<u>10 November 2020</u>

Binary response

<u>Data set</u>

1. Ungrouped data: Response - y_i for $i = 1, \ldots, n$

   Assume $y_i \sim Bernoulli(P_i)$

2. Grouped data: Response - (n_i, r_i) for $i = 1, \ldots, s$

   Assume $r_i \sim Binomial(n_i, P_i)$

<u>Problems if linear model is used</u>

Consider the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \quad \begin{cases} i = 1, 2, \ldots, n \\ y_i = 0, 1 \end{cases}$$

If we assume the usual $E(e_i) = 0$, we have $E(y_i) = P_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$ and $\mathrm{Var}(e_i) = P_i Q_i$

The other two problems on the above linear regression, i.e., fitting y_i on $x_1, \ldots, x_p$

1. Estimate of y_i is imprecise for ungrouped data.
2. $\hat{P}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}$ may not be within 0 and 1.

Link function

The link function provides the relationship between the linear predictor (linear combination of unknown parameter $\underset{\sim}{\beta}$) and the mean of the distribution function, i.e., $E(y_i) = \mu_i$. It is defined as $g(\mu)$.

For linear model,

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}$$

Then, $g(\mu_i) = \mu_i$ — Identity function

For binomial data,

1.

$$\begin{aligned} P_i &= \frac{e^{\underset{\sim}{x}_i^T \underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i^T \underset{\sim}{\beta}}} \\ \log\left[\frac{\mu_i}{1 - \mu_i}\right] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \end{aligned}$$

   Then, $g(\mu_i) = \log\left[\frac{\mu_i}{1-\mu_i}\right]$ — Logit function

2.

$$\begin{aligned} P_i &= \Phi(\underset{\sim}{x}_i^T \underset{\sim}{\beta}) \\ \Phi^{-1}(\mu_i) &= \underset{\sim}{x}_i^T \underset{\sim}{\beta} \end{aligned}$$

   Then, $g(\mu_i) = \Phi^{-1}(\mu_i)$ — Probit function

Weighted least squares

By minimizing

$$SS_{\text{Res},\boldsymbol{V}} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$

Then,

- $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$

- $\text{Var}(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$

As $\boldsymbol{V} = \text{Var}(\boldsymbol{\varepsilon}) = diag[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$, the weighted least squares estimator of $\boldsymbol{\beta}$ can be obtained by minimizing

$$SS_{\text{Res(weighted)}} = \sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$$

where $w_i = 1/\sigma_i^2$.

Weighted least squares for grouped data with large n_i

We fit the model as follows:

$$\log\left[\frac{\hat{P}_i}{1-\hat{P}_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + e_i \qquad i = 1, 2, \ldots, s$$

with $\hat{w}_i = \dfrac{1}{\hat{\sigma}_i^2} \approx n_i\hat{P}_i(1-\hat{P}_i)$, where $\hat{P}_i = r_i/n_i$.