Mnohonásobná lineární regrese a korelace

Mnohonásobná korelace

Mnohonásobná korelační závislost nám umožňuje sledovat, jak závisí proměnná *y* nejen na vysvětlující proměnné x_1, ale také na dalších proměnných x_2,x_3 …, x_k.

Koeficient párový

Koeficient vícenásobné (totální) korelace

Koeficient dílčí (parciální) korelace

Mnohonásobná korelace

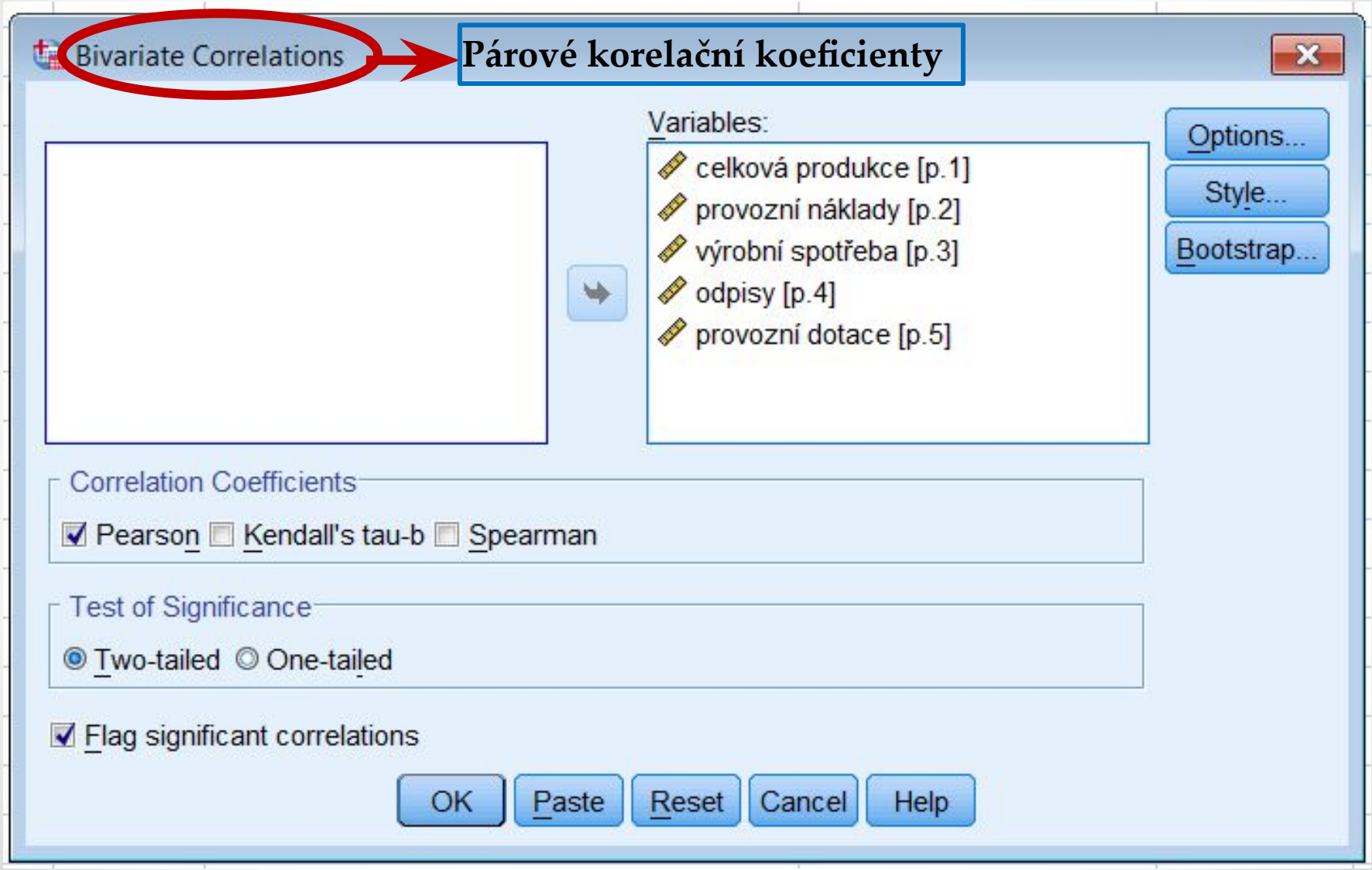
Sílu jednoduché lineární závislosti mezi jednou závisle proměnnou *y* a **jedou** vysvětlující proměnnou ***x*** udávají:

