10,2	0,537	0,053	0,524	0,013	28	31	18,9	0,610	0,032	0,533	0,077
7	20	8,9	0,445	0,050	0,525	-0,080	29	32	8,7	0,272	0,031	0,533	-0,261
8	20	10,5	0,525	0,050	0,525	0,000	30	32	16,2	0,506	0,031	0,533	-0,027
9	22	10,6	0,482	0,045	0,527	-0,045	31	32	21,1	0,659	0,031	0,533	0,126
10	22	16,4	0,745	0,045	0,527	0,219	32	33	11,4	0,345	0,030	0,533	-0,188
11	23	17,1	0,743	0,043	0,528	0,216	33	33	15,7	0,476	0,030	0,533	-0,058
12	23	12,5	0,543	0,043	0,528	0,016	34	33	20,1	0,609	0,030	0,533	0,076
13	25	8,6	0,344	0,040	0,529	-0,185	35	35	11,5	0,329	0,029	0,534	-0,206
14	25	16,3	0,652	0,040	0,529	0,123	36	35	19,7	0,563	0,029	0,534	0,029
15	25	11,5	0,460	0,040	0,529	-0,069	37	36	18,5	0,514	0,028	0,535	-0,021
16	26	8,1	0,312	0,038	0,530	-0,218	38	37	16,9	0,457	0,027	0,535	-0,078
17	26	17,1	0,658	0,038	0,530	0,128	39	37	22	0,595	0,027	0,535	0,060
18	26	19,9	0,765	0,038	0,530	0,236	40	38	32	0,842	0,026	0,535	0,307
19	27	15,1	0,559	0,037	0,530	0,029	41	39	11,5	0,295	0,026	0,536	-0,241
20	27	19,2	0,711	0,037	0,530	0,181	42	39	26,4	0,677	0,026	0,536	0,141
21	27	14,2	0,526	0,037	0,530	-0,005	43	40	19,2	0,480	0,025	0,536	-0,056
22	29	13,4	0,462	0,034	0,532	-0,070	44	40	29,4	0,735	0,025	0,536	0,199

Obr. 13.8 - ukázka transformace dat pro odstranění heteroskedasticity z modelu

Nový transformovaný regresní model má nyní tvar

$$y_i^{'} = 0{,}5469 - 0{,}443.x_i^{'}$$

Přesvědčíme se, zda byla z modelu opravdu odstraněna heteroskedasticita. Provedeme opět stejně jako u původního netransformovaného modelu Bartletův test heteroskedasticity pro dvě skupiny reziduí e_1 a e_2.

Dvouvýběrový F-test pro rozptyl		
	Soubor 1	*Soubor 2*
Stř. hodnota	0,016094408	-0,015369802
Rozptyl	0,01506189	0,023610488
Pozorování	22	22
Rozdíl	21	21
F	0,637932161	
P(F<=f) (1)	0,155382615	
F krit (1)	0,479803022	

Obr. 13.9 - F test o shodě dvou rozptylů (Výstup ze software MS Excel)

Z výstupu F testu pro Bartletův test heteroskedasticity vidíme, že po provedení navržené transformace v modelu již není statisticky významně porušen předpoklad homoskedasticity rezduí.

13.3 Závěr

Homoskedasticita je důležitým předpokladem lineárního regresního modelu a spočívá v tom, že rozptyl reziuí ε_i v regresním modelu je konstantní. Pokud podmínka není splněna, hovoříme o heteroskedasticitě. Odstranění heteroskedasticity z modelu je možno provést transformací s využitím metody vážených nejmenších čtverců MVNČ.

14 Výskyt autokorelace reziduí v modelu

14.1 Teoretická část

Dalším důležitým předpokladem klasického lineárního regresního modelu je nepřítomnost autokorelace reziduí v modelu. Autokorelaci je možno definovat jako porušení předpokladu o vzájemné nezávislosti náhodných složek z různých pozorování. V praktických úlohách je velmi často tento požadavek porušen.

Příčiny vzniku autokorelace

- Setrvačnost v datech vyvolaná např. způsobem měření, systematickými vlivy
- Chyby měření vysvětlované proměnné jsou zahrnuty do náhodné složky modelu
- Odhad modelu z dat obsahujících zpožděné proměnné
- Fyzikální důvody – např. vliv teploty, vliv okolního prostředí
- Volba nesprávného modelu

Důsledky autokorelace

- Vychýlené odhady rozptylu modelu a směrodatných chyb bodových odhadů
- Intervaly spolehlivosti nejsou směrodatné
- Statistické testy ztrácejí na síle

Možnosti odstranění autokorelace

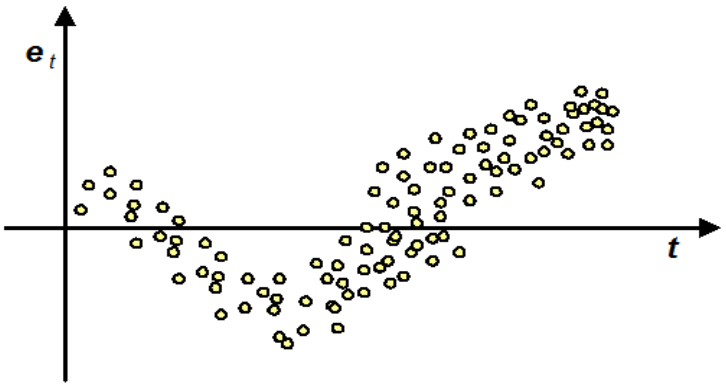
- Použití vhodnějšího trendu
- Použití relevantnějších zpoždění proměnných

Identifikace přítomnosti autokorelace v datech

Grafická analýza dle grafu vývoje reziduí v čase

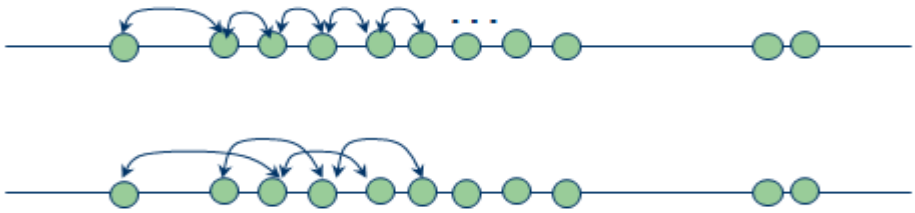
Přibližný odhad výskytu autokorelace v modelu je možno provést pomocí grafické analýzy grafu vývoje reziduí v čase. Nepřítomnost autokorelace je charakterizována náhodným kolísáním reziduí v čase. Pozitivní autokorelace je vytvářena dlouhými sekvencemi reziduí stoupajících nebo klesajících, tj. jdoucích ve stejném trendu.

Negativní autokorelace vzniká hlavně pravidelným střídáním reziduí (sekvence vyšší-nižší-vyšší-nižší). (14)



Obr. 14.1 - graf vývoje reziduí v čase, ukázka pozitivní autokorelace (15)

Podle délky zpoždění m mezi rezidui rozlišujeme autokorelaci prvního, druhého až m-tého řádu.



Obr. 14.2 - grafické znázornění autokorelace prvního a druhého řádu v čase (15)

Řád nejvyšší autokorelace je možno odhalit použitím reziduální autokorelační funkce (ACF).

Grafická analýza dle reziduální autokorelační funkce (ACF) (16)

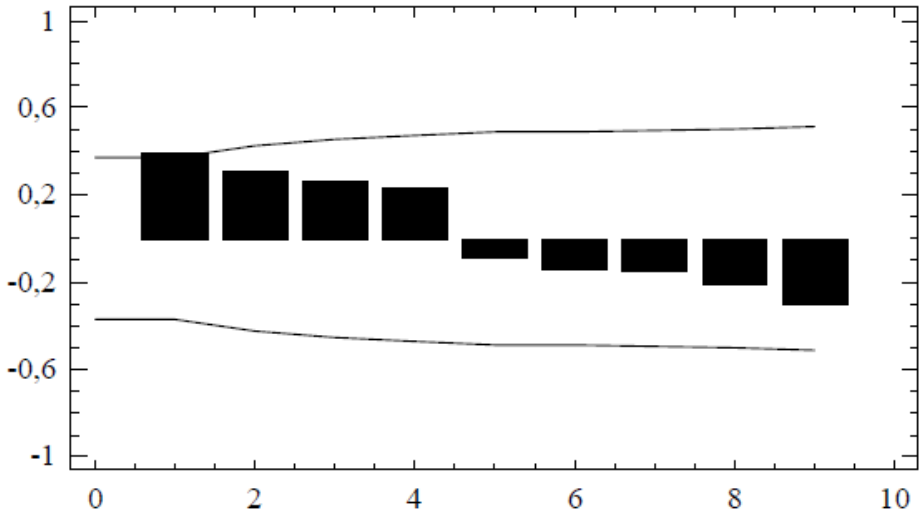
Mírou lineární závislosti časově zpožděných veličin a_t a a_{t-k} jsou koeficienty autokorelace reziduí definované vztahem

$$r_k = \hat{\rho}_k = \frac{\sum_{t=k+1}^{T} \hat{a}_t . \hat{a}_{t-k}}{\sum_{t=1}^{T} \hat{a}_t^2} \in \langle -1, 1 \rangle$$

Graf, ve kterém jsou na vodorovné ose časová zpoždění a na svislé ose koeficienty autokorelace reziduí r_k se nazývá reziduální autokorelační funkcí (ACF). Pokud žádný autokorelační koeficient r_k nepřekračuje meze 95% intervalu

$$\left(\frac{-2}{\sqrt{T}}, \frac{2}{\sqrt{T}}\right)$$

je možné předpokládat, že nesystematická složka není autokorelovaná.



Obr. 14.3 - ukázka reziduální autokorelační funkce (ACF) (10)

Testy autokorelace

Durbin-Watsonův test

Nekorelovanost v nesystematické složce (nepřítomnost autokorelace reziduí) můžeme testovat pomocí prvního koeficientu autokorelace

H_0: $\rho_1 = 0$ autokorelace v modelu není přítomna

H_1: $\rho_1 \neq 0$ autokorelace v modelu je přítomna

Durbinovo-Watsonovo kritérium má tvar

$$DW = \frac{\sum_{t=2}^{T}(\hat{a}_t - \hat{a}_{t-1})^2}{\sum_{t=1}^{T}\hat{a}_t^2}$$

14.2 Praktická část

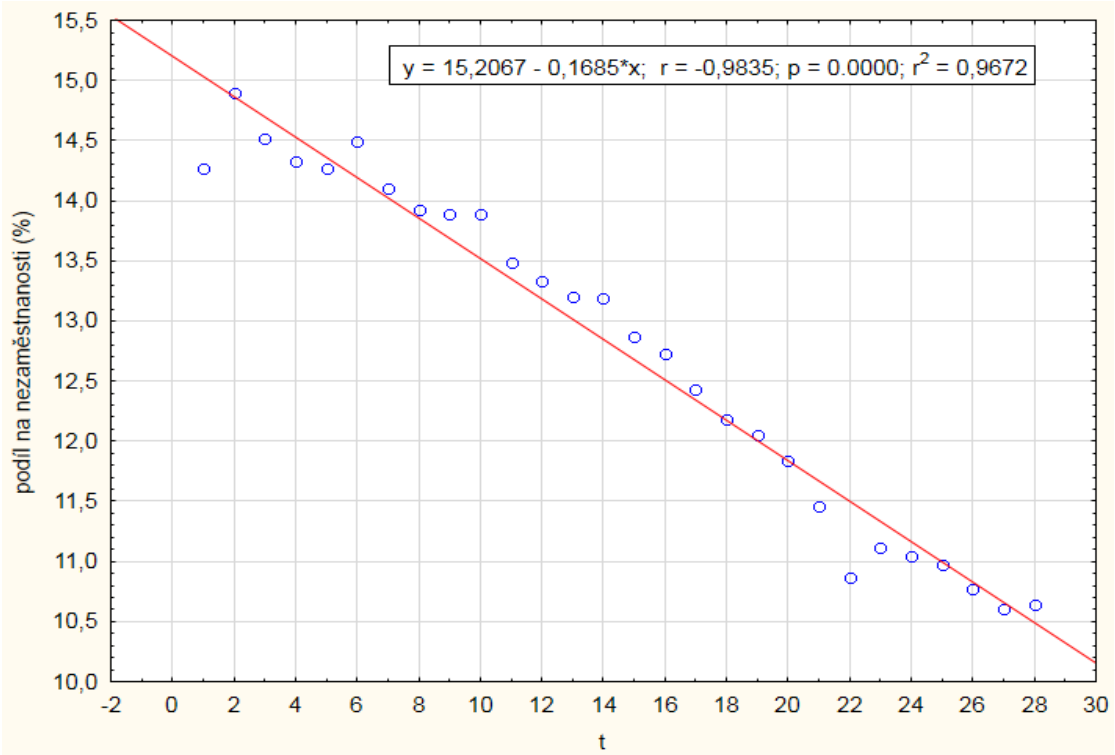
Identifikaci a odstranění autokorelace reziduí si můžeme ukázat na příkladu vývoje nezaměstnanosti použitém v (16). Jak již bylo uvedeno v teoretické části, v případě ekonomických veličin dochází v některých případech při vývoji v čase k setrvačnosti, která může způsobovat výskyt autokorelace reziduí.

Příklad 14.3: Máme k dispozici sezónně očištěné čtvrtletní údaje o podílu nezaměstnaných 30 – 34 letých na celkové nezaměstnanosti v letech 1994 až 2000 v %.

	A	B	C	D	E
1		1. čtvrtetí	2. čtvrtletí	3. čtvrtetí	4. čtvrtletí
2	1994	14,273	14,8985	14,5173	14,3222
3	1995	14,273	14,4958	14,1054	13,9243
4	1996	13,8846	13,8918	13,4877	13,3276
5	1997	13,2049	13,1872	12,8699	12,7308
6	1998	12,4282	12,1805	12,0462	11,8357
7	1999	11,4572	10,8719	11,1196	11,04
8	2000	10,9717	10,7712	10,6048	10,6422

Obr. 14.4 - sezónně očištěná čtvrtletní časová řada podílu nezaměstnaných 30 – 34 letých na celkové nezaměstnanosti v letech 1994 - 2000

Pro prvotní náhled na data nejdříve vytvoříme bodový graf závislosti podílu nezaměstnaných na čase.



Obr. 14.5 - bodový graf závislosti podílu nezaměstnaných na čase (Výstup ze sw Statistica 10)

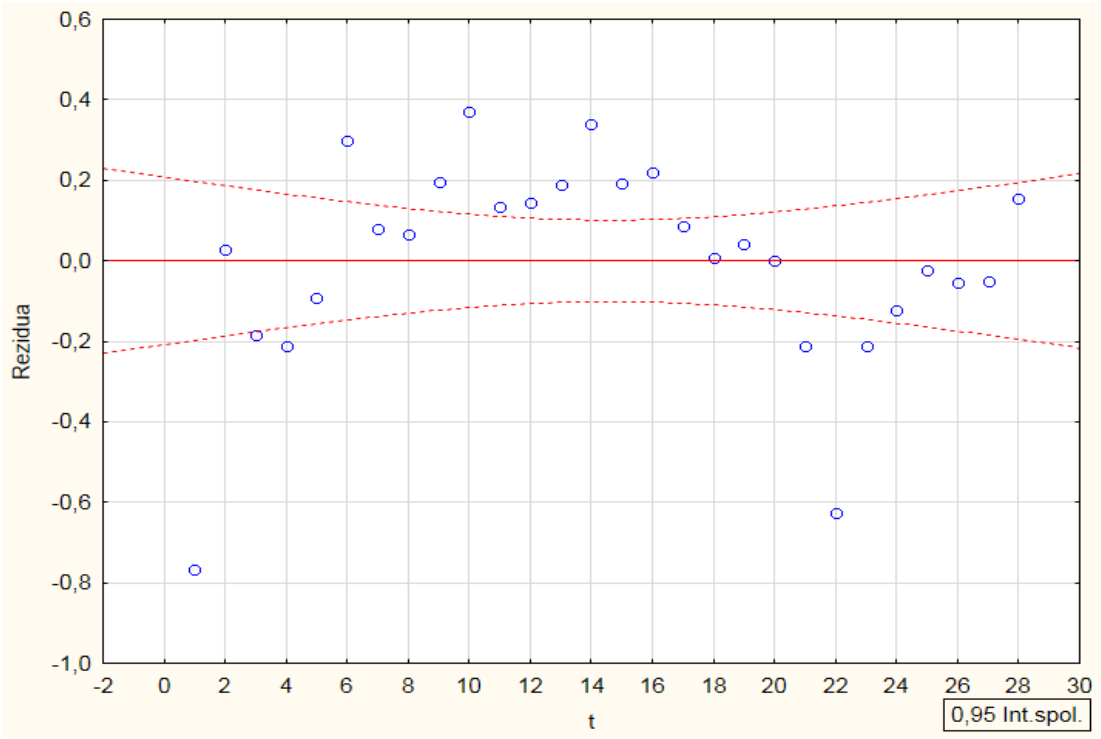
Pro proložení sezónně očištěné časové řady použijeme nejdříve lineární trend.

