Jednoduchá analýza kovariance

Osnova:

1. Výchozí situace v analýze kovariance
2. Příklady situací řešených analýzou kovariance
3. Model analýzy kovariance
4. Součty čtverců v analýze kovariance
5. Předpoklady v analýze kovariance a jejich ověřování
6. Testy hypotéz v analýze kovariance
7. Příklad na analýzu kovariance

1. Výchozí situace v analýze kovariance

Předpokládáme, že soubor n objektů se rozpadá do $r \geq 2$ skupin podle variant (úrovní) nějakého faktoru A, přičemž v i-té skupině je n_i objektů, $i = 1, \ldots, r$, $\sum_{i=1}^{r} n_i = n$. Na těchto objektech sledujeme hodnoty závisle proměnné veličiny Y a nezávisle proměnné veličiny X. Jejich hodnoty u j-tého objektu z i-té skupiny označíme y_{ij} a x_{ij}. Veličina X se nazývá doprovodná proměnná neboli kovariáta. Předpokládáme, že mezi Y a X existuje ve všech skupinách lineární závislost, přičemž regresní přímky, které tuto závislost modelují, jsou rovnoběžné.

Chceme zjistit, zda střední hodnoty proměnné Y jsou stejné ve všech skupinách odpovídajících úrovním faktoru A, pokud eliminujeme vliv kovariáty.

Úkoly tohoto typu řeší analýza kovariance (ANCOVA). Představuje spojení regresní analýzy a analýzy rozptylu (ANOVY). Při ANOVĚ uvažujeme v jednotlivých skupinách kolísání hodnot veličiny Y kolem průměru, zatímco při ANCOVĚ je to kolísání kolem regresní přímky.

2. Příklady situací řešených analýzou kovariance

Obor	Závisle proměnná veličina Y	Faktor A	Kovariáta X
Medicína	Systolický krevní tlak	Pohlaví pacienta	Věk pacienta
Pedagogika	Počet bodů v testu z matematiky	Metoda výuky	Skóre v testu čtenářské gramotnosti
Biologie	Počet zvukových pulsů, které vydá samec cvrčka za 1 s	Druh cvrčka	Teplota vnějšího prostředí
Pedologie	Obsah arzénu v půdě	Lokalita, odkud půda pochází	Obsah hliníku v půdě
Bankovnictví	Výše úvěru	Typ zaměstnání klienta	Věk klienta

3. Model analýzy kovariance

Vzájemný vztah závisle proměnné veličiny Y, kovariáty X a faktoru A popisuje model

$y_{ij} = \mu + \alpha_i + \beta\left(x_{ij} - \bar{x}\right) + \varepsilon_{ij}$, $i = 1,\ldots,r$, $j = 1,\ldots,n_i$, kde

