

Analýza datového souboru Studenti2004

Vzorové řešení včetně popisu, jak pracovat se s programem Statistica 6.0
(text v rámečcích)

Úkoly:

1. Spočítejte popisné (deskriptivní) statistiky
2. Záviseí výška na pohlaví? Sestavte model analýzy rozptylu, ověřte předpoklady, testujte zda očekávaná výška závisí na pohlaví.
3. Modelujte závislost výšky na hmotnosti a obvodu pasu, ověřte předpoklady, zjednodušte model.

0. Popis datového souboru

Studenti, kteří navštívili 1.12.2004 úvodní přednášku Základy stochastického modelování byli požádáni, aby uvedli své pohlaví a některé tělesné rozměry. Získané údaje jsou uloženy v datovém souboru Studenti2004.sta. Počet studentů byl 22 (8 mužů, 14 žen).

Soubor obsahuje následující veličiny

pohlavi - kategoriální (dichotomická) proměnná, kódování: F,M

pohlavi1 - kategoriální (dichotomická) proměnná, kódování: -1,1

vyska(m) - spojitá veličina, výška uvedená v metrech.

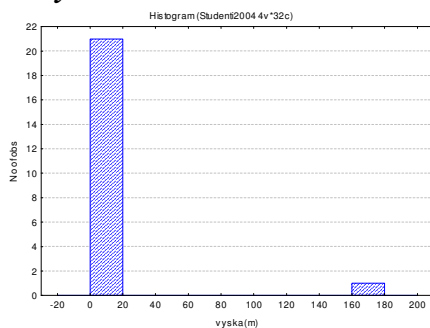
vaha(kg) - spojitá veličina, váha uvedená v kilogramech.

obvod_pasu(cm) - spojitá veličina, obvod pasu uvedený v centimetrech.

1.Popisné statistiky

1.histogramy, sumární statistiky

Výška



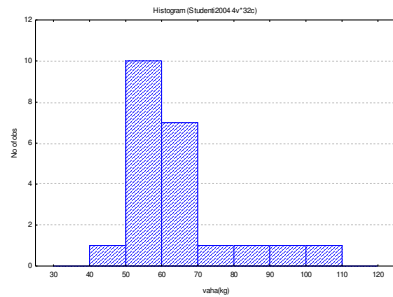
Graphs -> histograms

Histogram i maximum uvedené v tabulce deskriptivních statistik napovídají, že v datovém souboru je přítomno odlehle pozorování. Student č.22 nezadal svou výšku v metrech ale v centimetrech. **V pokračování příkladu budeme pracovat s opravenou výškou.**

Statistics -> Basic statistics and
tables -> Descriptive statistics

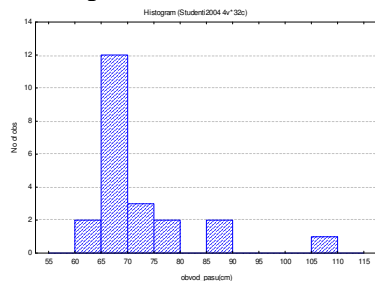
Variable	Descriptive Statistics (Studenti2004)							
	Valid N	Mean	Median	Minimum	Maximum	Percentile 10,00000	Percentile 90,00000	Std.Dev.
vyska(m)	22	9,496818	1,685000	1,560000	173,0000	1,610000	1,850000	36,51907

Váha



Variable	Descriptive Statistics (Studenti2004)							
	Valid N	Mean	Median	Minimum	Maximum	Percentile 10,00000	Percentile 90,00000	Std.Dev.
vaha(kg)	22	64,47727	60,50000	43,00000	110,0000	53,00000	83,00000	14,68760

