Cvičení 6.: Analýza kovariance

V jistém zdravotnickém zařízení v centru Prahy byly zaznamenány údaje o 233 pacientech stomatologického oddělení, které měly přispět k objasnění vlivu některých chorob na stav chrupu. Pacienti byli rozděleni do pěti skupin podle onemocnění na některou z vybraných chorob, šestá skupina byla kontrolní.
Sledované choroby:
1. vředová choroba žaludku
2. hypertenze
3. tuberkulóza
4. srdeční onemocnění
5. cukrovka
Stav chrupu je vyjádřen pomocí indexu KPE, který se počítá podle vzorce:
KPE = ((počet zubů s kazem+počet zubů s plombou+počet extrahovaných zubů)/32)100%
U každého pacienta byl rovněž zaznamenán jeho věk. Data jsou uložena v souboru stav_chrupu.sta. Proměnná ID obsahuje kódy pro jednotlivé choroby a kontrolní skupinu, proměnná Y je index KPE a proměnná X obsahuje věk pacienta. Pro tento datový soubor proveďte analýzu kovariance.

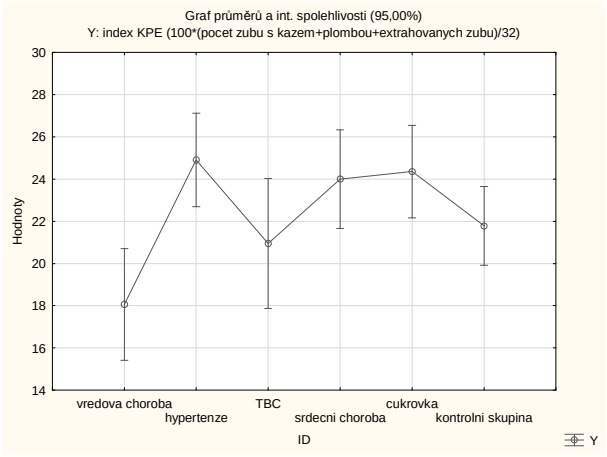
Řešení v systému STATISTICA

Nejprve vypočteme číselné charakteristiky indexu KPE v jednotlivých skupinách pacientů. Návod: Statistiky – Základní statistiky a tabulky – Rozklad & jednofakt. ANOVA – OK – Proměnné – Závislé – Y, Grupovací - ID – OK – Skupiny tabulek - zaškrtneme Rozptyly - Výpočet.

Rozkladová tabulka popisných statistik (stav_chrupu.sta)
N=233 (V seznamu záv. prom. nejsou ChD)

