MAXIMUM LIKELIHOOD ESTIMATION

• Given samples $(x_i, y_i) \in R^p \times \{0, 1\}$, $i = 1, \ldots n$, we let $p(x_i) = P(y_i = 1|x_i)$, and assume

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta^T x_i, \quad i = 1, \ldots n$$

• To construct an estimate $\hat{\beta}$ of the coefficients, we will use the principle of maximum likelihood. I.e., assuming independence of the samples, the likelihood (conditional on x_i, $i = 1, \ldots n$) is

$$\begin{aligned} L(\beta) &= \prod_{i\,:\,y_i=1} p(x_i) \cdot \prod_{i\,:\,y_i=0} (1 - p(x_i)) \\ &= \prod_{i=1}^{n} p(x_i)^{y_i} \big(1 - p(x_i)\big)^{1-y_i}. \end{aligned}$$

We will choose $\hat{\beta}$ to maximize this likelihood criterion

• Note that maximizing a function is the same as maximizing the log of a function (because log is monotone increasing). Therefore $\hat{\beta}$ is equivalently chosen to maximize the log likelihood

$$\ell(\beta) = \sum_{i=1}^{n} y_i \log p(x_i) + (1 - y_i) \log\big(1 - p(x_i)\big).$$

It helps to rearrange this as

$$\begin{aligned} \ell(\beta) &= \sum_{i=1}^{n} y_i \big[\log p(x_i) - \log\big(1 - p(x_i)\big)\big] + \log\big(1 - p(x_i)\big) \\ &= \sum_{i=1}^{n} y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right) + \log\big(1 - p(x_i)\big). \end{aligned}$$

Finally, plugging in for $\log(p(x_i)/(1 - p(x_i))) = x^T{}_i\, \beta$ and using $1 - p(x_i) = 1/(1 + \exp(x^T{}_i\, \beta))$, $i = 1, \ldots n$,

$$\ell(\beta) = \sum_{i=1}^{n} y_i(x_i^T\beta) - \log\left(1 + \exp(x_i^T\beta)\right). \tag{1}$$

You can see that, unlike the least squares criterion for regression, this criterion ℓ (β) does not have a closed-form expression for its maximizer (e.g., try taking its partial derivatives and setting them equal to zero). Hence we have to run an optimization algorithm to find $\hat{\beta}$

- Somewhat remarkably, we can maximize (1) by running repeated weighted least squares regressions! For those of you who have learned a little bit of optimization, this is actually just an instantiation of Newton's method. Applied to the criterion (1), we refer to it as iteratively reweighted least squares or IRLS
- In short: estimation of $\hat{\beta}$ in logistic regression is more involved than it is in linear regression, but it is possible to do so by iteratively using linear regression software.