Obvod pasu



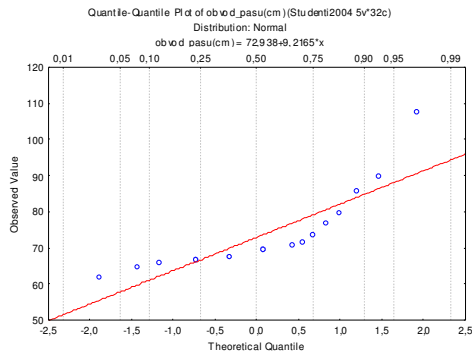
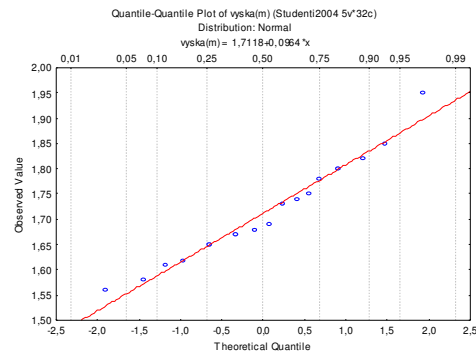
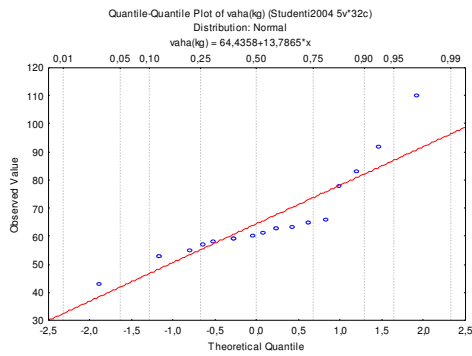
Variable	Descriptive Statistics (Studenti2004)							
	Valid N	Mean	Median	Minimum	Maximum	Percentile 10,00000	Percentile 90,00000	Std.Dev.
obvod_pasu(cm)	22	72,95455	70,00000	62,00000	108,0000	66,00000	86,00000	10,29784

b. Korelační matice

Variable	Correlations (Studenti2004)		
	Marked correlations are significant at $p < ,05000$ N=22 (Casewise deletion of missing data)		
	vyska(m)	vaha(kg)	obvod_pasu(cm)
vyska(m)	1,00	0,80	0,69
vaha(kg)	0,80	1,00	0,96
obvod_pasu(cm)	0,69	0,96	1,00

Statistics -> Basic statistics and tables
-> Correlation matrices -> Two lists(rect matrix)

Vidíme, že sledované veličiny jsou navzájem korelované. Pearsonův korelační koeficient obvodu pasu a váhy je dokonce 0,96.



Graphs-> 2D Graphs ->
Quantile-Quantile plots

Distribution: Normal

Pearsonův korelační koeficient je vhodný při normálním rozdělení náhodných veličin. Jak však vidíme na výše uvedených Q-Q plotech, předpoklad normality pro váhu a obvod pasu není splněn. Spočteme tedy navíc ještě Spearmanův korelační koeficient, který předpoklad normality nevyžaduje:

Spearman Rank Order Correlations (Studenti2004 5v'32c)			
MD pairwise deleted			
Marked correlations are significant at p < .05000			
Variable	vyska(m)	vaha(kg)	obvod_pasu(cm)
vyska(m)	1,000000	0,869467	0,679432
vaha(kg)	0,869467	1,000000	0,770080
obvod_pasu(cm)	0,679432	0,770080	1,000000

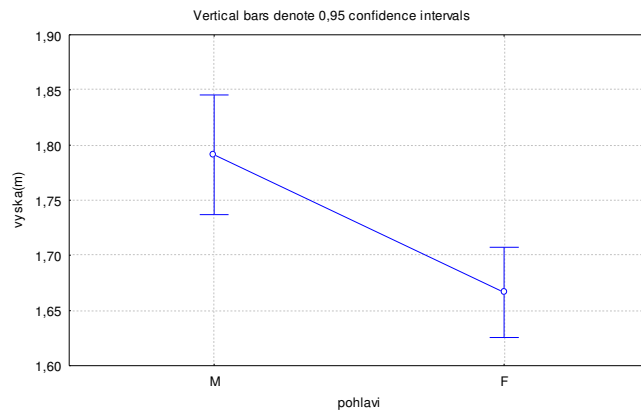
Statistics ->
Nonparametrics ->
Correlations
(Spearman...) ->
Spearman Rank R