N=28	Výsledky regrese se závislou proměnnou : nezaměstnanost R= ,98346539 R2= ,96720417 Upravené R2= ,96594279 F(1,26)=766,78 p<0,0000 Směrod. chyba odhadu : ,26015					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(26)	p-hodn.
Abs.člen			15,20673	0,101022	150,5293	0,000000
t	-0,983465	0,035516	-0,16854	0,006086	-27,6909	0,000000

Obr. 14.6 - statistické charakteristiky modelu pro lineární trend (Výstup ze sw Statistica 10)

Lineární trend má statisticky významné odhady obou parametrů na 5% hladině významnosti. Adjustovaný koeficient determinace má hodnotu 96,6 %.

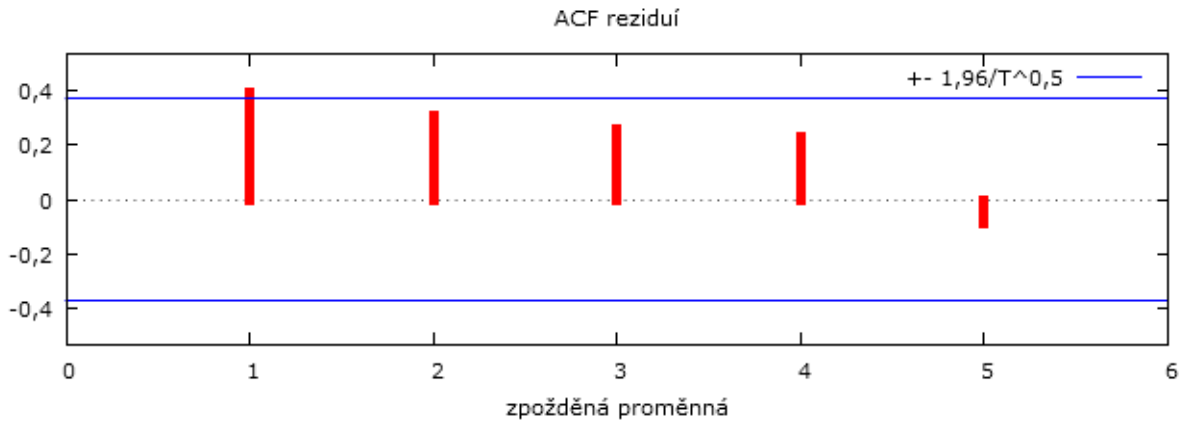
Ověříme případný výskyt autokorelace reziduí pro lineární trend grafickou analýzou grafu reziduí v závislosti na čase a grafickou analýzou reziduální autokorelační funkce.



Obr. 14.7 - graf reziduí v závislosti na čase – lineární trend (Výstup ze sw Statistica 10)

V grafu reziduí v závislosti na čase můžeme sledovat delší stoupající a klesající sekvence reziduí z čehož můžeme usuzovat na pravděpodobný výskyt pozitivní autokorelace v modelu.

O výskytu autokorelace se můžeme přesvědčit také z autokorelační funkce ACF.



Obr. 14.8 - graf autokorelační funkce ACF – lineární trend (Výstup ze sw Gretl)

Z grafu autokorelační funkce ACF vidíme, že první koeficient autokorelace reziduí je na 5% hladině statisticky významný (odhad prvního koeficientu autokorelace r_1 = 0,393888 přesahuje horní mez intervalu spolehlivosti), z čehož vyplývá, že rezidua lineárního trendu vykazují korelační závislost nesystematické složky a lineární trend je tedy nevhodný.

Odstranění problému autokorelace

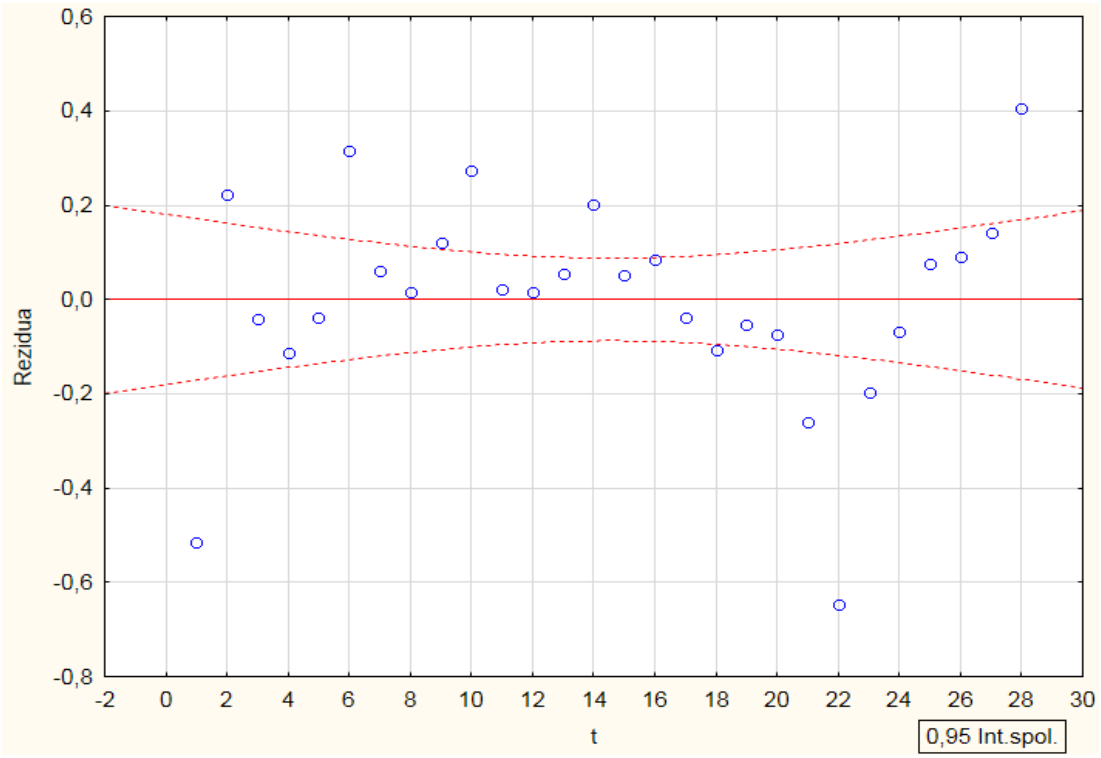
Jak již bylo uvedeno výše, jednou z možností odstranění autokorelace reziduí může být aplikace jiného modelu trendu. Zkusíme tedy aplikovat kvadratický trend a ověřit, zda se míra autokorelace reziduí v modelu snížila.

N=28	Výsledky regrese se závislou proměnnou : nezaměstnanost R= ,98757070 R2= ,97529589 Upravené R2= ,97331956 F(2,25)=493,49 p<0,0000 Směrod. chyba odhadu : ,23026					
	b*	Sm.chyba z b*	b	Sm.chyba z b	t(25)	p-hodn.
Abs.člen			14,89676	0,140458	106,0588	0,000000
t	-0,621709	0,130268	-0,10654	0,022324	-4,7725	0,000067
V1**2	-0,372773	0,130268	-0,00214	0,000747	-2,8616	0,008399

Obr. 14.10 - statistické charakteristiky modelu pro kvadratický trend (Výstup ze sw Statistica 10)

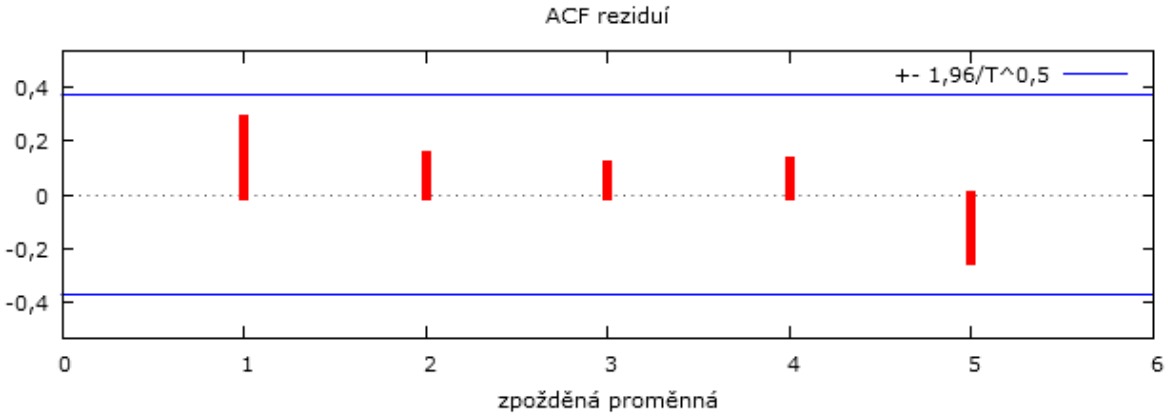
Kvadratický trend má statisticky významné odhady všech parametrů na 5% hladině významnosti. Adjustovaný koeficient determinace má hodnotu 97,33 %.

V grafu reziduí v závislosti na čase můžeme sledovat již méně výrazné stoupající či klesající sekvence reziduí než v případě lineárního trendu.



Obr. 14.11 - graf reziduí v závislosti na čase – kvadratický trend (Výstup ze sw Statistica 10)

O odstranění výskytu autokorelace se můžeme dále přesvědčit z autokorelační funkce ACF.



Obr. 14.12 - graf autokorelační funkce ACF – kvadratický trend (Výstup ze sw Gretl)

Z grafu autokorelační funkce ACF vidíme, že při aplikaci kvadratického trendu žádný z koeficientů autokorelace nepřesahuje meze 95% intervalu spolehlivosti, z čehož vyplývá, že rezidua kvadratického trendu nyní nevykazují dle grafu autokorelační funkce ACF

korelační závislost nesystematické složky a kvadratický trend je vhodnější než trend lineární.

14.3 Závěr

Dalším z neopominutelných předpokladů regresního modelu je nepřítomnost autokorelace reziduí, kterou je možno definovat jako porušení předpokladu o vzájemné nezávislosti náhodných složek z různých pozorování. Odstranění nebo alespoň zlepšení autokorelace reziduí v modelu je možno provést volbou vhodnějšího regresního trendu.

15 Výskyt odlehlých hodnot v modelu

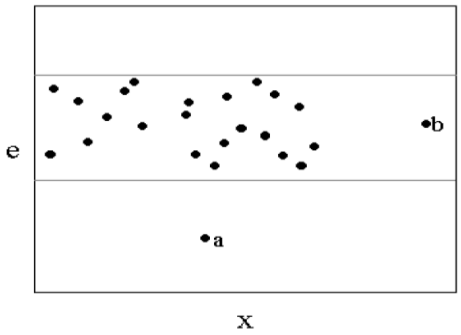
15.1 Teoretická část

Při posuzování kvality modelu je nutné pečlivě sledovat výskyt tzv. odlehlých bodů. Odlehlý bod je takový, který leží mimo základní konfiguraci bodů v grafu. I jedna jediná odlehlá hodnota (v případě nevelkého výběrového souboru), může způsobit problém při odhadu regresních parametrů a tedy i výpočtu předpovědí na základě modelu. Odlehlý bod nazýváme vlivný, pokud se po jeho odstranění podstatně změní poloha regresní přímky. (14)

Podle toho, kde se vlivné body vyskytují, lze provést dělení na:

1. *Vybočující pozorování* (outliers), které se liší v hodnotách vysvětlované (závisle) proměnné *y* od ostatních

2. E*xtrémy* (extrems), které se liší v hodnotách vysvětlujících (nezávisle) proměnných *x*. (14)



Obr 15.1: Ukázka vybočujícího pozorování (bod a) a extrému (bod b)

Důsledky výskytu odlehlých hodnot v modelu

- Zkreslení odhadů parametrů a příslušných intervalů spolehlivosti
- Snížení hodnoty koeficientu determinace R^2
- Zkreslené a nereálné predikce při použití regresního modelu

Identifikace výskytu odlehlých hodnot v modelu

- Na základě grafického znázornění
- Použití tabulky Cookových vzdáleností
- Použití jiných měr odlehlosti sledovaného bodu (např. Standardizovaná rezidua)

Odstranění problému výskytu odlehlých hodnot v modelu

- Prověření hodnoty dané proměnné, zdali při jejím přepisu nedošlo k chybě
- Statistická transformace dané proměnné
- Odstranění případu s odlehlou hodnotou
- Vymazání celé proměnné obsahující odlehlou hodnotu

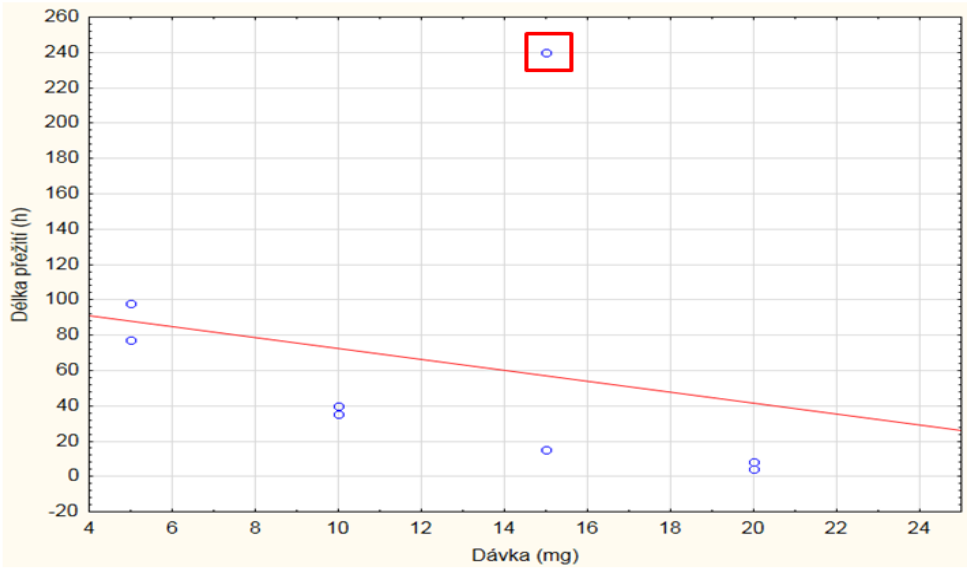
15.2 Praktická část

Příklad 15.1: V rámci laboratorního výzkumu účinnosti rodenticid byla sledována závislost délky přežití divokých potkanů na množství zkonzumované hubící látky. Mezi 8 sledovanými jedinci se objevil výjimečný jedinec, velmi odolný proti působení aplikované látky (řádek 6 – přežití 240 h při požití 15 mg látky). Tato situace se u divokých potkanů na rozdíl od geneticky homogenních laboratorních potkanů, kde je rozptyl délek přežití malý, občas vyskytuje. Na tomto příkladu si prakticky demonstrujeme, jak může jedna jediná odlehlá hodnota ovlivnit výsledky regresního modelu.

