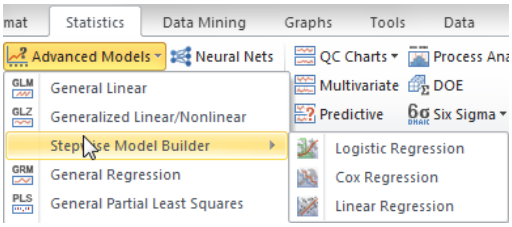
Manuální kroková regrese

Newsletter Statistica ACADEMY

Téma: Logistická regrese
Typ článku: Novinka verze 12, návody

Dnes si popíšeme funkcionalitu, která Vám pomůže při tvorbě regresního modelu (v našem případě modelu logistické regrese). Jedná se o krokovou výstavbu modelu, nicméně na rozdíl od klasických krokových algoritmů, dnes představovaný Vám dovolí v každém kroku vybrat manuálně, které proměnné vstoupí v daném kroku do modelu nebo které model opustí.

My se soustředíme na logistickou regresi, protože ta byla implementována jako první, nicméně v dalších verzích softwaru budou analogické možnosti i pro Coxovu regresi a regresi klasickou.



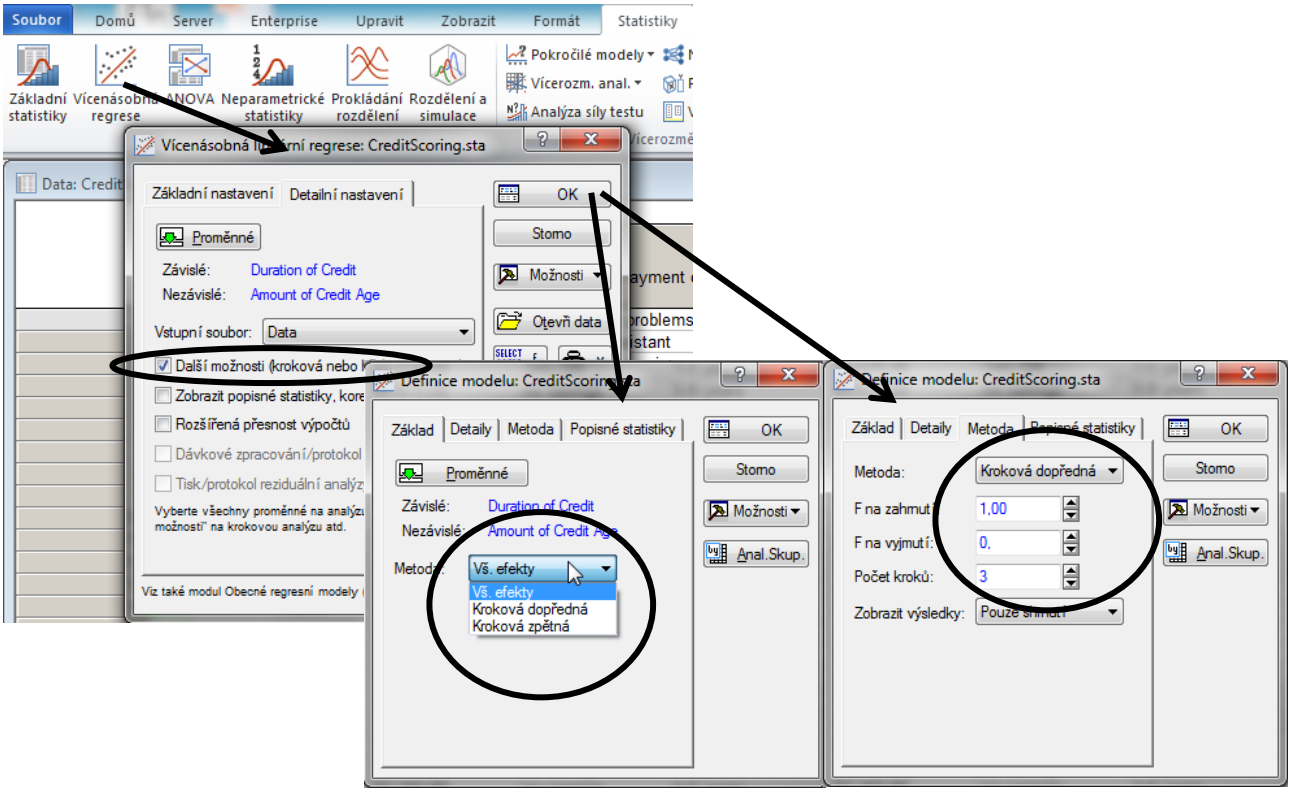
Kroková regrese