2. Analýza rozptylu

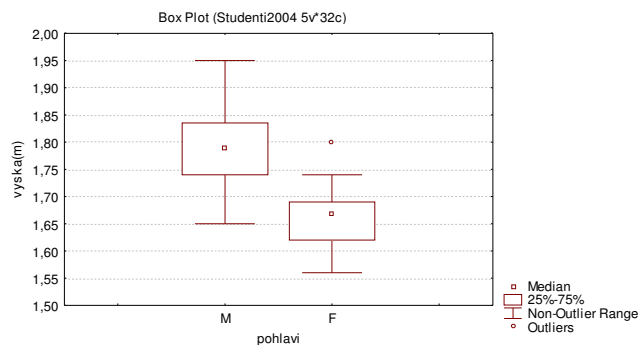
Statistics ->ANOVA ->
one-way ANOVA

dependent variable = vyska(m)
categorical predictor(factor)=pohlavi

Nejdříve si nakreslíme krabicové diagramy pro muže a ženy:



Statistics ->ANOVA ->
one-way ANOVA
->All_effects/Graphs
->OK



Graphs -> 2D Graphs ->
Box plots ... -> Quick ->
Box - Whiskers

*Dependent variable –
vyska(m)
Grouping variable - pohlavi*

Náš model analýzy rozptylu je

$$Y_{it} = \mu_i + \varepsilon_{it} = \mu + \alpha_i + \varepsilon_{it}$$

kde $i = 1, 2$ (M,F), $t = 1..22$. Ohady parametrů μ_i jsou uvedeny v následující tabulce (hodnoty Mean):

Effect	Descriptive Statistics (Studenti2004)						
	Level of Factor	N	vyska(m) Mean	vyska(m) Std.Dev.	vyska(m) Std.Err	vyska(m) -95,00%	vyska(m) +95,00%
Total		22	1,711818	0,094448	0,020136	1,669942	1,753694
pohlavi	M	8	1,791250	0,088711	0,031364	1,717086	1,865414
pohlavi	F	14	1,666429	0,063803	0,017052	1,629590	1,703268

Statistics ->ANOVA
-> one-way ANOVA
->Summary
->Cell statistics

Pro testování nulové hypotézy $H_0 : \alpha_i = 0$ sestrojíme následující tabulku analýzy rozptylu.

Effect	Univariate Results for Each DV (Studenti2004) Sigma-restricted parameterization Effective hypothesis decomposition				
	Degr. of Freedom	vyska(m) SS	vyska(m) MS	vyska(m) F	vyska(m) p
Intercept	1	60,86457	60,86457	11270,29	0,000000
pohlavi	1	0,07932	0,07932	14,69	0,001041
Error	20	0,10801	0,00540		
Total	21	0,18733			

Statistics ->ANOVA ->
one-way ANOVA
->All_effects

Zvolme hladinu testu 0,05. Vidíme, že p-hodnota je 0,00104, tedy je menší než hladina testu a proto na hladině 0,05 zamítáme hypotézu $H_0 : \alpha_i = 0$.

Lze tedy říci, že střední hodnota výšky závisí na pohlaví.

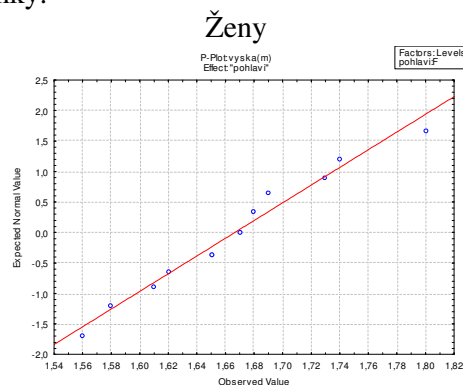
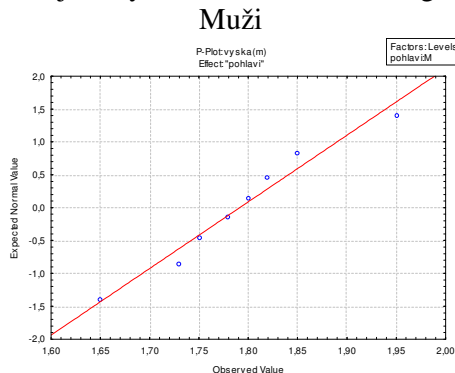
(Vzhledem k tomu, že máme pouze dvě skupiny – muže a ženy, můžeme testovat hypotézu o shodnosti středních hodnot výšek pro muže a ženy také pomocí t-testu. P-hodnota je 0,001041, tedy na hladině 0,05 zamítáme hypotézu o shodnosti středních hodnot výšek. Pro neparametrický Mann-Whitneyův test je p-hodnota 0,004149).

