Statistika v evaluacích

Mgr. Lubomíra Červová
21.6.2012

Program

- základní metody analýzy číselných proměnných (*průměr, kvantily, směrodatná odchylka, intervaly spolehlivosti, histogram, boxplot*)
- porovnání průměrů ve skupinách
- korelační analýza
- lineární regrese

Používaný software: IBM SPSS Statistics

Data

- **informace z databáze Albertina o firmách v letech 2008 a 2009**
  - základní informace z účetních výkazů firem
- **data z monitorovacího systému ESF Monit 7+**
  - informace o tom, zda firma získala podporu z grantu

+ další datové soubory pro doplnění …

Základní metody analýzy číselných proměnných

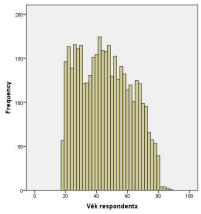
OPERAČNÍ PROGRAM LIDSKÉ ZDROJE A ZAMĚSTNANOST
PODPORUJEME VAŠI BUDOUCNOST
www.esfcr.cz

Popis souboru

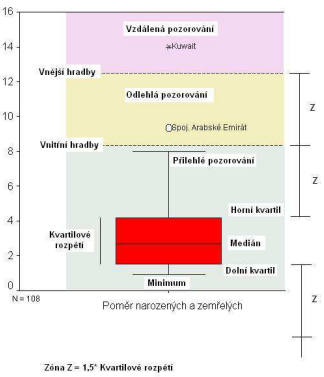
číselná proměnná

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Věk respondenta	4011	18	88	45,34	16,714
Valid N (listwise)	4011				



Boxplot



Zóna Z = 1,5* Kvartilové rozpětí

Komparace skupin

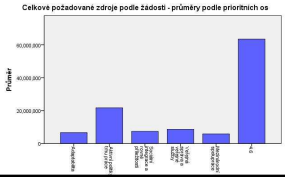
Porovnání jedné nebo více číselných proměnných ve skupinách

Průměrné hodnoty požadovaných finančních prostředků dle žádosti

Statistics: Průměr

Prioritní osa	Celkové zdroje - žádost	Soukromé financování - žádost	Veřejné fin. prostředky celkem - žádost
Adaptabilita	6603492	84014	6498078
Aktivní politiky trhu práce	21650899	9593	21640757
Sociální integrace a rovné příležitosti	7371197	5820	7351275
Veřejná správa a veřejné služby	8586513	28064	8557953
Mezinárodní spolupráce	5799177	77	5795820
4.6	63517095	0	63517095
Celkem	8261469	40913	8205719

Celkové požadované zdroje podle žádosti - průměry podle prioritních os



Testy hypotéz

Interval spolehlivosti

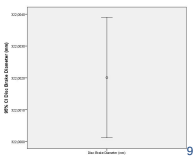
bodový odhad x intervalový odhad

interval spolehlivosti

- vyjadřuje oblast, kde by se s danou spolehlivostí (obvykle 95%) měla nacházet hodnota statistiky pro celou populaci
- předpokládá pravděpodobnostní výběr *(obvykle prostý náhodný výběr)*

Descriptives

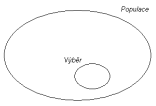
			Statistic	Std. Error
Disc Brake Diameter (mm)	Mean		322,002008	,0009566
	95% Confidence Interval for Mean	Lower Bound	322,000116	
		Upper Bound	322,003902	
	5% Trimmed Mean		322,001863	
	Median		322,001811	
	Variance		,000	
	Std. Deviation		,0108224	
	Minimum		321,9759	
	Maximum		322,0306	
	Range		,0547	
	Interquartile Range		,0133	
	Skewness		,200	,214
	Kurtosis		-,126	,425



Úvod do testování statistických hypotéz

Výběrová šetření

- Lze určitou vlastnost výběrového souboru zobecnit s danou spolehlivostí na celou populaci?
  (například průměrný plat zaměstnanců v určitém oboru)
- Mohou být zjištěné rozdíly na výběrovém souboru způsobené náhodou? *(například rozdíly mezi regiony)*



Princip testování statistických hypotéz (1)

H_0: nulová hypotéza (*na začátku předpokládáme, že platí*)
H_A: alternativní hypotéza

T … testová statistika (*kritérium pro rozhodování*)
Hladina spolehlivosti (*volíme předem, obvykle 95%*)

Pro testování použijeme testovou statistiku *T* a nalezneme vhodnou podmnožinu *W*, tzv. **kritický obor**

Rozhodování:
Pokud $\{T \in W\} \Rightarrow$ zamítneme H_0, v opačném případě H_0 nezamítneme

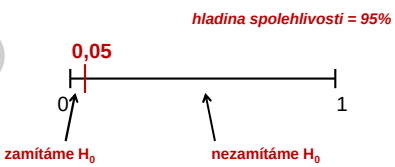
		rozhodnutí	
		H_0	**H_A**
platí	**H_0**	OK	**chyba I. druhu**
	H_A	**chyba II. druhu**	OK

Princip testování statistických hypotéz (2)

Significance (dosažená hladina významnosti)

- pravděpodobnost, že zamítneme nulovou hypotézu ačkoliv platí

Rozhodování na základě significance:



Princip testování statistických hypotéz (3)

síla testu: pravděpodobnost, že statistický test zamítne nulovou hypotézu v případě, že neplatí

síla testu = 1 - β , kde
β je pravděpodobnost chyby II. druhu

Sílu testu ovlivňuje:

- **použitý statistický test** (některé testy jsou silnější než jiné)
- **velikost výběru** (obecně čím větší je velikost výběru, tím větší je síla testu)
- zvolená **hladina spolehlivosti**
- **effect size** (zjištěný rozdíl od nulové hypotézy)
- jednostranná x oboustranná **alternativní hypotéza**
- **chyba měření** (jakákoliv chyba, která narušuje přesnost měření, snižuje sílu testu)

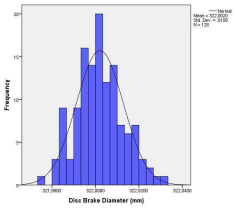
Testování normality

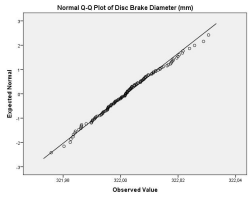
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Disc Brake Diameter (mm)	,051	128	,200*	,993	128	,785

a. Lilliefors Significance Correction
*. This is a lower bound of the true significance.