y_{ij} … j-tá hodnota závisle proměnné veličiny Y v i-té skupině

μ … společná část střední hodnoty veličiny Y

α_i … efekt i-té úrovně faktoru A na veličinu Y

β … směrnice regresní přímky popisující závislost Y na X v každé skupině

x_{ij} … j-tá hodnota kovariáty X v i-té skupině

$\bar{x}$ … průměr hodnot kovariáty X

ε_{ij} ... j-tá hodnota náhodné odchylky od modelu v i-té skupině

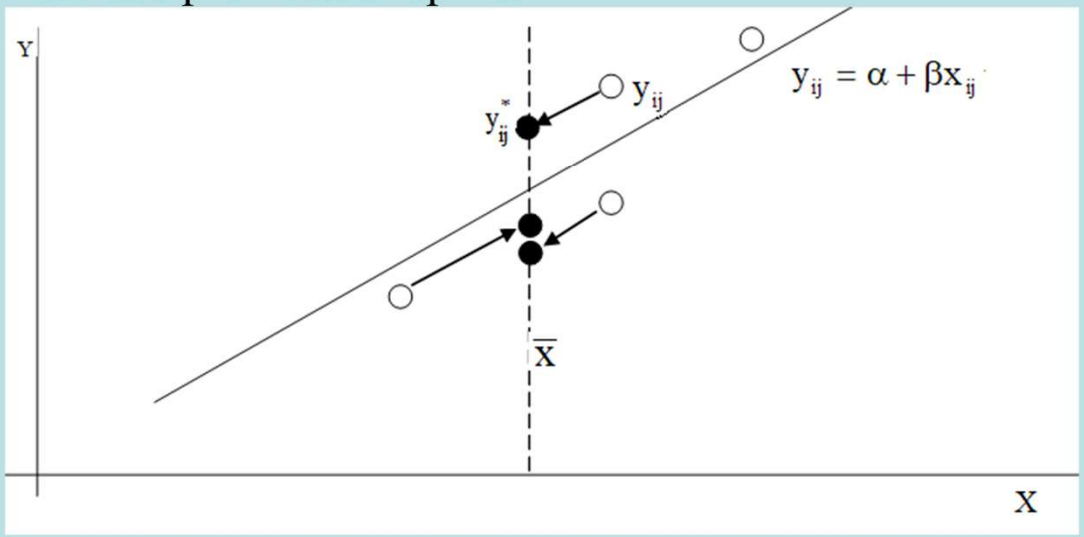
Předpokládáme, že $\varepsilon_{ij} \sim N\left(0, \sigma^2\right)$ a že platí reparametrizační rovnice $\sum_{i=1}^{r} n_i \alpha_i = 0$.

V modelu $y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$ člen α_i vyjadřuje analýzu rozptylu a člen $\beta(x_{ij} - \bar{x})$ vyjadřuje regresní část modelu.

Známe-li regresní parametr β, můžeme místo původních hodnot y_{ij} pracovat s upravenými hodnotami y_{ij}^*, kde $y_{ij}^* = y_{ij} - \beta(x_{ij} - \bar{x})$. Porovnáním této a původní rovnice získáme výsledný tvar modelu analýzy kovariance: $y_{ij}^* = \mu + \alpha_i + \varepsilon_{ij}$, $i = 1,\ldots,r, j = 1,\ldots,n_i$.

Odtud je vidět, že analýza kovariance je vlastně analýza rozptylu aplikovaná na upravené hodnoty závisle proměnné veličiny Y. Upravená hodnota je vlastně původní hodnota veličiny Y přepočítaná pomocí regresního vztahu mezi Y a X na průměrnou hodnotu kovariáty X.

Ilustrace pro i-tou skupinu:



4. Součty čtverců v analýze kovariance

Stejně jako v ANOVĚ je i v ANCOVĚ základem metody rozklad celkového součtu čtverců na složky. Jeho vyjádření je však složitější než v ANOVĚ, protože kromě různých součtů čtverců pro Y musíme uvažovat také součty čtverců pro kovariátu X a součty čtverců pro obě proměnné X a Y. Rozlišujeme tedy tři typy součtů čtverců:

a) pro celkovou variabilitu:

$S_T(y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2$ … variabilita jednotlivých pozorování veličiny Y kolem celkového průměru

$S_T(x) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$ … variabilita jednotlivých pozorování kovariáty X kolem celkového průměru

$S_T(x,y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})(y_{ij} - \bar{y})$ … společná variabilita jednotlivých pozorování veličin Y a X kolem jejich celkových průměrů

b) pro meziskupinovou variabilitu

$S_A(y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{r} n_i (\bar{y}_i - \bar{y})^2$ … variabilita skupinových průměrů veličiny Y kolem celkového průměru

$S_A(x) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{r} n_i (\bar{x}_i - \bar{x})^2$ … variabilita skupinových průměrů kovariáty X kolem celkového průměru

$S_A(x,y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y}) = \sum_{i=1}^{r} n_i (\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})$ … společná variabilita skupinových průměrů veličin Y a X kolem jejich celkových průměrů

c) pro vnitroskupinovou (reziduální) variabilitu

$S_E(y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ … variabilita jednotlivých pozorování veličiny Y kolem příslušných skupinových průměrů

$S_E(x) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ … variabilita jednotlivých pozorování kovariáty X kolem příslušných skupinových průměrů

$S_E(x,y) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$ … společná variabilita jednotlivých pozorování veličin Y a X kolem příslušných skupinových průměrů

Lze ukázat, že platí $S_T(.) = S_A(.) + S_E(.)$.

