Logistic Regression

Y_i is either 0 or 1 (2 classes)

$Y_i \sim \text{Bernoulli}(f_1(x_i))$

$$f_1(\underline{x}_i) = \frac{e^{\underline{x}_i' \underline{\beta}}}{1 + e^{\underline{x}_i' \underline{\beta}}}$$

Implicitly, $f_0(\underline{x}_i) = 1 - f_1(\underline{x}_i)$

$$= \frac{1}{1 + e^{\underline{x}_i' \underline{\beta}}}$$

Multinomial Logistic Reg.

K possible categories for the response.

Y_i is one of the integers $\{1, 2, \ldots, K\}$

$Y_i \sim \text{Categorical}(f_1(x_i), f_2(x_i), \ldots, f_K(x_i))$

$$f_1(x_i) = \frac{1}{1 + e^{\underline{x}_i' \underline{\beta}^{(2)}} + e^{\underline{x}_i' \underline{\beta}^{(3)}} + \cdots + e^{\underline{x}_i' \underline{\beta}^{(K)}}}$$

$$f_2(x_i) = \frac{e^{x_i' \underline{\beta}^{(2)}}}{1 + e^{\underline{x}_i' \underline{\beta}^{(2)}} + e^{\underline{x}_i' \underline{\beta}^{(3)}} + \cdots + e^{\underline{x}_i' \underline{\beta}^{(K)}}}$$

$$\vdots$$

$$f_K(x_i) = \frac{e^{x_i' \beta^{(K)}}}{1 + e^{\underline{x}_i' \underline{\beta}^{(2)}} + e^{\underline{x}_i' \underline{\beta}^{(3)}} + \cdots + e^{\underline{x}_i' \underline{\beta}^{(K)}}}$$

Note: for each class j, $0 \le f_j(x_i) \le 1$, and

$$f_1(x_i) + f_2(x_i) + \cdots + f_K(x_i) = 1$$

Suppose x_{i1} increases by 1 unit.

Relative to the baseline response category,

$$\frac{P(Y_i=2)}{P(Y_i=1)} = \frac{f_2(x_i)}{f_1(x_i)} = \frac{\left(\dfrac{e^{x_i'\beta^{(2)}}}{1+e^{x_i'\beta^{(2)}}+\dots+e^{x_i'\beta^{(K)}}}\right)}{\left(\dfrac{1}{1+e^{x_i'\beta^{(2)}}+\dots+e^{x_i'\beta^{(K)}}}\right)}$$

$$= e^{x_i'\beta^{(2)}}$$

$$= e^{\beta_0^{(2)}+\beta_1^{(2)}x_{i1}+\dots+\beta_p^{(2)}x_{ip}}$$

$x_{i1}^* = x_{i1}+1$ means this ratio of probabilities changes to

$$e^{\beta_0^{(2)}+\beta_1^{(2)}(x_{i1}+1)+\dots+\beta_p^{(2)}x_{ip}}$$

$$= e^{\beta_0^{(2)}+\beta_1^{(2)}x_{i1}+\dots+\beta_p^{(2)}x_{ip}}\, e^{\beta_1^{(2)}}.$$

The probability of class 2 <u>relative to</u> the probability of class 1 <u>changes</u> by being multiplied by $e^{\beta_1^{(2)}}$

AIC

↑ Akaike Information Criterion

$2k - 2\log(\text{Likelihood at maximum})$

- A good model will have a high likelihood (high probability of training data)
- A large value of k (many parameters) means we're at risk of overfitting training data.
- Best model has small AIC:
  - high prob. of training data
  - not many parameters

For logistic regression with p features:

$$2(p+1) - 2\left[\cancel{\sum_{i=1}^{n} \log\{} \sum_{i:y_i=0} \log\{f_0(x_i)\} + \sum_{i:y_i=1} \log\{f_1(x_i)\}\right]$$

BIC is like AIC but penalty $\log(n)k$ instead of $2k$.

$$BIC = \log(n)k - 2\log(\text{likelihood})$$

for linear regression with p features:

$$AIC = 2(\overset{\leftarrow \beta \text{ and } \sigma^2}{p+2}) - 2\sum_{i=1}^{n} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-1}{2\sigma^2}(y_i - x_i'\beta)^2\right\}\right]$$

$$= 2(p+2) - 2\sum_{i=1}^{n}\left\{-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - x_i'\beta)^2\right\}$$

$$= 2(p+2) + \sum_{i=1}^{n}\log(2\pi\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - x_i'\beta)^2$$

$$= 2(p+2) + n\log\left(2\pi\cdot\frac{RSS}{n}\right) + \frac{1}{\left(\frac{RSS}{n}\right)}\cdot RSS$$

$$= 2p + 4 + n\log(2\pi) + n\log(RSS) - n\log(n) + n$$

$$= \underbrace{2p}_{\text{choose a model with}} + \underbrace{n\cdot\log(RSS)}_{\to \text{ choose a model with low RSS}} + C$$

choose a model with
- low polynomial degree
- few explanatory variables

→ choose a model with low RSS