Párové korelační koeficienty

$$r_{yx_1} \qquad r_{yx_2} \qquad r_{x_1x_2}$$

$$-1 \leq r \leq 1$$

Mnohonásobná korelace



Párové korelační koeficienty

Correlations

		celková produkce	provozní náklady	výrobní spotřeba	odpisy	provozní dotace
celková produkce	Pearson Correlation	1	,857**	,851**	,293**	,107
	Sig. (2-tailed)		,000	,000	,000	,134
	N	197	197	197	197	197
provozní náklady	Pearson Correlation	,857**	1	,899**	,424**	,259**
	Sig. (2-tailed)	,000		,000	,000	,000
	N	197	197	197	197	197
výrobní spotřeba	Pearson Correlation	,851**	,899**	1	,080	,213**
	Sig. (2-tailed)	,000	,000		,265	,003
	N	197	197	197	197	197
odpisy	Pearson Correlation	,293**	,424**	,080	1	-,207**
	Sig. (2-tailed)	,000	,000	,265		,004
	N	197	197	197	197	197
provozní dotace	Pearson Correlation	,107	,259**	,213**	-,207**	1
	Sig. (2-tailed)	,134	,000	,003	,004	
	N	197	197	197	197	197

**. Correlation is significant at the 0.01 level (2-tailed).

Mnohonásobná korelace

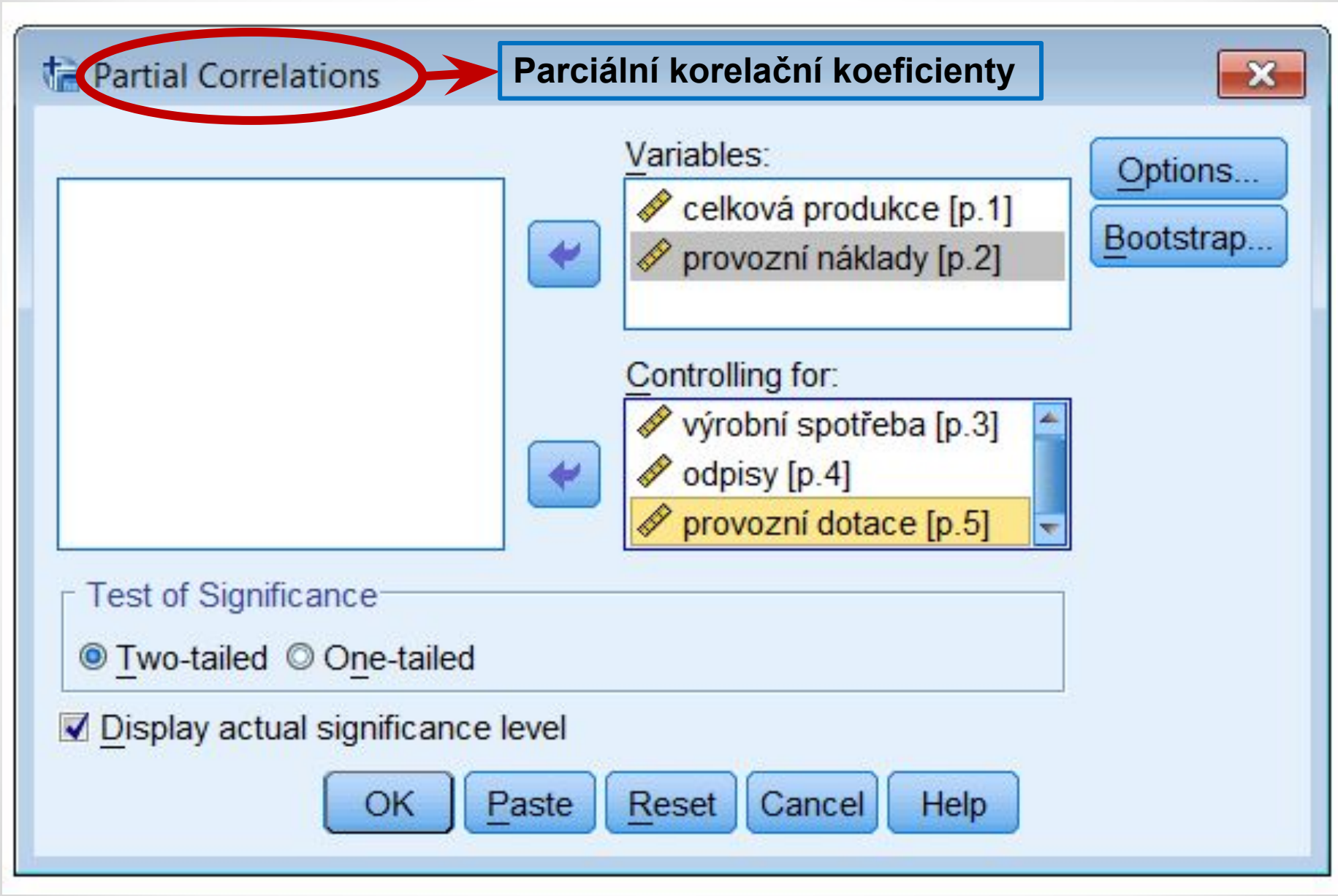
Koeficienty dílčí (parciální) korelace

charakterizuje sílu lineární závislosti mezi závisle proměnnou a jednou nezávisle proměnnou, jsou-li hodnoty zbývajících proměnných v modelu konstantní.

$$-1 \leq r \leq 1$$

$r_{yx_1 \cdot x_2}$ parciální korelační koeficient mezi y a x_1 s vyloučením vlivu x_2 (při konstantním vlivu x_2).