H_0: proměnná má normální rozložení





T-testy, ANOVA

Testy hypotéz o průměrech

- **Jednovýběrový t-test** – porovnání průměru s konstantou
- **Párový t-test** – porovnání průměrů dvojice proměnných
- **Dvouvýběrový t-test** – porovnání průměrů ve dvou skupinách
- **Jednoduchá ANOVA** (analýza rozptylu) – porovnání průměrů v několika skupinách

Předpoklady:

- nezávislost pozorování
- nezávislost skutečných hodnot a chyb
- **normální rozložení** (testovaná proměnná/rozdíl proměnných/testovaná proměnná ve skupinách)
- dvouvýběrový t-test, ANOVA: **skupiny se nepřekrývají**
- dvouvýběrový t-test, ANOVA: **shoda rozptylů ve skupinách**

Jednovýběrový t-test

H_0: průměr (střední hodnota) proměnné odpovídá testované konstantě, tj. $\mu = 322$

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Disc Brake Diameter (mm)	128	322,002009	,0108224	,0009566

One-Sample Test

	Test Value = 322					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
Disc Brake Diameter (mm)	2,101	127	,038	,0020094	,000116	,003902

Párový t-test

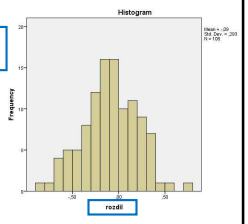
H_0: průměry (střední hodnoty) proměnných se rovnají, tj. $\mu_X = \mu_Y$

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	první měření	17,0119	108	1,05382	,10140
	druhé měření	17,0982	108	1,06566	,10254

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	první měření & druhé měření	108	,962	,000



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	první měření - druhé měření	-,08624	,29259	,02815	-,14205	-,03043	-3,063	107	,003

Dvouvýběrový t-test

H_0: průměry (střední hodnoty) ve dvou skupinách se rovnají, tj. $\mu_1 = \mu_2$

Group Statistics

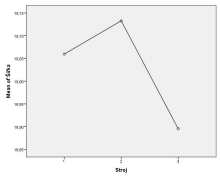
	Měřicí přístroj	N	Mean	Std. Deviation	Std. Error Mean
Šířka	A	37	19,0192	,32840	,05399
	B	35	19,0411	,23192	,03920
Průměr vývodu	A	37	1,5055	,09345	,01536
	B	35	1,4889	,11545	,01952

Independent Samples Test

		Levene's Test for Equality of Variances F	Sig.	t-test for Equality of Means t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Šířka	Equal variances assumed	3,009	,087	-,340	70	,735	-,02291	,06735	-,15724	,11142
	Equal variances not assumed			-,343	64,874	,732	-,02291	,06672	-,15616	,11034
Průměr vývodu	Equal variances assumed	2,677	,106	,672	70	,504	,01659	,02469	-,03266	,06583
	Equal variances not assumed			,668	65,457	,507	,01659	,02484	-,03301	,06618

Jednoduchá analýza rozptylu (ANOVA)

H_0: průměry (střední hodnoty) ve dvou nebo více skupinách se rovnají, tj. $\mu_1 = \mu_2 = \ldots = \mu_K$



Test of Homogeneity of Variances

	Levene Statistic	df1	df2	Sig.
Šířka	,152	2	69	,859
Průměr výrobu	1,044	2	69	,357

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Šířka	Between Groups	,709	2	,354	4,880	,010
	Within Groups	5,012	69	,073		
	Total	5,721	71			
Průměr výrobu	Between Groups	,062	2	,031	3,020	,055
	Within Groups	,710	69	,010		
	Total	,773	71			

Typy testů

Parametrické testy

- testy o parametrech známých pravděpodobnostních modelů, předpokládají určitý typ rozdělení, ze kterého výběr pochází (obvykle normální)
- drobné odchylky od předpokladů většinou nevadí
- při výrazném porušení předpokladů můžeme získat nekorektní výsledky

Neparametrické testy

- určeny pro situaci, kdy nejsou splněny předpoklady parametrických testů
- nemají předpoklady o typu rozdělení dat
- vyváženo menší citlivostí (vyšší pravděpodobnost nezamítnutí testované hypotézy)
- testují jinou hypotézu o rozdělení základního souboru než je hypotéza o jeho parametru

Klasifikace neparametrických testů

- jednovýběrové testy
- dvouvýběrové testy
  - nezávislé výběry
  - závislé výběry
- vícevýběrové testy
  - nezávislé výběry
  - závislé výběry

Neparametrické testy pro 2 závislé výběry

Wilcoxonův test (*Wilcoxon*)
- test hypotézy o shodě mediánů ve dvou závislých výběrech
- silnější než znaménkový test
- předpokládá stejný tvar rozdělení
- princip: celkové pořadí na základě absolutních hodnot rozdílů testovaných proměnných, průměrné pořadí + i - by měla být podobná

Znaménkový test (*Sign test*)
- test hypotézy o shodě mediánů ve dvou závislých výběrech
- princip: podobný jako u W. testu, ale bere v úvahu pouze znaménka rozdílů, aplikuje binomický test

McNemarův test (*McNemar*)
- pouze pro dichotomické proměnné
- zjišťuje, zda po určité události došlo ke změně výskytu určitého jevu

Neparametrické obdoby párového t-testu

Neparametrické testy pro dva nezávislé výběry

Mann-Whitneyův test (*Mann-Whitney U)*
- testuje hypotézu, že obě skupiny pocházejí ze stejného základního souboru
- předpokládá stejný tvar rozložení dat ve skupinách
- princip: společné pořadí, průměrné pořadí v obou skupinách by měla být podobná

Dvouvýběrový Kolmogorov-Smirnovův test (*Kolmogorov-Smirnov Z*)
- obecnější a slabší než Mann-Whitneyův test
- založen na porovnání empirických distribučních funkcí
- testová statistika vychází z jejich maximální absolutní odchylky
- citlivý na polohu a škálu měření

Wald-Wolfowitzův test (*Wald-Wolfowitz runs*)
- nemá žádné dodatečné předpoklady
- princip: společné pořadí, v setříděné posloupnosti by se měla náhodně střídat pozorování z obou skupin → Runs test