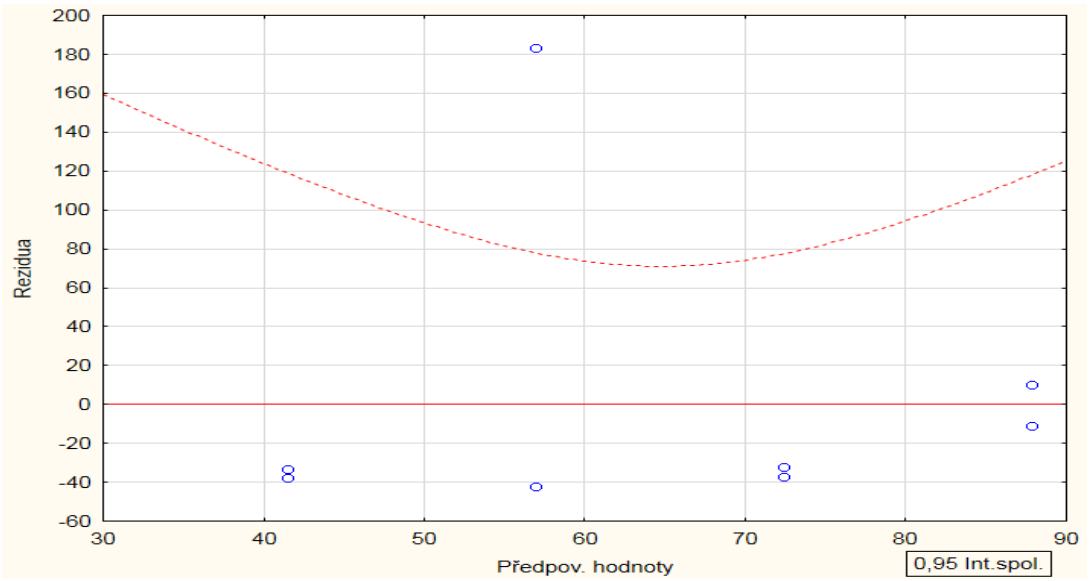
	1 Dávka (mg)	2 Délka přežití (h)
1	5	98
2	5	77
3	10	40
4	10	35
5	15	15
6	15	240
7	20	8
8	20	4

Obr 15.2 - data pro závislost délky přežití divokých potkanů (h) na množství zkonzumované hubící látky (mg). Odlehlá hodnota se nachází v řádku 6.

Již při prvotním seznámení s daty vidíme podezřele vysokou hodnotu v 6. řádku datové matice. Rozhodnutí, zda budeme danou hodnotu považovat za odlehlou, provedeme na základě grafické analýzy dat, grafické analýzy reziduí a tabulky Cookových vzdáleností.



Obr 15.3 - identifikace odlehlé hodnoty z řádku 6 na základě grafické analýzy (Výstup ze sw Statistica 10)



Obr 15.4 - identifikace odlehlé hodnoty z řádku 6 na základě grafické analýzy reziduí (Výstup ze sw Statistica 10)

Případ	Cookovy vzdálen. ,006 . . . . . ,516	Cookovy vzdál. Setříděno						
		Pozorovaná hodnota	Předpovězená hodnota	Reziduum	Stand. Rezid.	Mahalanobis. vzdálen.	Odstran. rezidua	Cookova vzdálen.
6	. . . | . . *	240,0000	56,90000	183,1000	2,229590	0,175000	215,4118	0,516028
8	. * . | . . .	4,0000	41,45000	-37,4500	-0,456025	1,575000	-57,6154	0,086137
7	. *. | . . .	8,0000	41,45000	-33,4500	-0,407317	1,575000	-51,4615	0,068719
5	. * . | . . .	15,0000	56,90000	-41,9000	-0,510212	0,175000	-49,2941	0,027022
4	.* . | . . .	35,0000	72,35000	-37,3500	-0,454807	0,175000	-43,9412	0,021472
3	.* . | . . .	40,0000	72,35000	-32,3500	-0,393923	0,175000	-38,0588	0,016108
2	.* . | . . .	77,0000	87,80000	-10,8000	-0,131511	1,575000	-16,6154	0,007164
1	.* . | . . .	98,0000	87,80000	10,2000	0,124204	1,575000	15,6923	0,006390
Minimum	.* . | . . .	4,0000	41,45000	-41,9000	-0,510212	0,175000	-57,6154	0,006390
Maximum	. . . | . . *	240,0000	87,80000	183,1000	2,229590	1,575000	215,4118	0,516028
Průměr	. * . | . . .	64,6250	64,62500	0,0000	-0,000000	0,875000	-3,2353	0,093630
Medián	. * . | . . .	37,5000	64,62500	-32,9000	-0,400620	0,875000	-41,0000	0,024247

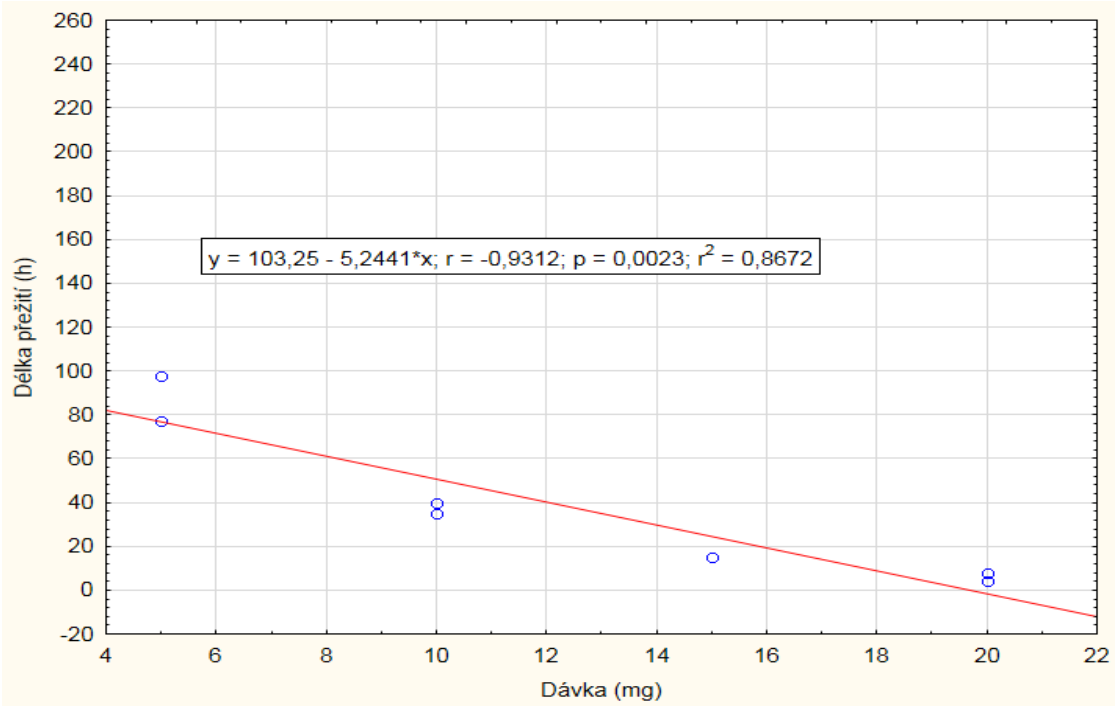
Obr 15.5 - identifikace odlehlého pozorování z řádku 6 na základě tabulky Cookových vzdáleností (Výstup ze sw Statistica 10)

Cookova vzdálenost je často užívanou mírou váhy sledovaného bodu při výpočtu MNČ a používá se pro identifikaci odlehlých pozorování. Čím je tato hodnota vyšší, tím více daný bod ovlivňuje výslednou hodnotu odhadnutých regresních parametrů. Z tabulky Cookových vzdálenosti vidíme, že Cookova vzdálenost pro hodnotu z 6. řádku významně překračuje hodnoty vzdáleností ostatních bodů.

Pro demonstraci míry ovlivnění výsledků regresního modelu jednou odlehlou hodnotou si nyní vytvoříme 2 regresní modely:

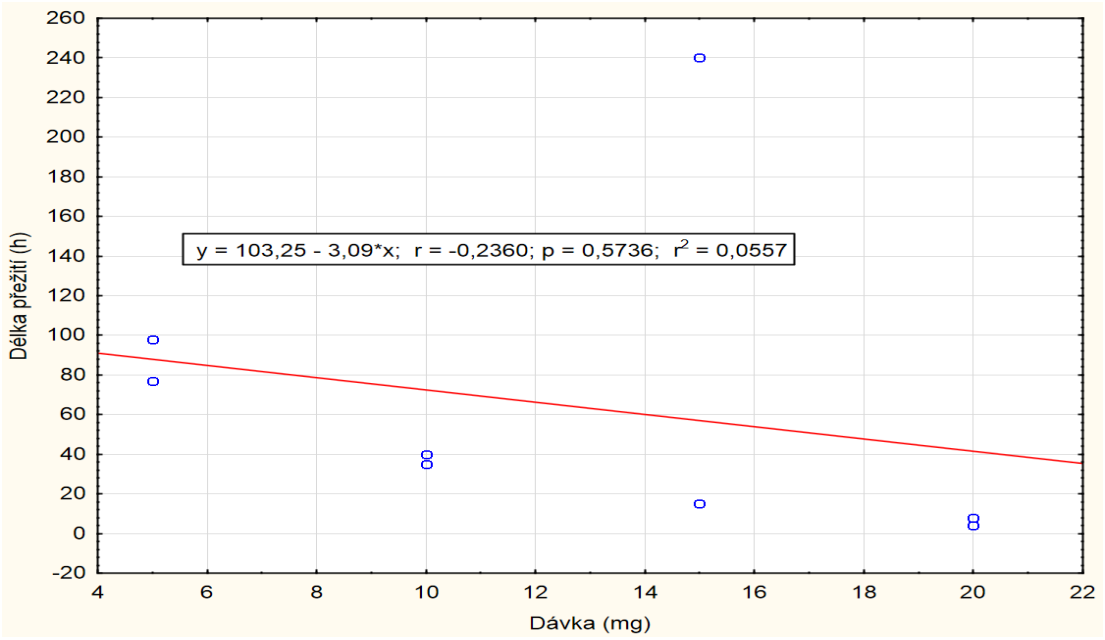
- model bez sledované odlehlé hodnoty (Obr. 15.6)
- model se zahrnutím sledované odlehlé hodnoty (Obr. 15.7)

Z modelu bez sledované odlehlé hodnoty vidíme, že regresní rovnice má tvar y = 103,25-5,2441.x a koeficient determinace je roven R^2 = 86,72 %.



Obr. 15.6 - lineární regresní model pro závislost délky přežití divokých potkanů (h) na množství zkonzumované hubící látky (mg) - bez zahrnutí sledované odlehlé hodnoty (Výstup ze sw Statistica 10)

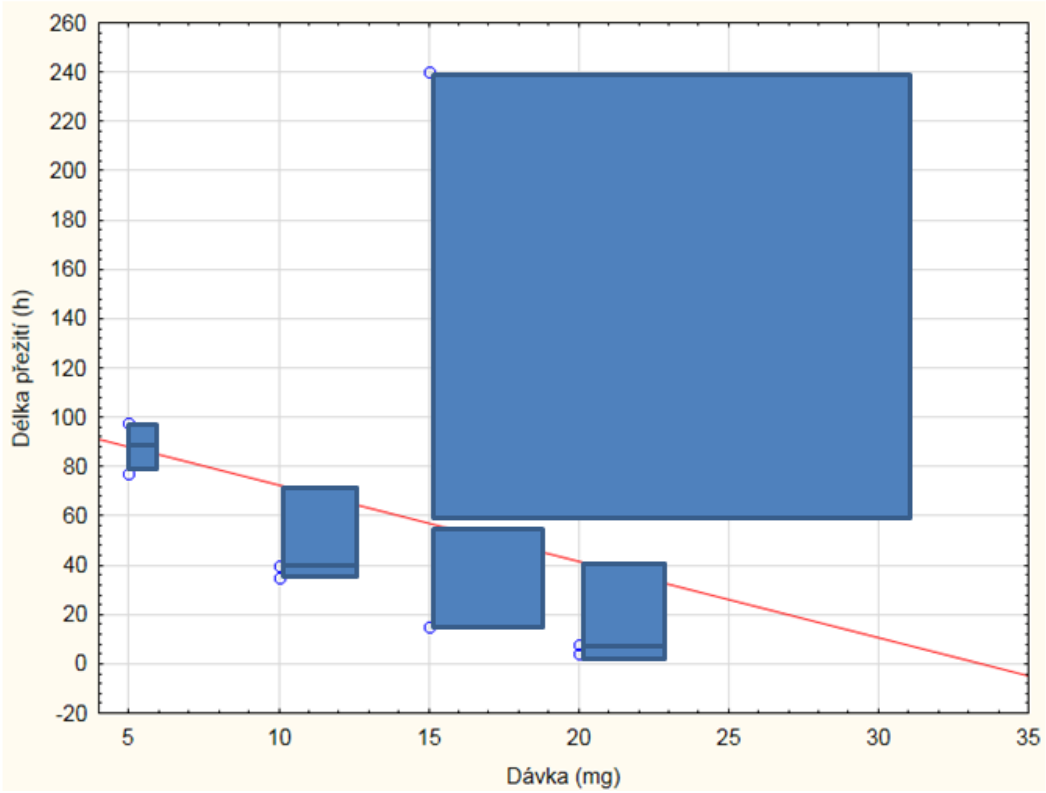
Z modelu se zahrnutím sledované odlehlé hodnoty vidíme, že regresní rovnice má tvar y = 103,25-3,09.x a koeficient determinace je roven pouze R^2 = 5,57 %.



Obr. 15.7 - lineární regresní model pro závislost délky přežití divokých potkanů (h) na množství zkonzumované hubící látky (mg) - se zahrnutím sledované odlehlé hodnoty (Výstup ze sw Statistica 10)

Při srovnání obou modelů vidíme, že rovnice regresní přímky má při zahrnutí sledované odlehlé hodnoty velmi odlišný tvar od původního modelu. Zejména odhad parametru b_1 (směrnice přímky) je velmi rozdílný. Také koeficient determinace R^2 je v případě modelu se zahrnutím sledované odlehlé hodnoty podstatně nižší.

Důvodem velmi odlišného odhadu parametrů regresního modelu při výskytu jedné jediné odlehlé hodnoty je velmi zásadní přínos reziduálního čtverce této odlehlé hodnoty při minimalizaci součtu čtverců v rámci Metody nejmenších čtverců, která je základem pro výpočet regresních parametrů modelu (viz Kapitola 1). Přínos reziduálního čtverce odlehlé hodnoty je možno demonstrovat na Obr. 15.8.



Obr. 15.8 - přínos reziduálního čtverce odlehlé hodnoty v modelu závislosti délky přežití divokých potkanů na množství zkonzumované hubící látky (upravený výstup ze sw Statistica 10)

Finální dopad zahrnutí odlehlé hodnoty do modelu si můžeme demonstrovat na příkladu predikce délky přežití jedince při množství zkonzumované hubící látky např. 18 mg:

V případě modelu bez sledované odlehlé hodnoty je výsledná odhadovaná hodnota přežití y = 8,86 dní. V případě modelu se zahrnutím jedné jediné odlehlé hodnoty je výsledná (velmi nereálná) odhadovaná hodnota přežití y = 47,63 dní.

15.3 Závěr

Při tvorbě regresního modelu je velmi důležité identifikovat případné odlehlé body. Odlehlý bod může v případě nevelkého souboru způsobit problém při odhadu regresních parametrů a výpočtu předpovědí na základě modelu. Nejčastějším způsobem vyřešení problému je odstranění dané odlehlé hodnoty či transformace proměnné s danou odlehlou hodnotou.

16 Nesplnění předpokladu normality reziduí

16.1 Teoretická část

Dalším z důležitých předpokladů regresního modelu je normalita reziduí. Podstatou předpokladu je, že rezidua e_i získaná ze vzorce $e_i = y_i - y_{iTEOR}$, by měla splňovat předpoklad normálního rozdělení.

Příčiny vzniku nenormality reziduí

Nenormalita reziduí může být způsobena různými příčinami. Jednou z příčin je přímo nenormalita vstupní vysvětlující či vysvětlované proměnné. Nenormalita chyb je však také často důsledkem porušeného jiného z předpokladů regresního modelu. Před samotným ověřováním normality je tedy žádoucí zajistit splnění ostatních předpokladů regresního modelu. (4)

Důsledky nenormality reziduí

- **Důsledky pro bodové odhady regresních koeficientů** - pokud potřebujeme pouze odhad parametrů regresní funkce, není třeba předpoklad normality ověřovat. Odhady regresních koeficientů metodou nejmenších čtverců zůstávají při splnění ostatních předpokladů nestrannými a konzistentními bez ohledu na rozdělení chybových členů. (4)
- **Důsledky pro metody statistického úsudku** - normalitu potřebujeme k platnosti metod statistického úsudku (t-testy, predikční a konfidenční interval, F test). Pracujeme-li však s daty o vyšším rozsahu počtu pozorování, zůstávají díky centrální limitní větě zmíněné metody statistického úsudku v platnosti. (4)

Identifikace přítomnosti nenormality reziduí

Použití testu

Předpoklad normality reziduí je možno ověřit pomocí vhodného testu normality aplikovaného na získaná rezidua e_i. Jednou z možností je často používaný Shapiro-Wilkův test normality. Jak již bylo uvedeno dříve v této práci, v případě velkého

počtu dat v modelu může být u Shapiro-Wilkova testu považována i velmi malá odchylka od normálního rozdělení za významnou. Závěr o výskytu normality reziduí je potom nesprávný.