Mnohonásobná korelace



Koeficienty dílčí korelace

Příklad ⇒ vyjadřuje závislost celkové produkce na provozních nákladech za předpokladu, že výrobní spotřeba, odpisy a provozní dotace jsou neměně.

Correlations

Control Variables			celková produkce	provozní náklady
výrobní spotřeba & odpisy & provozní dotace	celková produkce	Correlation	1,000	,155
		Significance (2-tailed)	.	,031
		df	0	192
	provozní náklady	Correlation	,155	1,000
		Significance (2-tailed)	,031	.
		df	192	0

Konstantní proměnné

Mnohonásobná korelace

Sílu vztahu závisle proměnné *y* na **všech** vysvětlujících proměnných ***x*** udává:

Koeficient vícenásobné (totální) korelace *R*

$$0 \leq \boldsymbol{R} \leq 1$$

(**1** znamená úplnou **závislost** a hodnota **0 nezávislost**).

Koeficient totální korelace

Příklad ⇒ vyjadřuje závislost celkové produkce na všech prediktorech (nezávisle proměnných).

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,884[a]	,782	,778	6744,344

Koeficient mnohonásobné korelace R

Koeficient mnohonásobné determinace R^2

Opravená hodnota R^2 (adjusted R^2) nebere v úvahu stupně volnosti, proto je vždy v modelu s větším počtem vysvětlujících proměnných vyšší hodnota R^2. Potřebujeme-li porovnat kvalitu modelů s různým počtem vysvětlujících proměnných pro stejnou vysvětlovanou proměnnou *y*, použijeme opravenou hodnotu.

