IBM SPSS MISSING VALUES

Budujte přesnější modely s odhadem chybějících hodnot

Profesionálové z oblasti výzkumu trhu a marketingu, sociálních věd, data miningu i z mnoha dalších oborů se spoléhají na IBM® SPSS® Missing Values, aby zlepšili kvalitu svých dat.

Jestliže při analýze dat ignorujete či vynecháváte chybějící hodnoty, vaše analýzy, přehledy a zprávy mohou být zkreslené, zavádějící, nebo dokonce bezcenné. IBM SPSS Missing Values vám pomůže odhadnout chybějící data, a vy tak obdržíte hodnotnější závěry.

IBM SPSS Missing Values je nepostradatelným nástrojem každého, pro koho je přesnost dat a validita výsledků klíčovým požadavkem. Prozkoumá vaše data a odhalí možné závislosti a příčiny vzniku chybějících hodnot. Odhadne sumární statistiky a doplní váš soubor o optimální odhady skutečných hodnot na chybějících místech datové matice. Vhodně zvolené statistické algoritmy vám pomohou nahradit chybějící informaci a zkvalitnit vaše závěry.

Pokuste se například zlepšit ve výzkumu kvalitu otázek, které jste identifikovali jako problematické vzhledem ke struktuře chybějících hodnot. Tabulka Percent Mismatch of Patterns ukazuje, zda chybějící hodnoty jedné proměnné jsou v nějakém vztahu s chybějícími hodnotami jiných proměnných. Zjistíte například, že respondenti, kteří neodpovídají na otázky týkající se výše jejich příjmu, rovněž častěji odmítají dotaz na vzdělání. Využijte tedy tuto znalost pro zlepšení kvality vašich budoucích výzkumů a upravte tyto otázky.

Nejdůležitější:

Počítejte již dopředu s chybějícími hodnotami a udělejte z IBM SPSS Missing Values součást přípravy dat.

- **snadno prověříte data z různých úhlů pohledu**
- **určíte, jaké problémy způsobují vynechané hodnoty**
- **nahradíte vynechané hodnoty reálnými odhady**
- **zobrazíte chybějící data a extrémní hodnoty**
- **odhalíte skrytá vychýlení**

Analyzujte chybějící data rychle, jednoduše a efektivně

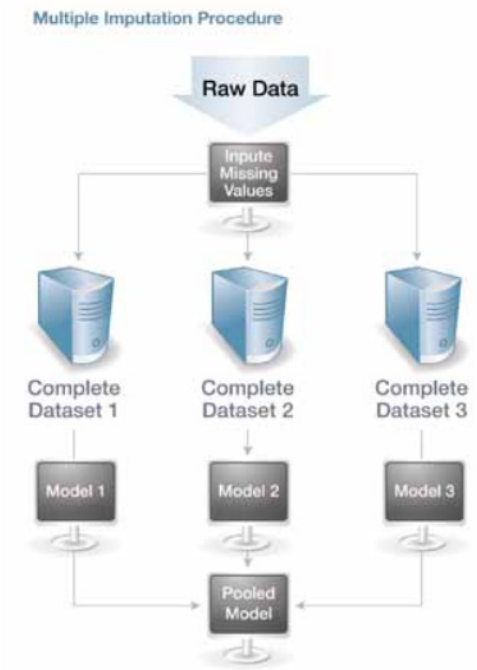
Pro rychlou diagnostiku závažnosti problému chybějících hodnot využijte report, shrnující celkovou informaci o chybějících hodnotách. Report zobrazí strukturu vynechaných hodnot pro všechny případy, které obsahují alespoň jednu vynechanou hodnotu. Tak lépe odhalíte rozsah a typ chybějících hodnot. Současně s přehledem chybějících dat získáte také informace o extrémních hodnotách.

Pomocí t-testu nebo kontingenčních tabulek určíte významné závislosti mezi platnými a vynechanými daty a rozhodnete, zda mohou chybějící data způsobit problémy ve vaší další analýze.

Získejte sumární statistiky vzorů chybějících hodnot a zaměřte se na skupiny proměnných, které se vyskytují v reportovaných vzorech.