Grafická analýza

Při kontrole splnění předpokladu normality reziduí je tedy vhodnější použít grafické znázornění pomocí normálního p-grafu reziduí, případně pomocí histogramu reziduí.

Grafická analýza pomocí normálního p-grafu reziduí

Podstatou normálního p-grafu reziduí je grafické znázornění kvantilů analyzovaných dat (v tomto případě reziduí) proti kvantilům předpokládaného normovaného normálního rozdělení. Pokud mají rezidua přibližně normální rozdělení, nebudou se významně odchylovat od čáry v grafu. V normálním p-grafu reziduí je také možné dobře identifikovat zjevné odlehlé hodnoty.

Grafická analýza pomocí histogramu reziduí

Druhou možností pro grafické posouzení případné nenormality reziduí je histogram reziduí. Pokud mají rezidua přibližně normální rozdělení, měl by tvar histogramu odpovídat tvaru Gaussovy křivky.

16.2 Praktická část

Demonstrace vlivu porušení některých předpokladů regresního modelu na normalitu reziduí

Pro demonstraci vlivu porušení některých předpokladů regresního modelu na normalitu reziduí modelu bylo z práce vybráno pět, v předchozích kapitolách již uvedených, příkladů s různými typy porušení předpokladů regresního modelu.

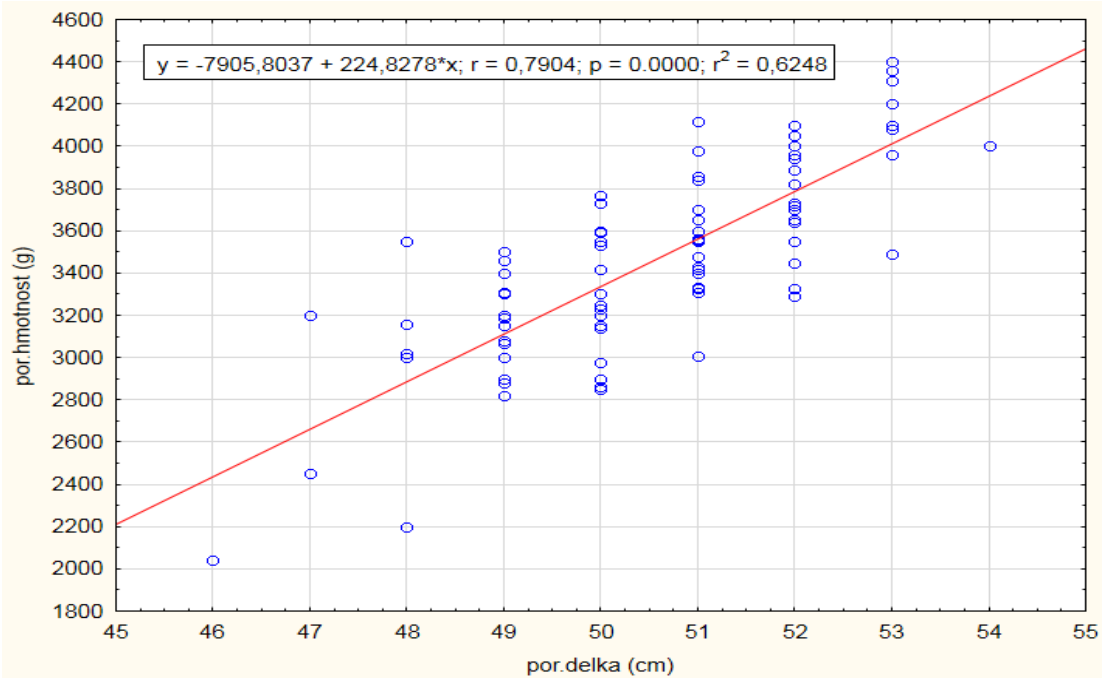
- Všechny předpoklady splněny – Závislost porodní hmotnosti na porodní délce
- Nevhodný tvar funkční závislosti – Závislost tržeb na investicích do reklamy
- Výskyt heteroskedasticity v modelu - Závislost měsíčního příjmu na věku
- Výskyt autokorelace v modelu - Závislost nezaměstnanosti na vývoji v čase

- Výskyt extrémně odlehlé hodnoty - Závislost délky přežití potkanů na aplikované dávce rodenticid

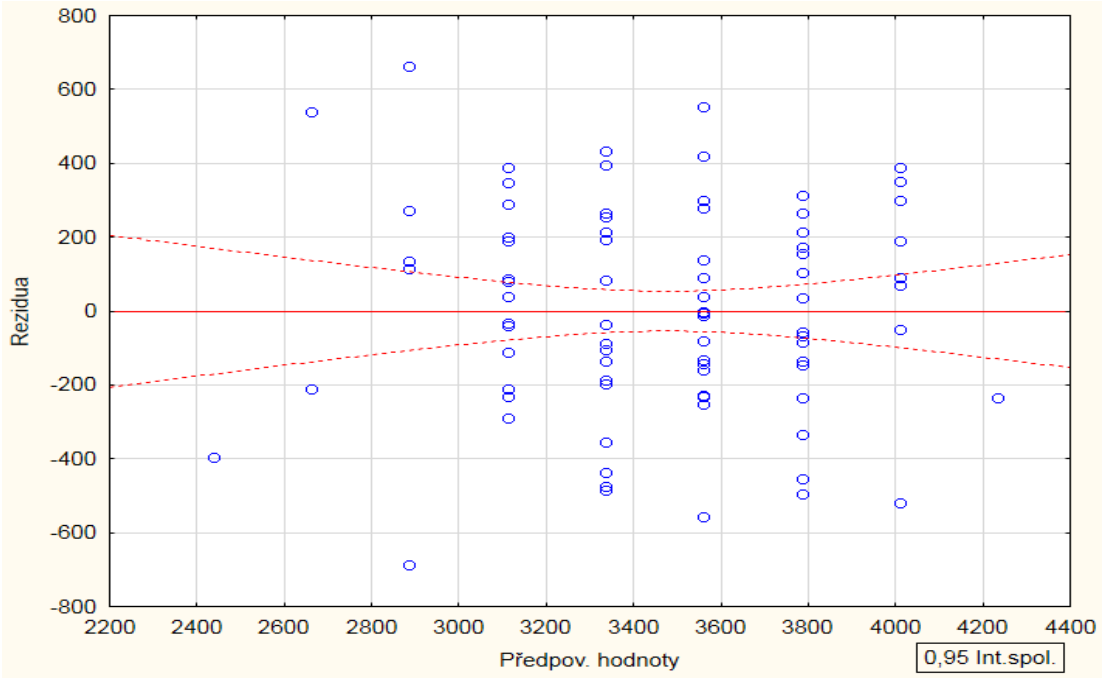
Na těchto konkrétních příkladech bylo následně demonstrováno, jak se porušení daného předpokladu promítne do grafů pro identifikaci nenormality reziduí (normální p-graf reziduí, histogram reziduí). Pro každý příklad porušení předpokladu jsou vždy uvedeny komparativně 4 grafy:

- Bodový graf závislosti y na x, znázorňující průběh dané závislosti
- Graf reziduí proti předpovězeným hodnotám, ukazující na (ne)splnění daného předpokladu
- Normální p-graf reziduí pro posouzení normality reziduí
- Histogram pro posouzení normality reziduí

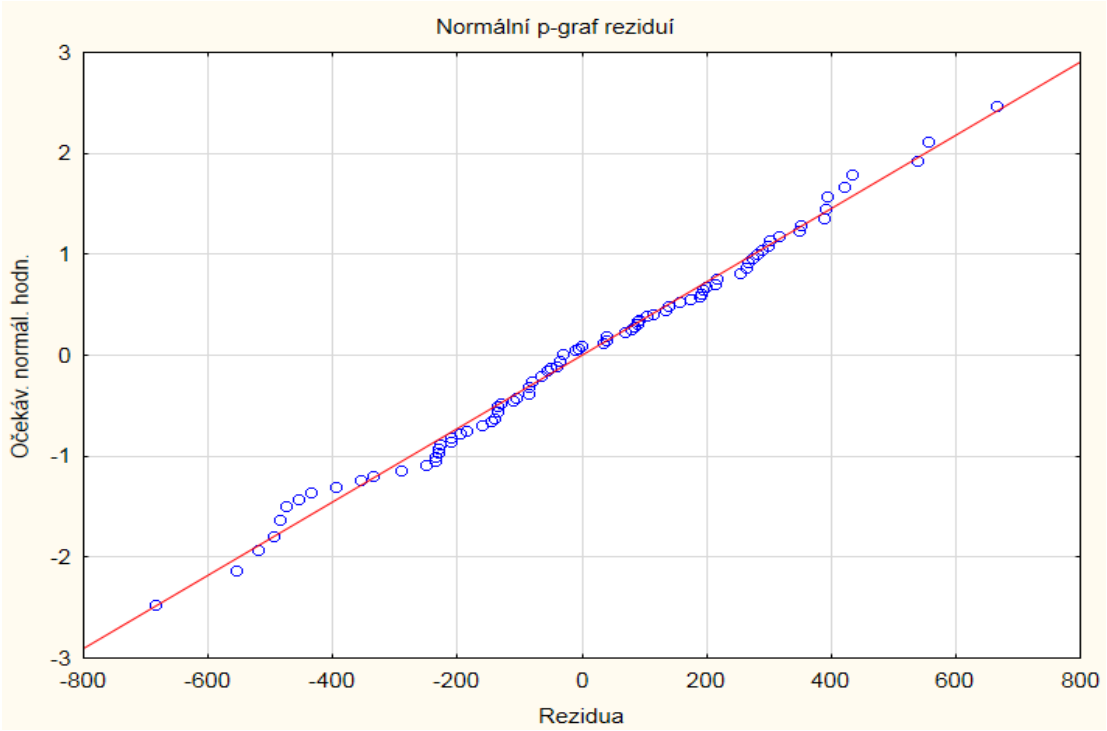
Příklad 16.1: Všechny předpoklady splněny – Závislost porodní hmotnosti na porodní délce



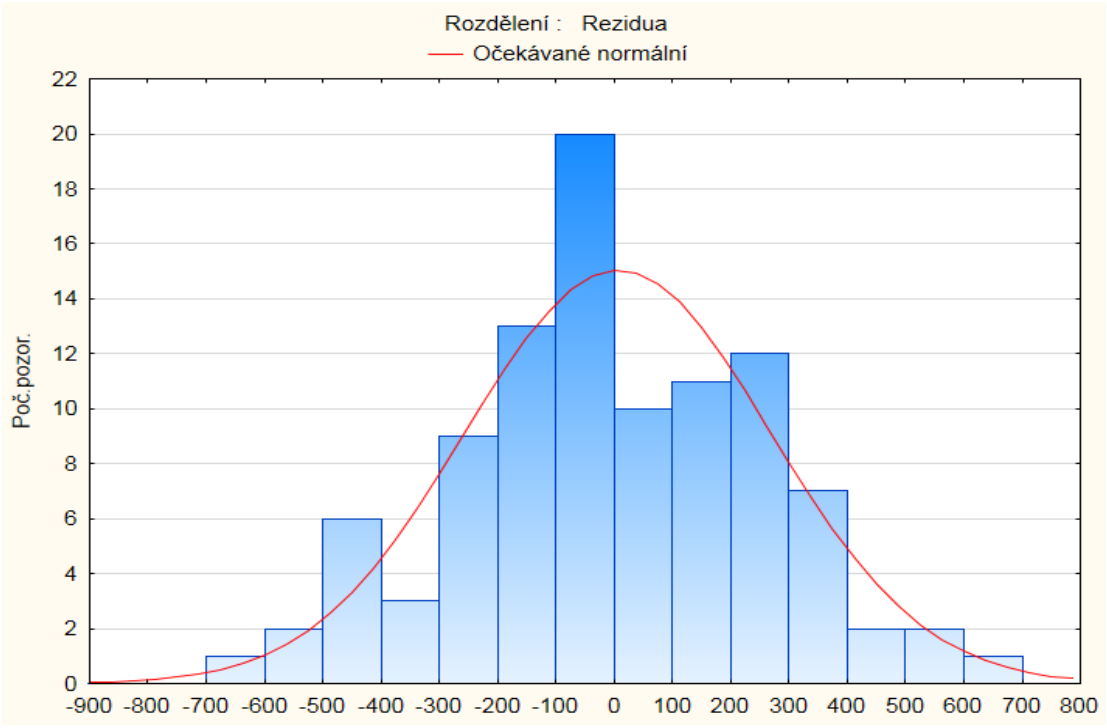
Obr. 16.1 - bodový graf závislosti y na x, znázorňující danou závislost - Všechny předpoklady splněny (Výstup ze sw Statistica 10)



Obr. 16.2 - graf reziduí proti předpovězeným hodnotám - Všechny předpoklady splněny (Výstup ze sw Statistica 10)

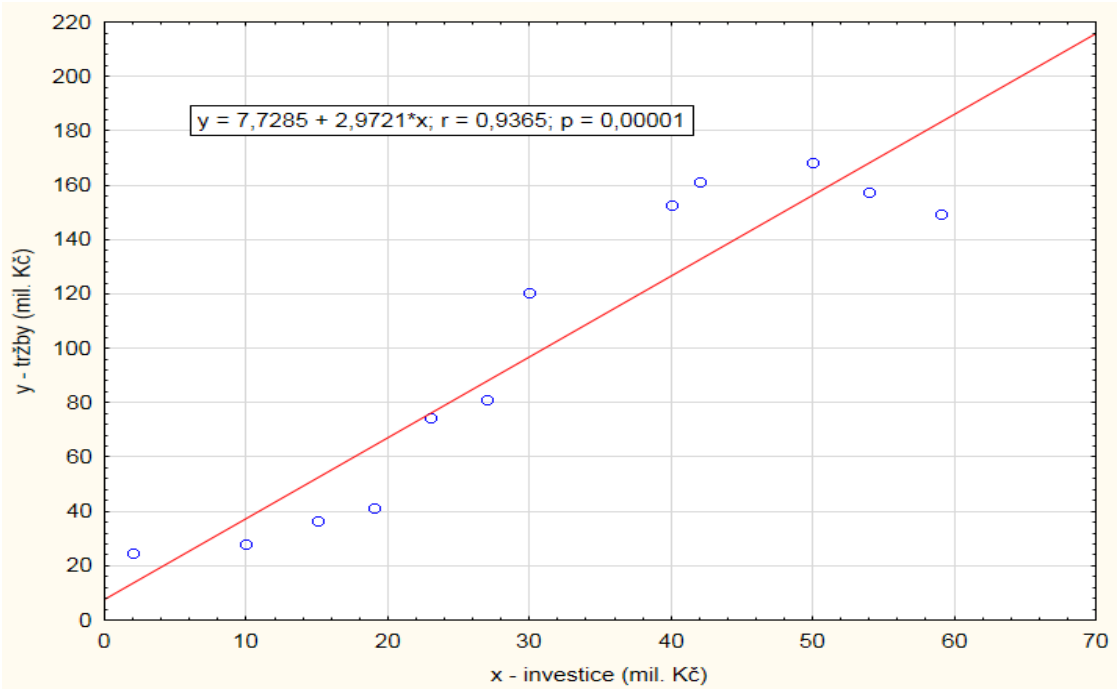


Obr. 16.3 - normální p-graf reziduí pro posouzení normality reziduí - Všechny předpoklady splněny (Výstup ze sw Statistica 10)

