Center for Statistik
Handelshøjskolen i København

MPAS
Tue Tjur
December 2004

Notat 2:

Hosmer–Lemeshow's test for goodness–of–fit i en logistisk regressionsmodel

Vi betragter en logistisk regressionsmodel for binomialfordelte observationer $y_1, \ldots, y_n$ specificeret på formen

$$y_i \sim \text{bin}(m_i, p_i)\,, \quad p_i = \frac{\exp(\alpha + \beta x_i + \ldots)}{1 + \exp(\alpha + \beta x_i + \ldots)}$$

hvor $\alpha + \beta x_i + \ldots$ står for et sædvanligt lineært udtryk i regressionsvariable, faktorer osv. Under forudsætning af at alle antalsparametrene er tilpas store, og ingen fittede værdier $m_i\hat{p}_i$ ligger for tæt på de ekstreme værdier 0 og m_i, kan man til modelkontrol benytte *Pearson's test for goodnes–of–fit,*

$$\sum_{i=1}^{n} \frac{(y_i - m_i\hat{p}_i)^2}{m_i\hat{p}_i(1-\hat{p}_i)}$$

Denne teststørrelse vil, under forudsætning af at modellen er korrekt, være approksimativt χ^2–fordelt med $n - d$ frihedsgrader, hvor d er antal parametre i modellen. En (streng) tommelfingerregel for hvornår denne approksimation kan bruges er, at alle de fittede værdier og deres komplementære skal være ≥ 5, altså

$$5 \leq m_i\hat{p}_i \leq m_i - 5 \text{ for alle } i\,.$$

Asymptotisk er dette test ækvivalent med kvotienttestet (som man lige så godt kan bruge) for modellen imod den fulde model (dvs. modellen som har en frit varierende sandsynlighedsparameter for hver observation).

Hvis antalsparametrene er små, og i særdeleshed hvis de alle sammen er 1 (som de er i logistiske regressionsmodeller for egentlige binære data) kan dette test overhovedet ikke bruges. For at forstå hvor galt det kan gå hvis man gør det alligevel, kan man overveje følgende. Betragt de tre kvotienttesstørrelser

X_1: kvotienttesstørrelsen for reduktion af den fulde model til den model vi er interesseret i.

X_2: kvotienttesstørrelsen for reduktion af den model vi er interesseret i til den trivielle model givet ved $p_i = \frac{1}{2}$ for alle i.

X: kvotienttestst\o rrelsen for reduktion af den fulde model til den trivielle model ($p_i = \frac{1}{2}$ for alle i).

X_2 er en særdeles respektabel teststørrelse, som man uden tøven kan benytte i situationer hvor antallet af parametre i modellen er væsentligt mindre end antallet af observationer. Under hypotesen om at den trivielle model er korrekt (den er sjældent relevant, men det kan være ligemeget i dette tankeeksperiment) vil X_2 være approksimativt χ^2–fordelt med d (= antal parametre i modellen) frihedsgrader.

Hvis alle antalsparametrene er 1 gælder der imidlertid, som man let ser,

$$X_1 + X_2 = X = 2n\log(2)$$

hvoraf umiddelbart følger, at X_1's fordeling ikke har den fjerneste lighed med en χ^2–fordeling med $n - d$ frihedsgrader, og i øvrigt heller ikke kan fortolkes som et mål for "goodness–of–fit".

I øvrigt kan man også bemærke at "teststørrelsen" X — som altså slet ikke er stokastisk, men altid lig med $2n\log(2)$ — i sig selv er et eksempel på en kvotienttestst\o rrelse for en model (den sædvanlige "møntkastmodel") imod den fulde model. Den tilsvarende Pearson teststørrelse

$$\sum_{i=1}^{n} \frac{(y_i - \frac{1}{2})^2}{\frac{1}{2} \times (1 - \frac{1}{2})}$$

ses også at være konstant — men med anden værdi, nemlig n (!)

Modelkontrol i en logistisk regressionsmodel for binære data må således foretages ved hjælp af andre metoder. En lettere forbedret udgave af *Hosmer–Lemeshow's test* går ud på at gruppere observationerne, sædvanligvis i 10 grupper, efter værdierne af de fittede værdier $\hat{p}_i$. Vi vil her vil nøjes med at forklare testet i tilfældet hvor alle antalsparametre er 1. Det er ikke svært at omskrive formlerne, så de også kan bruges når antalsparametrene kan være > 1. Betragt summerne

$$\begin{aligned} s_1 &= \sum_{i:\hat{p}_i \in [0.0,0.1[} y_i \\ &\vdots \qquad\qquad \vdots \\ s_{10} &= \sum_{i:\hat{p}_i \in [0.9,1.0]} y_i \end{aligned}$$

Disse antal er åbenbart stokastisk uafhængige, med fordelinger som næsten (men ikke helt) er binomialfordelinger. Disse variables middelværdier og varianser kan umiddelbart beregnes (idet vi ignorerer at selve de mængder der summeres over i virkeligheden er stokastiske):

$$m_j = \mathrm{E}(s_j) = \sum_{i:\hat{p}_i \in [\frac{j-1}{10}, \frac{j}{10}[} p_i \,,$$

$$v_j = \mathrm{var}(s_j) = \sum_{i:\hat{p}_i \in [\frac{j-1}{10}, \frac{j}{10}[} p_i(1-p_i)\,.$$

Hvis vi med $\hat{m}_i$ og $\hat{v}_i$ betegner de tilsvarende estimerede størrelser (altså samme udtryk som ovenfor, bare med $\hat{p}_i$ indsat alle vegne i stedet for p_i) er det nærliggende at benytte størrelsen

$$\sum_{j=1}^{10} \frac{(s_j - \hat{m}_j)^2}{\hat{v}_j}$$

som surrogat for Pearson's teststørrelse. Det er ikke helt klart hvor mange frihedsgrader dens χ^2–fordeling skal forsynes med, og det jo heller ikke på forhånd givet, at den overhovedet er approksimativt χ^2–fordelt. Men simulationsstudier viser, at den i pæne tilfælde (idéelt set når s_j'erne og deres komplementære allesammen er ≥ 5, i praksis under noget mere beskedne krav) har en fordeling, der kan approksimeres godt med en χ^2–fordeling med $10 - 2 = 8$ frihedsgrader.

Hvis de fittede værdier fordeler sig i enhedsintervallet på en så ujævn måde, at nogle af de 10 intervaller indeholder meget få eller slet ingen værdier, må man modificere testet ved at ændre inddelingen af enhedsintervallet, så alle intervaller kommer til at indeholde et rimeligt antal fittede værdier. I så fald bliver det antal frihedsgrader man skal bruge i tabelopslaget naturligvis antallet af intervaller minus 2.

I ISU kan Hosmer–Lemeshows teststørrelse (i tilfælde af binære observationer) for eksempel beregnes på følgende måde. Efter

```
FITLOGITLINEAR Y= ...
```

skrives

```
SAVEFITTED p
FACTOR g n 10 { n = antal obs. }
COMPUTE g=INT(1+10*p)
COMPUTE v=p*(1-p)
TABULATE pp=p|vv=v|yy=y g
tt=sqr(yy-pp)/vv
HLtest=sum(tt)
LIST HLtest
```

I SAS kan man, ifølge Agresti (reference nedenfor) benytte en option ved navn `LACKFIT` til `PROC LOGISTIC`.

Reference:

Alan Agestri: *An Introduction to Categorical Data Analysis.*
Wiley 1996.