Ověření předpokladů pro jednoduché třídění (one-way ANOVA):

- nezávislá pozorování** - splněno
- normální rozdělení** – histogramy pro obě skupiny, normální diagramy
- rozptyly shodné pro obě skupiny** – Levenův test atd.

Statistics ->ANOVA -> one-way ANOVA -> more results -> Assumptions

Ad b. Normální diagramy (normal probability plot) – data pocházející z normálního rozdělení, mají body soustředěné kolem diagonální přímky.



Ad c. Hypotézu homogenity rozptylu nelze na hladině 0,05 zamítnout (p-hodnota Levenova testu je 0,4315). Předpoklad o konstantním rozptylu je splněn.

Levene's Test for Homogeneity of Variances (Studenti2004)				
Effect: "pohlaví"				
Degrees of freedom for all F's: 1, 20				
	MS Effect	MS Error	F	p
vyska(m)	0,001439	0,002233	0,644456	0,4315

Post-hoc testy

Hypotézu o shodnosti středních hodnot výšek pro obě pohlaví jsme zamítli. V případě, že bychom porovnávali více skupin a hypotézu zamítli, zajímalo by nás např. pro které dvě skupiny se střední hodnoty liší. V tom případě bychom použili některý z post-hoc testů.

Statistics ->ANOVA -> one-way ANOVA -> more results -> Post-hoc

3. Regrese

Statistics -> Multiple regression
->Quick->Summary: Regression results

dependent variable = vyska(m)
independent variables(list)=vaha(kg),obvod_pasu(cm)

Regression Summary for Dependent Variable: vyska(m) (Studenti2004)						
R= ,84164034 R2= ,70835846 Adjusted R2= ,67765935 F(2,19)=23,074 p<,00001 Std.Error of estimate: ,05362						
N=22	Beta	Std.Err. of Beta	B	Std.Err. of B	t(19)	p-level
Intercept			1,623051	0,128086	12,67155	0,000000
vaha(kg)	1,657191	0,429084	0,010656	0,002759	3,86216	0,001050
obvod_pasu(cm)	-0,894220	0,429084	-0,008201	0,003935	-2,08402	0,050897

Regesní model je tvaru: $Vyska = 1.62 + 0.01vaha - 0.008obvod$

Odhady parametru jsou v tabulce uvedeny ve sloupci B (ve sloupci Beta jsou uvedeny odhady parametrů v případě, že jsou všechny veličiny standartizovány – tj. mají nulovou střední hodnotu a jednotkový rozptyl). Parametr 0,01 u váhy je znamená, že střední přírůstek výšky při změně váhy o jeden kilogram a nezměněném obvodu pasu je 0,01 metru.

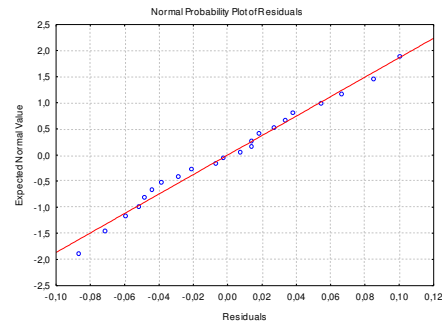
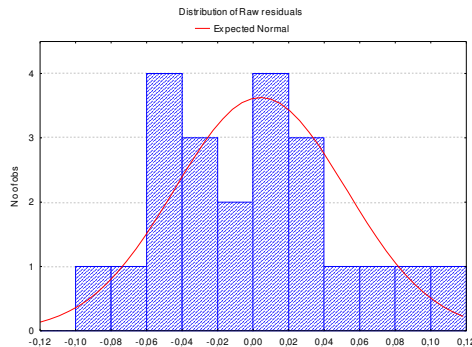
Koeficient determinace R^2 (R2) je roven 0,708. Tedy 70.8% variability v datech jsme schopni vysvětlit naším modelem.

Ověření předpokladů

- nezávislost pozorování** - splněno
- normalita reziduí** – histogram reziduí, normální diagram reziduí
- konstatní rozptyl** – bodový diagram, závislost reziduí na odhadnutých hodnotách.

