EGRET — STATISTICKÝ SOFTWARE PRO EPIDEMIOLOGII

MAREK MALÝ

Státní Zdravotní Ústav, Praha

Při statistickém vyhodnocování epidemiologických studií se setkáváme i se speciálními úlohami, které jsou poměrně obtížně řešitelné pomocí běžných statistických programových balíků (jako SPSS, BMDP, atp.). Proto jsou vedle obecnějších programů, které umožňují vyhodnocení méně komplikovaných epidemiologických modelů (např. GLIM), vyvíjeny i speciální programy a balíky přímo zaměřené na zpracování epidemiologických dat. V posledních letech zaznamenal široké rozšíření zejména program EGRET, kterému je věnován tento příspěvek. Závěrečné odstavce textu shrnují informace o některých dalších programech s příbuznou tématikou.

EGRET (Epidemiological GRaphics, Estimation, and Testing package; nyní ve verzi 0.26.06) je statistický programový systém ne zcela tradičního zaměření vhodný (nejen) pro analýzu dat v epidemiologii a zdravotnictví vůbec. Je vyvíjen od roku 1984 organizací "Statistics and Epidemiology Research Corporation" v Seattle, USA. Prvotní projekt, který si kladl za cíl vytvořit PC-program pro analýzu přežívání, byl posléze výrazně rozšířen, takže když se v roce 1988 s EGRETem seznamovali první uživatelé, byl to již ucelený programový balík. V současnosti je široce používaným a oblíbeným nástrojem.

Program se řídí pomocí menu a i když se při prvním spuštění nejeví jako nejpřátelštější, lze do jeho tajů poměrně snadno proniknout, mj. také proto, že je k dispozici podrobný manuál a téměř v každém okamžiku výpočtu také bohatá nápověda (help). EGRET zahrnuje zejména modely pro analýzu přežívání, pro stratifikované i nestratifikované kontingenční tabulky typu 2xK (K2), podmíněnou i nepodmíněnou logistickou regresi, Coxovu regresní analýzu, Kaplanův-Meierův odhad a další speciální typy regrese (Poissonovu, exponenciální, Weibullovu). Výsledky vystupující na obrazovku, jsou zároveň ukládány i do souboru na disk. Výstupy jsou doplněny grafikou (lze instalovat i verzi bez grafiky).

Proces zpracování dat probíhá vlastně dvoustupňově, protože EGRET je tvořen dvěma základními, na sebe úzce navazujícími moduly DEF a PECAN, které se spouštějí přímo z prostředí DOSu.

Modul DEF (Data Definition Module) musí být spuštěn před vlastní analýzou pro každý datový soubor, který chceme zpracovávat. DEF vytvoří z původních dat (ve formě ASCII souboru obsahujícího jenom číselné údaje bez identifikace proměnných) dva nové soubory, jeden obsahující data v binární formě (označení *.BDF) a druhý, definiční, s hlavičkami (*.HDR; ten obsahuje nejen základní informaci o datech, ale i informace

vložené až v rámci běhu modulu DEF, jako jsou jména proměnných, kód pro chybějící pozorování a informace o modelu, který má být pro zpracování dat použit).

V rámci DEFu se provádějí následující činnosti:

1. Výběr úlohy (Vytvoření nového definičního souboru či úprava již existujícího, import z Lotusu)
2. Specifikace jména dat a počtu řádků, které v datech zabírá jedno pozorování
3. Specifikace formátu pro čtení dat (buď jsou hodnoty odděleny mezerou, čárkou či tabelátorem nebo lze definovat libovolný vlastní formát symbolikou jazyka FORTRAN; na obrazovce se objeví prvních 8 řádků dat)
4. Vložení jmen proměnných (na základě různých zabudovaných aritmetických funkcí, např. MIN, LOG, LT, EQ, RECODE)
5. Specifikace kódů pro chybějící pozorování (kód může být pro každou proměnnou jiný)
6. Případné vytvoření nových proměnných
7. Volba modelu pro analýzu a označení klíčových proměnných určujících expozici, atd.
8. Vytvoření binárního datového a definičního souboru.

V bodu 7 je možno volit z této základní nabídky modelů:

- Logistic regression;
- Conditional logistic regression;
- Cox proportional hazards survival analysis;
- Poisson regression;
- Logistic regression with random effects;
- Traditional 2xK table analysis;
- Kaplan – Meier estimation;
- Parametric survival regression.