Parametr β v modelu $y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$ odhadneme metodou nejmenších čtverců:

$$\hat{\beta} = \frac{S_E(x,y)}{S_E(x)}.$$

Dále zavedeme následující označení, které využijeme při testování hypotéz:

$S_{E,i}(x,y) = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(y_{ij} - \bar{y}_i)$ … společná variabilita pozorování X a Y v i-té skupině kolem průměrů v i-té skupině,

$S_{E,i}(x) = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ … variabilita pozorování kovariáty X v i-té skupině kolem průměru v i-té skupině.

Platí tedy, že

$$S_E(x,y) = \sum_{i=1}^{r} S_{E,i}(x,y),$$

$$S_E(x) = \sum_{i=1}^{r} S_{E,i}(x).$$

5. Předpoklady v analýze kovariance a jejich ověřování

Analýzu kovariance lze korektně použít, jsou-li splněny následující předpoklady.

a) Jednotlivé skupiny jsou navzájem nezávislé.
Tento předpoklad musí být zajištěn vhodnou organizací experimentu, kterým se získají data.

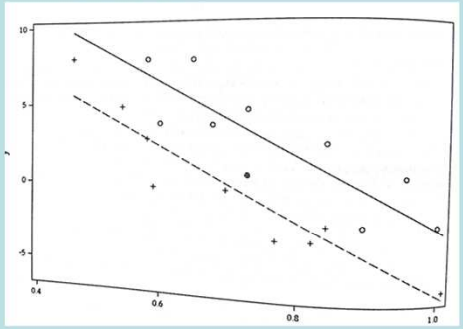
b) Hodnoty veličiny Y jsou ve všech skupinách normálně rozloženy.
Ověření provedeme pomocí diagnostických grafů a testů normality. Při větších rozsazích výběrů není ANCOVA citlivá na mírné porušení normality.

c) Hodnoty veličiny Y mají ve všech skupinách stejný rozptyl (předpoklad homoskedasticity).
Ověření provedeme pomocí krabicových diagramů a testů homogenity rozptylu. Mírné porušení homoskedasticity příliš nevadí.

Poznámka: První tři předpoklady jsou shodné s předpoklady pro ANOVU. Další předpoklad je specifický pro ANCOVU.

d) Regresní přímky modelující závislost Y na X jsou ve všech r skupinách rovnoběžné.

Ilustrace:



Nulová hypotéza H_0: $\beta_1 = \beta_2 = \ldots = \beta_r := \beta$, alternativní hypotéza H_1: aspoň jedna dvojice regresních koeficientů se liší.
Pro testování rovnoběžnosti regresních přímek použijeme statistiku

$$T_0 = \frac{n-2r}{r-1} \cdot \frac{\sum_{i=1}^{r} \frac{S_{E,i}(x,y)^2}{S_{E,i}(x)} - \frac{S_{E,i}(x,y)^2}{S_{E,i}(x)}}{S_E(y) - \sum_{i=1}^{r} \frac{S_{E,i}(x,y)^2}{S_{E,i}(x)}}$$

, která se za platnosti H_0 řídí rozložením F(r-1, n-2r). H_0 tedy zamítáme na hladině významnosti α, když $T_0 \geq F_{1-\alpha}(r-1, n-2r)$.
V případě, že hypotézu o rovnoběžnosti regresních přímek zamítneme, nelze použít standardní ANCOVU.

6. Testy hypotéz v analýze kovariance

Pomocí ANCOVY testujeme dvě hypotézy. První se týká regrese Y na X, druhá vztahu Y a faktoru A.

a) Test nulovosti regrese

Nulová hypotéza $H_0 : \beta = 0$, alternativní hypotéza $H_1 : \beta \neq 0$.

