Helsingin yliopisto

Otanta-aineistojen analyysi

Kevät 2012 Periodi IV

Risto Lehtonen

Teema 4
Asetelmaperusteinen monimuuttuja-analyysi

Logistinen ANOVA ja GWLS-estimointi

- Binäärinen tulosmuuttuja
- Diskreetit selittäjät
- Laskenta: SAS / IML

Logistinen ANCOVA ja PML-estimointi

- Binäärinen tulosmuuttuja
- Diskreetit ja jatkuvat selittäjät
- Laskenta: SAS / SURVEYLOGISTIC

Esimerkki: OHC-aineisto

Materiaali:

Lehtonen-Pahkinen (2004)

- Section 8.3
- Section 8.4

The SURVEYLOGISTIC Procedure

Overview

Categorical responses arise extensively in survey research. Common examples of responses include

- binary: e.g., attended graduate school or not
- ordinal: e.g., mild, moderate, and severe pain
- nominal: e.g., ABC, NBC, CBS, FOX TV network viewed at a certain hour

Logistic regression analysis is often used to investigate the relationship between such discrete responses and a set of explanatory variables. See Binder (1981, 1983), Roberts, Rao, and Kumar (1987), Skinner, Holt, and Smith (1989), Morel (1989), and Lehtonen and Pahkinen (1995) for papers that describe logistic regression for sample survey data.

For binary response models, the response of a sampling unit can take a specified value or not (for example, attended graduate school or not). Suppose ***x*** is a row vector of explanatory variables and π is the response probability to be modeled. The linear logistic model has the form

$$\mathbf{logit}(\pi) \equiv \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \mathbf{x}\boldsymbol{\beta}$$

where α is the intercept parameter and $\boldsymbol{\beta}$ is the vector of slope parameters.

The logistic model shares a common feature with the more general class of generalized linear models, namely, that a function $g = g(\mu)$ of the expected value, μ, of the response variable is assumed to be linearly related to the explanatory variables. Since μ implicitly depends on the stochastic behavior of the response, and since the explanatory variables are assumed to be fixed, the function *g* provides the link between the random (stochastic) component and the systematic (deterministic) component of the response variable. For this reason, Nelder and Wedderburn (1972) refer to *g*(·) as a link function. One advantage of the logit function over other link functions is that differences on the logistic scale are interpretable regardless of whether the data are sampled prospectively or retrospectively (McCullagh and Nelder 1989, Chapter 4). Other link functions that are widely used in practice are the probit function and the complementary log-log function. The SURVEYLOGISTIC procedure enables you to choose one of these link functions, resulting in fitting a broad class of binary response models of the form

$$g(\pi) = \alpha + \mathbf{x}\boldsymbol{\beta}$$

For ordinal response models, the response *Y* of an individual or an experimental unit may be restricted to one of a usually small number of ordinal values, denoted for convenience by 1, ... ,*D*, *D*+1 $(D \geq 1)$. For example, the pain severity can be classified into three response categories as 1=mild, 2=moderate, and 3=severe. The SURVEYLOGISTIC procedure fits a common slopes cumulative model, which is a parallel lines regression model based on the cumulative probabilities of the response categories rather than on their individual probabilities. The cumulative model has the form

$$g(\Pr(Y \leq d \,|\, \mathbf{x})) = \alpha_d + \mathbf{x}\boldsymbol{\beta}, \quad 1 \leq d \leq D$$

where $\alpha_1, \ldots, \alpha_k$ are *k* intercept parameters and $\boldsymbol{\beta}$ is the vector of slope parameters. This model has been considered by many researchers. Aitchison and Silvey (1957) and Ashford

(1959) employ a probit scale and provide a maximum likelihood analysis; Walker and Duncan (1967) and Cox and Snell (1989) discuss the use of the log-odds scale. For the log-odds scale, the cumulative logit model is often referred to as the *proportional odds* model.