Pro každou položku základní nabídky existuje ještě nabídka speciální vážící se k této konkrétní volbě. Například pro logistickou regresi je třeba dále zvolit či označit

- typ rizika (aditivní, multiplikativní);
- proměnnou určující velikost skupin;
- proměnnou obsahující výsledné hodnoty měření (outcome variable);
- počet opakování (váha udávající kolikrát má být které pozorování uvažováno).

Pro Kaplanův – Meierův odhad je nutno označit proměnnou obsahující doby poruch, proměnnou indikující, zda pozorování je cenzorováno, a časy, kdy subjekty vstupují do studie. Pro kontingenční tabulky umožňuje počet opakování rozlišit, zda se zadávají individuální údaje či data již rozdělená do tříd.

Při volbě poslední položky pro Kaplanův–Meierův odhad jsou na obrazovce informace v následující podobě (v dolní řádce je menu):

UPDATING DEFINITION OF c:\egret\data\rats

Variables			
1. days2cancr	2. censoring	3. group	4. DaysLes100

DEFAULT ANALYSIS MODEL: Kaplan-Meier estimation
DEFAULT FAILURE TIMES: Failure Time Variable: days2cancr
DEFAULT CENSORING: Censoring Indicator Variable: censoring
DEFAULT ENTRY TIMES: – All Subjects Enter at Time Zero
– Entry-time variable

{btwn options} ■ ENTER{choose}|END ■ ESC ■ PgUp, PgDn {scroll vars} ■ ?{help}

Druhý základní modul, PECAN (Parameter Estimation through Conditional probability ANalysis) zajišťuje vlastní analýzu dat a výstup výsledků. Umožňuje již jen drobné korekce a menší doplnění základního modelu (např. stratifikaci), který uživatel zvolil v rámci DEF. Po zadání jména datového souboru se v dolní řádce objeví nabídka hlavního menu. V jeho pravé části jsou položky společné všem modelům, položky v levé části jsou poněkud odlišné pro regresní modely, kontingenční tabulky, Kaplanův–Meierův odhad, atd.

Nejobsažnější oblastí zpracovanou v EGRETu jsou regresní modely. Jejich stručný přehled poskytuje následující tabulka:

Model	S PEVNÝMI EFEKTY		S NÁHODNÝMI EFEKTY
Data	PODMÍNĚNÝ	NEPODMÍNĚNÝ	MARGINÁLNÍ
NEPÁROVANÁ	Coxův	Logistický	Beta-binomický
		Poissonův	Logisticko-normální
		Exponenciální	Logisticko-binomický
		Weibullův	
PÁROVĚ	Podmíněně		Logisticko-binomický
VYVÁŽENÁ	logistický		s rozlišitelnými daty

Pro regresní, ale i jiné modely lze v manuálu nalézt jejich popis včetně vzorců pro výpočet a odkazů na literaturu, z níž byly modely převzaty. Zajímavou kapitolu představuje regresní diagnostika založená na koncepci tzv. delta–beta (viz [8]). Hlavní menu pro regresi má tvar

New|Regrp| ■ Fit|Score|stepWise ■ Varscrol|auX cmds ■ ?help ■ >

Z této nabídky lze zvolit vytvoření nového modelu (vlastně je to nová volba proměnných zahrnutých do modelu daného nastavením v DEF); přidání parametrů (včetně interakcí) ke stávajícím; vlastní odhad parametrů modelu; test vhodnosti modelu; postupné přidávání proměnných do modelu včetně diagnostiky. Po provedení odhadu (volba Fit) se v nabídce objevuje ještě regresní diagnostika a intervaly spolehlivosti.

Pomocné příkazy (auX cmds) zahrnují výpočet popisných statistik pro specifikované proměnné, faktorizaci, výběr dat, různé grafické výstupy (bodový graf, histogram) s možností úprav textů, měřítka, atd. Je také možno vpisovat poznámky do automaticky vytvářeného výstupního souboru.

Pro porovnání uvádím ještě základní nabídkový řádek pro Kaplanův – Meierův odhad:

Stratify ■ Display|Plot ■ Matched sets file ■ Varscrol|auX cmds ■ ?help >

Je možno stratifikovat a zobrazit číselně i graficky (s mnoha možnostmi editace obrázku) pravděpodobnosti přežití. Zatím zřejmě není zabudován test na porovnání skupin.