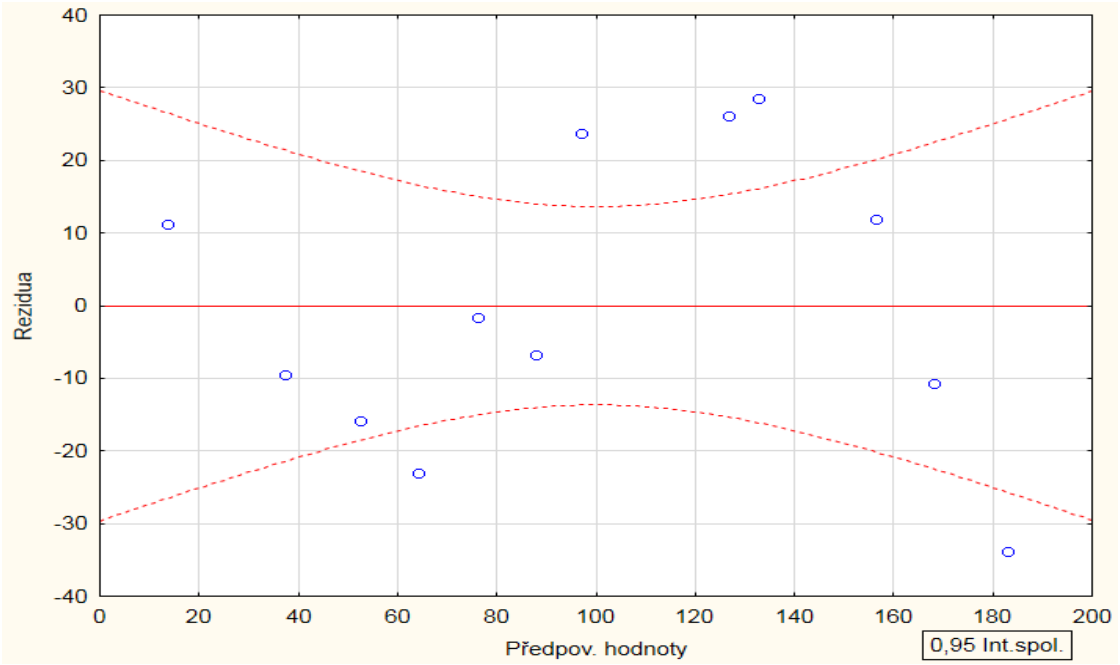


Obr. 16.4 - histogram pro posouzení normality reziduí - Všechny předpoklady splněny (Výstup ze sw Statistica 10)

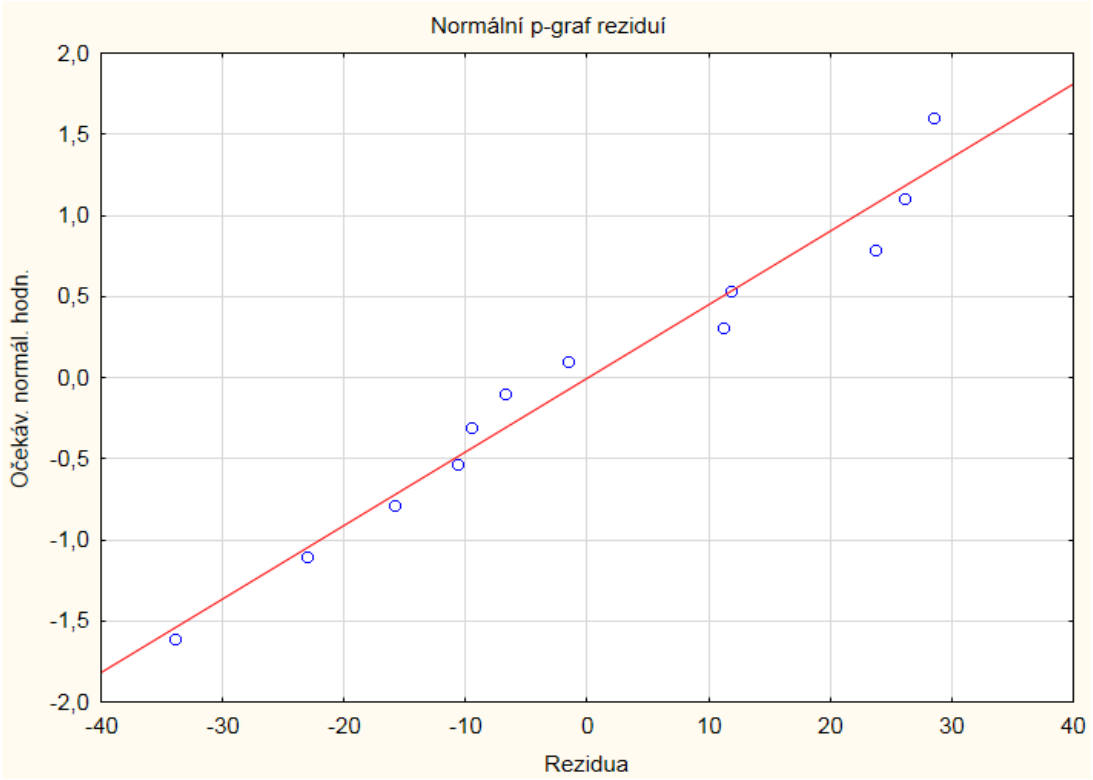
Příklad 16.2: Nevhodný tvar funkční závislosti – Závislost tržeb na investicích do reklamy



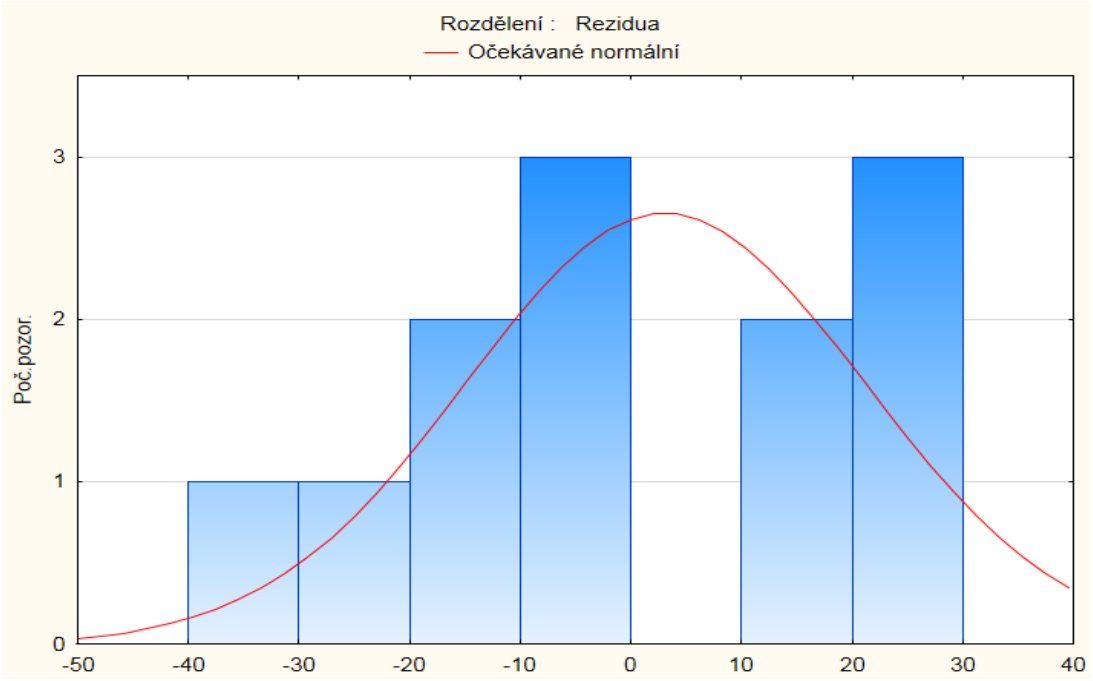
Obr. 16.5 - bodový graf závislosti y na x, znázorňující danou závislost - Nevhodný tvar funkční závislosti (Výstup ze sw Statistica 10)



Obr. 16.6 - graf reziduí proti předpovězeným hodnotám - Nevhodný tvar funkční závislosti (Výstup ze sw Statistica 10)

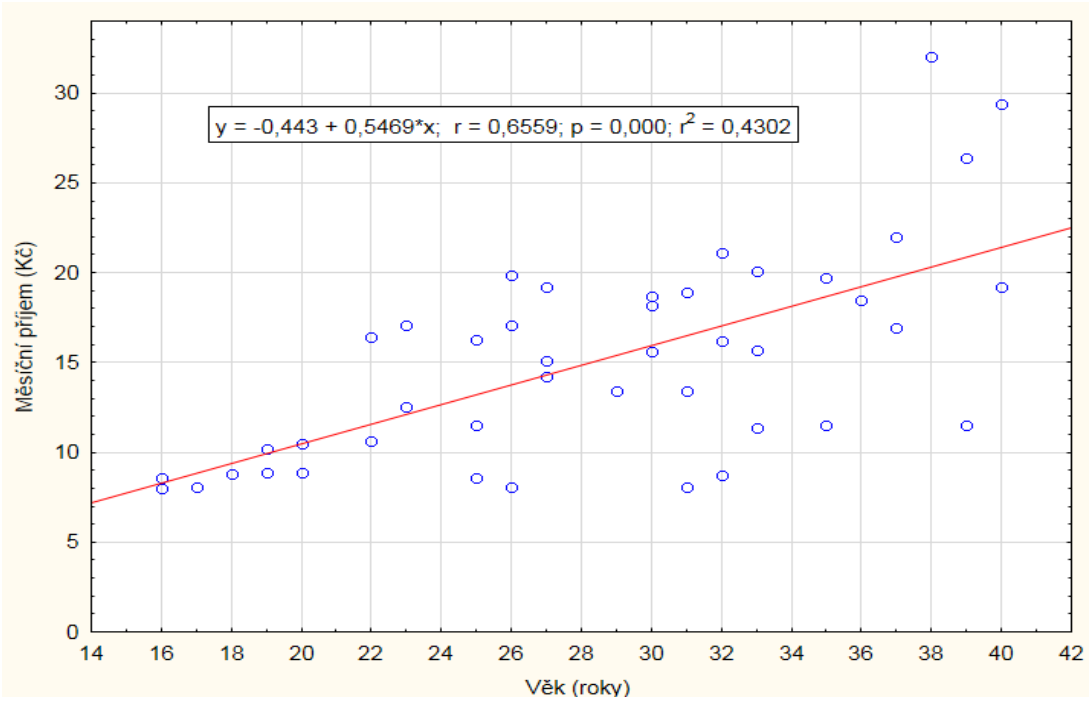


Obr. 16.7 - normální p-graf reziduí pro posouzení normality reziduí – Nevhodný tvar funkční závislosti (Výstup ze sw Statistica 10)

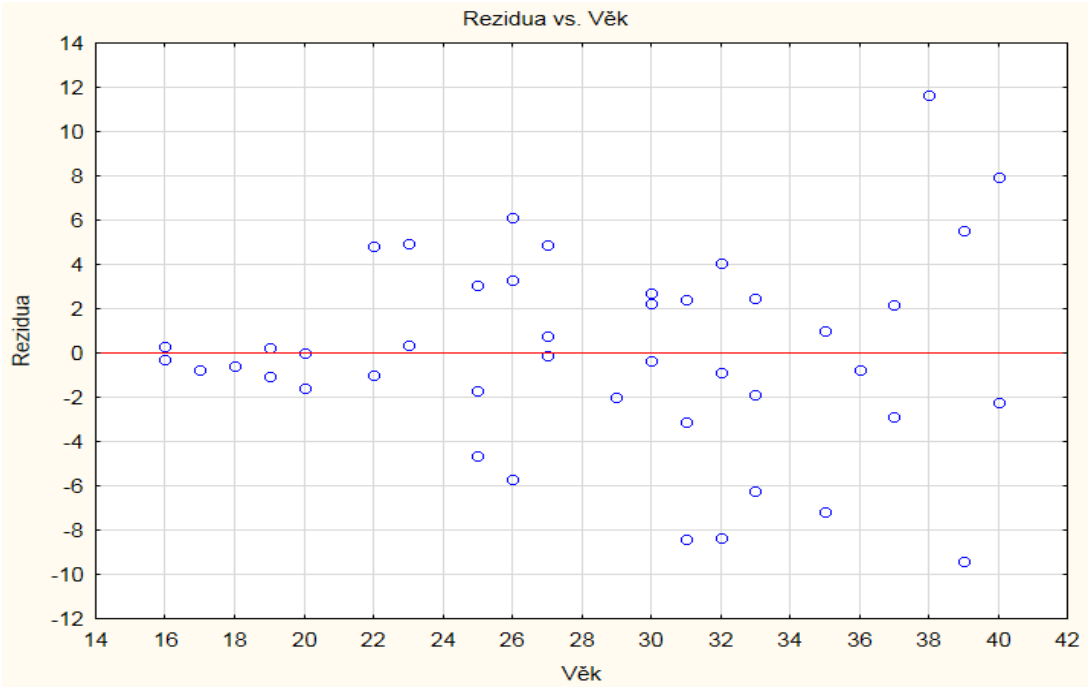


Obr. 16.8 - histogram pro posouzení normality reziduí - Nevhodný tvar funkční závislosti (Výstup ze sw Statistica 10)

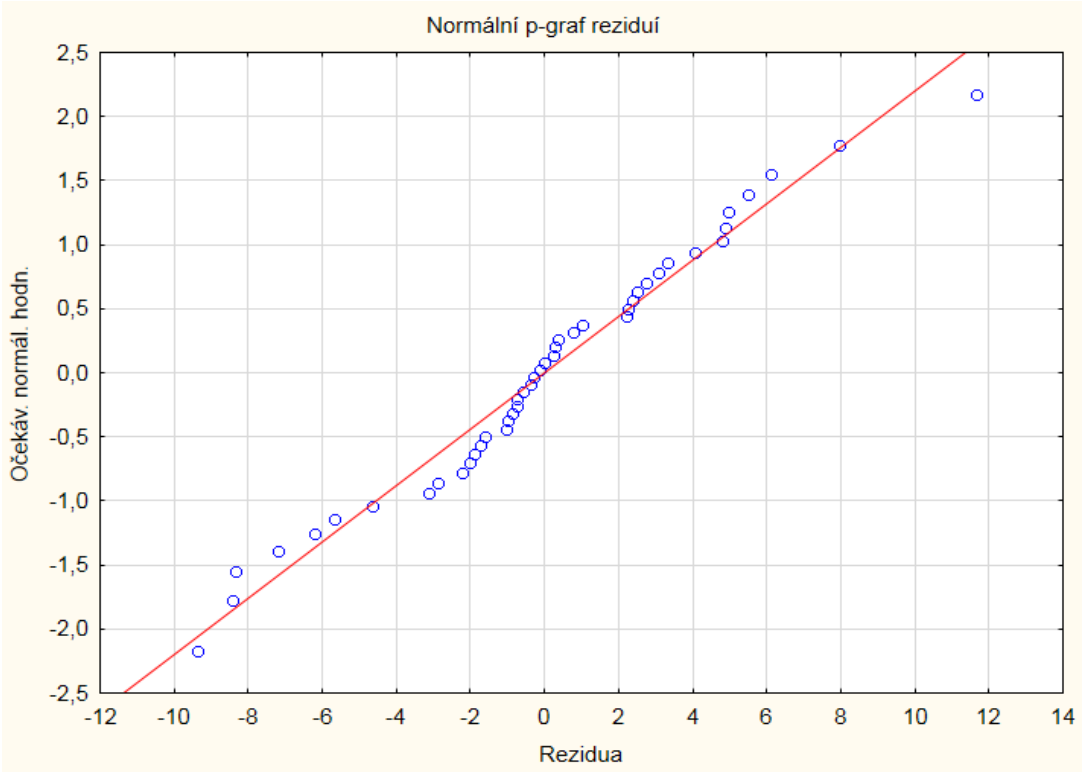
Příklad 16.3: Výskyt heteroskedasticity v modelu - Závislost měsíčního příjmu na věku