Využijte mnohonásobné imputace k nahrazení chybějících hodnot

Procedura pro mnohonásobné imputace v IBM SPSS Missing Values vám pomůže porozumět struktuře chybějících hodnot ve vašich datech a nahradit chybějící hodnoty věrohodnými odhady. Procedura poskytuje plně automatický imputační mód, který vyhledává nejvhodnější imputační metodu nan základě charakteristiky dat, nebo lze model přizpůsobit požadavkům uživatele.



Je vygenerováno několik kompletních datových souborů (obvykle tři až pět), každý s jinak imputovanými hodnotami Pro tyto datové soubory jsou na základě obvyklých technik, například lineární regrese, vytvořeny modely a odhadnuty jejich parametry. Konečné odhady parametrů se získají jejich kombinací (sdružením) a respektováním variability uvnitř a mezi imputacemi.

Analýza jednotlivých souborů a sdružení jejich výsledků jsou podporovány zvolenou statistickou procedurou z IBM SPSS Statistics, například REGRESSION. Při práci s datovými soubory s mnohonásobnými imputacemi poskytují existující procedury automaticky sdružené odhady parametrů.

Získejte věrohodnější závěry

Neomezujte se pouze na kompletní případy. Chybějící hodnoty snadno nahradíte na základě existujících dat, a tím zvýšíte šanci získat statisticky průkazné závěry. Do analýz potom můžete zahrnout i méně zastoupené kategorie, a tak odstranit případná skrytá vychýlení.

Modul IBM SPSS Missing Values je součástí softwarové řady IBM SPSS Statistics. Tento software je poskytován také ve formě tří balíčků: IBM SPSS Statistics Standard, IBM SPSS Statistics Professional a IBM SPSS Statistics Premium. Seskupením podstatných funkcí do jednotlivých balíčků získáte pro celý váš tým nebo oddělení metody a funkce, které jsou potřebné pro provedení analýz vedoucích k úspěchu vaší firmy.

IBM SPSS Missing Values je k dispozici pro instalaci v režimu lokální desktopové aplikace, avšak při požadavku na větší výkon a škálovatelnost lze provést také instalaci v režimu klient/server.

SPECIFIKACE

Analýza struktury

- *tabulka Data Patterns: vynechaná data a extrémní hodnoty pro všechny případy a všechny proměnné*
  - *zobrazení systémových vynechaných hodnot a tří typů uživatelem definovaných vynechaných hodnot*
  - *vzestupné či sestupné třídění*
  - *zobrazení hodnot specifikovaných proměnných*
- *tabulka Missing Patterns: rozložení vynechaných hodnot pro všechny případy, které obsahují alespoň jednu vynechanou hodnotu*
  - *seskupení podobných struktur vynechaných hodnot*
  - *řazení podle struktury vynechaných hodnot a proměnných*
  - *zobrazení hodnot specifikovaných proměnných*

- *tabulka Separate Variance: t-testy pro porovnání skupin vynechaných a platných hodnot (za předpokladun různých rozptylů)*
  - *t-test, stupně volnosti, průměr, hladina spolehlivosti, četnost*
- *tabulka Distribution of Categorical Variables: zobrazí rozdíly mezi platnými a vynechanými daty kategorizovaných proměnných*
  - *kontingenční tabulky porovnávající rozložení ostatních kategorizovaných proměnných ve skupinách vynechaných a platných hodnot zvolené kategorizované proměnné*
- *tabulka Percent Mismatch of Patterns: vyhodnocení závislosti chybějících dat mezi dvěma různými proměnnými*
  - *setříděné matice podle struktury vynechaných hodnot nebo podle proměnných*
- *tabulka Tabulated Patterns: identifikuje strukturu vynechaných hodnot spolu s četností a průměry všech proměnných*
  - *tabulka zobrazující souhrnné informace pro četnosti a průměry chybějících dat*

Statistiky

- *jednorozměrné statistiky: četnost, průměr, směrodatná odchylka a standardní chyba průměru všech hodnot kromě vynechaných, relativní četnost vynechaných hodnot a extrémy*
- *metoda práce s chybějícími hodnotami Listwise: průměr, kovarianční a korelační matice pro proměnné a případy neobsahující vynechané hodnoty*
- *metoda práce s chybějícími hodnotami Paiwise: četnosti, průměr, rozptyl, kovarianční a korelační matice*