Praktický každá z různých typů regrese má v softwaru Statistica implementovánu krokovou verzi odhadu. Princip je jednoduchý, software v každém kroku algoritmu přidá či odebere jednu proměnnou na základě přednastaveného kritéria. Pokud již nemá co přidat či odebrat, pak algoritmus končí a výstupem je model, který použije některé nebo všechny proměnné, které jsme zadali v množině proměnných, které mají být použity pro vysvětlení odezvy (závislou proměnnou). Vše probíhá automaticky na základě nastavených kritérií.

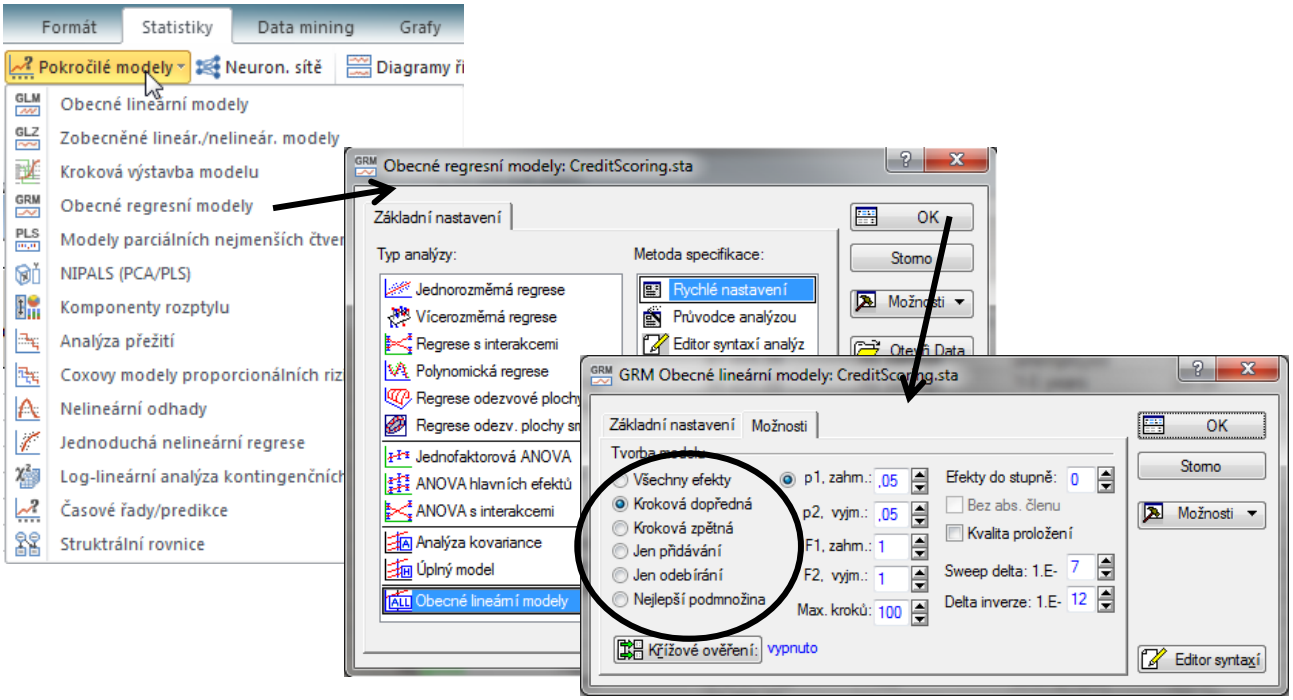
Jinak řečeno, z množství možných modelů (rozdílnost modelů je v tom, které proměnné zařadíme, aby vysvětlovali odezvu a které ne). Kroková regrese tedy vybere ze všech možných modelů jakéhosi kandidáta pro vhodný model.