For nominal response logistic models, where the *D*+1 possible responses have no natural ordering, the logit model can also be extended to a *generalized logit* model, which has the form

$$\log\left(\frac{\Pr(Y=i\mid\mathbf{x})}{\Pr(Y=D+1\mid\mathbf{x})}\right)=\alpha_i+\boldsymbol{\beta}_i'\mathbf{x},\quad i=1,\ldots,D$$

where the $\alpha_1,\ldots,\alpha_D$ are *D* intercept parameters, and the $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_D$ are *D* vectors of parameters. These models were introduced by McFadden (1974) as the *discrete choice* model, and they are also known as *multinomial* models.

The SURVEYLOGISTIC procedure fits linear logistic regression models for discrete response survey data by the method of maximum likelihood. For statistical inferences, PROC SURVEYLOGISTIC incorporates complex survey sample designs, including designs with stratification, clustering, and unequal weighting.

The maximum likelihood estimation is carried out with either the Fisher-scoring algorithm or the Newton-Raphson algorithm. You can specify starting values for the parameter estimates. The logit link function in the ordinal logistic regression models can be replaced by the probit function or the complementary log-log function.

Odds ratio estimates are displayed along with parameter estimates. You can also specify the change in the explanatory variables for which odds ratio estimates are desired. Variances of the regression parameters and odds ratios are computed using the Taylor expansion approximation; see Binder (1983).

The SURVEYLOGISTIC procedure enables you to specify categorical variables (also known as CLASS variables) as explanatory variables. It also enables you to specify interaction terms in the same way as in the LOGISTIC procedure.

Like many procedures in SAS/STAT software that allow the specification of CLASS variables, the SURVEYLOGISTIC procedure provides a CONTRAST statement for specifying customized hypothesis tests concerning the model parameters. The CONTRAST statement also provides estimation of individual rows of contrasts, which is particularly useful for obtaining odds ratio estimates for various levels of the CLASS variables.

Previous | Next | Top of Page

SAS PROC SURVEYLOGISTIC

Syntax

The following statements are available in PROC SURVEYLOGISTIC:

PROC SURVEYLOGISTIC < options >**;**

BY variables **;**

CLASS variable <(v-options)> <variable <(v-options)>... >
< / v-options >**;**

CLUSTER variables **;**

CONTRAST 'label' effect values <,... effect values>< /options >**;**

FREQ variable **;**

MODEL events/trials = < effects > < / options >**;**

MODEL variable < (variable_options) > = < effects > < / options >**;**

STRATA variables < / options > **;**
< *label:* > **TEST** equation1 < , ... , < equationk >> < /option >**;**

UNITS independent1 = list1 < ... independentk = listk > < /option > **;**

WEIGHT variable </ option >**;**

The PROC SURVEYLOGISTIC and MODEL statements are required.

The CLASS, CLUSTER, STRATA, and CONTRAST statements can appear multiple times.

You should only use one MODEL statement and one WEIGHT statement.

The CLASS statement (if used) must precede the MODEL statement, and the CONTRAST statement (if used) must follow the MODEL statement.

* TILASTOLLINEN MALLI

Logitmalli

$$\log\left(\frac{\boldsymbol{p}}{1-\boldsymbol{p}}\right)=\boldsymbol{X}\boldsymbol{b}$$

missä
$\boldsymbol{p}$ on tuntematon osuusparametrien (u×1) vektori
u on osajoukkojen lukumäärä

$\boldsymbol{X}$ on (u×s) mallimatriisi, s on mallin estimoitavien parametrien lkm

$\boldsymbol{b}$ on estimoitavien parametrien (s×1) vektori

HUOM: Malliyhtälön vasen puoli: Epälineaarinen logitfunktio

Oikea puoli: Lineaarinen rakenne, jossa on termit selittäjien päävaikutuksia ja yhdysvaikutuksia varten

Osuusparametrin *p* tarkentuva estimaattori

$\hat{\boldsymbol{p}}=(\hat{p}_1,\hat{p}_2,...,\hat{p}_u)'$, missä kukin $\hat{p}_j$ koostuu

osuusestimaattoreista $\hat{p}_j=\frac{y_j}{x_j}, \quad j=1,...,u$

(vrt. 1. teema)