Mnohonásobné imputace

- *specifikace proměnných, jejichž hodnoty budou imputovány, omezení pro imputované hodnoty (minimum, maximum), specifikace prediktorů*
- *imputace hodnot pro kategorizované i číselné proměnné. Pro kategorizované proměnné je užita logistická regrese, pro číselné lineární regrese. Pro číselné proměnné lze rovněž využít metodu Predictive mean matching, která zajistí, že imputované hodnoty leží v rozpětí hodnot původních dat*
- *detekce vzorů chybějících hodnot pomůže určit vhodnou imputační metodu*
- *k dispozici jsou tři imputační metody:*
  - *monotónní: optimální pro data, která mají monotónní strukturu chybějících hodnot*
  - *Fully conditional specification (FCS): iterativní metoda patřící mezi algoritmy Markov Chain Monte Carlo (MCMC), která je vhodná v případě, že data mají libovolnou strukturu chybějících hodnot (monotónní i nemonotónní)*
  - *automatická: na základě vlastností dat nalezne nejvhodnější imputační metodu (monotónní nebo FCS)*
- *specifikace*
  - *počet imputací*
  - *rozpětí imputovaných hodnot*
  - *užití interakcí při imputacích*
  - *možnost neprovádět imputace u proměnných, které mají vysoký podíl chybějících hodnot*
  - *hladiny tolerance pro kontrolu singularit*
- *lze specifikovat proměnnou určující váhy. Procedura zahrne váhy do výpočtu regrese a klasifikačních modelů určených pro imputaci chybějících hodnot. Váhy jsou užity i při sumarizaci imputovaných hodnot (průměr, směrodatná odchylka, standardní chyba).*
- *zobrazení celkových statistik chybějících hodnot v datech a imputačních modelů pro všechny proměnné, jejichž hodnoty jsou imputovány. Dále lze získat analýzy chybějících hodnot podle proměnných, přehled o struktuře chybějících hodnot nebo popisné statistiky imputovaných hodnot*
- *grafické zobrazení informace o chybějících hodnotách pro případy, proměnné i jednotlivé datové buňky*
- *uložení imputovaných hodnot a/nebo průběhu iterací při použití metody FCS do určených datových souborů IBM SPSS Statistics*
- *datové soubory mnohonásobných imputací mohou být analyzovány s využitím podporovaných analytických procedur tak, aby byly získány konečné sdružené odhady parametrů, které vycházejí z míry nejistoty imputovaných hodnot v jednotlivých souborech*

Analýzy

- *analytické procedury podporující mnohonásobné imputace (je nutné mít zakoupen vlastní modul, který obsahuje danou proceduru)*
- *popisné statistiky: četnosti, popisné statistiky, kontingenční tabulky, korelace, neparametrické korelace, parciální korelace*
- *porovnání průměrů: průměry a další popisné statistiky ve skupinách, t testy, neparametrické testy, jednoduchá analýza rozptylu, univariate ANOVA*
- *modely: obecný lineární model, zobecněný lineární model, lineární regrese, multinomická logistická regrese, binární logistická regrese, diskriminační analýza, ordinální regrese, lineární smíšené modely*
- *analýza délky života: Coxova regrese*

Sdružování

- *sdružování výstupů: výstupy se sdružují pomocí jednoho ze dvou typů sdružování, které odvodí sdružené parametry*
- *diagnostika sdružování*
  - *relativní nárůst rozptylu: vyjadřuje relativní variabilitu odhadu parametru mezi imputacemi*
  - *podíl chybějící informace: relativní nárůst rozptylu vyjádřený jako poměrná část, míra nejistoty způsobená chybějícími hodnotami*
  - *relativní účinnost: účinnost odhadu při M imputacích vzhledem k nekonečnému počtu imputací*
- *PMML modely pro kombinované odhady parametrů: lineární regrese, zobecněný lineární model, multinomická logistická regrese, binární logistická regrese, diskriminační analýza, Coxova regrese*