Neparametrické obdoby dvouvýběrového t-testu

Neparametrické testy pro K nezávislých výběrů

Kruskal-Wallisův test (*Kruskal-Wallis H*)
- testuje hypotézu, že všechny skupiny pocházejí ze stejného základního souboru
- předpokládá shodu tvaru rozdělení ve skupinkách
- test neříká nic o tom, které skupiny se od sebe odlišují
- princip: společné pořadí všech skupin, průměrná pořadí ve skupinách by měla být podobná

Mediánový test (*Median*)
- testuje hypotézu o shodě mediánů ve dvou a více nezávislých výběrech
- vhodný v situacích, kdy předpokládáme ve výběrech různá rozdělení
- princip: společný medián, kontingenční tabulka kategorií × kategorie pod/nad mediánem, test Chí-kvadrát

Neparametrické obdoby ANOVA

Neparametrické testy – ukázka (1)

Mann-Whitneyův test

Ranks

	Měřicí přístroj	N	Mean Rank	Sum of Ranks
Hloubka	A	37	36,73	1359,00
	B	35	36,26	1269,00
	Total	72		
Výška	A	37	36,82	1355,00
	B	35	36,37	1273,00
	Total	72		
Šířka	A	37	34,86	1290,00
	B	35	36,23	1338,00
	Total	72		
Průměr vývodu	A	37	38,46	1423,00
	B	35	34,43	1205,00
	Total	72		

H_0: obě skupiny pocházejí ze stejného základního souboru

Test Statistics[a]

	Hloubka	Výška	Šířka	Průměr vývodu
Mann-Whitney U	639,000	643,000	587,000	575,000
Wilcoxon W	1269,000	1273,000	1290,000	1205,000
Z	-,096	-,051	-,682	-,817
Asymp. Sig. (2-tailed)	,924	,960	,495	,414

a. Grouping Variable: Měřicí přístroj

Neparametrické testy – ukázka (2)

Kruskal-Wallisův test

Ranks

	Stroj	N	Mean Rank
Hloubka	1	24	34,58
	2	24	36,13
	3	24	38,79
	Total	72	
Výška	1	24	25,42
	2	24	28,79
	3	24	55,29
	Total	72	
Šířka	1	24	39,25
	2	24	44,13
	3	24	26,13
	Total	72	
Průměr vývodu	1	24	29,08
	2	24	44,33
	3	24	36,08
	Total	72	

H_0: všechny skupiny pocházejí ze stejného základního souboru

Test Statistics[a,b]

	Hloubka	Výška	Šířka	Průměr vývodu
Chi-Square	,497	29,336	9,498	6,386
df	2	2	2	2
Asymp. Sig.	,780	,000	,009	,041

a. Kruskal Wallis Test
b. Grouping Variable: Stroj

Korelační analýza

Statistika

- **statistika** zkoumá variabilitu dat:
  - popisuje ji
  - vysvětluje ji
  - predikuje ji
- **korelační analýza** zkoumá společnou variabilitu (kovariabilitu):
  - popisuje ji
  - používá ji pro vysvětlení
  - používá ji pro predikci

Úlohy a otázky:

A) Souvisí spolu výskyt proměnné X a proměnné Y tak, že s vyššími hodnotami X se pojí vyšší hodnoty Y a (a nižšími nižší), či naopak s vyššími hodnotami X se pojí nižší hodnoty Y (a s nižšími X vyšší Y)?

B) Můžeme v datech zjistit souběžnost resp. protiběžnost hodnot dvou číselných proměnných?

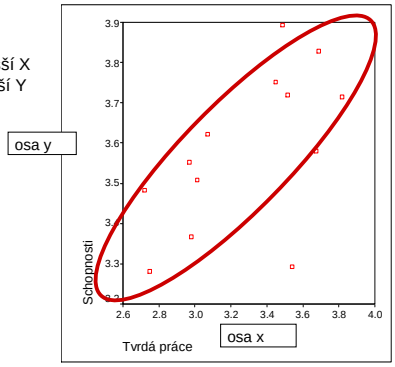
C) Je hodnota Y důsledkem hodnoty X? Reprezentuje proměnná X příčinu pro důsledek Y?

D) Jsou X a Y nositeli (částečně) stejné informace?

E) Vylučují se (resp. doplňují se) X a Y nebo naopak jedno předpokládá druhé?

Souběh variabilit

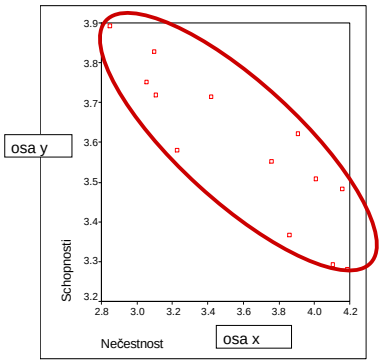
čím vyšší X
tím vyšší Y



Protiběh variabilit

čím vyšší X
tím nižší Y

čím nižší X
tím vyšší Y



Korelační analýza

- **korelace = vztah dvou *číselných* proměnných**
- **vztah:**
  - ***empirický*:** korelace je zachycení vztahu mezi vzniklými čísly, statistickými řadami, bez ohledu na to, co znamenají
  - ***kauzální:*** korelace je odraz geneze příčinného procesu

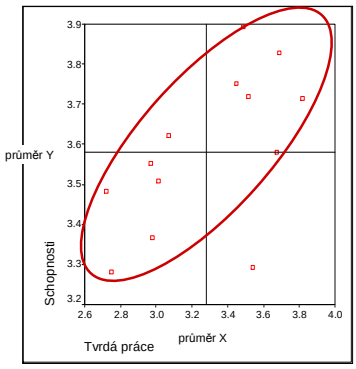
Popis kovariability

- **kovariance**: souběh variabilit dvou proměnných

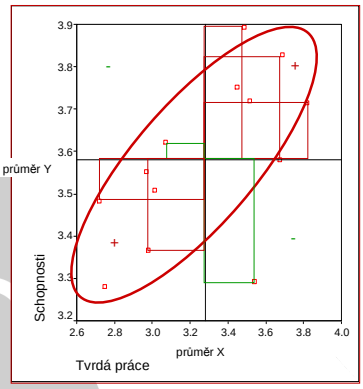
$$cov(X,Y) = \frac{1}{(N-1)} \sum (X_i - \overline{X}) * (Y_i - \overline{Y})$$

$$cov(X,Y) = \frac{1}{N*(N-1)/2} \sum_{i \neq j} (X_i - X_j) * (Y_i - Y_j)$$

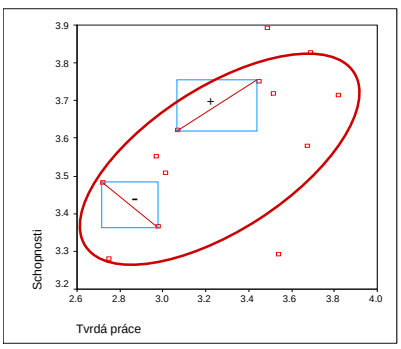
Kovariance = součet orientovaných plošek