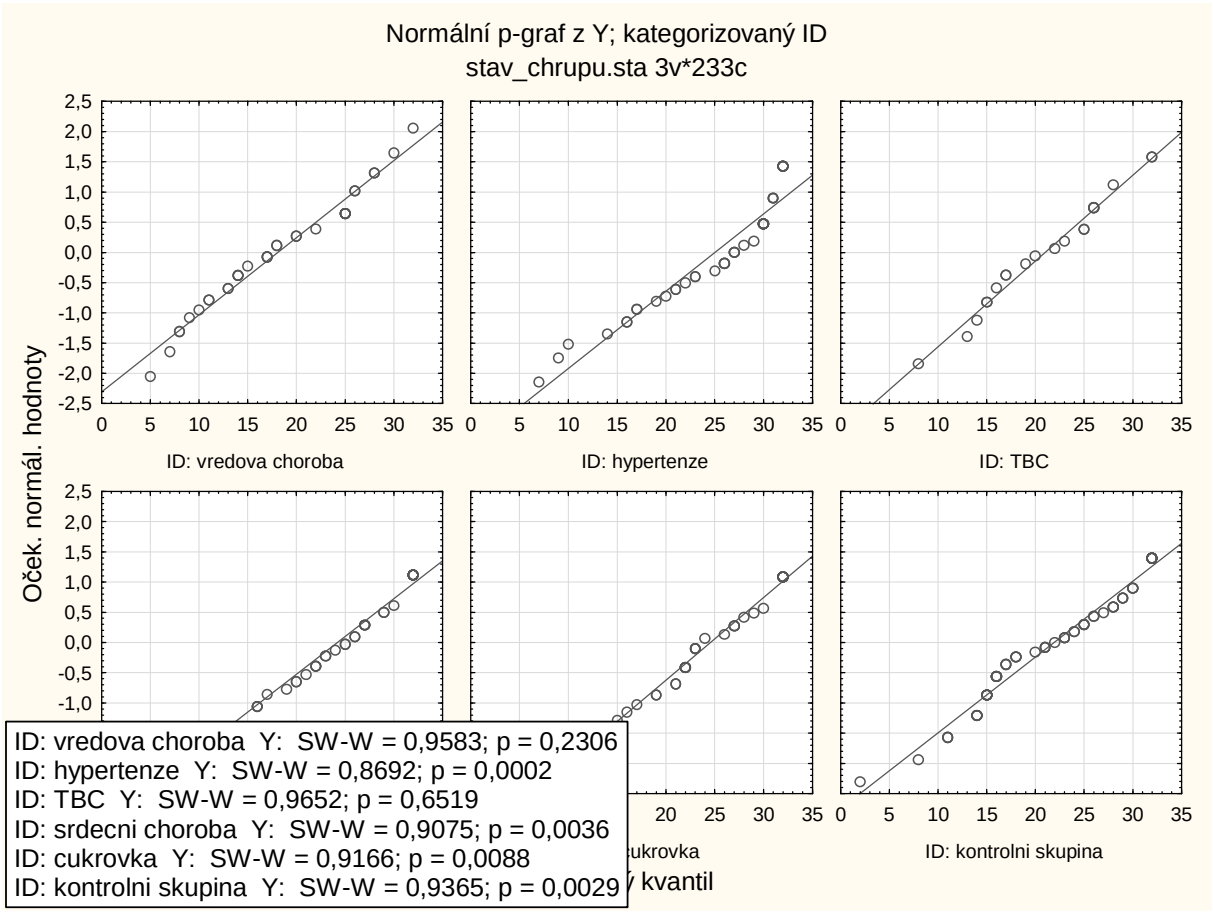
ID	Y průměr	Y N	Y Sm.odch.
vredova choroba	18,06061	33	7,457963
hypertenze	24,90244	41	7,013576
TBC	20,95000	20	6,565259
srdecni choroba	24,00000	39	7,211103
cukrovka	24,35135	37	6,587936
kontrolni skupina	21,77778	63	7,408462
Vš.skup.	22,51073	233	7,395893

Vidíme, že nejvyšší průměrnou hodnotu indexu KPE mají pacienti s hypertenzí, naopak nejnižší s vředovou chorobou. Směrodatné odchylky kolísají v úzkém rozmezí od 6,57 po 7,46.
Vytvoříme ještě graf průměrů s 95% intervaly spolehlivost.
Návrat do Statistiky dle skupin – na záložce Základní výsledky zvolíme Graf interakcí.



Z tohoto grafu je patrné, že vzhledem k nepřekrývajícím se intervalům spolehlivosti by metoda mnohonásobného porovnávání zřejmě odhalila rozdíl mezi třemi dvojicemi skupin: (vředová choroba, hypertenze), (vředová choroba, srdeční choroba) a (vředová choroba, cukrovka).

Normalitu hodnot veličiny Y v daných šesti náhodných výběrech posoudíme pomocí NP plotu a S-W testu.
Návod: Grafy – 2D Grafy – Normální pravděpodobnostní grafy – zaškrtneme S-W test a odškrtneme Neurčovat průměrnou pozici svázaných pozorování - Proměnné Y – OK – na záložce Kategorizovaný vybereme u Kategorie X Zapnuto, zaškrtneme Změnit proměnnou – Proměnná ID - OK – OK.