Pro jistotu uvedeme místa v softwaru, kde je možné krokovou regresi najít:

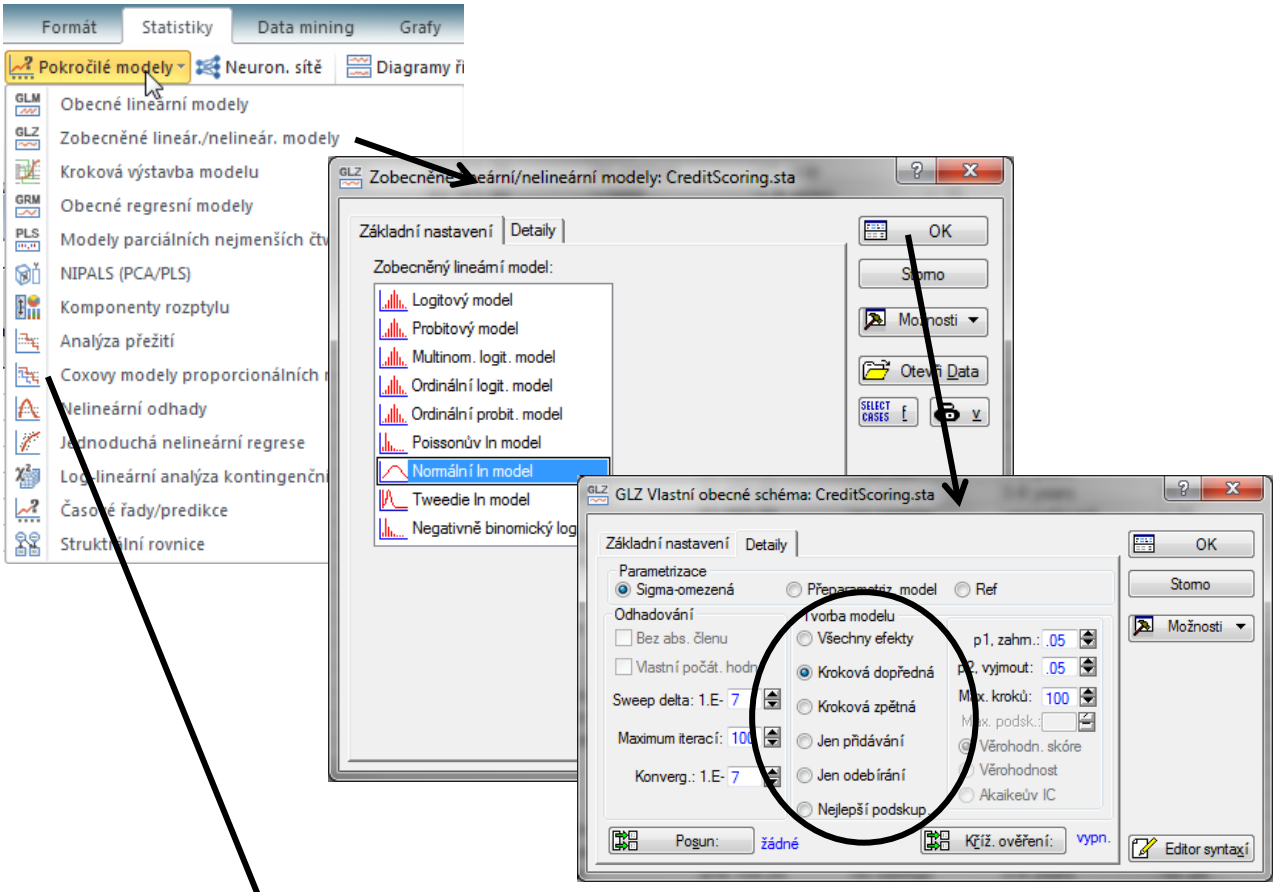
Lineární regrese:



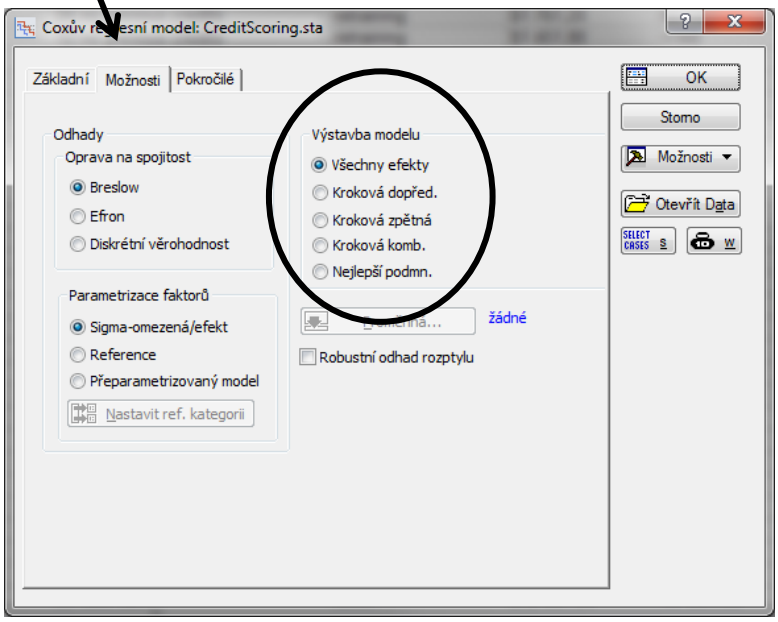
Další regresní modely (včetně kategorických regresorů):



Logistická regrese a další zobecněné modely:



Coxovy modely proporcionálních rizik:



Kroková výstavba modelu (Stepwise Model builder)

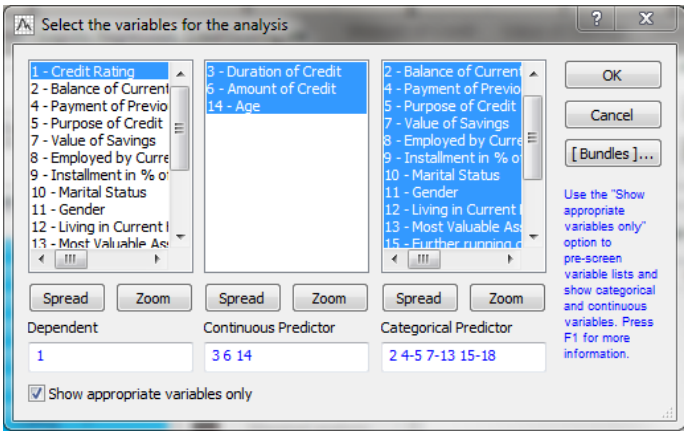
Výše popsané metody jsou automatické a výstup je možný ovlivnit pouze volbou kritérií či typu krokové regrese.

Nový modul, který si nyní představíme, je k dispozici pro situace, kdy potřebujeme dostat z krokové regrese něco více. Například můžeme chtít každý krok algoritmu kontrolovat manuálně sami, sami si chceme vybrat, který regresor zařadit či naopak vyloučit. Právě k tomu je představovaná funkcionalita ***Kroková výstavba modelu (Stepwise Model Builder)***. Samozřejmě, jak uvidíte níže, funkcionalita poskytuje výstupy a charakteristiky, které Vám s výběrem pomohou.

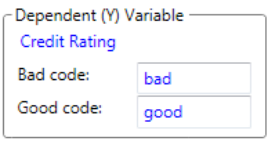
Příklad

Jak jsme již zmínili, vybrali jsme funkcionalitu pro logistickou regresi, která je dostupná ve verzi 12 a vyšších (snímky obrazovky jsou dělány v anglické verzi 12.7).

1. Otevřeme si nějaký soubor vhodný pro logistickou regresi, například soubor o uzavřených spotřebitelských úvěrech ***CreditScoring.sta*** – najdete jej v příkladech softwaru: ***Soubor->Otevřít příklady->Datasets***
2. Otevřeme modul ***Krokové výstavby modelu*** v menu ***Pokročilých modelů.***
3. Vybereme proměnné následovně:



4. Vybereme kódy pro dobré a špatné (modul a celá jeho implementace je inspirována problematikou kreditního rizika, proto je ponechána i takováto terminologie. Pro jakýkoli model logistické regrese to tedy znamená vložit kód, který nás zajímá do kategorie ***Bad code*** a druhý do kategorie ***Good code***. Kódy lze vybrat vepsáním či dvojklikem do kolonky pro kód.