Kovariance = součet orientovaných plošek



Kovariance = součet orientovaných plošek



Korelace – kovariance v poměru k rozptylům

korelace = vztah dvou proměnných

= kovariance standardizovaná k rozptylům obou proměnných

= měří ***vztah*** dvou variabilit, nikoliv jejich velikost

Korelační koeficient

$$r = \frac{cov(X,Y)}{\sqrt{varX * varY}} \qquad r = \frac{cov(X,Y)}{s_X \times s_Y}$$

- *r* je vypočten jako kovariance v poměru ke geometrickému průměru rozptylů
- jmenovatel je také součin směrodatných odchylek $s_x * s_Y$

Pearsonův lineární korelační koeficient

Vlastnosti:

- **r definován, pro n > 1**
- **r definován pro nenulové variability;**
  nesmí platit $s_X = 0$ nebo $s_Y = 0$
- r = 1, právě když body jsou seřazeny v nějaké přímce s nenulovým kladným spádem
- r = -1, právě když body jsou seřazeny v nějaké přímce s nenulovým záporným spádem
- čím více se r blíží k +1, tím více se body shlukují kolem stoupající přímky; čím více se r blíží k –1, tím více se body shlukují kolem klesající přímky
- jestliže v mraku bodů nelze vystopovat žádný *lineární* trend, r = 0

Pearsonův lineární korelační koeficient

Další vlastnosti:

- **r se nezmění, když se**

  - **posune** škála jedné nebo obou proměnných o **libovolnou konstantu** (změna počátku)

  - změní škála jedné nebo obou proměnných **násobkem** libovolnými činiteli (změna měřítka)

Nulové r:

- mrak bodů tvoří pravidelný kruhový útvar
- přímka, kolem které se shlukují body, je vodorovná nebo svislá
- body leží symetricky kolem osy procházející průměrem X a to i když odpovídají úplné závislosti Y na X,
  např. $Y = (X - 4)^2$
- silné shlukování kolem rostoucí/klesající přímky je zkresleno bodem vzdáleným od mraku
- kříží se kladný a záporný trend – překrytí dvou bodových mraků

Zkreslení koeficientu korelace

- **vzdálený bod** - mrak bodů ukazuje na silnou/slabou korelaci, ale vzdálený bod ji uměle sníží/zvýší
- **dvě skupiny** nulové korelace umístěné v rovině vykazují vyšší korelaci
- číselné proměnné mají **diskretní** povahu (škála celých čísel od 1 do K) a přesné seskupení hodnot kolem přímky není plně možné
- jsou-li **rozložení** X a/nebo Y výrazně **šikmá** s dlouhým koncem

Další míry korelace

- jiná data (pořadí)
- šikmá rozložení
- vzdálená pozorování
- zvyklosti oboru
- komparace s jinými výstupy

Pořadová korelace

- koeficienty pořadové korelace vycházejí při výpočtu koeficientu z pořadí a vzájemných pozic hodnot
- nejsou citlivé na vzdálená pozorování
- neměří linearitu vztahu, ale shodu pořadí obou statistických řad
- mohou se použít pro číselná data a pro data, která vyjadřují pouze pořadí případů z hlediska X a Y

Pořadová korelace - Spearmanovo ρ

- Spearmanův koeficient pořadové korelace ρ vznikne tak, že se **do vzorečku pro Pearsonův lineární korelační koeficient dosadí místo hodnot X a Y jejich pořadí v řadě**

ρ = 1, pokud jsou řady zcela shodné

ρ = -1, pokud jsou řady zcela protichůdné

ρ = 0, pokud mezi řadami není žádná tendence ke shodě či protichůdnosti, ale pořadí jsou k sobě zcela náhodně

Pořadová korelace - Kendallovo τ

- vypočte se z **počtu souhlasných a nesouhlasných dvojic** (shod a neshod v pořadí pro dvojice případů):
  - shoda dvojice: případy jsou u obou proměnných orientovány ve stejném směru, tj. když první je větší než druhý u X je větší i u Y (a naopak)
  - neshoda dvojice – případy opačně orientovány v pořadí, tj. když první je větší než druhý u X je menší u Y (a naopak)
- **shody a neshody jsou znaménka při výpočtu kovariance vycházejícího ze všech dvojic**

τ = (počet shod – počet neshod) / (počet shod + počet neshod)

- vlastnosti stejné jako u Spearmanova koeficientu

Testy hypotéz

u korelačních koeficientů je základní dvojicí hypotéz, které testujeme:

H_0: korelační koeficient je nulový
H_A: korelační koeficient je nenulový

- prokazujeme, že naše spočtená míra je signifikantně nenulová, tedy, že v datech se projevuje nějaký vztah
- platí pro Pearsonův, Spearmanův i Kendallův koeficient

Použití korelací – konfirmační přístup

testování hypotéz a teorií: korelovanost prokazuje představu o kauzálním procesu

nebezpečí:

•korelovanost vznikne jiným procesem, než je předpokládáno

•data jsou nepřesná (nebo je jich málo) a korelovanost je proto rozšumnělá a nepřesvědčivě nízká

•kauzální proces existuje, ale projevuje se jinými závislostmi než lineárními

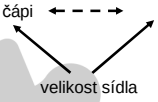
•kauzální procesy se projevují jen na části souboru a korelace je tak překryta jinými vztahy

Parciální korelace - společná příčina

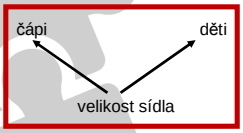
čápi → děti r = .56

čápi ← → děti r = .56

čápi ← - - → děti r = .56



r(v,č) = .81 r(v,d) = .69



r(č,d/v) = 0.0

r(v,č) = .81 r(v,d) = .69

Parciální korelace

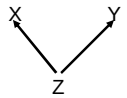
- X, Y, Z (tři známé proměnné v datech, jejichž vzájemné korelace jsou známy):

$$r(X,Y/Z) = \frac{r(X,Y) - r(X,Z) \times r(Y,Z)}{\sqrt{(1 - r(X,Z)^2) \times (1 - r(Y,Z)^2)}}$$