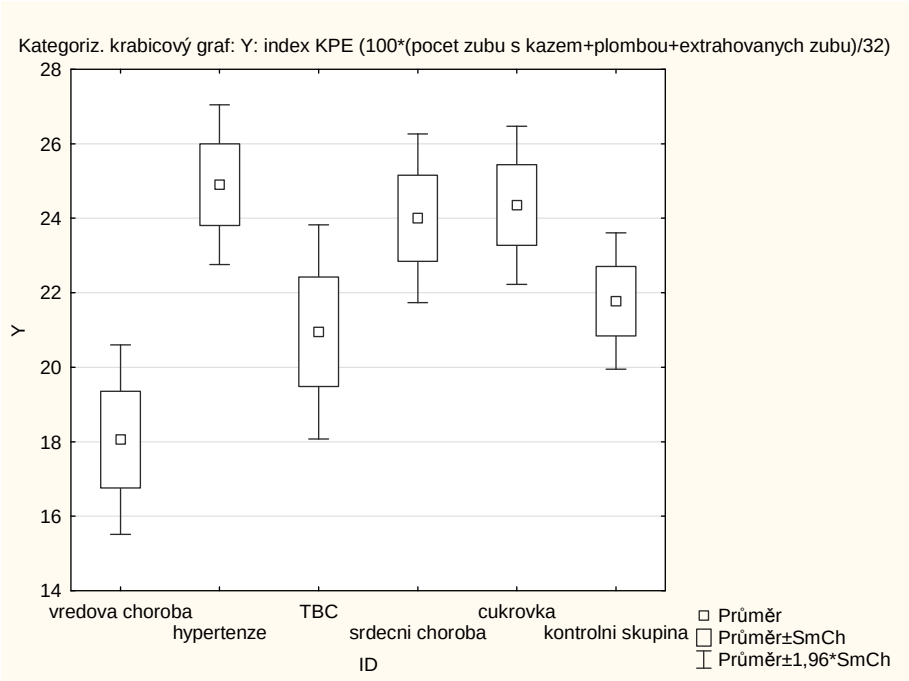


S výjimkou 1. a 3. skupiny S-W test zamítá hypotézu o normalitě na hladině významnosti 0,05, ale vzhledem k dostatečně velkým rozsahům skupin a jen mírným odchylkám od normality, které vidíme v NP plotech, budeme data považovat za normálně rozložená.

Nyní ověříme předpoklad shody rozptylů.
Návod: Návrat do Statistiky dle skupin - na záložce ANOVA&testy zaškrtneme Levenův test – Výpočet.

	Leveneův test homogenity rozpylů (stav_chrupu.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Y	33,68967	5	6,737935	3197,865	227	14,08751	0,478291	0,792276

Hypotézu o shodě rozptylů nezamítáme na hladině významnosti 0,05, protože p-hodnota je 0,7923.
Vykreslíme ještě krabicové diagramy.
Návod: Aktivujeme Statistiky dle skupin – na záložce Základní výsledky vybereme Kategoriz. krabicový graf.