Na ploše tohoto příspěvku není možno popsat podrobně všechny možnosti EGRETu, ale uvedené informace příklady by měly posloužit k základní orientace v jeho vlastnostech. Algoritmy EGRETu vycházejí z knižních i časopiseckých publikací známých osobností jako jsou J.J.Gart, J. Lubin, N.E. Breslow, N. Mantel, R.L. Prentice, D.C. Thomas, C. Mehta, které reprezentují zřejmě skutečně nejlepší dostupné podklady pro epidemiologicko – statistické zpracování dat.

K programovému souboru EGRET je přiloženo asi 30 datových souborů včetně jejich souborů definičních a udání zdroje dat. Uživatel si tak může snadno jednotlivé metody prakticky ověřit. Domnívám se, že EGRET je užitečným a přiměřeně obsáhlým pomocníkem všech, kteří zpracovávají data především s lékařskou a biologickou tématikou a že rozhodně není neužitečné se s ním seznámit.

V závěru se velmi stručně zmíním o některých dalších programech se zaměřením příbuzným EGRETu.

EPICURE 1.8 je soubor interaktivních příkazově řízených programů pro modelování a analýzu dat z oblasti medicíny, veřejného zdravotnictví, epidemiologie, hodnocení spolehlivosti. Je tvořen následujícími čtyřmi základními moduly:

GMBO –	pro binomickou regresi v kohortových a nepárovaných case – control studiích
PECAN –	pro analýzu párově vyvážených case – control studií pomocí podmíněné logistické regrese
PEANUTS –	pro analýzu přežívání na základě individuálních (ne skupinových) dat, která mohou být cenzorována.
AMFIT –	pro Poissonovu regresi v kohortových studiích a klinických pokusech na základě skupinových dat či kontingenčních tabulek.

Pomocný pátý modul DATAB slouží pro nezbytnou přípravu dat před zpracováním. Program je podrobně dokumentován (viz [9]).

EpiInfo 5.01b je soubor programů pro zpracování epidemiologických dat v dotazníkové formě (viz [3]). Je velmi užitečným nástrojem pro ukládání dat do databáze spojené s kontrolou správnosti a případným dopočítáváním. Databáze je převoditelná do formátu dBase III. V rámci modulu ANALYSIS lze provádět jednodušší statistické hodnocení, zejména lze vytvářet a vyhodnocovat kontingenční tabulky (i stratifikované), aplikovat lineární regresi, analýzu rozptylu jednoduchého třídění, počítat četnosti. EPIINFO obsahuje vlastní jednoduchý textový editor. V nejbližší době se má objevit verze 6.

LogXact je software pro přesný (exaktní) výpočet logistické regrese založený na algoritmech dr. Cyruse Mehty [7].

Vlastní EGRET je nyní možno doplnit produktem EGRET SIZ 1.00, který je určen pro odhady rozsahu výběru a síly testů pro skupinu nelineárních regresních modelů užívaných v epidemiologii a veřejném zdravotnictví.

LITERATURA

[1] Breslow N.E., Day N.E., *Statistical Methods in Cancer Research*, The Analysis of Case-control Studies, Vol. 1, IARC, Lyon, 1980.

[2] Breslow N.E., Day N.E. paper Statistical Methods in Cancer Research, The Design and Analysis of Cohort Studies, Vol. 2, IARC, Lyon, 1987.

[3] Dean A.G. et al., *EpiInfo, Version 5: a word processing, database and statistics program for epidemiology on microcomputers*, USD, Inc., Stone Mountain, Georgia, 1990.

[4] *EGRET, Reference Manual*, Statistics and Epidemiology Research Corporation and Cytel Software Corporation, Seattle, 1991.

[5] Hirji K.F., Mehta C.R., *Computing exact distributions for logistic regression*, JASA 82 (1987), 1110-1117.

[6] Kalbfleisch J.D., Prentice R.L., *The Statistical Analysis of Failure Time Data*, J. Wiley, New York, 1980.

[7] Mehta C.R., Patel N.R., Gray R., *Computing an Exact Confidence Interval for the Common Odds Ratio in Several 2x2 Contingency Tables*, JASA 80 (1986), 969-973.

[8] Pregibon D., *Logistic regression diagnostics*, Ann. Stat. 9 (1981), 705-724.

[9] Preston D.L., Lubin J.H., Pierce D.A., *EPICURE, User's Guide*, HiroSoft International Co., Seattle, 1992.