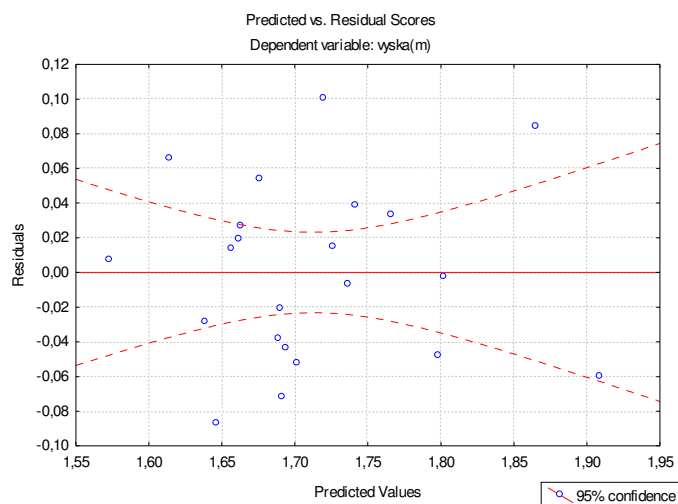
Statistics -> Multiple regression -
>Residuals/Assumptions/Prediction->perform residual
analysis->Residuals + Probability plots

Ad b. Normalita reziduí



Ad c. Konstatní rozptyl

Statistics -> Multiple regression -
 >Residuals/Assumptions/Prediction->perform residual analysis->
 Scatterplots -> Predicted vs.residuals



Vidíme, že rezidua v závislosti na výšce predikované modelem nevykazují žádný systematický trend (jsou náhodně rozmístěna kolem 0).

Další možnost jak ověřit homogenitu rozptylu, je rozdělit data do několika skupin dle hodnoty výšky predikované modelem, dále pak udělat Analýzu rozptylu a ověřit předpoklad shodnosti rozptylu ve všech skupinách – např. pomocí Levenova testu.

Zjednodušení modelu – závislost

Naší snahou je vytvořit co nejjednodušší matematický model, který bude obsahovat co nejméně nezávislých proměnných, ale přesto bude dostatečně dobře popisovat realitu.

Náš současný model je tvaru $Vyska_i = \beta_0 + \beta_1 vaha_i + \beta_2 obvod_i + \varepsilon_i, i = 1, \dots, 22$

Podívejme se opět na následující tabulku:

Regression Summary for Dependent Variable: vyska(m) (Student)						
R= ,84164034 R2= ,70835846 Adjusted R2= ,67765935 F(2,19)=23,074 p<,00001 Std.Error of estimate: ,05362						
N=22	Beta	Std.Err. of Beta	B	Std.Err. of B	t(19)	p-level
Intercept			1,623051	0,128086	12,67155	0,000000
vaha(kg)	1,657191	0,429084	0,010656	0,002759	3,86216	0,001050
obvod_pasu(cm)	-0,894220	0,429084	-0,008201	0,003935	-2,08402	0,050897

V posledním sloupci je uvedena p-hodnota t-testu, pomocí kterého testujeme nulovou hypotézu $H_0: \beta_i = 0$, $i = 0, 1, 2$.

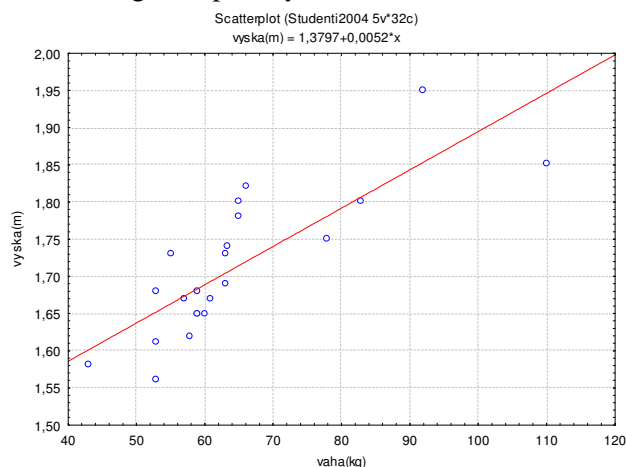
Nulovou hypotézu zamítáme na hladině 0,05 pro β_0 a β_1 . Pro β_2 nelze nulovou hypotézu zamítnout, což je možné interpretovat tak, že výška nezávisí na veličině obvod pasu.

