Center for Statistik
Handelshøjskolen i København

MPAS
Tue Tjur
Marts 2006

Notat 2:

Fortolkning af logistiske regressionsmodeller i forbindelse med projekt 2.

Som nævnt i databeskrivelsen er der "foretaget en oversampling af data", hvilket (iflg. Christian Haxholdt) betyder at nogle af "ikke–churnerne" er udeladt. Man må gå ud fra, at de ikke–churnere vi har med i datasættet er tilfældigt udvalgt blandt ikke–churnerne i det oprindelige datasæt. Det angives, at "a priori sandsynligheden" for churn i perioden er 0.04, medens den faktiske hyppighed af churn i vores datasæt er 0.134161. Det skulle altså betyde, at antallet 17367 af ikke–churnere i datasættet er fremkommet som

$$17367 = \pi_0 \times N_0$$

hvor N_0 betegner antallet af ikke–churnere i den oprindelige medlemsdatabase og π_0 er sandsynligheden for udvælgelse. Til bestemmelse af π_0 og N_0 har vi endvidere relationen

$$\frac{2691}{N_0 + 2691} = 0.04$$

som udtrykker at den relative hyppighed af churnere i det oprindelige datasæt netop er 0.04. Løsning af disse to ligninger giver

$$N_0 = 64584 \quad \text{og} \quad \pi_0 = 0.268906.$$

Det er en eksklusiv egenskab ved den logistiske regressionsmodel, at denne form for skævvridning af et datasæt kan indbygges i modellen på en meget simpel måde.

Betragt følgende lidt mere generelle situation. Antag at vi har et (typisk stort) datasæt med en binær variabel y, som kan beskrives ved en logistisk regressionsmodel

$$p_i = P(Y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

hvor $\eta_i = \log \frac{p_i}{1-p_i}$ på sædvanlig måde kan skrives som et konstantled plus en linearkombination af baggrundsvariable. Vi betragter nu et "udtyndet" datasæt, som er fremkommet ved at vi for hver observation i det oprindelige datasæt har medtaget denne med en sandsynlighed der

afhænger af om y_i er 0 eller 1. Indikatorerne U_i for medlemsskab af det udtyndede datasæt får således fordelinger givet ved

$$P(U_i = 1 \mid Y_i = y) = \begin{cases} \pi_1 \text{ for } y = 1 \\ \pi_0 \text{ for } y = 0 \end{cases}$$

For en observation fra det udtyndede datasæt kan sandsynligheden for en positiv respons nu udregnes som den betingede sandsynlighed

$$p_i' = P(Y_i = 1 \mid U_i = 1) = \frac{P(Y_i = 1 \text{ og } U_i = 1)}{P(U_i = 1)}$$

$$= \frac{P(Y_i = 1)P(U_i = 1 \mid Y_i = 1)}{P(Y_i = 1)P(U_i = 1 \mid Y_i = 1) + P(Y_i = 0)P(U_i = 1 \mid Y_i = 0)}$$

$$= \frac{p_i \pi_1}{p_i \pi_1 + (1 - p_i)\pi_0}$$

Heraf følger at

$$\log \frac{p_i'}{1 - p_i'} = \log \frac{p_i \pi_1}{(1 - p_i)\pi_0} = \eta_i + \log \pi_1 - \log \pi_0.$$

Det udtyndede datasæt er således igen beskrevet ved en logistisk regressionsmodel, hvor alle de interessante parametre (koefficienter til regressionsvariable, hovedvirkninger af faktorer osv.) er de samme som i modellen for det store datasæt, kun konstantleddet er ændret ved addition af $\log \pi_1 - \log \pi_0$.

I det konkrete tilfælde betyder dette, at vi efter estimation af parametrene i en logistisk regressionsmodel for det udtyndede datasæt kan fortolke disse som estimater for parametre i modellen for det oprindelige datasæt, når blot konstantleddet ændres ved subtraktion af

$$\log \pi_1 - \log \pi_0 = \log(1) - \log(0.268906) = 1.3134.$$

Tilsvarende kan fittede værdier $\hat{p}_i'$, estimeret i det udtyndede datasæt, ved en simpel transformation føres over i størrelser $\hat{p}_i$ der kan fortolkes som estimerede sandsynligheder for churn i det oprindelige datasæt:

$$\hat{p}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)} = \frac{\exp\left(\log \frac{\hat{p}_i'}{1 - \hat{p}_i'} - 1.3134\right)}{1 + \exp\left(\log \frac{\hat{p}_i'}{1 - \hat{p}_i'} - 1.3134\right)}.$$