Logistic regression for case-control studies (Prentice & Pyke, 1979, Biometrika)

Setup: random sample of size n_1 from case population ($Y=1$)
random sample of size n_0 from control population ($Y=0$)

Observations on an explanatory variable X (assume only one X):

$$X_{11}, \cdots, X_{1n_1} \quad \text{iid} \quad \Pr(X=x \mid Y=1)$$
$$X_{01}, \cdots, X_{1n_0} \quad \text{iid} \quad \Pr(X=x \mid Y=0).$$

Assume for simplicity that X has a finite number of possible levels $x_0, \cdots, x_K$.

x_0 = reference level

1) Prospective model vs retrospective model

Prospective model

$$\text{logit}\ \Pr(Y=1 \mid X=x) = \alpha + \beta x.$$

Now consider $X=x$ vs $X=x_0$. From above

$$(1) \quad \frac{\Pr(Y=1 \mid X=x)/\Pr(Y=0 \mid X=x)}{\Pr(Y=1 \mid X=x_0)/\Pr(Y=0 \mid X=x_0)} = e^{\beta(x-x_0)}.$$

This implies

$$(2) \qquad \frac{Pr(X=x \mid Y=1) / Pr(X=x_0 \mid Y=1)}{Pr(X=x \mid Y=0) / Pr(X=x_0 \mid Y=0)} = e^{\beta(x-x_0)}$$

(previously showed that LHS(1) = LHS(2)).

Define $\gamma_k = \log [Pr(X=x_k \mid Y=0) / Pr(X=x_0 \mid Y=0)]$.

(note $\gamma_0 \equiv 0$)

Then we can write

$$Pr(X=x_k \mid Y=0) = e^{\gamma_k} / \sum_{\ell=0}^{K} e^{\gamma_\ell}$$

$$Pr(X=x_k \mid Y=1) = e^{\gamma_k + \beta x_k} / \sum_{\ell=0}^{K} e^{\gamma_\ell + \beta x_\ell} .$$

2) Approaches to inference : a) likelihood inference based directly on above

b) likelihood inference based on rewritten form of model

3) Rewrite model

Define $q_k = (1-\rho)\,Pr(X=x_k|Y=0) + \rho\,Pr(X=x_k|Y=1)$,

where $\rho = \frac{n_1}{n}$ $(n = n_0 + n_1)$.

Also define $\delta_0 = \log\left[(1-\rho) / \sum_{\ell=0}^{k} e^{\gamma_\ell}\right]$

$\delta_1 = \log\left[\rho / \sum_{\ell=0}^{k} e^{\gamma_\ell + \beta x_\ell}\right]$

Then we can write

$$Pr(X=x_k|Y=0) = \left(\frac{1}{1-\rho}\right) e^{\gamma_k + \delta_0}$$

$$= q_k (1-\rho)^{-1} \left[e^{\gamma_k+\delta_0} / (e^{\gamma_k+\delta_0} + e^{\gamma_k+\delta_1+\beta x_k})\right]$$

$$= q_k (1-\rho)^{-1} \left[1 / (1 + e^{\alpha+\beta x_k})\right]$$

$$(\alpha = \delta_1 - \delta_0)$$

and $Pr(X=x_k|Y=1) = q_k \rho^{-1} \left[e^{\alpha+\beta x_k} / (1+e^{\alpha+\beta x_k})\right]$.