Testová statistika pro test nulovosti regrese má tvar

$$T_0 = (n - r - 1) \cdot \frac{\frac{S_E(x,y)^2}{S_E(x)}}{S_E(y) - \frac{S_E(x,y)^2}{S_E(x)}}$$

a za platnosti H_0 se řídí rozložením F(1, n-r-1). H_0 tedy zamítáme na hladině významnosti α, když $T_0 \geq F_{1-\alpha}(1, n - r - 1)$.

V případě, že hypotézu o nulovosti regrese nezamítneme, nemá smysl provádět ANCOVU, stačí ANOVA.

b) Test hypotézy o nevýznamnosti faktoru A

Nulová hypotéza H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_r = 0$, alternativní hypotéza H_1: aspoň jeden efekt faktoru A je nenulový.

Pro testování této hypotézy slouží statistika

$$T_0 = \frac{n-r-1}{r-1} \cdot \frac{S_T(y) - \frac{S_T(x,y)^2}{S_T(x)} - S_E(y) + \frac{S_E(x,y)^2}{S_E(x)}}{S_E(y) - \frac{S_E(x,y)^2}{S_E(x)}},$$

která se za platnosti H_0 řídí rozložením F(r-1, n-r-1). H_0 tedy zamítáme na hladině významnosti α, když $T_0 \geq F_{1-\alpha}(r-1, n-r-1)$.

Pokud zamítneme hypotézu o nevýznamnosti faktoru A, provedeme mnohonásobné porovnávání, abychom zjistili, které dvojice skupin se liší na dané hladině významnosti.

7. Příklad na analýzu kovariance

Popis situace: Chovatel ústřic chtěl zjistit, zda poloha ústřic v různé výšce vodního sloupce uměle ohřáté vody ovlivňuje jejich růst. Za tímto účelem provedl pilotní studii, která byla organizována takto:
Náhodně vybral 200 ústřic a náhodně je rozdělil do 20 sáčků po 10 ústřicích. V nádrži, kam přitéká chladicí voda z elektrárny, si vybral 5 lokací a do každé umístil 4 sáčky s ústřicemi. Lokace jsou následující:
1 – chladné dno,
2 – chladný povrch,
3 – teplé dno,
4 – teplý povrch,
5 – střední hloubka, střední teplota (lokace 5 je považována za kontrolní).
Experiment probíhal po dobu jednoho měsíce. Na jeho počátku a na konci byl každý sáček s ústřicemi zvážen a byla zaznamenána jeho hmotnost v gramech.
Roli faktoru A tedy hraje lokace (má 5 variant, v datovém souboru je označena jako proměnná ID), závisle proměnnou veličinou Y je koncová hmotnost sáčku s deseti ústřicemi a doprovodnou proměnnou – kovariátou X – je počáteční hmotnost.

Datový soubor:

	1 ID	2 Y	3 X
1	chladne dno	924,21	771,12
2	chladne dno	1037,61	907,2
3	chladne dno	1068,795	935,55
4	chladne dno	878,85	759,78
5	chladny povrch	958,23	810,81
6	chladny povrch	898,695	759,78
7	chladny povrch	870,345	751,275
8	chladny povrch	861,84	759,78
9	teple dno	997,92	810,81
10	teple dno	824,985	635,04
11	teple dno	819,315	657,72
12	teple dno	856,17	691,74
13	teply povrch	992,25	830,655
14	teply povrch	765,45	618,03
15	teply povrch	1031,94	859,005
16	teply povrch	864,675	688,905
17	kontrola	697,41	578,34
18	kontrola	663,39	555,66
19	kontrola	859,005	711,585
20	kontrola	618,03	513,135