* LOGITMALLIN PARAMETRIVEKTORIN *b* ESTIMOINTI

a) GWLS-estimointi / Yleistetty painotettu PNS

Generalized Weighted Least Squares
Ei-iteratiivinen menetelmä (laskennallisesti kevyt)

$$\hat{\boldsymbol{b}} = (\boldsymbol{X}'(\boldsymbol{H}\hat{\boldsymbol{V}}_{\boldsymbol{des}}\boldsymbol{H})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{H}\hat{\boldsymbol{V}}_{\boldsymbol{des}}\boldsymbol{H})^{-1}F(\hat{\boldsymbol{p}})$$

missä $\hat{\boldsymbol{V}}_{\boldsymbol{des}}$ on osuusestimaattorivektorin $\hat{\boldsymbol{p}}$ asetelmaperusteinen kovarianssimatriisiestimaattori (SAS: Linearisointimenetelmä tai jacknife)

$\boldsymbol{H}\hat{\boldsymbol{V}}_{\boldsymbol{des}}\boldsymbol{H}$ on funktiovektorin $F(\hat{\boldsymbol{p}})$ kov.matriisin asetelman suhteen tarkentuva estimaaattori

Matriisin $\boldsymbol{H}$ diagonaalialkiot ovat $h_j = 1/(\hat{p}_j(1-\hat{p}_j))$

Kerroinvektorin $\hat{\boldsymbol{b}}$ asetelmaperusteinen kovarianssimatriisiestimaattori on:

$$\hat{\boldsymbol{V}}_{\boldsymbol{des}}(\hat{\boldsymbol{b}}) = (\boldsymbol{X}'(\boldsymbol{H}\hat{\boldsymbol{V}}_{\boldsymbol{des}}\boldsymbol{H})^{-1}\boldsymbol{X})^{-1}$$

Laskenta: Helppo ohjelmoida itse (esim. SAS/IML-matriisialgebran kielellä)

b) Pseudo-uskottavuusestimointi (PML)

Suurimman uskottavuuden estimointiyhtälöiden täydentäminen alkiotason painomatriisilla

PML on **iteratiivinen menetelmä** (laskennallisesti raskaampi kuin GWLS)

Parametrivektorin *b* asetelman suhteen tarkentuva estimointi
Ratkaistaan iteratiivisesti PML-estimointiyhtälöt

$$\boldsymbol{X}'\boldsymbol{W}\boldsymbol{f}(\hat{\boldsymbol{b}}) = \boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{p}}$$

missä $\boldsymbol{W}$ on diagonaalinen $u \times u$ painomatriisi, jonka diagonaalialkiot ovat $\hat{n}_j$ (analyysipainoilla painotetut osajoukkojen otoskoot)

$\boldsymbol{f} = \exp(\boldsymbol{X}\boldsymbol{b})/(1+\exp(\boldsymbol{X}\boldsymbol{b}))$ on logitfunktion käänteisfunktio

Mallilla sovitetut osuudet $\hat{f}_j$ estimoidaan lausekkeella

$$\hat{\boldsymbol{f}} = \exp(\boldsymbol{X}\hat{\boldsymbol{b}})/(1+\exp(\boldsymbol{X}\hat{\boldsymbol{b}}))$$

PML-estimaattorivektorin $\hat{\boldsymbol{b}}$ asetelman suhteen tarkentuva kovarianssimatriisiestimaattori on muotoa (“Sandwich form”, “Robust”, “Empirical”)

$$\hat{\boldsymbol{V}}_{des}(\hat{\boldsymbol{b}}) = \hat{\boldsymbol{V}}_{bin}(\hat{\boldsymbol{b}})\boldsymbol{X}'\boldsymbol{W}\hat{\boldsymbol{V}}_{des}\boldsymbol{W}\boldsymbol{X}\hat{\boldsymbol{V}}_{bin}(\hat{\boldsymbol{b}})$$

missä

$$\hat{\boldsymbol{V}}_{bin}(\hat{\boldsymbol{b}}) = (\boldsymbol{X}'\boldsymbol{W}\hat{\Delta}\boldsymbol{W}\boldsymbol{X})^{-1}$$

on ML-estimointia vastaava binominen estimaattori, jossa $u \times u$ matriisin $\hat{\Delta}$ diagonaalialkiot ovat binomisia varianssiestimaatteja $\hat{f}_j(1-\hat{f}_j)/\hat{n}_j$