Above subject to: $\Sigma_k q_k = 1$

$$\left.\begin{aligned} \Sigma_k\, Pr(X=x_k|Y=0) &= 1 \\ \Sigma_k\, Pr(X=x_k|Y=1) &= 1 \end{aligned}\right\} (*)$$

4) Likelihood-based inference : n_{ik} = # obs in grp i with $X = x_k$

$$L = \left[\prod_{k=0}^{K} \Pr(X = x_k | Y = 0)^{n_{0k}} \right] \times \left[\prod_{k=0}^{K} \Pr(X = x_k | Y = 1)^{n_{1k}} \right]$$

$$= \prod_{k=0}^{K} q_k^{n_{0k}+n_{1k}} (1-\rho)^{-n_{0k}} \rho^{-n_{1k}} \left(\frac{1}{1+e^{\alpha+\beta x_k}} \right)^{n_{0k}} \times \left(\frac{e^{\alpha+\beta x_k}}{1+e^{\alpha+\beta x_k}} \right)^{n_{1k}}.$$

$$\mathcal{L} = \log L = \text{const} + \sum_{k=0}^{K} (n_{0k}+n_{1k}) \log q_k + \sum_{k=0}^{K} n_{1k} [\alpha + \beta x_k] - \sum_{k=0}^{K} (n_{0k}+n_{1k}) \log (1+e^{\alpha+\beta x_k}).$$

Ignoring the constraints (*), $\mathcal{L}$ is maximized by:

i) $\hat{q}_k = \frac{n_{0k}+n_{1k}}{n}$

ii) solving

$$\frac{\partial \mathcal{L}}{\partial \alpha} = \sum_{k=0}^{K} n_{1k} - \sum_{k=0}^{K} (n_{0k}+n_{1k}) \frac{e^{\alpha+\beta x_k}}{1+e^{\alpha+\beta x_k}} \overset{\text{set}}{=} 0$$

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{k=0}^{K} n_{1k} x_k - \sum_{k=0}^{K} (n_{0k}+n_{1k}) x_k \frac{e^{\alpha+\beta x_k}}{1+e^{\alpha+\beta x_k}} \overset{\text{set}}{=} 0.$$

Same equations as prospective logistic regression .....

Now, in the development just above we ignored the constraints (*), but the solution the results happens to satisfy the constraints automatically. Indeed, we have the estimating equation

$$\sum_{k=0}^{K} n_{1k} - \sum_{k=0}^{K} (n_{0k} + n_{1k}) \frac{e^{\hat{\alpha}+\hat{\beta}x_k}}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = 0 \qquad (1)$$

and we also have $\hat{q}_k = \frac{n_{0k} + n_{1k}}{n}$.

Combining these we get

$$\sum_{k=0}^{K} n_{1k} - \sum_{k=0}^{K} n\hat{q}_k \frac{e^{\hat{\alpha}+\beta x_k}}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = 0$$

$$\Leftrightarrow \quad n \sum_{k=0}^{K} \hat{q}_k \frac{e^{\hat{\alpha}+\beta x_k}}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = n_1$$

$$\Leftrightarrow \quad \sum_{k=0}^{K} \hat{p}^{-1} \hat{q}_k \frac{e^{\hat{\alpha}+\beta x_k}}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = 1$$

$$\Leftrightarrow \quad \sum_{k=0}^{K} \hat{Pr}(X = x_k \mid Y = 1) = 1$$

Further, going back to (1) and using the relation

$$\frac{1}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = 1 - \frac{e^{\hat{\alpha}+\hat{\beta}x_k}}{1+e^{\hat{\alpha}+\hat{\beta}x_k}}$$

we get

$$\sum_{k=0}^{K} n_{1k} - \sum_{k=0}^{K} (n_{0k} + n_{1k}) + \sum_{k=0}^{K} (n_{0k}+n_{1k}) \frac{1}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = 0$$

$$\Leftrightarrow \quad \sum_{k=0}^{K} n\hat{q}_k \frac{1}{1+e^{\hat{\alpha}+\hat{\beta}x_k}} = n_0$$

$$\Leftrightarrow \quad \sum_{k=0}^{K} \hat{Pr}(X = x_k \mid Y = 0) = 1$$

It turns out that the analysis based on the retrospective likelihood leads not only to the same estimates as an analysis based on the prospective likelihood, but also the asymptotic $Cov(\hat{\beta})$ is the same under the two analyses. The proof of this latter fact is more complex; see Prentice and Pyke (1979, Biometrika) for details.