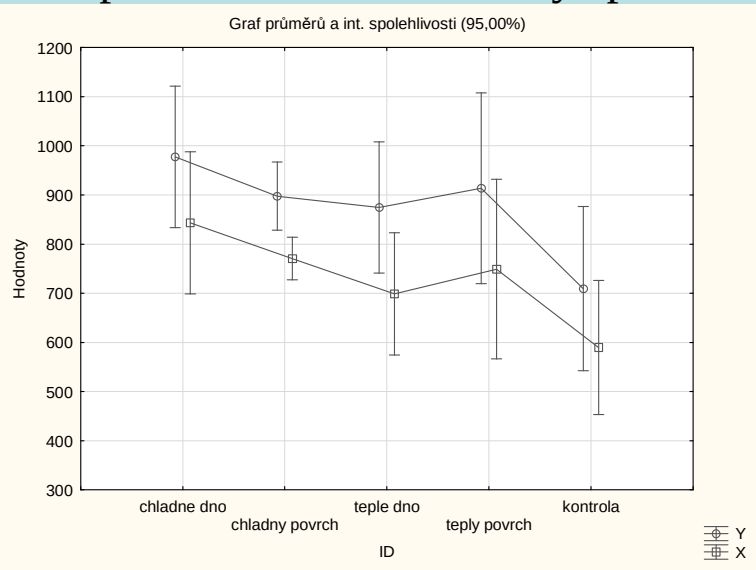
Úkolem je posoudit, zda lokace má vliv na koncovou hmotnost při eliminaci vlivu počáteční hmotnosti. Problém budeme řešit pomocí analýzy kovariance.

7.1. Výpočet číselných charakteristik datového souboru

Rozkladová tabulka popisných statistik (chov_ustric.sta)
N=20 (V seznamu záv. prom. nejsou ChD)

ID	Y průměr	Y N	Y Sm.odch.	X průměr	X N	X Sm.odch.
chladne dno	977,37	4	90,41	843,41	4	90,88
chladny povrch	897,28	4	43,58	770,41	4	27,23
teple dno	874,60	4	83,80	698,83	4	78,21
teply povrch	913,58	4	121,84	749,15	4	114,79
kontrola	709,46	4	104,87	589,68	4	85,65
Vš.skup.	874,46	20	123,18	730,30	20	114,47

Graf průměrů s 95% intervaly spolehlivosti

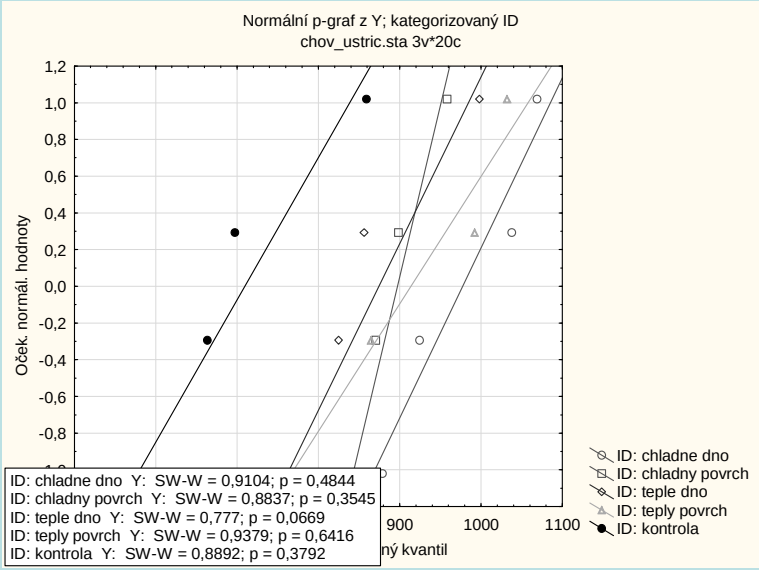


Nejvyšších hodnot nabývá hmotnost na chladném dně, naopak nejnižší hodnoty vykazuje kontrolní lokace.