5. Zmáčknete tlačítko ***Full Sample*** (kdy se používá celý soubor) nebo ***Subsample*** (kde se používá jen podmnožina) a tím se vygeneruje první výsledky (přesněji podklady pro první rozhodnutí). V každém řádku vidíte výsledek logistické regrese, jak by vypadal, kdyby regresor byla jen daná proměnná. V oblasti ***Marginal Results Table*** tedy vidíte tolik různých modelů, kolik je nezávislých proměnných (jeden model je označen jedním číslem v prvním sloupci tabulky, modely pro kategorické proměnné jsou na více řádcích, aby byly vidět odhady pro jednotlivé hladiny proměnné, parametrizaci lze nastavit v ***Model parameters*** při výběru proměnných).

6. V tabulce marginálních výsledků můžete označit jakýkoli počet proměnných a z nich vytvořit model, čímž provedete první krok Vašeho manuálního algoritmu. O zařazení či nezařazení se můžete rozhodovat na základě kritérií v tabulce jako je Somersovo D nebo p-hodnota. Pomocí tlačítka ***Correlations*** nebo ***Marginal Analysis*** se můžete podívat na přesný tvar modelů i na korelace parametrů daného modelu. Kategorické proměnné mají typicky více řádků, do modelu budou zahrnuty všechny řádky, nehledě na to, jestli vybereme jeden nebo všechny. Marginální tabulku je možné seřadit podle jednotlivých sloupců jednoduše kliknutím na záhlaví daného sloupce.

7. Po označení vybrané proměnné (proměnných) pro přidání do modelu klikneme na tlačítko ***Add Variable(s)***. V okně pro výsledný model se objeví proměnné, které jsme vybrali, všechny ostatní položky marginálních výsledků se přepočítají. Máme za sebou první krok naší manuální krokové regrese – zařadili jsme do modelu první proměnnou (proměnné). V sekci ***Model Results*** se můžeme podívat na shrnutí modelu. Další informace získáte v sekci ***Model Analysis*** - například ROC křivku či lift chart kliknutím na tlačítko ***Graphs,*** stabilitu modelu si můžete ověřit pomocí validační množiny (aktivní, pokud máte vybrán validační vzorek – ***Validation Sample***)nebo bootstrapu.

8. Po kliknutí na tlačítko ***Marginal analysis*** zjistíme, co znamenají nyní jednotlivé přepočítané řádky ***Marginal Results Table***, podobně jako před prvním výběrem ukazují a radí, jak může vypadat další krok. Každý řádek ukazuje odhad, jak by vypadal, kdybychom vybrali tento parametr a přidali k aktuálnímu modelu v sekci ***Model Results Table***. Ukazuje tedy možnou situaci v dalším kroku.

9. Takto můžeme pokračovat dalšími kroky dále a dále, můžeme přidávat či odebírat proměnné z modelu tak dlouho, až s ním budeme spokojeni. Celá historie toho, jak proměnné přidáváme a ubíráme, se ukládá a je možné ji získat kliknutím na tlačítko ***Summary***. To je dobrá možnost, pokud chcete či potřebujete někomu ukázat, jak byl model postaven, ke každému kroku je možné ve verzi 12.7 navíc přidat komentář, čímž například vysvětlíte postup v daném kroku.

Poznámka: ve verzi 12.7 je možné kliknout pravým tlačítkem na řádek v tabulce marginálních výsledků a zjistit vyvolat vlastnosti daného modelu, jako například ROC křivku atd.