Mnohonásobná regrese

Mnohonásobná regresní analýza je metoda, pro modelování závislostí několika vysvětlovaných náhodných veličin (závisle proměnných) Y_1, Y_2, .. Y_G na jedné nebo několika vysvětlujících veličinách (nezávisle proměnných) X_1, X_2, .. X_K.

Mnohonásobná regrese

Cíle mnohonásobné regrese jsou stejné jako u regrese jednoduché:

1. **vysvětlit rozptyl** v závisle proměnné *Y* (pomocí R^2);

2. **odhadnout** (vypočítat) **vliv** každé z nezávisle proměnných ***X*** na proměnnou závislou *Y* (pomocí parciálních regresních koeficientů *b*);

3. **predikovat** pomocí sestavené regresní rovnice pro jednotlivé případy hodnoty závisle proměnné.

Mnohonásobná regrese

Před vlastní regresní analýzou je potřeba **ověřit kvalitu dat**.

Samotné analýze tedy musí předcházet podrobná diagnostika (analýza) vstupních proměnných (viz. 4. přednáška)

Mnohonásobná regrese

Model vyjadřující závislost veličiny ***Y*** na veličinách $X_1, X_2, \ldots, X_k$ lze zapsat ve tvaru:

$$y_i = f(x_{i1}, x_{i2}, \ldots, x_{ik}) + \varepsilon$$

kde: $f(x_{i1}, \ldots, x_{ik})$ … regresní funkce ($i = 1, 2, \ldots, n$)

ε …… náhodná chyba.

Mnohonásobná regrese

Lineární vícenásobný regresní model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, + \varepsilon$$

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ …..jsou neznámé parametry,

$x_1, \ldots, x_k$ ……..jsou vysvětlující proměnné,

ε …… náhodné chyby.

Koeficienty $\beta_0, \beta_1, \ldots \beta_K$ jsou obecně neznámé parametry, které je třeba z výběru odhadnout pomocí MMČ.

Mnohonásobná regrese

Odhadnutou regresní funkci lze zapsat ve tvaru (MMČ)

$$y` = b_0 + b_1x_1 + b_2x_2 + ... + b_kx_k$$

b_0 …….. je absolutní člen,

b_1,..,b_k... jsou **dílčí parciální regresní koeficienty**, které udávají změnu závisle proměnné *y* odpovídající jednotkové změně jedné nezávisle proměnné *x*, za předpokladu, že hodnoty zbývající nezávisle proměnných v modelu jsou konstantní.

(vyjadřují pouze část z vlivu, působících na vysvětlovanou proměnnou *y*)

Mnohonásobná regrese

Předpoklady modelu (viz. 4. přednáška)

Vysvětlující proměnné musí být **vzájemně nezávislé** – nesmí být korelované.

Náhodné chyby ε jsou nezávislé, normálně rozdělené náhodné veličiny s nulovými středními hodnotami a stejným rozptylem (**homoskedascita**).

Hodnocení mnohonásob. modelu z hlediska testů významnosti

Test významnosti dílčích výběrových regresních koeficient (parametrů ***b***) provádíme pomocí ***t* – testů**.

Test významnosti celého regresního modelu se provádí pomocí **upravené jednoduché ANOVY ⇒ *F* – testů**

Hodnocení mnohonásob. modelu z hlediska testů významnosti

Výsledek F - testu	Výsledek t - testů	Hodnocení modelu
nevýznamný	všechny nevýznamné	Posuzované proměnné jsou lineárně nezávislé; model je nevhodný nevystihuje variabilitu závisle proměnné.
významný	všechny významné	Model se považuje za vhodný k vystižení variability proměnné y, to však neznamená, že je optimálně navržen.
významný	některé nevýznamné	Model je vhodný, ale provádí se zpravidla vypuštění nevýznamných parametru modelu.
významný	všechny nevýznamné	Zvláštní případ způsobený multikolinearitou; paradox - model je nutné upravit a nebo zcela změnit.