7.2. Ověření předpokladů ANCOVY

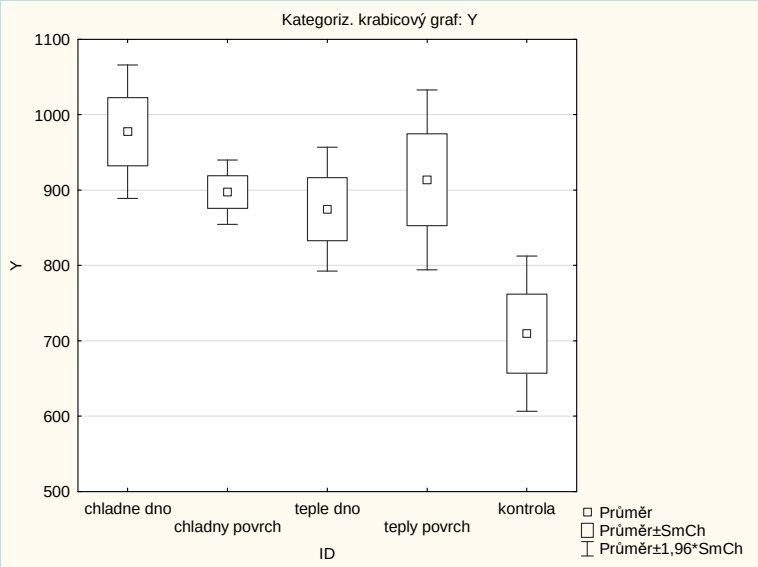
a) Nezávislost daných pěti náhodných výběrů – splněno, zajištěno organizací experimentu.

b) Normalita hodnot veličiny Y v daných pěti náhodných výběrech – posoudíme pomocí NP plotu a S-W testu.



Ani v jednom případě nezamítáme normalitu na hladině významnosti 0,05.

c) Homogenita rozptylu hodnot veličiny Y v daných pěti náhodných výběrech – posoudíme pomocí krabicového grafu a pomocí Levenova testu.



Proměnná	Leveneův test homogenity rozpylů (chov_ustric.sta) Označ. efekty jsou význ. na hlad. p < ,05000							
	SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Y	9733,08	4	2433,270	25517,69	15	1701,179	1,430343	0,272145

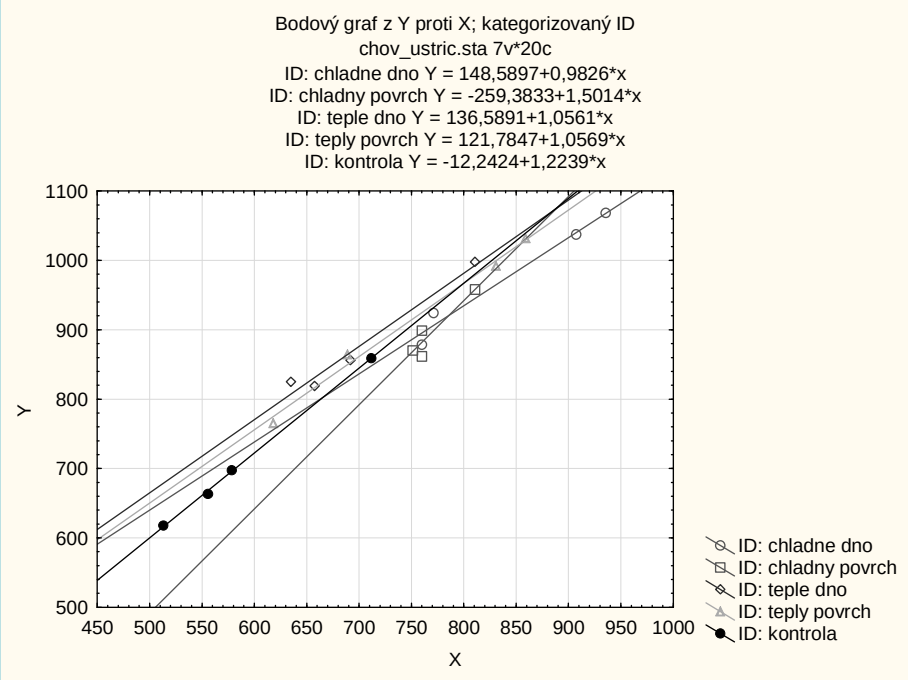
Na hladině významnosti 0,05 nezamítáme hypotézu o homogenitě rozptylu.

