

Notat 5:

**Fortolkning af logistiske regressionsmodeller
i forbindelse med opgave 2.**

Antag, for det (meget store) fuldstændige datasæt bestående af samtlige lån, at

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

hvor Y_i er indikatoren for tab (1 hvis der er tab i forbindelse med lån nr. i , 0 ellers), og $\eta_i = \gamma + \dots$ på sædvanlig måde er en linearkombination af baggrundsvARIABLE hørende til lån nr. i . Datasættet antages, kort sagt, beskrevet ved en passende logistisk regressionsmodel.

Nu betragter vi det "udtyndede" datasæt (altså det vi faktisk har til rådighed). Indikatorerne U_i for medlemsskab af det udtyndede datasæt har fordelinger givet ved

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

Parametrene π_1 og π_0 er i det konkrete tilfælde angivet til 1 og 0.008, svarende til at alle med respons 1 er udtaget, medens lån med respons 0 er udtaget med sandsynlighed 0.8%.

For et "tilfældigt" lån fra det udtyndede datasæt har vi nu

$$\begin{aligned} p'_i = P(Y_i = 1 \mid U_i = 1) &= \frac{P(Y_i = 1 \text{ og } U_i = 1)}{P(U_i = 1)} \\ &= \frac{P(Y_i = 1)P(U_i = 1 \mid Y_i = 1)}{P(Y_i = 1)P(U_i = 1 \mid Y_i = 1) + P(Y_i = 0)P(U_i = 1 \mid Y_i = 0)} \\ &= \frac{p_i \pi_1}{p_i \pi_1 + (1 - p_i) \pi_0} \end{aligned}$$

Heraf følger let at

$$\eta'_i = \log \frac{p'_i}{1 - p'_i} = \log \frac{p_i}{1 - p_i} + \log \frac{\pi_1}{\pi_0} = \eta_i + \log \pi_1 - \log \pi_0.$$

I ord: Det udtyndede datasæt er igen beskrevet ved en logistisk regressionsmodel, hvor alle de interessante parametre er de samme, bortset fra at konstantleddet er ændret ved addition af $\log \pi_1 - \log \pi_0$. I det

konkrete tilfælde betyder det, 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 “store” datasæt, når blot konstantleddet ændres ved subtraktion af $-\log 0.008 = 4.8283$.