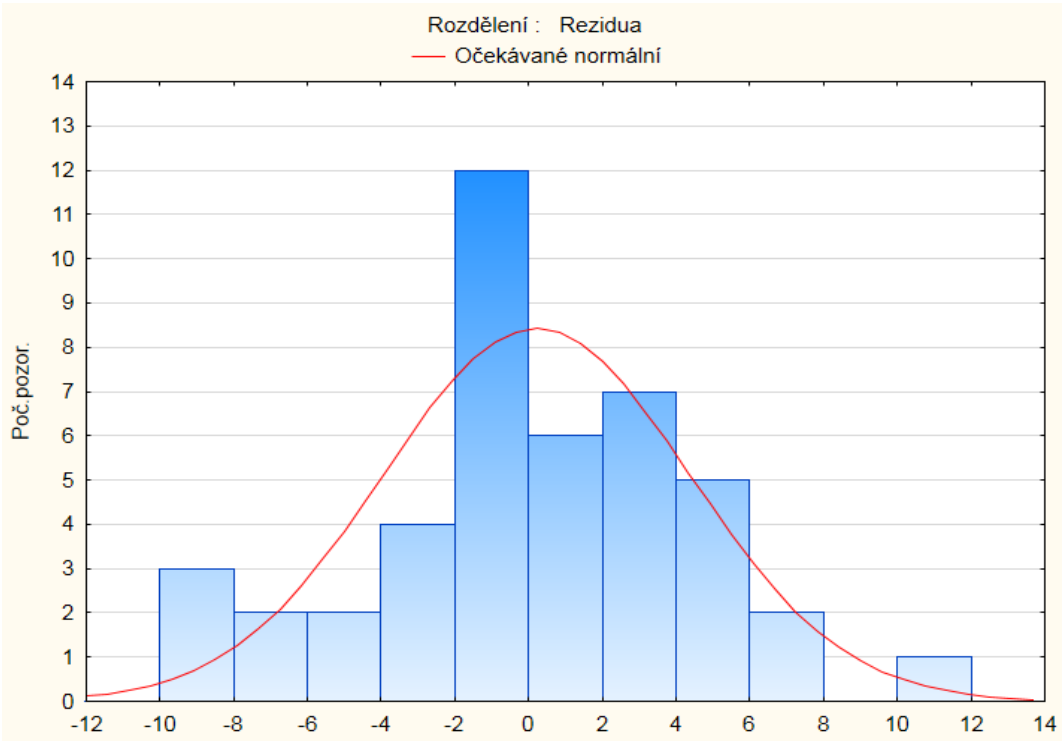
Obr. 16.9 - bodový graf závislosti y na x, znázorňující danou závislost – Výskyt heteroskedasticity v modelu (Výstup ze sw Statistica 10)



Obr. 16.10 - graf reziduí proti předpovězeným hodnotám - Výskyt heteroskedasticity v modelu (Výstup ze sw Statistica 10)

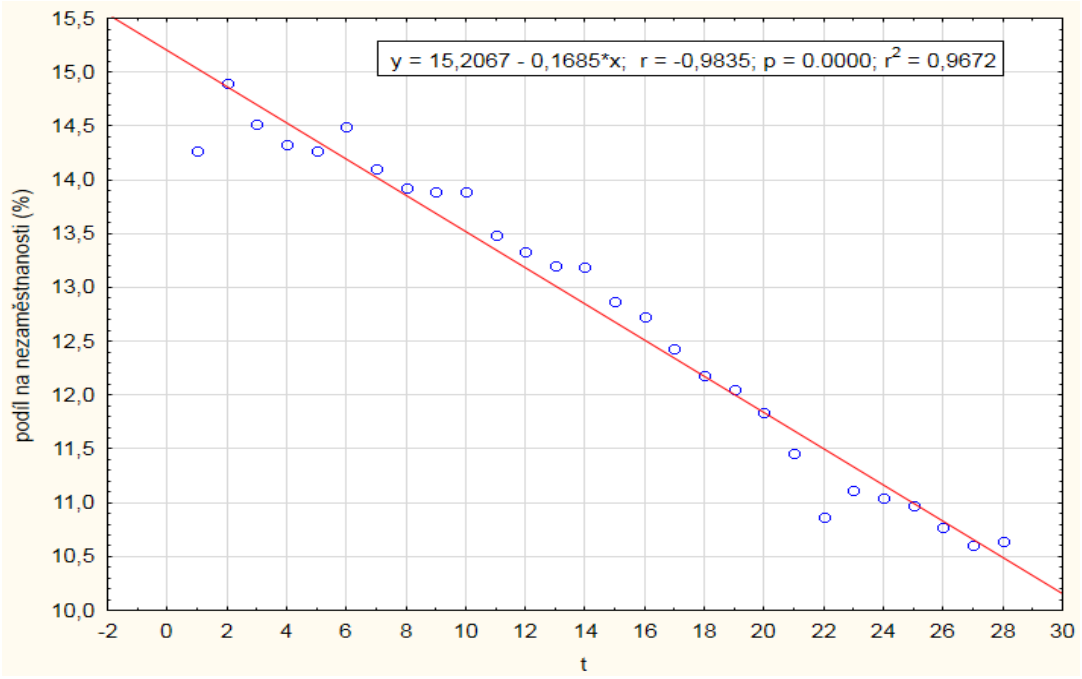


Obr. 16.11 - normální p-graf reziduí pro posouzení normality reziduí – Výskyt heteroskedasticity v modelu (Výstup ze sw Statistica 10)

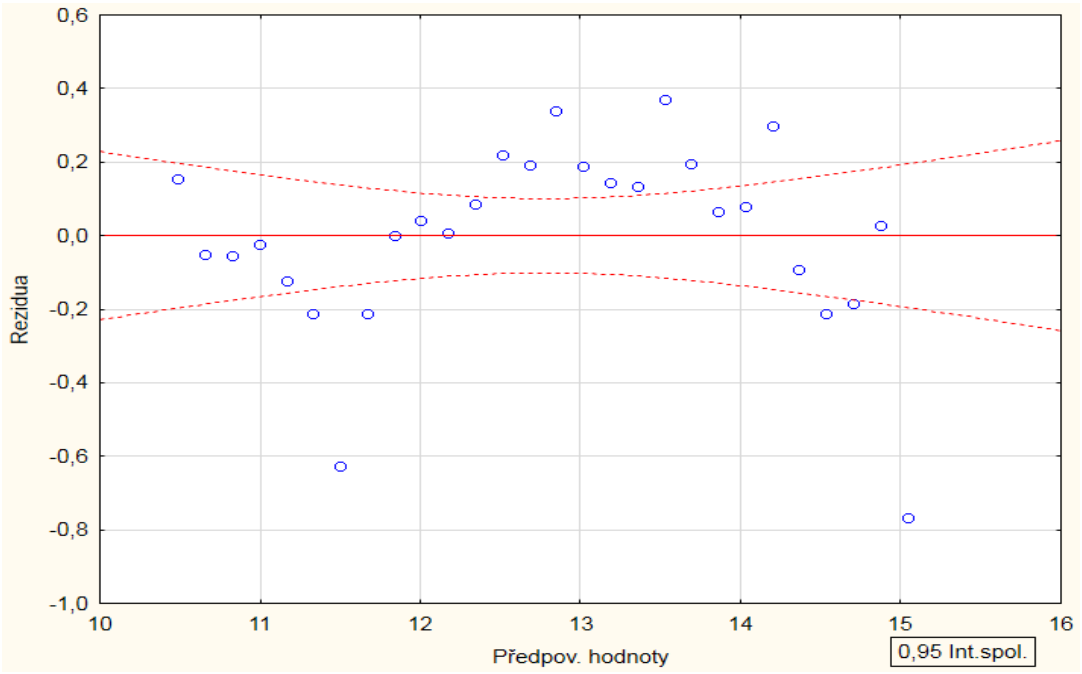


Obr. 16.12 - histogram pro posouzení normality reziduí - Výskyt heteroskedasticity v modelu (Výstup ze sw Statistica 10)

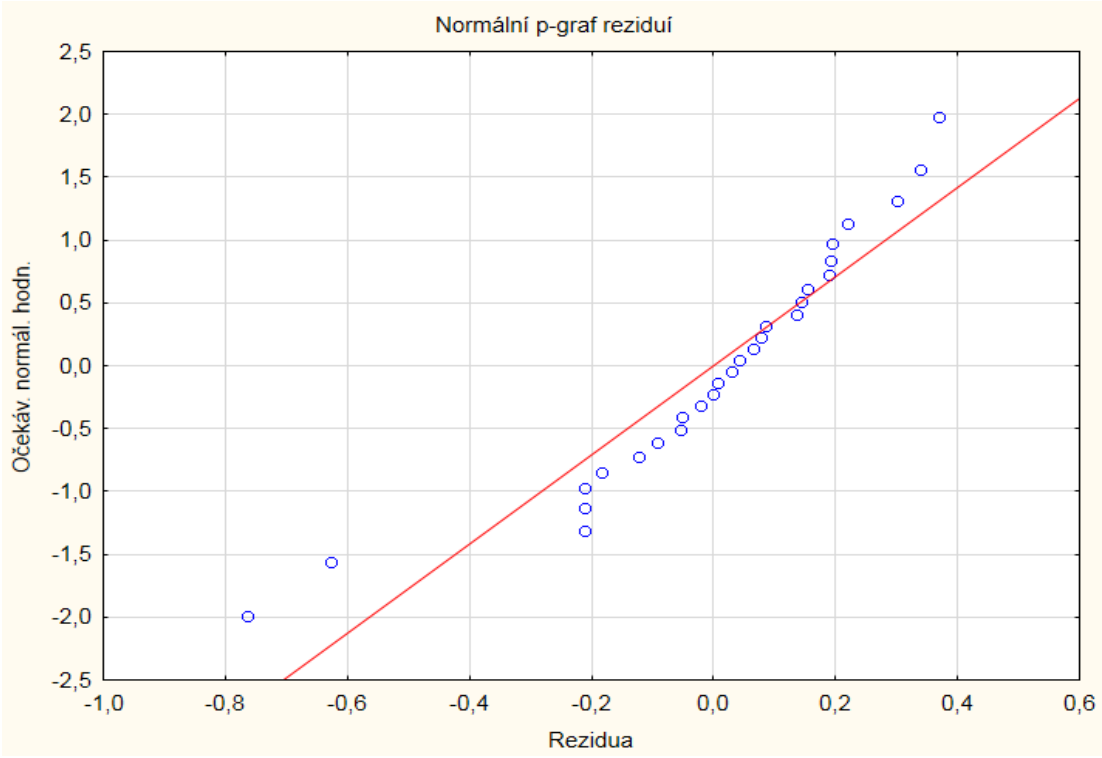
Příklad 16.4: Výskyt autokorelace v modelu - Závislost nezaměstnanosti na vývoji v čase



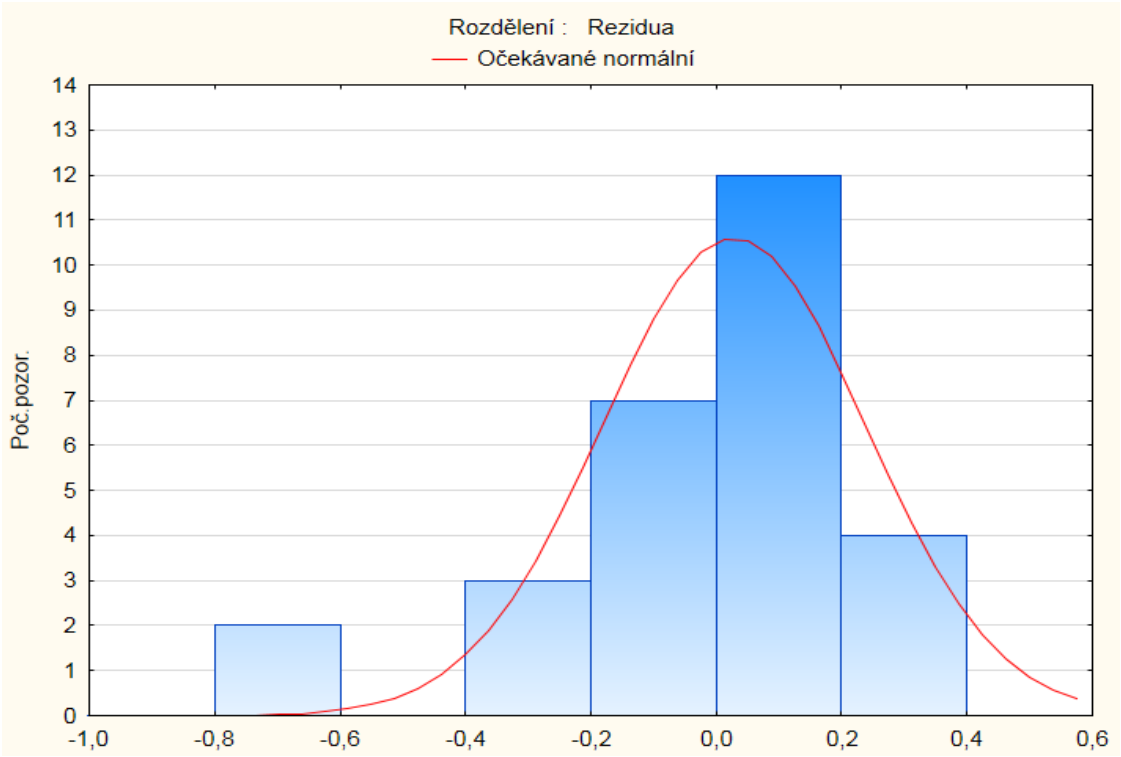
Obr. 16.13 - bodový graf závislosti y na x, znázorňující danou závislost – Výskyt autokorelace v modelu (Výstup ze sw Statistica 10)



Obr. 16.14 - graf reziduí proti předpovězeným hodnotám - Výskyt autokorelace v modelu (Výstup ze sw Statistica 10)

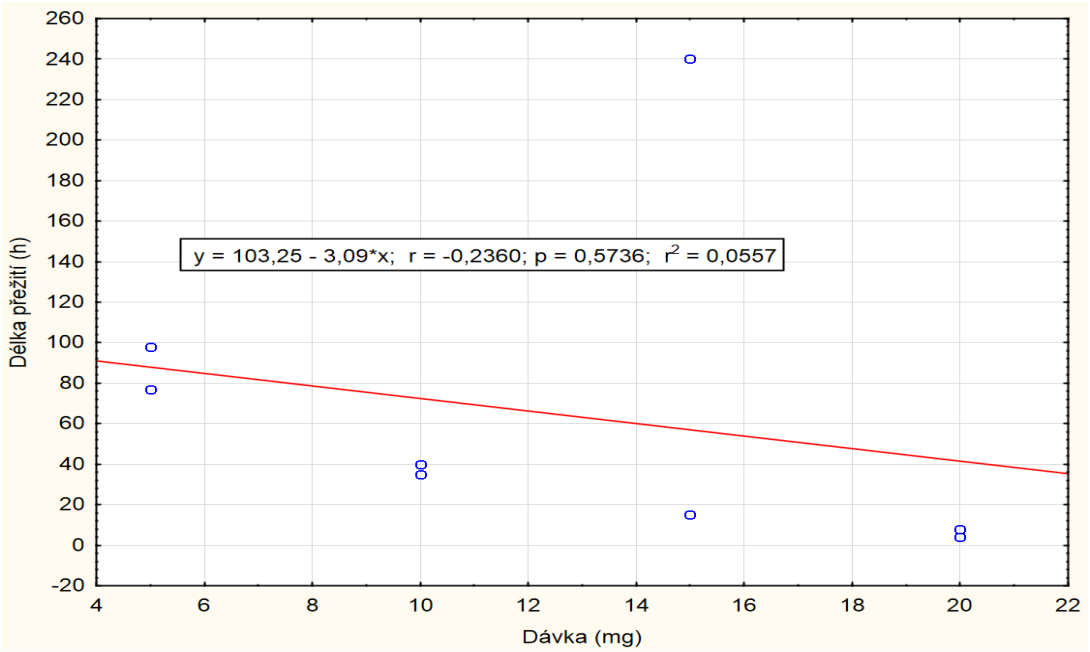


Obr. 16.15 - normální p-graf reziduí pro posouzení normality reziduí – Výskyt autokorelace v modelu (Výstup ze sw Statistica 10)