d) Rovnoběžnost regresních přímek ve všech pěti lokacích

Efekt	Stupně volnosti	Y SČ	Y PČ	Y F	Y p
Jednorozm. výsledky pro každou záv. proměnnou (chov_ustric.sta) Sigma-omezená parametrizace Dekompozice typu II					
Abs. člen	1	15293457	15293457	67142,62	0,000000
ID	4	1363	341	1,50	0,275182
X	1	125413	125413	550,60	0,000000
ID*X	4	1116	279	1,22	0,360175
Chyba	10	2278	228		
Celkem	19	288271			

Zajímá nás řádek ID*X. Příslušná p-hodnota je 0,3602, tedy na hladině významnosti 0,05 nezamítáme hypotézu o shodě směrnic daných pěti regresních přímek.

Test můžeme ještě doplnit grafem:



Vidíme, že směrnice regresních přímek popisujících závislost konečné hmotnosti na počáteční hmotnosti v jednotlivých lokacích se pohybují od 0,9826 na chladném dně po 1,5014 na chladném povrchu. Rozdíly směrnic regresních přímek však nejsou prokazatelné na hladině významnosti 0,05.

7.3. Provedení ANCOVY

Nyní budeme testovat dvě hypotézy. První se týká regrese Y na X a tvrdí, že ve všech pěti skupinách je směrnice regresní přímky nulová (test nulovosti regrese). Druhá se týká vztahu Y a faktoru ID a tvrdí, že faktor ID je nevýznamný.

Efekt	Jednorozm. výsledky pro každou záv. proměnnou (chov_ustric.sta) Sigma-omezená parametrizace Dekompozice typu II Stupně volnosti	Y SČ	Y PČ	Y F	Y p
Abs. člen	1	15293457	15293457	63092,26	0,000000
X	1	125413	125413	517,38	0,000000
ID	4	9716	2429	10,02	0,000482
Chyba	14	3394	242		
Celkem	19	288271			

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 0,000482, tedy vliv lokace na koncovou hmotnost ústřic je prokázán na hladině významnosti 0,05.

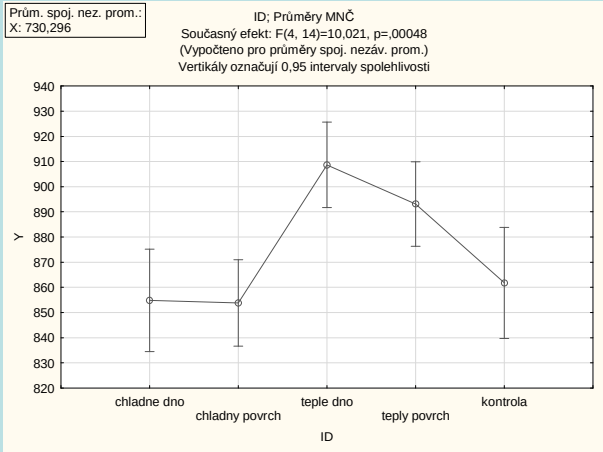
Upozornění: Při provádění ANCOVY je zapotřebí zvolit Součet čtverců Typ II (parciální), protože ne všechny úrovně faktoru jsou zastoupeny ve všech hodnotách kovariáty.

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

Efekt	Odhady parametrů (chov_ustric.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	83,4139	34,95089	2,38660	0,031671	8,4517	158,3761				
X		2	1,0832	0,04762	22,74608	0,000000	0,9810	1,1853	1,006667	0,044257	0,911745	1,101588
ID	chladne dno	3	-19,6150	8,80317	-2,22818	0,042776	-38,4959	-0,7341	-0,103332	0,046375	-0,202796	-0,003867
ID	chladny povrch	4	-20,6303	7,22004	-2,85736	0,012665	-36,1157	-5,1448	-0,108680	0,038035	-0,190257	-0,027103
ID	teple dno	5	34,2278	7,12217	4,80581	0,000279	18,9523	49,5033	0,180312	0,037519	0,099840	0,260783
ID	teply povrch	6	18,7021	7,02037	2,66397	0,018517	3,6449	33,7593	0,098522	0,036983	0,019201	0,177843

Na řádku X, ve sloupci Y Param. najdeme odhad 1,0832.