- je-li parciální korelační koeficient roven nebo blízko k nule (**nesignifikantní**), znamená to, že proměnná Z plně vysvětluje korelaci mezi X a Y
- pokud se parciální koeficient jen **podstatně redukuje**, znamená to, že Z ovlivňuje vztah X a Y , ale není samo

Parciální korelace

Model 1: Společná příčina X, Y, Z:



Z = společná příčina

r(X,Y) ≠ 0, r(X,Z) ≠ 0, r(Y,Z) ≠ 0,
r(X,Y/Z) = 0

Model 2: Zprostředkující vlastnost

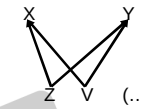
X → Z → Y

r(X,Y) ≠ 0, r(X,Z) ≠ 0, r(Y,Z) ≠ 0,
r(X,Y/Z) = 0

Modely 1 a 2 se nedají statisticky odlišit

Parciální korelace
model 3 – společné příčiny

X, Y, Z, V



Z = společná příčina

(….W, T, …)

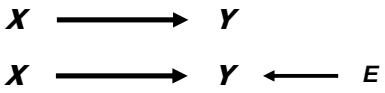
$r(X,Y) \neq 0$, $r(X,Z) \neq 0$, $r(Y,Z) \neq 0$, $r(X,V) \neq 0$, $r(Y,V) \neq 0$,
$r(X,Y/Z) \neq 0$, $r(X,Y/V) \neq 0$
$r(X,Y/Z,V) = 0$

Korelační analýza

- **korelační analýza je první stupeň analýzy vztahů, po ní následují:**
  - regresní analýza,
  - faktorová analýza,
  - analýza kovariančních struktur a kauzálních sítí
  - další speciální analýzy …

Lineární regrese

Regresní analýza: jednosměrný vztah



- nezávislá proměnná –> závislá proměnná
- příčina –> následek
- prediktor –> predikant

u korelace: jednosměrný ***nebo*** symetrický vztah
u regrese: ***jednosměrný*** vztah

Popis vztahu rovnicí

- rovnice

model vztahu

forma převodu

chyba rovnice

$$Y = f(X) + \varepsilon$$

směr vztahu

rovnice rozloží hodnotu Y na dvě části:

a) model = převod z X

b) residuum/zbytek, které se neúčastní převodu

Dva významy rovnic: model vztahu/procesu a predikce

lineární model

residuum

$Y = a + bX + \varepsilon$

rovnice datového procesu

predikční rovnice

$\tilde{Y}(X) = a + bX$

očekávaná hodnota Y pro dané X

$\tilde{Y} = E(Y/X)$

očekáno pro dané X

není obsaženo v X

$Y = \tilde{Y} + \varepsilon$

očekáváme, že chyba je v průměru nulová

a = konstanta

posunutí

b = regresní koeficient

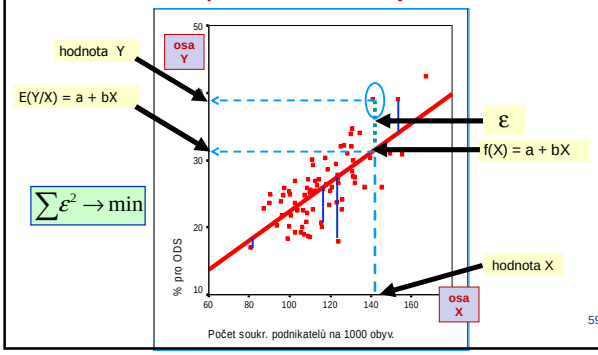
směrnice přímky

Významy parametrů: rovnoměrné posunutí a individuální změna

- **b** = regresní koeficient – koeficient úměry vlivu X na Y u každého jednoho případu
  - b > 0 … přímka má růstový/stoupavý trend
    kladný trend
    s rostoucím X roste Y
  - b < 0 … přímka má ztrátový/klesavý trend
    záporný trend
    s rostoucím X klesá Y
  - b = 0 … přímka je rovnoběžná s osou X, absence trendu
    s rostoucím X se Y nemění: nulový trend
    hodnota Y na X nezávisí
- **a** = posunutí – hodnota Y pro nulové X nebo koeficient rovnoměrné změny pro každý případ bez ohledu na jeho X hodnotu

Lineární regrese: zachycení souběhu variabilit přímkou

hledáme vhodnou přímku – metoda nejmenších čtverců



Dekompozice variability závislé (vysvětlované, cílové) proměnné Y

$$Y = a + bX + \varepsilon$$

var(Y) = var(E(Y/X)) + var ε

TSS = MSS + ESS

celkový (korigovaný k průměru) součet čtverců Y = součet čtverců očekávaných/predikovaných modelových hodnot + součet čtverců odchylek (residuí, chyb)

TSS $= \sum(Y_i - \overline{Y})^2$
MSS $= \sum(a + bX_i - \overline{Y})^2$
ESS $= \sum(Y_i - (a + bX_i))^2$

Testování významnosti modelu: tabulka ANOVA

F - test – kritérium pro zjištění existence vztahu

F(1, n-2) = MSS / (ESS /n-2)

df = (1,n-2)

dosažená významnost F = alfa

$F = R^2 / (1- R^2) * (n-2)$

n = počet případů

Dekompozice variability Y: koeficient determinace

variabilita Y je dekomponována do dvou částí korespondujících rozdělení hodnoty Y na dvě složky modelem:

Y = očekávání/predikce + chyba/residuum

$var\ Y = var(E(Y/X)) + var\ \varepsilon$

z toho logicky odvodíme míru determinace:

koeficient determinace:

$R^2 = var(E(Y/X))/var(Y)$

$= 1 – (var\ \varepsilon / var(Y))$

$= 1 – ESS / TSS = MSS/TSS$

Koeficient determinace

koeficient determinace – míra explanační síly rovnice, intenzita působení X na Y lineárním vztahem

vyjádřeno %: $100*R^2\%$

procento var Y vysvětlené modelovou rovnicí

R^2 = čtverec korelačního koeficientu $r(X,Y)$

$R^2 = r(X,Y)^2$

$F = R^2 / (1- R^2) * (n-2)$

n = počet případů

Testy významnosti parametrů

H_0: b = 0 vs. H_A: b ≠ 0

t - test – kriterium významnosti *b*

t (n-2) = b/ *sterr*(b)
df = n-2
dosažená významnost t = alfa*

obdobně pro parametr *a*

pro rovnici s jedním prediktorem:

$t^2 = F$

Konfidenční intervaly pro *b* a *a*

konfidenční interval pro parametr – *kriterium přesnosti odhadu*

interval pro spolehlivost *gamma* (= 1 – *alpha*)
b = odhadnuté b ± t(n-2;alfa/2)*sterr(b)
gama = 1 –*alfa*
analogicky pro *a*

skutečná hodnota parametru není intervalem zachycena s rizikem alfa

Konfidenční pás pro přímku (očekávané hodnoty k danému *X*)

konfidenční pás pro přímku – kriterium pro přesnost určení přímky

pás spolehlivosti pro E(Y/X):
*odhad (a +b*X)= odhad* a + odhad *b**X
± t(n-2;alfa/2)*s*√(1/n + (X-aveX)²/ΣX_i²)
gama = 1 –*alfa*

skutečná přímka vztahu není obsažena v pásu spolehlivosti s rizikem alfa

Predikční pás pro jednotlivá pozorování (odhad *Y* pro dané známé *X*)

predikční pás – kriterium přesnosti predikce

predikční pás pro jednotlivé pozorování:

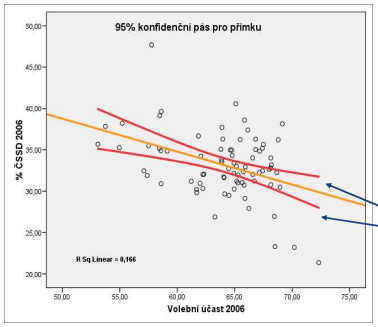
odhad Y = a + bX

$$\pm t(n-2;alfa/2)*s*\sqrt{(1 + 1/n + (X-aveX)^2/\Sigma x_i^2)}$$

gama = 1 – *alfa*

predikovaná hodnota je mimo v hranice pásu s rizikem alfa

Regresní pás spolehlivosti



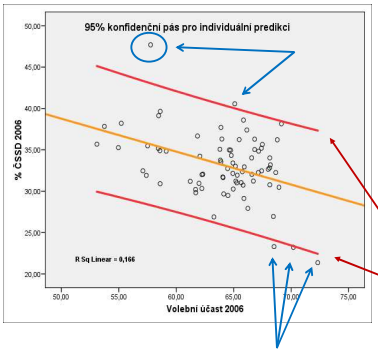
intervaly spolehlivosti pro parametry:

a = (45.5; 71.89)

b = (-0.604; -0.193)

pás spolehlivosti pro regresní přímku

Predikční pás pro jednotlivá pozorování



intervaly spolehlivosti pro parametry:

a = (-5.09; 6.31)

b = (0.169; 0.266)

pás predikce pro jednotlivá X

Chyba rovnice

- náhodné vlivy
  - při měření, zjišťování
- náhodné vlivy
  - při chování
- tvar funkce
- vlivy chybějících proměnných

regresní rovnice vyhlazuje tyto vlivy a očišťuje vztah o nepodstatné a nahodilé

Rozptyl reziduí - *var ε*

nevychýlený odhad chybové variance – kriterium predikční přesnosti

$s_r^2 = ESS/(n-2) = \sum(Y_i - \text{očekávané } Y_i)^2/(n-2)$

směrodatná odchylka s_r

dosahuje minimální možné hodnoty pro lineární model – je to kriterium pro výběr přímky

Předpoklady modelu

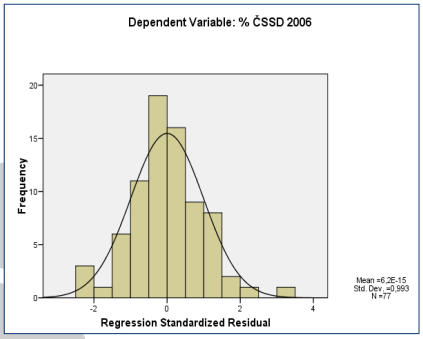
pro odhad metodou nejmenších čtverců:

- **předpoklad tvaru – linearita a aditivita chyby**
- **nezávislost reziduí na prediktorech**

pro testování:

- **nezávislost reziduí mezi sebou**
- **homoscedasticita – rozptyl reziduí je konstantní vzhledem k *X***
- **normalita reziduí**

Normalita reziduí



Vychýlená pozorování

- box plot standardizovaná rezidua
- standardizovaná rezidua (pokud jsou normálně rozložena) :

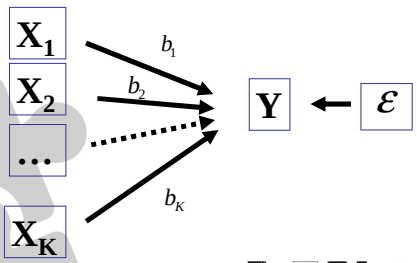
  hodnoty přes ±3.00 (or ±2.00, or ±c)

- (seznamy extrémních příkladů)
- vizuální analýza – histogram reziduí

- Cookova distance (prah = 4/(n-p))
- Leverage (práh = 2p/n; p = počet odhadovaných parametrů)
- dfbeta
- dfFit

Mnohorozměrná lineární regresní analýza

schéma lineárního regresního modelu:



Mnohorozměrná lineární regresní analýza

- je nalezení **predikční rovnice regresního modelu:**

$$Y = f(X_1, X_2, ..., X_K) + \varepsilon$$

$$E(Y) = \tilde{Y} = f(X_1, X_2, ..., X_K)$$

- hodnoty *Y* predikujeme pomocí proměnných $X_1, ..., X_K$ až na chybu

Mnohorozměrná lineární regresní analýza

- **lineární (nejčastější tvar) regrese**

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_K X_K + e$$

převodní koeficient X_1 na Y

odhady z dat

chyba rovnice šum

Mnohorozměrná lineární regrese: rovnice s *K* nezávislými proměnnými

$$Y = a + b_1 X_1 + b_2 X_2 + ... + b_K X_K + \varepsilon$$

$$Y = \tilde{Y} + \varepsilon$$

$$\tilde{Y} = a + b_1 X_1 + b_2 X_2 + ... + b_K X_K$$

hodnota Y, kterou očekáváme pro dané hodnotu X_k

$$\tilde{Y} = E(Y / X_1, X_2, ..., X_K)$$

Y = predikční model + residuum

predikční model = lineární kombinace prediktorů + konstanta

Vlastnosti rovnice

- *Y* je číselná proměnná, *X* jsou číselné proměnné
- koeficienty *beta* jsou modelové hodnoty, koeficienty b jsou jejich odhady (pozor na dvojznačnost značení !!!)