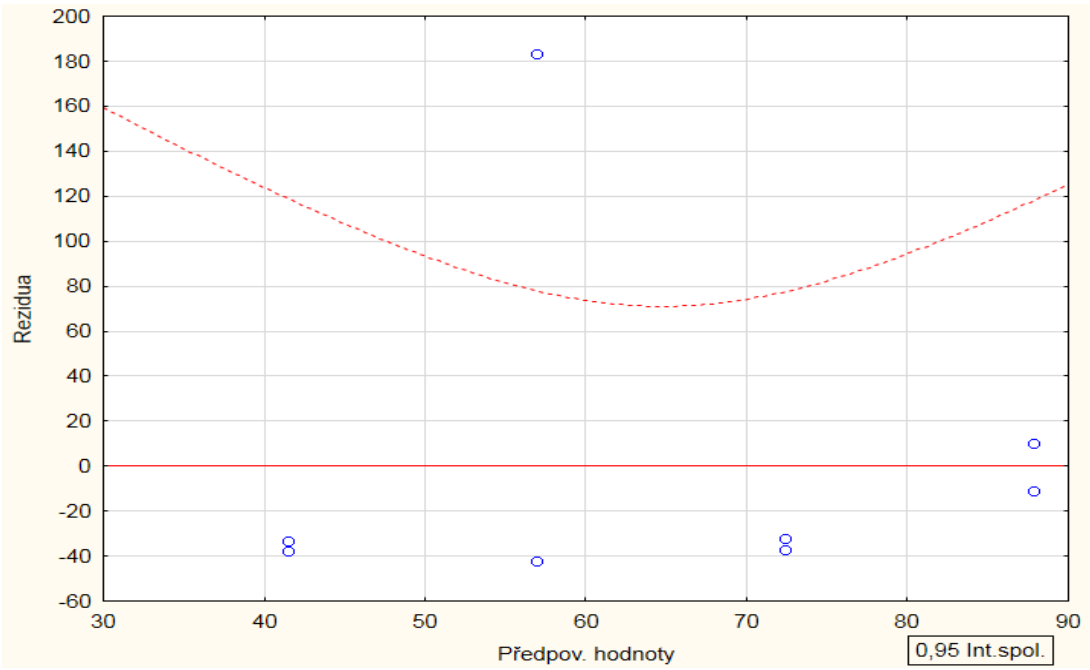


Obr. 16.16 - histogram pro posouzení normality reziduí - Výskyt autokorelace v modelu (Výstup ze sw Statistica 10)

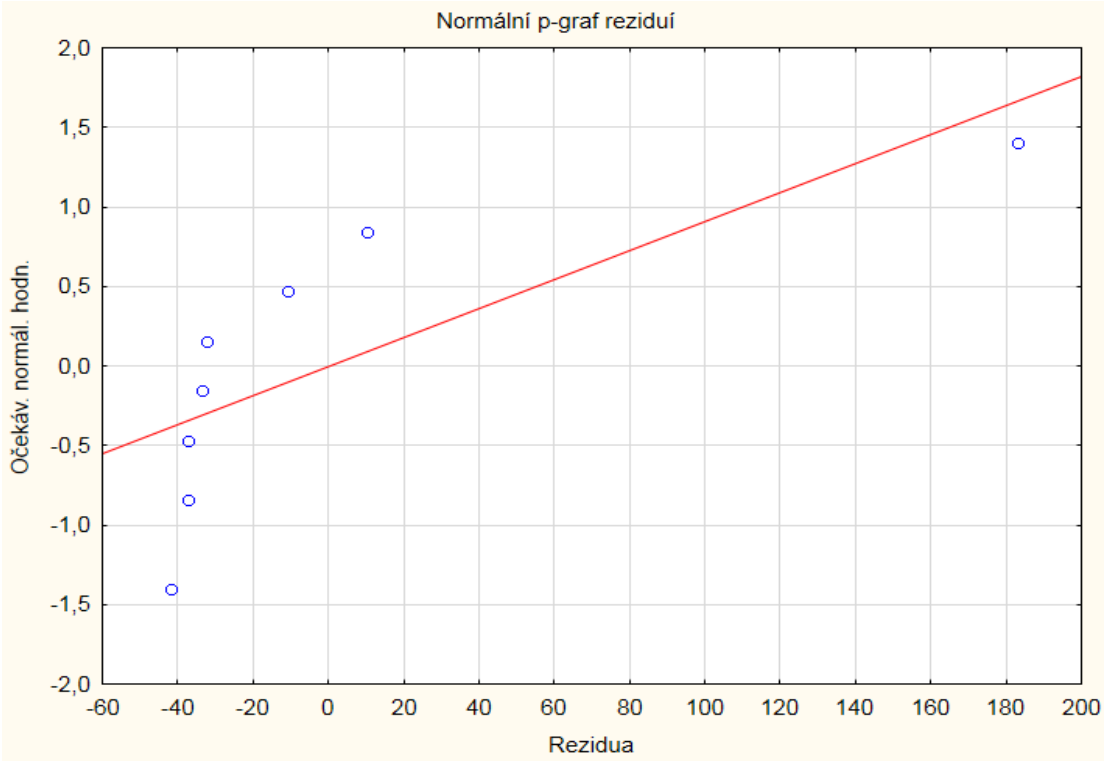
Příklad 16.5: Výskyt extrémně odlehlé hodnoty - Závislost délky přežití potkanů na aplikované dávce



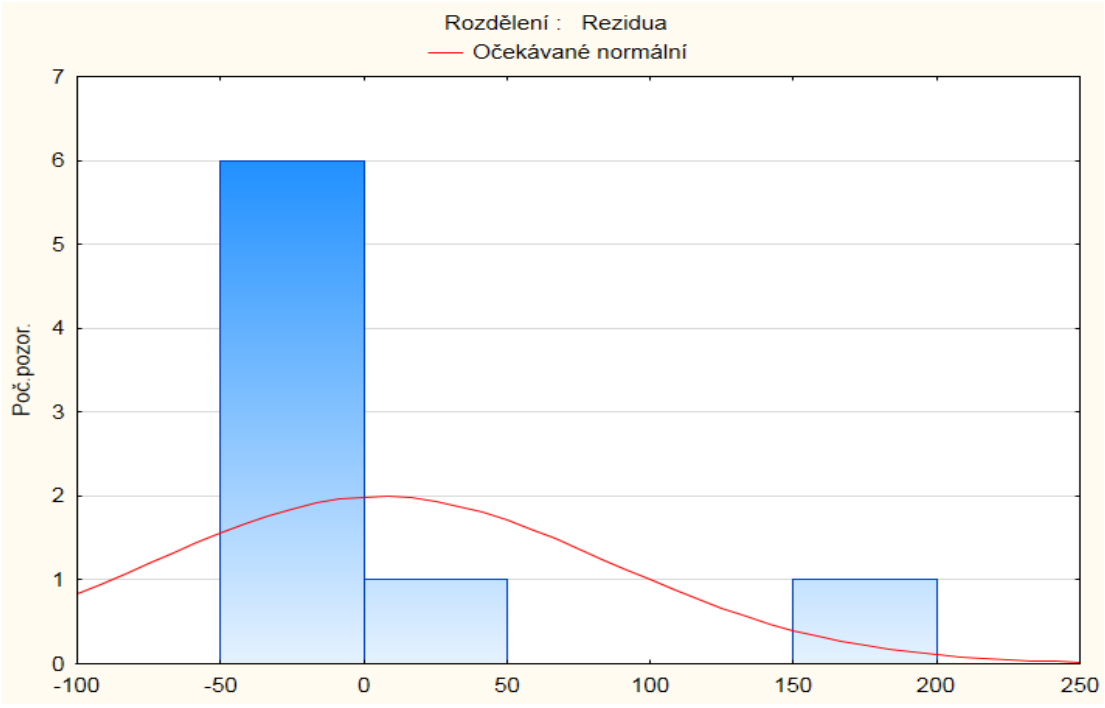
Obr. 16.17 - bodový graf závislosti y na x, znázorňující danou závislost – Výskyt extrémně odlehlé hodnoty (Výstup ze sw Statistica 10)



Obr. 16.18 - graf reziduí proti předpovězeným hodnotám - Výskyt extrémně odlehlé hodnoty (Výstup ze sw Statistica 10)



Obr. 16.19 - normální p-graf reziduí pro posouzení normality reziduí – Výskyt extrémně odlehlé hodnoty (Výstup ze sw Statistica 10)



Obr. 16.20 - histogram pro posouzení normality reziduí - Výskyt extrémně odlehlé hodnoty (Výstup ze sw Statistica 10)

Z uvedených obrázků jsme mohli na konkrétních příkladech vidět, jak se porušení jednotlivých předpokladů regresního modelu promítne do porušení předpokladu normality a projeví na normálním p-grafu reziduí či histogramu reziduí.

Odstranění nenormality reziduí z modelu

Jak již bylo uvedeno výše v této kapitole, nenormalita chyb bývá často důsledkem porušení jiného z předpokladů regresního modelu. Odstranění nenormality reziduí tedy nejčastěji spočívá primárně v zajištění splnění jiných předpokladů regresního modelu (nejčastěji výskytu odlehlých pozorování, případně heteroskedasticity).

16.3 Závěr

Podstatou splnění předpokladu o normalitě reziduí je, že rezidua e_i získaná ze vzorce $e_i = y_i - y_{iTEOR}$ by měla splňovat předpoklad normálního rozdělení. Jednou z příčin nenormality reziduí je přímo nenormalita vstupní vysvětlující či vysvětlované proměnné. Nenormalita chyb je také často důsledkem porušeného jiného z předpokladů regresního modelu. Před samotným ověřováním normality je tedy žádoucí zajistit splnění ostatních předpokladů regresního modelu.

17 Důsledky nepoužití zpožděné korelace v regresních modelech

17.1 Teoretická část

Při hodnocení závislosti vysvětlované proměnné y na uvažovaných možných vysvětlujících proměnných $x_1,\ldots,x_p$ jsou často uvažovány pouze klasické (nezpožděné) korelační koeficienty uspořádané v korelační matici. V reálných situacích však mnoho ekonomických, medicínských i dalších jevů reaguje až s určitým zpožděním na podnět. Hodnota vysvětlované proměnné y potom nemusí záviset nutně s hodnotou vysvětlované proměnné x_i ve stejném čase t, ale v nějakém čase t-j, kde j vyjadřuje míru zpoždění.

Důsledkem použití pouze klasického (nezpožděného) korelačního koeficientu potom může být neobjevení důležité funkční závislosti v analyzovaných datech a tedy nesestavení optimálního modelu závislosti y na $x_1,\ldots,x_p$.

Řešením problému je použití zpožděných korelačních koeficientů v případě jakýchkoliv analyzovaných dat, kde připadá ze znalosti problematiky v úvahu existence zpožděné závislosti. Některé specializované statistické software pro analýzu časových řad (např. Eviews, Gretl) nabízejí možnost automatického zpožďování analyzovaných proměnných. Pokud není toto softwarové zázemí k dispozici, je možno jednoduše např. v software MS Excel data posunout postupně vždy o 1 časové období dozadu a pro výpočet zpožděných korelačních koeficientů použít klasický korelační koeficient dostupný ve většině statistických software.

Volba maximálního uvažovaného zpoždění závisí na znalosti daného konkrétního problému. Optimální zpoždění daných vysvětlujících proměnných získáme dle nejvyšších zpožděných korelačních koeficientů, případně z grafického zobrazení časových řad vysvětlované a vysvětlující proměnné. Při finální stavbě zpožděného vícenásobného regresního modelu potom musíme některá zpoždění ještě přizpůsobit splnění některých nezbytných předpokladů regresního modelu, zejména autokorelace reziduí.

17.2 Praktická část

Příklad 17.1: Úkolem projektu pro soukromou společnost zabývající se výrobou mléčných výrobků z vykupovaného mléka byla předpověď ceny mléka (CENA) na základě známých předpovědí cen komodit (SMP, BUTTER, CHEESE, VSB, VCH). Odhadované ceny komodit na 12 měsíců dopředu jsou na trhu odhadovány pro celou EU na základě burzovních futures. K dispozici byla historická data vývoje ceny mléka a sledovaných komodit za období 1/2005-12/2009 a předpovědi komodit (prediktorů) na období 1/2010-12/2010.

Zavedení proměnných

- CENA – odhadovaná cena mléka v CZK/kg
- SMP – cena sušeného mléka v CZK/kg
- BUTTER– cena másla v CZK/kg
- CHEESE– cena sýra v CZK/kg
- VSB – valorizace SMP+BUTTER
- VCH– valorizace CHEESE

Pokud bychom pro předběžnou analýzu závislosti ceny mléka (CENA) na cenách komodit uvažovali v modelu pouze klasickou nezpožděnou korelaci, vypadala by tabulka korelačních koeficientů dle Obr. 17.1.

	SMP	BUTTER	CHEESE	VSB	VCH
CENA	0,468	0,619	0,856	0,533	0,796

Obr. 17.1 - ukázka síly závislosti při použití pouze nezpožděných korelačních koeficientů

Na základě vypočtených korelačních koeficientů by se mohlo mylně zdát, že cena mléka závisí na cenách komodit poměrně volně. Problém však spočívá v tom, že není uvažována možná zpožděná reakce farmářů na ceny komodit na trhu při stanovování ceny mléka.

Pokud rozšíříme úvahy i o možné zpoždění ceny mléka a použijeme pro analýzu závislosti zpožděné korelační koeficienty, budou korelační matice pro jednotlivé vysvětlující proměnné vypadat dle Obr. 17.2. Čím je hodnota korelačního koeficientu mezi zpožděnou řadou a cenou mléka vyšší, tím více daná komodita v daném

časovém zpoždění cenu mléka ovlivňuje. Červenou barvou jsou zvýrazněny nejvyšší zpožděné korelační koeficienty pro sledované vysvětlující proměnné.

Zpoždění	t	t-1	t-2	t-3	t-4	t-5	t-6	t-7	t-8	t-9	t-10	t-11	t-12
SMP	0,468	0,575	0,682	0,775	0,851	0,899	0,908	0,881	0,818	0,728	0,615	0,482	0,334
BUTTER	0,619	0,742	0,837	0,896	0,914	0,892	0,830	0,742	0,636	0,512	0,377	0,243	0,113
CHEESE	0,856	0,903	0,911	0,887	0,838	0,762	0,656	0,522	0,376	0,219	0,052	-0,099	-0,237
VSB	0,533	0,647	0,752	0,837	0,897	0,924	0,911	0,863	0,784	0,679	0,553	0,412	0,259
VCH	0,796	0,868	0,910	0,925	0,916	0,886	0,832	0,750	0,645	0,517	0,370	0,214	0,050

Obr. 17.2 - matice korelačních koeficientů pro nalezení optimálního časového zpoždění vlivu jednotlivých proměnných na cenu mléka (Výstup ze sw MS Excel)

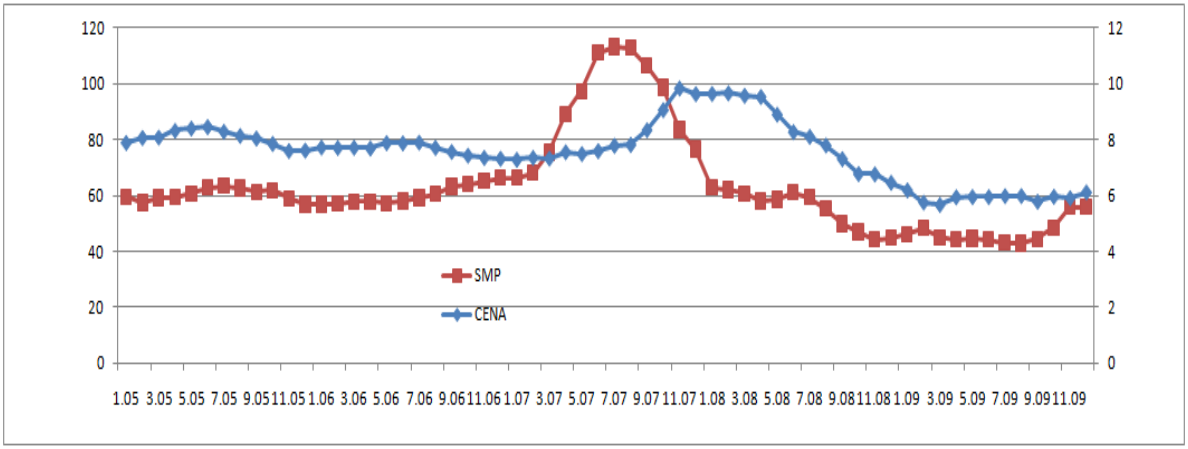
Optimální zpoždění pro jednotlivé proměnné tedy je:

- SMP - 6 měsíců
- BUTTER - 4 měsíce
- CHEESE - 2 měsíce
- VSB - 5 měsíců
- VCH- 3 měsíce

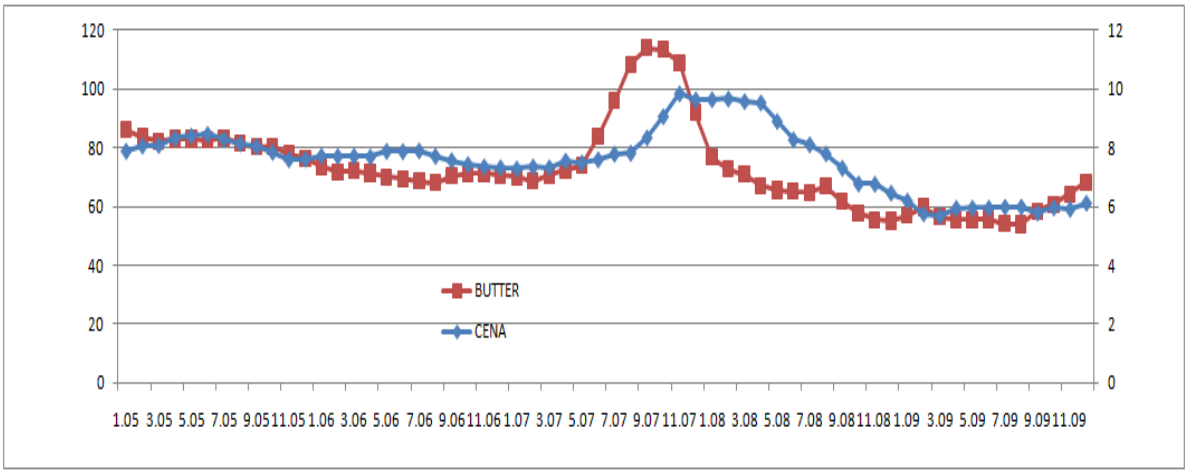
Výše uvedená zpožděná závislost ceny mléka na sledovaných komoditách je také dobře demonstrovatelná pomocí spojnicových grafů v Obr 17.3. V grafech můžeme sledovat, že posun grafu vývoje ceny mléka oproti grafu vývoje dané komodity poměrně dobře odpovídá zpoždění určenému použitím zpožděných korelačních koeficientů.