Č. buňky	ID; Průměry MNČ (chov_ustric.sta) Současný efekt: F(4, 14)=10,021, p=,00048 (Vypočteno pro průměry spoj. nezáv. prom.) ID	Y Průměr	Y Sm.Ch.	Y -95,00%	Y +95,00%	N
1	chladne dno	854,8407	9,46656	834,5370	875,1445	4
2	chladny povrch	853,8255	8,01554	836,6339	871,0171	4
3	teple dno	908,6835	7,92750	891,6808	925,6863	4
4	teply povrch	893,1578	7,83617	876,3509	909,9647	4
5	kontrola	861,7712	10,26834	839,7478	883,7946	4



Nejnižší upravený průměr konečné hmotnosti pozorujeme na lokaci chladný povrch, naopak nejvyšší na lokaci teplé dno.

Na závěr provedeme mnohonásobné porovnávání, abychom identifikovala dvojice lokací, které se liší na hladině významnosti 0,05.

Č. buňky	Tukeyův HSD test; proměnná Y (chov_ustric.sta) Přibližné pravděpodobnosti pro post hoc testy Chyba: Between MSE = 242,40, sv = 14,000 ID	{1} 977,37	{2} 897,28	{3} 874,60	{4} 913,58	{5} 709,46
1	chladne dno		0,000179	0,000151	0,000504	0,000151
2	chladny povrch	0,000179		0,289479	0,590163	0,000151
3	teple dno	0,000151	0,289479		0,022974	0,000151
4	teply povrch	0,000504	0,590163	0,022974		0,000151
5	kontrola	0,000151	0,000151	0,000151	0,000151	

7.4. Srovnání výsledků ANCOVY a ANOVY

Pokud budeme testovat hypotézu o nevýznamnosti vlivu lokace na konečnou hmotnost sáčku ústřic pomocí jednofaktorové ANOVY, bez eliminace vlivu počáteční hmotnosti, dostaneme tabulku

Proměnná	Analýza rozptylu (chov_ustric.sta) Označ. efekty jsou význ. na hlad. p < ,05000 SČ efekt	SV efekt	PČ efekt	SČ chyba	SV chyba	PČ chyba	F	p
Y	159464,2	4	39866,04	128806,6	15	8587,105	4,642547	0,012239

Testová statistika nabývá hodnoty 4,6426, odpovídající p-hodnota je 0,0122, tedy na hladině významnosti 0,05 považuje ANOVA vliv lokace za prokázaný. Při použití ANCOVY je však testová statistika 10,02 a p-hodnota pouze 0,0005.

Tukeyova metoda mnohonásobného porovnávání poskytne tabulku p-hodnot:

ID	Tukeyův HSD test; proměn.:Y (chov_ustric.sta) Označ. rozdíly jsou významné na hlad. p < ,05000 {1} M=977,37	{2} M=897,28	{3} M=874,60	{4} M=913,58	{5} M=709,46
chladne dno {1}		0,739248	0,537759	0,862937	0,007414
chladny povrch {2}	0,739248		0,996604	0,999099	0,075213
teple dno {3}	0,537759	0,996604		0,973812	0,138138
teply povrch {4}	0,862937	0,999099	0,973812		0,047617
kontrola {5}	0,007414	0,075213	0,138138	0,047617	

Vidíme, že se liší pouze dvojice lokací (chladné dno, kontrola) a (teplý povrch, kontrola), zatímco ANCOVA ukázala, že se neliší pouze dvojice lokací (chladný povrch, teplé dno) a (chladný povrch, teplý povrch).