Výběr proměnných

Proměnné, které mohou být přidány do analýzy

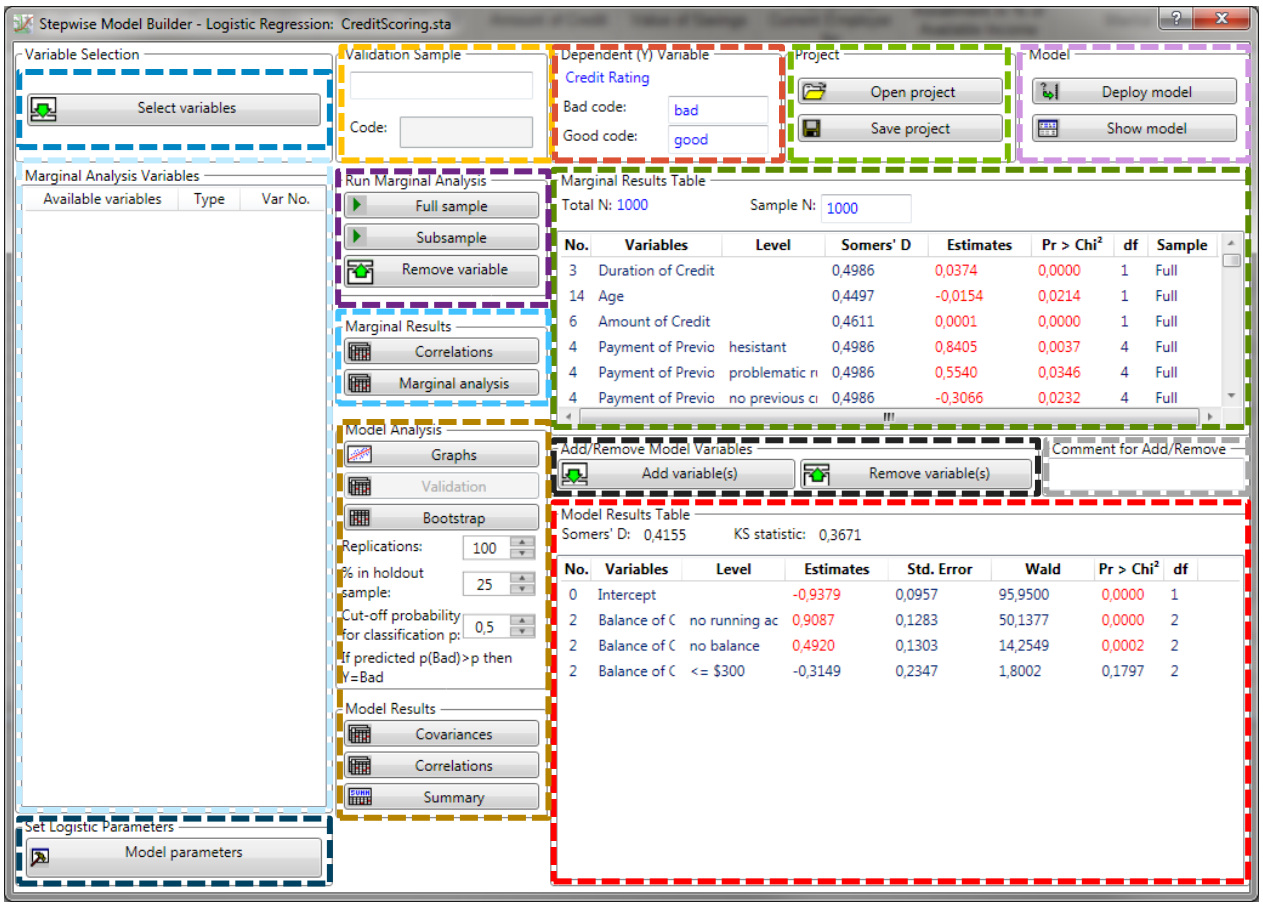
Parametry modelu, například jaká parametrizace bude použita

Výběr validační množiny.

Místo pro kódy závislé proměnné – to, co nás zajímá či to, co označuje, že nastala událost.

Rozdělanou analýzu a její nastavení lze uložit či načíst

Model, který je zrovna aktuální v oblasti pro výsledky modelu, lze uložit do prostředí Enterprise nebo vyexportovat jako PMML kód.



Výsledky marginálních analýz, jak by vypadal odhad pro danou proměnnou, kdyby byla zařazena do modelu spolu s proměnnými v aktuálně vybraném modelu.

Tlačítka pro přesunutí proměnných do sekce marginálních výsledků, tedy na vybraných proměnných jsou napočítány modely. Tlačítko Remove Variable je pro vyjmutí z oblastí marginálních výsledků, tedy zrušení těchto proměnných jako kandidátů na přidání v dalším kroku.

Výsledky marginálních modelů.

Tlačítka pro přidání či odebrání vybraných proměnných do/z modelu.

Místo pro komentář, který se uloží v případě přidání či odebrání proměnné(proměnných) do výsledků.

Aktuálně budovaný model

Výsledky pro aktuálně budovaný model i se souhrnem provedených kroků. Možnost provést další výpočty jako bootstrap.

Dell | Information Management | Software Group
V Parku 2325/16, 148 00 Praha 4
t +420 233 325 006, e info@statistica.cz
www.dell.cz/statsoft, www.statsoft.cz