Příklad

Sestavte nejvhodnější lineární regresní model pro závislost celkové produkce na provozních nákladech, výrobní spotřebě, odpisech a provozních dotacích.

$$y` = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$$

y …… celková produkce

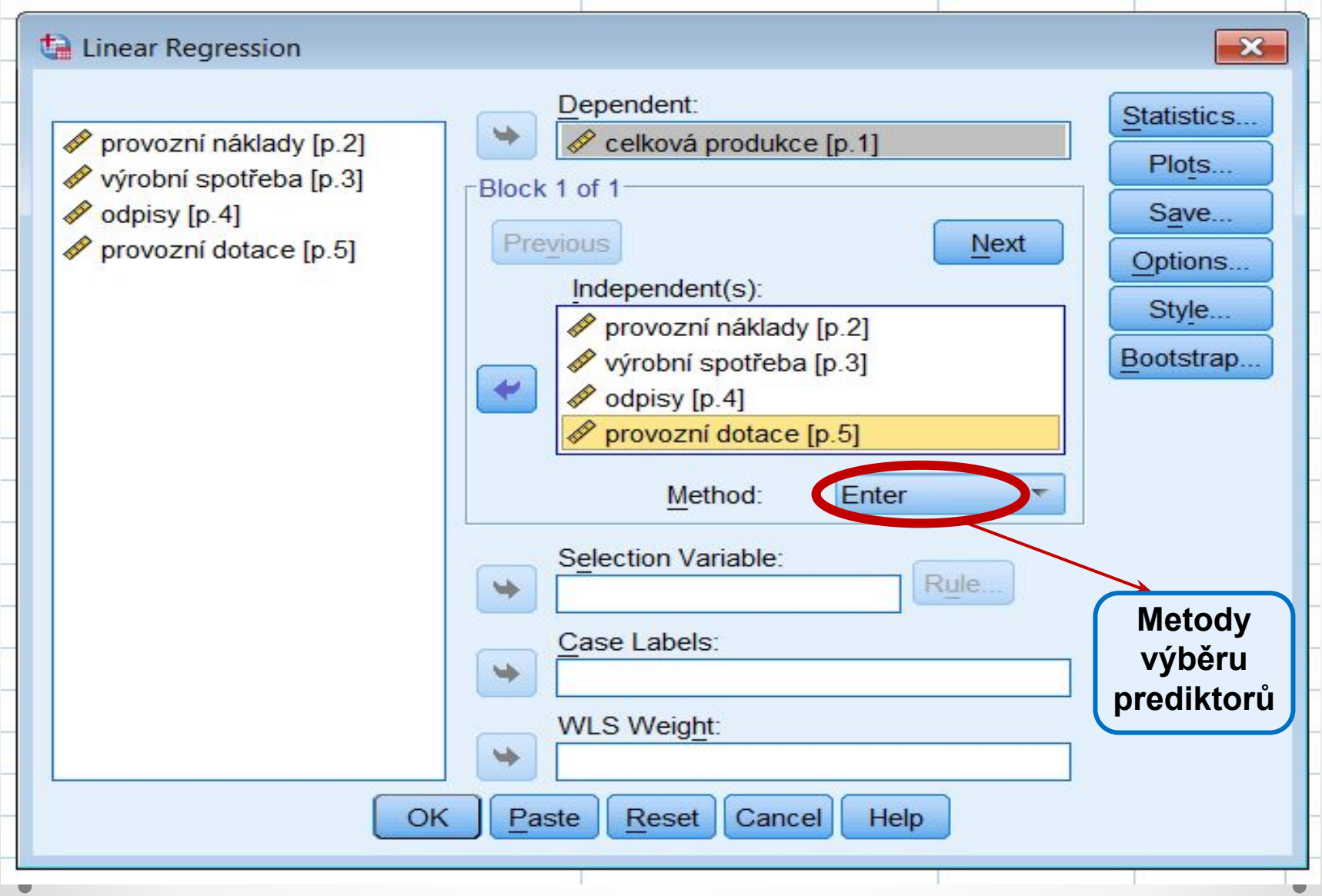
x_1 …. .provozní náklady

x_2 ….. výrobní spotřeba

x_3……odpisy

x_4 ……provozní dotace

Příklad


Linear Regression
provozní náklady [p.2]
výrobní spotřeba [p.3]
odpisy [p.4]
provozní dotace [p.5]
Dependent:
celková produkce [p.1]
Block 1 of 1
Previous
Next
Independent(s):
provozní náklady [p.2]
výrobní spotřeba [p.3]
odpisy [p.4]
provozní dotace [p.5]
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Metody výběru prediktorů

Metody výběru prediktorů (*x*)

ENTER – všechny prediktory vstoupí do rovnice (rozhodnutí uživatele).

1. metoda **FORWARD** – postupné zařazování prediktorů;
2. metoda **BACKWARD** – postupné vyřazování prediktorů;
3. metoda **STEPWISE** – kombinace obou, je založena na postupném vstup bloků proměnných (prediktorů).

Příklad

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,884[a]	,782	,778	6744,344

a. Predictors: (Constant), provozní dotace, odpisy, výrobní spotřeba, provozní náklady

Totální korelační koeficient - kvalita regresního odhadu; hodnocení volby vysvětlujících proměnných.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,135E+10	4	7837246532	172,300	,000[b]
	Residual	8733345903	192	45486176,58		
	Total	4,008E+10	196			

a. Dependent Variable: celková produkce

b. Predictors: (Constant), provozní dotace, odpisy, výrobní spotřeba, provozní náklady

Model jako celek je statistický významný vyplývá to z *F*-testu.

Příklad

Coefficients[a]