Laskenta: SAS / SURVEYLOGISTIC

Binäärinen tulosmuuttuja

Logit ANOVA
Logistinen regressioanalyysi
Logistinen ANCOVA

Asetelmaperusteiset Waldin ja F-testit
Rao-Scott-menetelmällä korjatut SRS-testit

c) GEE-estimointi Generalized Estimating Equations / **GWEE-estimointi** (painotettu GEE)

ML-estimoinnin ja PML-estimoinnin yleistyksiä

Perustuu **monimuuttujaiseen kvasi-uskottavuusestimointiin** (multivariate quasilikelihood), Iteratiivinen menetelmä

GEE/GWEE-variantit survey-analyysissa:

GEE/GWEE-IND - “Independent working correlation”
Vastaa ML/PML-estimointia

GEE/GWEE-EXCH - “Exchangeable working corr.”

Laskenta:
SUDAAN / REGRESS/LOGISTIC/MULTILOG
SAS/GENMOD (Yleistetyt lineaariset mallit)

Logit ANOVA
Logistinen regressioanalyysi
Logistinen ANCOVA

Asetelmaperusteiset Waldin ja F-testit
Rao-Scott-menetelmällä korjatut SRS-testit

* LOGITMALLIN PARAMETREJA KOSKEVIEN LINEAARISTEN HYOPTEESIEN TESTAUS

Nollahypoteesi

$$H_0: \boldsymbol{Cb} = 0$$

Asetelmaperusteinen Waldin testisuure

$$X^2_{des}(\boldsymbol{b}) = (\boldsymbol{C}\hat{\boldsymbol{b}})'(\boldsymbol{C}\hat{\boldsymbol{V}}_{\boldsymbol{des}}(\hat{\boldsymbol{b}})\boldsymbol{C}')^{-1}(\boldsymbol{C}\hat{\boldsymbol{b}})$$

missä $\boldsymbol{C}$ on $c \times s$ kontrastimatriisi

Asymptoottisesti χ^2-jakautunut vapausastein df=c

Vaihtoehtoinen testisuure:

Toisen kertaluvun Rao-Scott-korjattu SRS-perusteinen testisuure

ks: Lehtonen and Pahkinen (2004) ss. 272-275

Mallin yksittäisiä parametreja koskevat testit

Ho: $b_k = 0$, $k = 1,...,s$

Asetelmaperusteinen Wald testisuure

$$X^2_{des}(b_k) = \frac{\hat{b}_k^2}{\hat{v}_{des}(\hat{b}_k)} \qquad k = 1,...,s$$

joka on asymptoottisesti χ^2-jakautunut vapausastein df=1

HUOM:

Vastaava t-testisuure on:

$$t_{des}(b_k) = \frac{\hat{b}_k}{s.e_{des}(\hat{b}_k)} \qquad k = 1,...,s$$

(Vastaavan Waldin testisuureen merkkinen neliöjuuri)

NOTE: SAS: Mallien parametrien t-testisuureet ovat F-korjatun asetelmaperusteisen Waldin testisuureen merkkisiä neliöjuuria

YHTEENVETO: Ryvästymiseen reagointi eri estimointimenetelmissä

	Reagointi ryvästymiseen	
	Mallin parametrien estimoinnissa	Varianssien estimoinnissa
Least squares		
LS	Ei	Ei
GWLS	Kyllä	Kyllä
Likelihood		
ML	Ei	Ei
PML	Ei	Kyllä
GEE-IND	Ei	Kyllä
GEE-EXCH	Kyllä	Kyllä

LS	Tavanomainen PNS-estimointi (Least Squares)
GWLS	Painotettu PNS (Generalized Weighted Least Squares)
ML	Tavanomainen suurimman uskottavuuden estimointi (Standard Maximum Likelihood)
PML	Pseudo-uskottavuusestimointi (Pseudo Maximum Likelihood)
GEE	Yleistetyt estimointiyhtälöt (Generalized Estimating Equations)
GEE-IND	Vastaa PML-menetelmää
GEE-EXCH	Monimuuttujainen kvasi-uskottavuusmenetelmä (Multivariate quasilikelihood)