Ověříme rovnoběžnost regresních přímek ve všech šesti skupinách pacientů

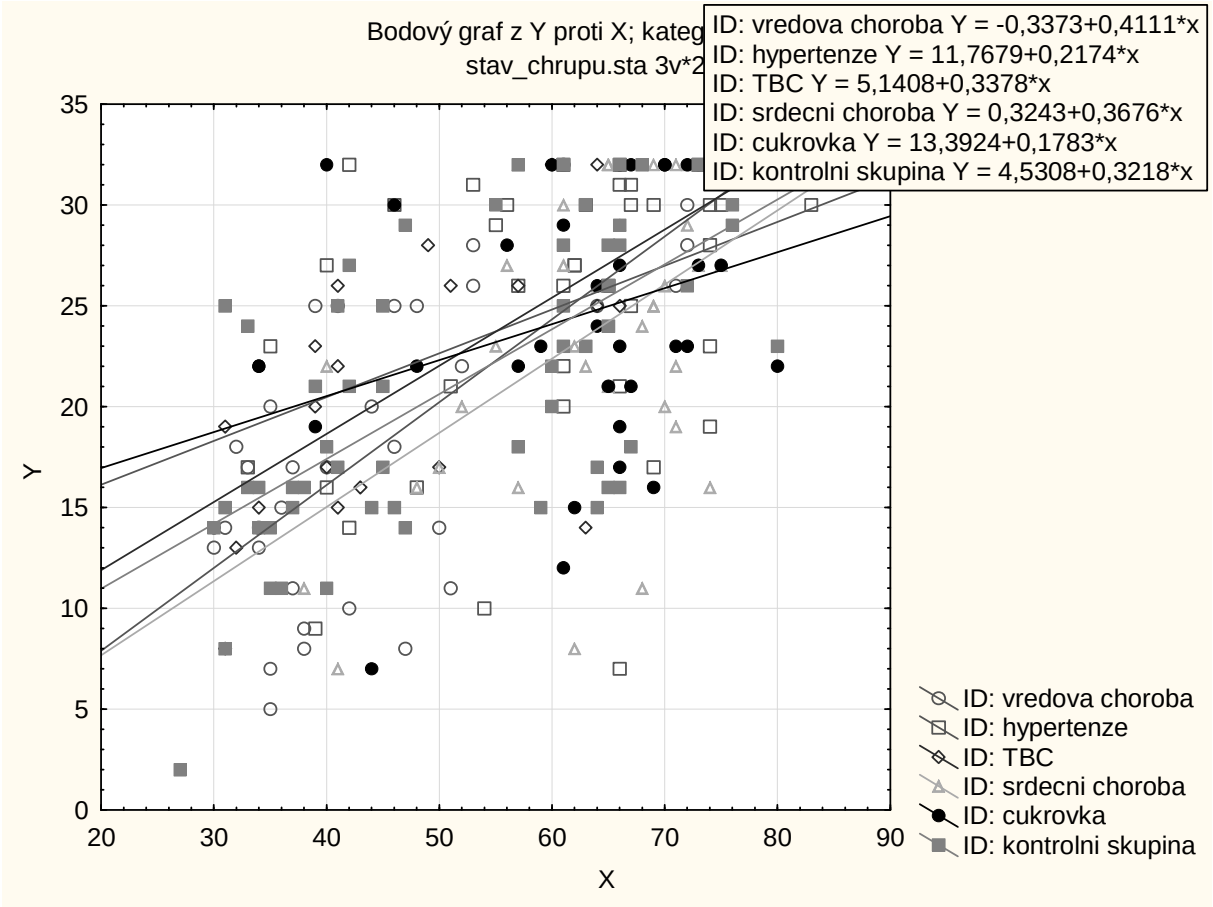
Návod: Statistiky – Pokročilé lineární/nelineární modely – Obecné lineární modely – Typ analýzy Obecné lineární modely – Metoda specifikace Rychlé nastavení – OK – Proměnné – Závisle proměnné: Y, Kateg. nez. prom.: ID, Spoj. nezáv. prom.: X – OK – Meziskupinové efekty – zaškrtneme Vlastní efekty meziskupinového schématu – přidáme ID, X a interakci ID*X – OK – na záložce Možnosti zvolíme Součet čtverců Typ II (parciální) – OK - Více výsledků – Jednorozm. výsledky

Efekt	Stupně volnosti	Y SČ	Y PČ	Y F	Y p
Jednorozm. výsledky pro každou záv. proměnnou (stav_chrupu.sta) Sigma-omezená parametrizace Dekompozice typu II					
Abs. člen	1	118068,8	118068,8	3405,851	0,000000
ID	5	280,3	56,1	1,617	0,156595
X	1	3614,2	3614,2	104,257	0,000000
ID*X	5	232,3	46,5	1,340	0,248431
Chyba	221	7661,3	34,7		
Celkem	232	12690,2			

Zajímá nás řádek ID*X. Příslušná p-hodnota je 0,2484, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě směrnic daných šesti regresních přímek.