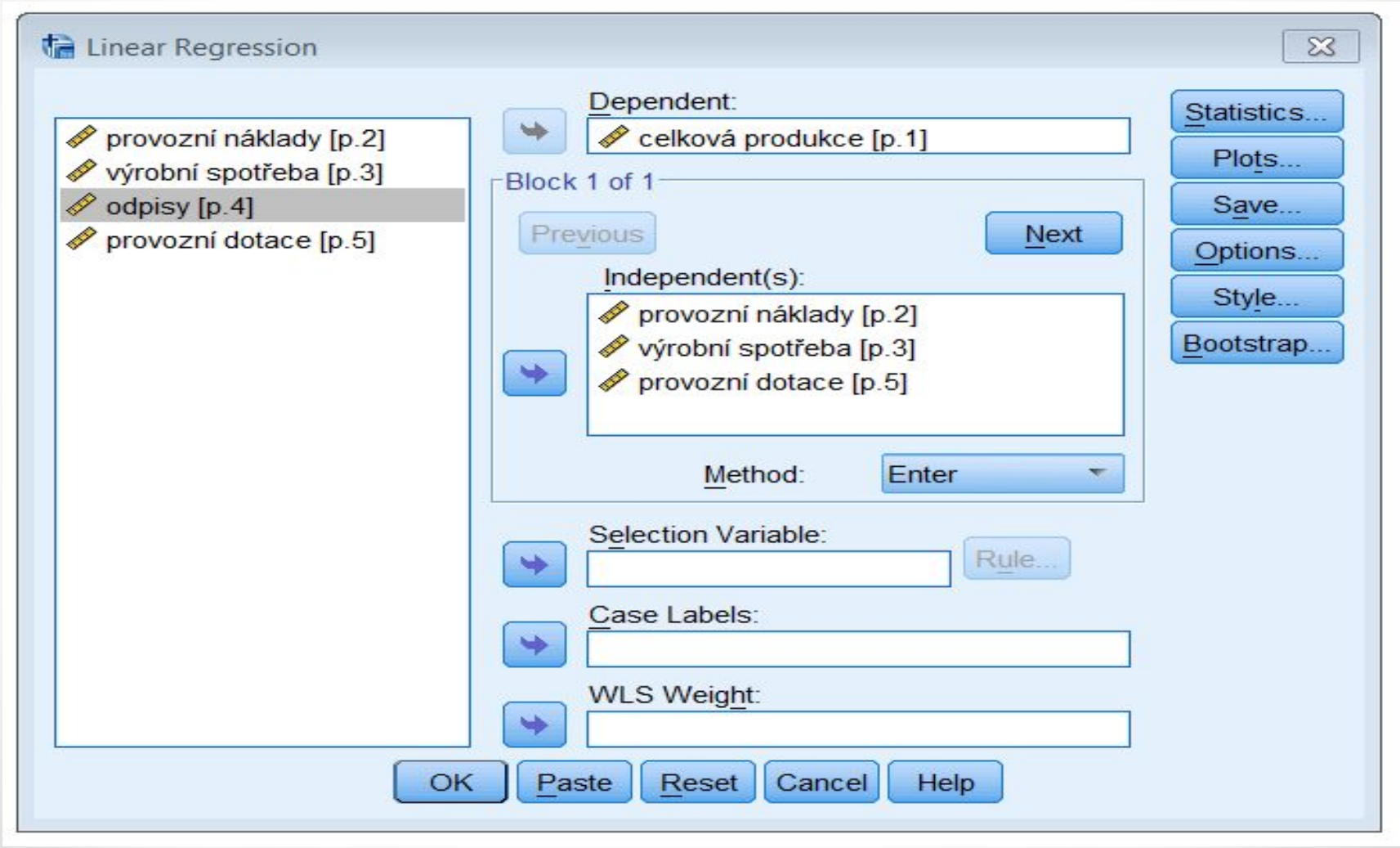
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3928,587	2593,368		1,515	,131
	provozní náklady	,396	,182	,351	2,171	,031
	výrobní spotřeba	,899	,228	,547	3,934	,000
	odpisy	,199	,172	,084	1,159	,248
	provozní dotace	-,801	,427	-,083	-1,875	,062

a. Dependent Variable: celková produkce

Z *t*-testů vyplývá, že některé regresní koeficienty jsou nevýznamné. I přesto, že je model vhodný jako celek budeme pokračovat v modelování vztahu mezi proměnnými ⇒ provedeme korigaci modelu ⇒ vypuštění nevýznamných proměnných.

Příklad

Z úvodního posouzení modelu vyplynulo, že budeme provádět vypuštění proměnných. V našem případě – *odpisy* x_3.



Příklad

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,884[a]	,781	,777	6750,336

a. Predictors: (Constant), provozní dotace, výrobní spotřeba, provozní náklady

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3,129E+10	3	1,043E+10	228,878	,000[b]
	Residual	8794436852	193	45567030,32		
	Total	4,008E+10	196			

a. Dependent Variable: celková produkce

b. Predictors: (Constant), provozní dotace, výrobní spotřeba, provozní náklady

Příklad

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5750,066	2064,658		2,785	,006
	provozní náklady	,581	,088	,515	6,621	,000
	výrobní spotřeba	,678	,126	,413	5,365	,000
	provozní dotace	-1,104	,338	-,114	-3,264	,001

a. Dependent Variable: celková produkce

$$y` = 5750{,}066 + 2064{,}658\, x_1 + 0{,}678\, x_2 - 1{,}104\, x_4$$

Po analýze hodnocení modelu a dílčích regresních koeficientů byl sestaven regresní model pro danou závislost, u kterého byla provedena redukce počtu vysvětlujících proměnných z původních 4 na 3 proměnné.