(zajímavé však je, že v regresním modelu závislosti výšky pouze na obvodu pasu nulovou hypotézu zamítáme). Veličinu obvod pasu nezahrneme do zjednodušeného modelu.

Další důvod pro vypuštění této veličiny je, že váha a obvod pasu jsou silně korelované (korelační koeficient je 0,96 – tzv. problém multikolinearity), takže obvod pasu nepřináší do modelu žádnou novou informaci.

Sestrojíme zjednodušený regresní model, do kterého nezahrneme veličinu Obvod pasu.

Rovnice regresní přímky : $Vyska = 1.38 + 0.005vaha$



Graphs -> Scatterplots->
Variables

$X - vaha(kg)$
 $Y - vyska(m)$

Regression Summary for Dependent Variable: vyska(m) (Student)						
R= ,80105748 R2= ,64169308 Adjusted R2= ,62377774 F(1,20)=35,818 p<,00001 Std.Error of estimate: ,05793						
N=22	Beta	Std.Err. of Beta	B	Std.Err. of B	t(20)	p-level
Intercept			1,379687	0,056853	24,26741	0,000000
vaha(kg)	0,801057	0,133848	0,005151	0,000861	5,98482	0,000000

Koeficient determinace R^2 (R2) je 0,64 (což je trochu méně než pro model, do kterého jsme zahrnuli i obvod pasu). Nulovou hypotézu $H_0: \beta_1 = 0$ zamítáme na hladině 0,05. Samozřejmě i v tomto případě je nutné ověřit předpoklady.

Odlehlá pozorování (outliers)

Statistics -> Multiple regression -
>Residuals/Assumptions/Prediction->perform residual
analysis-> Outliers -> Casewise plot of outliers

Deleted residuals

Deleted residual je reziduum, které vypočteme následujícím způsobem: Vytvoříme regresní model ze všech pozorování vyjma jednoho konkrétního. Pak spočteme reziduum pro toto pozorování (skutečně naměřená hodnota výšky minus hodnota predikovaná modelem – který jsme vytvořili bez tohoto pozorování). Takto spočteme rezidua pro všechna pozorování. Jestliže se hodnota *deleted residual* pro konkrétní pozorování liší od standartizované hodnoty rezidua, pak toto pozorování může být odlehlé, neboť jeho vyloučení způsobilo velkou změnu regresní rovnice.

Deleted residuals						
Case	- ,19	,122
1	.	*
4	*
11	*
20	.	.	*	.	.	.
5	*
22	*
8	*
10	.	.	.	*	.	.
18	.	.	.	*	.	.
9	.	.	.	*	.	.
2	.	.	.	*	.	.
19	.	.	.	*	.	.
17	*	.
6	.	.	.	*	.	.
14	*	.
21	*	.
15	.	.	.	*	.	.
13	.	.	.	*	.	.
3	.	.	.	*	.	.
7	*	.
16	*	.
12	*	.
Minimum	.	*
Maximum	*
Mean	*	.
Median	*	.

V této tabulce jsou seřazena *deleted residua* podle své velikosti. Zdá se, že největší vliv na regresní přímku má první pozorování. Podívejte se na polohu tohoto pozorování do bodového diagramu na předchozí stránce nahoře (scatterplot).

Deleted residuals						
Case	-18,	171,
22	*
13	.	*
1	.	.	*	.	.	.
20	.	*
18	.	*
14	.	*
12	.	*
10	.	*
19	.	*
2	.	*
16	.	*
9	.	*
15	.	*
3	.	*
21	.	*
17	.	*
8	.	*
5	.	*
11	.	*
6	.	*
4	.	*
7	.	*
Minimum	.	*
Maximum	*
Mean	.	*
Median	.	*

Vraťme se, pro zajímavost, zpět k původním datům (22. student uvedl svou výšku v centimetrech místo v metrech.). Výsledky analýzy reziduí jsou uvedeny v tabulce nahoře. *Deleted residuum* pro pozorování č.22 je extrémně velké (171), toto pozorování je tedy odlehlé.