- b_k je převodní koeficient X_k na *Y*
  nazývá se *parciální regresní koeficient*
  b_k = přírůstek *Y* při jednotkové změně X_k, jsou-li ostatní *X* beze změny
  = čistý vliv X_k na *Y*, parciální, částečný
  (vliv očištěný od korelací X_k s jinými prediktory)

- obvyklé kriterium hledání modelu je *metoda nejmenších čtverců*:

$$\sum \varepsilon^2 \rightarrow \min$$

Mnohorozměrná lineární regrese: R^2 a R

Koeficient determinace = podíl/procento z var *Y*, které je vysvětleno množinou prediktorů pomocí lineárního modelu

$$R^2 = \text{var}(\tilde{Y}) / \text{var}(Y)$$
$$= MSS / TSS$$

$$R^2 = (\text{var}(Y) - \text{var}(\varepsilon)) / \text{var}(Y)$$

Koeficient vícenásobné korelace R = korelační koeficient mezi *Y* a hodnotami predikcí/očekávání odvozených z modelové funkce (tj. z nalezené lineární kombinace nezávislých proměnných *X*).

Lineární kombinace modelu **maximalizuje** korelační koeficient s *Y*.

$$R = r(Y, a + b_1X_1 + b_2X_2 + \ldots + b_KX_K)$$

Testování významnosti modelu

F - test – kriterium pro zjištění existence vztahu
kriterium toho, zda v model obsahuje významnou vztahovou informaci mezi množinou nezávislých proměnných a *Y*
F(k, n-k-1) = (ESS/k)/(MSS/n-k-1)
dosažená významnost F = alfa*
n = počet případů
k = počet regresních koeficientů

$$F = R^2 / (1 - R^2) * (n-k-1)/k$$

n = počet případů

k = počet regresních koeficientů

Kolinearita

$$\hat{Y} = a + b_1 X_1 + b_2 X_2$$

- **kolinearita** – vysoká korelace mezi X_1 a X_2 (obecně vysoká korelovanost mezi nezávislými proměnnými) vede **k nestabilitě** odhadu koeficientů
- je **obtížné separovat** vlivy vysoce korelovaných proměnných
- přelévání dat mezi X_1 a X_2 pro nová pozorování a pro jiné datové soubory k témuž problému a z toho plynoucí změny v b_1 a v b_2

- **singularita** – kompletní korelovanost mezi skupinou prediktorů, tj. jeden lze plně vyjádřit jako lineární kombinaci ostatních
  - krajní případ kolinearity – ***regresi nemůžeme počítat, jednu proměnnou je nutno vynechat***

Metody pro automatický výběr prediktorů

metody výběru prediktorů:

- **prediktory jsou určeny - ENTER – všechny vstoupí do rovnice (rozhodnutí uživatele)**
1. **metoda FORWARD – postupné zařazování prediktorů**
2. **metoda BACKWARD – postupné vyřazování prediktorů**
3. **metoda STEPWISE – kombinace obou**
- **postupný vstup určených bloků proměnných (prediktorů)**

Prokládání křivky

Prokládání křivky v IBM SPSS Statistics

Obecný model vztahu závislé proměnné *y* na nezávislé proměnné *x* :

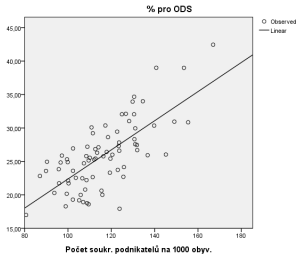
$$y = f(x;b) + \varepsilon$$

b … vektor neznámých parametrů

ε … náhodná chyba

- modely **vycházející z jednoduché lineární regrese**: funkce jsou lineární v parametrech nebo je lze na funkce lineární v parametrech převést vhodnou transformací (např. logaritmováním)
- vektor b odhadujeme **metodou nejmenších čtverců**, tj. na základě požadavku, aby součet čtverců chyb (reziduí) byl co nejmenší
- procedura *Curve Estimation* v *IBM SPSS Statistics Base* nabízí celkem 11 takových modelů

Model: Linear (přímka)



rovnice:

$$y = b_0 + b_1 x + \varepsilon$$

b_0: posunutí ve směru osy y

b_1: sklon přímky

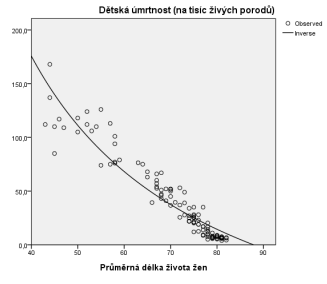
Model Summary and Parameter Estimates

Dependent Variable:% pro ODS

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	,515	79,696	1	75	,000	,609	,218

The independent variable is Počet soukr. podnikatelů na 1000 obyv..

Model: Inverse (hyperbola)



rovnice:

$$y = b_0 + \frac{b_1}{x} + \varepsilon$$

b_0: posunutí ve směru osy y

b_1: tvar křivky

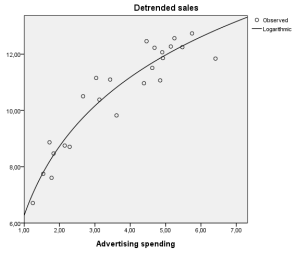
Model Summary and Parameter Estimates

Dependent Variable:Dětská úmrtnost (na tisíc živých porodů)

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Inverse	,877	762,560	1	107	,000	-146,938	12908,889

The independent variable is Průměrná délka života žen.

Model: Logarithmic (logaritmus)



rovnice:

$$y = b_0 + b_1 \ln(x) + \varepsilon$$

posunutí ve směru osy y

tvar křivky

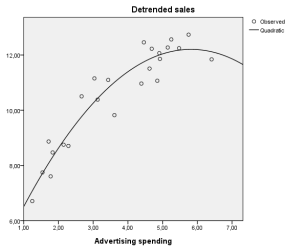
Model Summary and Parameter Estimates

Dependent Variable:Detrended sales

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Logarithmic	,901	199,831	1	22	,000	6,274	3,539

The independent variable is Advertising spending.