Obr. 17.3: Spojnicové grafy vývoje ceny mléka proti vývoji sledovaných komodit SMP, BUTTER, CHEESE, VSB, VCH (Výstup ze sw MS Excel)

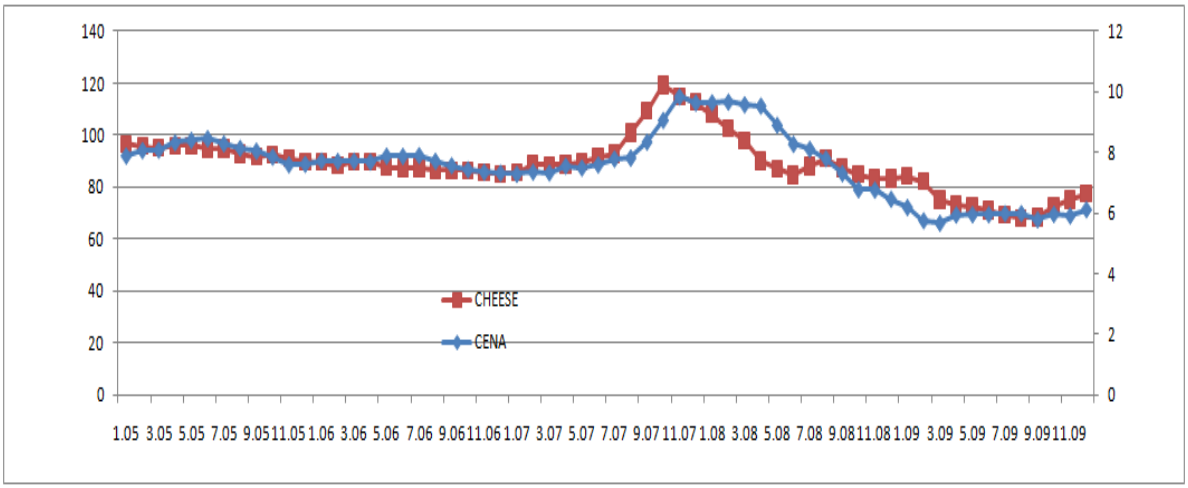
SMP



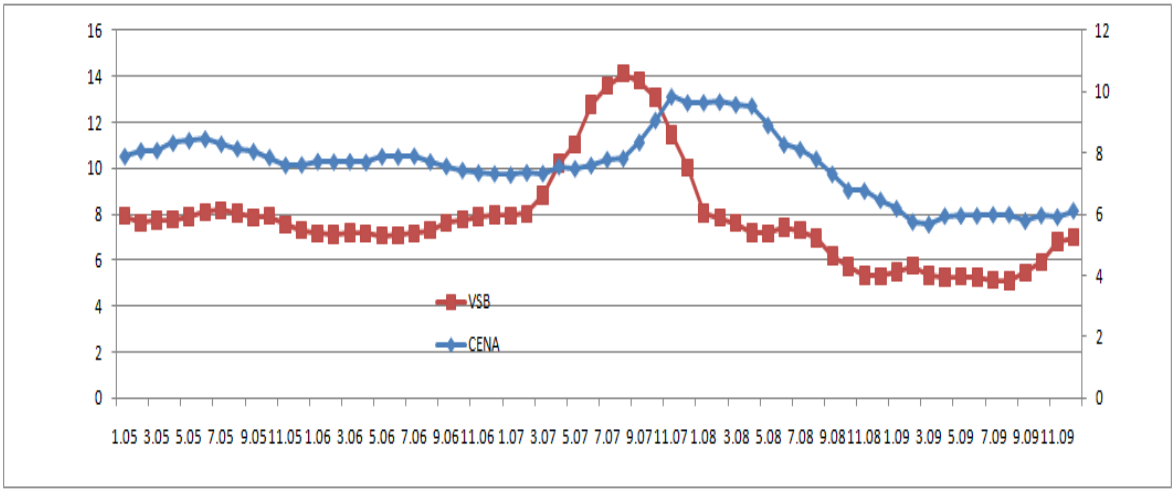
BUTTER



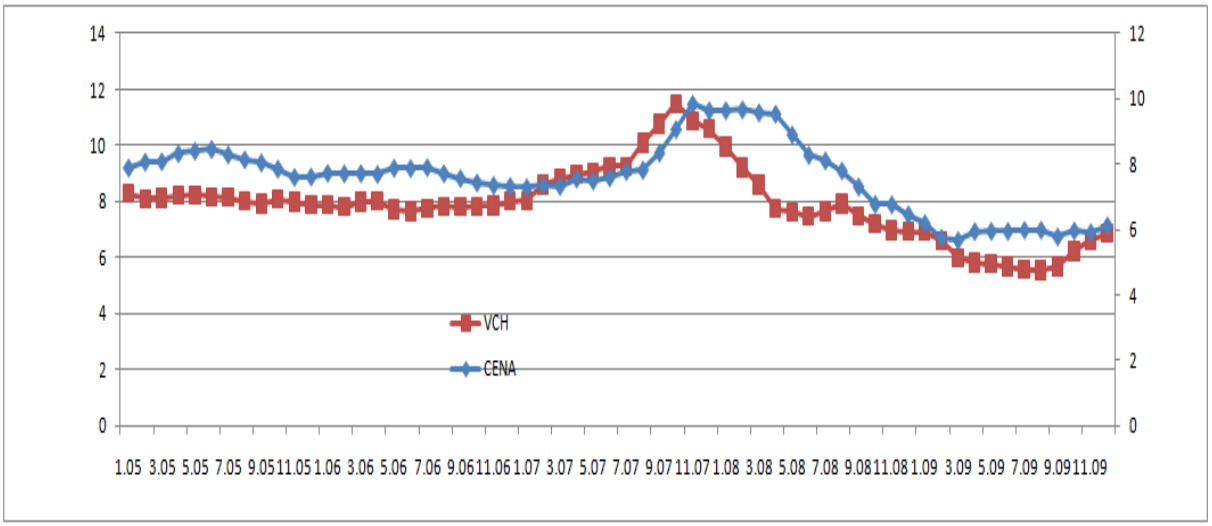
CHEESE



VSB



VCH



Výpočet odhadu ceny mléka na základě použití zpožděných proměnných oproti současným proměnným má následující výhody:

- Větší vysvětlující hodnotu pro cenu mléka než u nezpožděných proměnných.
- Možnost místo odhadů cen komodit na daný měsíc použít pro predikci ceny mléka již přesné reálné ceny komodit známé z předchozích měsíců.

Na základě zjištěných optimálních zpoždění jednotlivých komodit je možno následně sestavit regresní model pro odhad ceny mléka y na základě zpožděných regresorů SMP, BUTTER, CHEESE, VSB, VCH.

Výsledná rovnice regresního modelu pro predikci ceny mléka na základě **zpožděných** cen komodit může pro představu vypadat následovně (koeficient za názvem proměnné indikuje zpoždění dané proměnné v modelu.

CENA=8,708085253+0,027154115*SMP6+0,285307051*BUTTER4+0,09888345 *CHEESE2+1,188737389*VSB5+0,211761762*VCH3

17.3 Závěr

Pouhé použití obvyklých nezpožděných korelací může být zavádějící v případech, kdy vysvětlovaná proměnná y závisí na vysvětlujících proměnných $x_1,\ldots,x_p$ s nějakým časovým zpožděním. V takové situaci vychází nezpožděné korelační koeficienty oproti očekávání nízké, protože ve stejném čase spolu proměnné reálně příliš nekorelují. Použitím zpožděných korelačních koeficientů je potom možné v datech najít nečekané závislosti, které po podrobnější analýze často odpovídají reálné situaci. Velké části ekonomů a vědců, užívajících statistiku amatérsky bez spolupráce se statistikem, však existence zpožděných korelací bohužel není známa.

Závěr

Cílem této rigorózní práce byla snaha vytvořit portfolio příkladů využitelných při statistických školeních, které by absolventům nematematických oborů zjednodušily pochopení a práci s regresními modely. Regresní analýza dává s dostupným statistickým softwarem poměrně snadno a lehce cenné výstupy, ale nedodržení důležitých zásad při tvorbě modelu a jeho finálním použití může vést k velmi zavádějícím výsledkům, které mohou mít zejména v lékařství, ale i v jiných oborech negativní důsledky.

Účelem bylo vytvořit materiál, který by byl co nejlepším kompromisem mezi plně korektním matematickým přístupem k vysvětlení regresních modelů a zjednodušeným přístupem směřovaným k absolventům nematematických oborů, kteří však často regresní modely ve své praxi aktivně využívají.

V práci byla za tímto účelem věnována pozornost zejména didaktickému vysvětlení podstaty regresního modelování. V Kapitole 1 „Návrh optimálního didaktického postupu při vysvětlení regresní analýzy“ si může čtenář osvojit na konkrétním praktickém příkladu závislosti tržeb na investicích do reklamy danou problematiku a teorii jen s použitím minimálního množství matematických odvození. Vysvětlení teorie s použitím matematických odvození a symboliky jsem se snažil nahradit ukázkami podstaty a interpretace na konkrétních výstupech a výsledcích daného příkladu.

V Kapitole 2 „Praktické použití vytvořeného regresního modelu pro analýzu nákladů a tržeb a optimalizaci investic ve firmě“ si pozorný čtenář může uvědomit další aplikaci regresního modelu než je klasická predikce a kvantifikace vlivu jednotlivých vysvětlujících proměnných na vysvětlovanou proměnnou. Přestože tato aplikace regresních modelů v dostupné statistické literatuře téměř není zmiňována, patří dle mého názoru zvláště v manažerských a ekonomických odvětvích k velmi praktickým nástrojům, které statistika a matematika uživatelům poskytuje.

V Kapitolách 3-17 jsem se pokusil demonstrovat na jednoduchých příkladech dopady nejčastějších chyb, kterých se laici při použití regresních modelů dopouštějí a které mohou znehodnotit práci s regresním modelem.

Pro nejčastější chyby při aplikaci regresních modelů by bylo možno na základě této práce navrhnout následující rozdělení podle míry fatálnosti na finální výsledek:

a) Závažné – vždy velmi zásadním způsobem ovlivňují finální podobu modelu, predikci z něj, nebo další statistické úsudky

- Výskyt výrazně odlehlé hodnoty v modelu (Kapitola 15)
- Extrapolace pro hodnoty mimo obor, na kterém byl model konstruován (Kapitola 10)
- Nepoužití zpožděných korelací v regresním modelu v případě nutnosti (Kapitola 17)
- Použití korelačního koeficientu pro nelineární závislost (Kapitola 3)

b) Středně závažné – mohou v některých případech ovlivnit finální podobu modelu, predikci z něj, nebo další statistické úsudky

- Neodstranění multikolinearity z regresního modelu (Kapitola12)
- Použití univariantní analýzy namísto multivariantní regrese (Kapitola 11)
- Neodhalení zdánlivé korelace v modelu (Kapitola 5)
- Odhad očekávané hodnoty x z původní regresní funkce namísto sdružené regresní funkce (Kapitola 6)
- Nesplnění předpokladu homoskedasticity (Kapitola 13)
- Výskyt autokorelace reziduí v modelu (Kapitola 14)
- Nesplnění předpokladu normality reziduí (Kapitola 16)

c) Méně závažné - mají spíše menší dopad na finální podobu modelu, predikci z něj, nebo další statistické úsudky

- Záměna koeficientu determinace a adjustovaného koeficientu determinace (Kapitola 8)
- Nesprávná interpretace korelačního koeficientu a směrnice přímky (Kapitola 4)
- Interpretace korelačního koeficientu bez současného přihlédnutí k hodnotě p (Kapitola 7)
- Záměna predikčního a konfidenčního intervalu (Kapitola 9)

V práci jsem se snažil použít zejména konkrétní příklady, se kterými jsem se setkal při své spolupráci s lékaři, vědci a ekonomy, používajícími regresní modely aktivně při svém zaměstnání. Z této spolupráce také vyplynuly nejčastější chyby, kterých se laický uživatel regresních modelů dopouští.

Dovoluji si tedy doufat, že tato práce přispěje k lepšímu pochopení regresních modelů ze strany nematematických uživatelů nejen při školeních mých, ale případně i školeních jiných kolegů, kteří by mohli výstupů této práce volně využít.

Seznam použité literatury

1. **Zvára, K.** *Regrese.* Praha : MATFYZPRESS, 2008.

2. **Anděl, J.** *Statistické metody.* Praha : MATFYZPRESS, 1998.

3. **Hebák, P.** *Regrese I.a II. část.* Praha : Vysoká škola ekonomická v Praze, 1998.

4. **Komárek A., Komárková L.** *Statistická analýza závislostí s příklady v R.* Praha : VŠE v Praze, 2007.

5. **Clarke, B.** *Linear Models.* New Jersey (USA) : John Wiley and Sons, Inc., 2008.

6. **Draper, N.** *Applied Regression Analysis.* New York (USA) : John Wiley and Sons, Inc., 1998.

7. **Harrell, F.** *Regression Modeling Strategies.* New York (USA) : Springer, Inc., 2001.

8. **Faraway, J.** *Linear Models with R.* New York (USA) : Chapman and Hall/CRC, 2005.

9. **Hald, A.** *A History of Mathematical Statistics from 1750 to 1930.* New York (USA) : John Wiley and Sons, 1998.

10. **Tvrdoň, J.** *Ekonometrie.* Praha : ČZU Praha, 2012.

11. http://www.oocities.org/qecon2002/founda10.html. *Problems in Regression Analysis and their corrections.*

12. **Kubíková, J.** *Základní kurz Statistiky 1.* Praha : STATSOFT, 2009.

13. **Ramík, J.** *Statistika pro navazující magisterské studium.* Karviná : Slezská univerzita v Opavě, 2007.

14. **Meloun M., Militký J.** *Kompedium statistického zpracování dat.* Praha : ACADEMIA, 2002.

15. **Řičař, M.** *Základy ekonometrie - cvičení.* Praha : VŠE, 2013.

16. **Arlt J., Arltová M., Rublíková E.** *Analýza ekonomických časových řad s příklady.* Praha : VŠE v Praze, 2002.