Test můžeme ještě doplnit grafem:

Návod: Grafy – Bodové grafy – Proměnné – X, Y – OK – na záložce Kategorizovaný zapneme kategorii X – Změnit proměnnou – ID – OK – zaškrtneme Rozložení Přes sebe – OK.



Vidíme, že směrnice regresních přímek kolísají od 0,1783 u pacientů s cukrovkou až po 0,4111 u pacientů s vředovou chorobou žaludku. Rozdíly mezi směrnicemi však nejsou průkazné na hladině významnosti 0,05.

Po ověření předpokladů přistoupíme k provedení samotné analýzy kovariance, s jejíž pomocí budeme testovat jednak hypotézu o nulovosti regrese Y na X a jednak hypotézu o nevýznamnosti vlivu skupiny na veličinu Y.

Návod: Statistiky – Pokročilé lineární/nelineární modely – Obecné lineární modely – Typ analýzy Obecné lineární modely – Analýza kovariance – OK - Proměnné – Závisle proměnné: Y, Kateg. nez. prom.: ID, Spoj. nezáv. prom.: X – OK – na záložce Možnosti zvolíme Součet čtverců Typ II (parciální) – OK - Více výsledků – Jednorozm. výsledky

	Jednorozm. výsledky pro každou záv. proměnnou (stav_chrupu.sta) Sigma-omezená parametrizace Dekompozice typu II				
Efekt	Stupně volnosti	Y SČ	Y PČ	Y F	Y p
Abs. člen	1	118068,8	118068,8	3380,428	0,000000
X	1	3614,2	3614,2	103,479	0,000000
ID	5	157,2	31,4	0,900	0,481862
Chyba	226	7893,5	34,9		
Celkem	232	12690,2			

Vidíme, že p-hodnota na řádku X je blízká 0, tedy hypotézu o nulovosti regrese zamítáme na hladině významnosti 0,05.
Na řádku ID je p-hodnota 0,4819, tedy vliv skupiny na stav chrupu nelze prokázat na hladině významnosti 0,05.
Pokud bychom provedli analýzu rozptylu bez eliminace vlivu věku, dostali bychom tabulku

	Analýza rozptylu (stav_chrupu.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
Proměnná	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Y	1182,463	5	236,4927	11507,76	227	50,69498	4,665012	0,000450

Zde je p-hodnota 0,0005, tedy vliv skupiny na hodnoty indexu KPE se ukázal jako významný. Je to však klamný závěr, protože rozdíly mezi skupinami neplynou z chorob, ale jsou způsobeny nestejným věkem pacientů ve skupinách, což je vidět z tabulky popisných statistik veličiny X:

Rozkladová tabulka popisných statistik (stav_chrupu.sta) N=233 (V seznamu záv. prom. nejsou ChD)			
ID	X průměr	X N	X Sm.odch.
vredova choroba	44,75758	33	12,08312
hypertenze	60,41463	41	13,04411
TBC	46,80000	20	11,91903
srdecni choroba	64,41026	39	10,94416
cukrovka	61,45946	37	13,44866
kontrolni skupina	53,60317	63	15,81500
Vš.skup.	56,02146	233	14,91830

Provedené testy ještě doplníme o odhad regresního koeficientu β a výpočet upravených průměrů.