Model: Quadratic (parabola)



rovnice:

$$y = b_0 + b_1 x + b_2 x^2 + \varepsilon$$

souřadnice vrcholu:

$$V = \left[-\frac{b_1}{2b_2}, b_0 - \frac{b_1^2}{4b_2} \right]$$

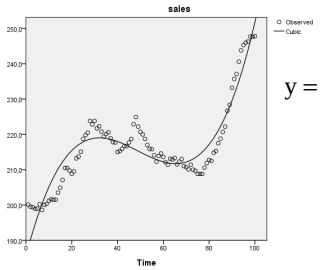
Model Summary and Parameter Estimates

Dependent Variable:Detrended sales

Equation	Model Summary					Parameter Estimates		
	R Square	F	df1	df2	Sig.	Constant	b1	b2
Quadratic	,908	104,213	2	21	,000	3,903	2,954	-,245

The independent variable is Advertising spending.

Model: Cubic (kubická křivka)



rovnice:

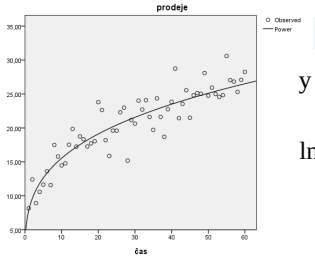
$$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3 + \varepsilon$$

Model Summary and Parameter Estimates

Dependent Variable:sales

Equation	Model Summary					Parameter Estimates			
	R Square	F	df1	df2	Sig.	Constant	b1	b2	b3
Cubic	,877	227,800	3	96	,000	185,507	2,495	-,058	,000

Model: Power (obecná mocnina)



rovnice:

$$y = b_0 x^{b_1} * (e^{\varepsilon})$$

nebo

$$\ln(y) = \ln(b_0) + b_1 \ln(x) + \varepsilon$$

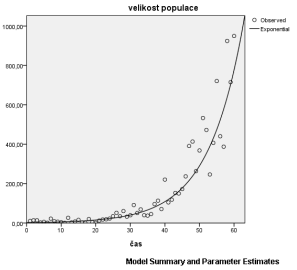
Model Summary and Parameter Estimates

Dependent Variable:prodeje

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Power	,854	340,453	1	58	,000	7,832	,298

The independent variable is čas.

Model: Exponential (exponenciála)



rovnice:

$$y = b_0 \exp(b_1 x) * (e^{\varepsilon})$$

nebo

$$\ln(y) = \ln(b_0) + b_1 x + \varepsilon$$

obecná exponenciála, růstová křivka a exponenciála se od sebe neliší typem průběhu křivky, ale pouze numerickými hodnotami a významem odhadovaných parametrů.

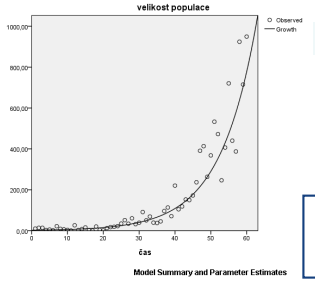
Model Summary and Parameter Estimates

Dependent Variable:velikost populace

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Exponential	,787	214,095	1	58	,000	2,105	,099

The independent variable is čas.

Model: Growth (růstová křivka)



rovnice:

$$y = \exp(b_0 + b_1 x) * (e^{\varepsilon})$$

nebo

$$\ln(y) = b_0 + b_1 x + \varepsilon$$

obecná exponenciála, růstová křivka a exponenciála se od sebe neliší typem průběhu křivky, ale pouze numerickými hodnotami a významem odhadovaných parametrů.

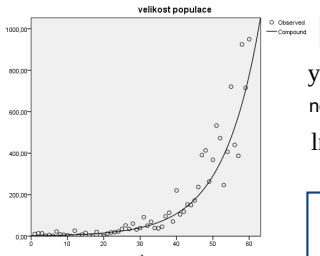
Model Summary and Parameter Estimates

Dependent Variable:velikost populace

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Growth	,787	214,095	1	58	,000	,744	,099

The independent variable is čas.

Model: Compound (obecná exponenciála)



Model Summary and Parameter Estimates

Dependent Variable:velikost populace

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Compound	,787	214,095	1	58	,000	2,105	1,104

The independent variable is čas.

rovnice:

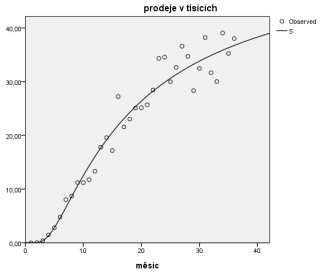
$$y = b_0 b_1^x * (e^{\varepsilon})$$

nebo

$$\ln(y) = \ln(b_0) + \ln(b_1)x + \varepsilon$$

obecná exponenciála, růstová křivka a exponenciála se od sebe neliší typem průběhu křivky, ale pouze numerickými hodnotami a významem odhadovaných parametrů.

Model: S (S-křivka)



Model Summary and Parameter Estimates

Dependent Variable:prodeje v tisících

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
S	,998	21324,039	1	34	,000	4,019	-14,915

The independent variable is měsíc.

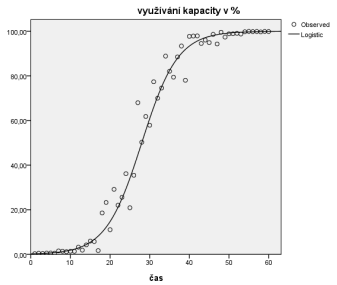
rovnice:

$$y = \exp(b_0 + \frac{b_1}{x}) * (e^{\varepsilon})$$

nebo

$$\ln(y) = b_0 + \frac{b_1}{x} + \varepsilon$$

Model: Logistic (logistická křivka)



Model Summary and Parameter Estimates

Dependent Variable:využívání kapacity v %

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Logistic	,973	2093,795	1	58	,000	4,744	,802

The independent variable is čas.

rovnice:

$$y = \frac{1}{\frac{1}{u} + b_0 b_1^x * (e^{\varepsilon})}$$

nebo

$$\ln(\frac{1}{y} - \frac{1}{u}) =$$

$$= \ln(b_0) + \ln(b_1)x + \varepsilon$$

parametr u vyjadřuje hodnotu, kterou je funkce omezena shora (zde 100%)