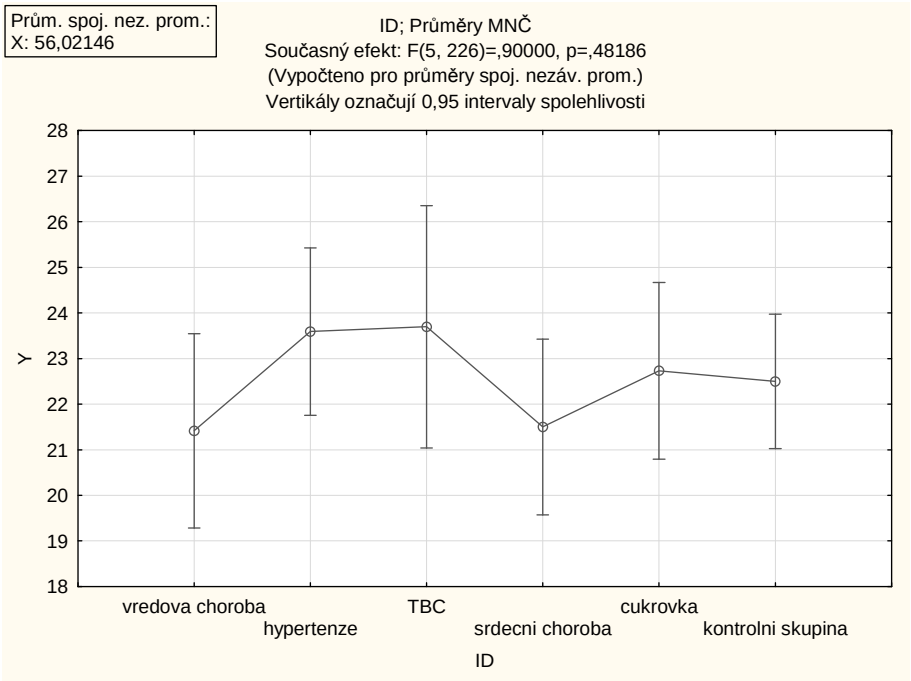
Odhad β: Návrat do GLM – na záložce Detaily zvolíme Koeficienty

Efekt	Odhady parametrů (stav_chrupu.sta) Sigma-omezená parametrizace Úroveň Efekt	Sloupec	Y Param.	Y Sm.Ch.	Y t	Y p	-95,00% LmtSpol.	+95,00% LmtSpol.	Y Beta (ß)	Y Sm.Ch. ß	-95,00% LmtSpol.	+95,00% LmtSpol.
Abs. člen		1	5,88830	1,668466	3,52917	0,000505	2,60056	9,176045				
X		2	0,29782	0,029278	10,17245	0,000000	0,24013	0,355516	0,600743	0,059056	0,484372	0,717113
ID	vredova choroba	3	-1,15758	0,983806	-1,17664	0,240577	-3,09619	0,781021	-0,098636	0,083829	-0,263822	0,066550
ID	hypertenze	4	1,02120	0,871171	1,17221	0,242345	-0,69546	2,737855	0,091519	0,078074	-0,062326	0,245364
ID	TBC	5	1,12353	1,180416	0,95181	0,342213	-1,20250	3,449556	0,086410	0,090786	-0,092484	0,265305
ID	srdecni choroba	6	-1,07123	0,914979	-1,17077	0,242923	-2,87421	0,731749	-0,094868	0,081031	-0,254541	0,064804
ID	cukrovka	7	0,15894	0,911339	0,17440	0,861708	-1,63687	1,954745	0,013903	0,079717	-0,143181	0,170987

Na řádku X, ve sloupci Y Param. najdeme odhad 0,2978.

Výpočet upravených průměrů a vykreslení grafu: Návrat do GLM – na záložce Průměry zvolíme Průměry MNČ a poté Graf

Č. buňky	ID; Průměry MNČ (stav_chrupu.sta) Současný efekt: F(5, 226)=,90000, p=,48186 (Vypočteno pro průměry spoj. nezáv. prom.) ID	Y Průměr	Y Sm.Ch.	Y -95,00%	Y +95,00%	N
1	vredova choroba	21,41526	1,080349	19,28642	23,54411	33
2	hypertenze	23,59405	0,931894	21,75773	25,43036	41
3	TBC	23,69637	1,348795	21,03855	26,35420	20
4	srdecni choroba	21,50161	0,977696	19,57505	23,42818	39
5	cukrovka	22,73178	0,984543	20,79172	24,67184	37
6	kontrolni skupina	22,49800	0,747939	21,02418	23,97183	63



Nejnižší upravený průměr věku pozorujeme u pacientů s vředovou chorobou, naopak nejvyšší u pacientů s cukrovkou.