Motivating Examples

- South African Heart Disease Data

- Challenger Disaster Data

- Data: $(y_i, \mathbf{x}_i)$ where $y_i \in \{0, 1\}$, or $(y_i, m_i, \mathbf{x}_i)$ where y_i denotes the number of 1's among m_i cases whose x-value $= \mathbf{x}_i$. Here we merge the intercept into $\mathbf{x}$.

- The linear model, $y_i \sim \mathsf{N}(\mathbf{x}_i^t \boldsymbol{\beta}, \sigma^2)$, is not appropriate. Instead we should model $y_i \sim \mathsf{Bin}\big(m_i, p(\mathbf{x}_i)\big)$.

The Binomial Distribution

- Bernoulli distribution: $Z = 1$ (success) or 0

$$\mathbb{P}(Z = 1) = p, \quad \mathbb{P}(Y = 0) = 1 - p.$$

- $Y =$ number of successes in m iid Bernoulli trials

$$Y \sim \mathsf{Bin}(m, p)$$

$$\begin{aligned} \mathbb{P}(Y = j) &= \binom{m}{j} p^j (1-p)^{m-j} \\ &= \frac{m!}{j!(m-j)!} p^j (1-p)^{m-j}, \quad j = 0, 1, \ldots, m. \end{aligned}$$

$$\mathbb{E}(Y) = mp, \quad \mathsf{Var}(Y) = mp(1-p).$$

Logistic Regression Model

Recall that for linear models, we assume the conditional mean of the response variable Y is a linear function of the covariates $\mathbf{x}$,

$$\mathbb{E}(Y \mid \mathbf{x}) = \mathbf{x}^t \boldsymbol{\beta}.$$

When Y is binary, 0 or 1, the conditional mean is

$$\mathbb{E}(Y \mid \mathbf{x}) = \mathbb{P}(Y = 1 | \mathbf{x}) = p(\mathbf{x}).$$

Since $p(\mathbf{x})$ is constrained to be between 0 and 1, it is not realistic to assume $p(\mathbf{x})$ takes a linear form. Instead we assume its transformation (or referred to as a link function) is a linear function,

$$g(p(\mathbf{x})) = \mathbf{x}^t \boldsymbol{\beta}.$$

Define the logit function (i.e., the odds)

$$\text{log(odds)=logit}(p) = \log \frac{p}{1-p}.$$

Write

$$p_i = p(\mathbf{x}_i) = \mathbb{P}\big(Y_i = 1 | X = \mathbf{x}_i\big).$$

With the logistic model, we assume the odds at a given $\mathbf{x}_i$ is a linear function of $\mathbf{x}_i$:

$$\text{logit}(p_i) = \mathbf{x}_i^t \boldsymbol{\beta}, \qquad \text{i.e.,} \qquad p_i = \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}}.$$

Parameter Estimation: MLE

- Likelihood:

$$f(y_1, \ldots, y_n; \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \text{ or}$$

$$f(y_1, \ldots, y_n; \boldsymbol{\beta}) \propto \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{m_i-y_i}.$$

- Log-likelihood:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[y_i \log \frac{e^{\mathbf{x}_i^t\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i^t\boldsymbol{\beta}}} + (1-y_i)\log\frac{1}{1+e^{\mathbf{x}_i^t\boldsymbol{\beta}}}\right] \\
&= \sum_{i=1}^{n}\left[y_i\mathbf{x}_i^t\boldsymbol{\beta} - \log(1+e^{\mathbf{x}_i^t\boldsymbol{\beta}})\right]
\end{aligned}$$

The NewtonRaphson method: to solve $\ell'(\boldsymbol{\beta}) = 0$, we start with some initial value $\boldsymbol{\beta}^0$, and then repeatedly update

$$\boldsymbol{\beta} \Leftarrow \boldsymbol{\beta}^0 - \ell''(\boldsymbol{\beta}^0)^{-1}\ell'(\boldsymbol{\beta}^0),$$

where ℓ' is a vector and ℓ'' is a matrix.

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_i \left[y_i \mathbf{x}_i^t \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}}) \right] \\
\ell'(\boldsymbol{\beta}^0) &= \sum_i y_i \mathbf{x}_i - \frac{e^{\mathbf{x}_i^t \boldsymbol{\beta}^0}}{1 + e^{\mathbf{x}_i^t \boldsymbol{\beta}^0}} \mathbf{x}_i \\
&= \sum_i \mathbf{x}_i (y_i - p_i^0) \\
\ell''(\boldsymbol{\beta}) &= \sum_i p_i^0 (1 - p_i^0) \mathbf{x}_i \mathbf{x}_i^t
\end{aligned}$$

The MLE $\hat{\boldsymbol{\beta}}$ can be obtained by the following Reweighted LS Algorithm:

- Start with some initial values $\boldsymbol{\beta}^0$
- Calculate the corresponding p_i^0 (based on $\boldsymbol{\beta}^0$) for $i = 1, \ldots, n$; define $W = \text{diag}(p_i^0(1 - p_i^0))_{i=1}^n$.
- Calculate

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta}^0 + W^{-1}(\mathbf{y} - \mathbf{p}^0).$$

- Update $\boldsymbol{\beta}^0$ with

$$\boldsymbol{\beta} = (\mathbf{X}^t W \mathbf{X})^{-1} \mathbf{X}^t W \mathbf{z}.$$

And iterative the above steps until convergence.

- In R, use the `glm` command.

- For each $\hat{\beta}_j$, we have the Z-score

$$Z = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim \text{N}(0, 1), \quad \text{approximately},$$

where `se` is calculated based on the iteratively reweighted least squares approximation. Hypothesis testing (e.g., the p-value) and CI for β_j can be obtained based on the Z-score.

- How to interpret $\hat{\beta}_j$?

- Model Selection: AIC or BIC (stepwise, backward or forward).

Deviance

- We have data $(y_i, \mathbf{x}_i, m_i)$, where

$$y_i \sim \text{Bin}(m_i, p_i), \quad p_i = p(\mathbf{x}_i),$$

  and logit $p(\mathbf{x}_i) = \mathbf{x}_i^t \boldsymbol{\beta}$.

- In logistic regression, we do not measure the residual as the difference between $y_i - m_i \hat{p}_i$, as what we did in linear regression. Instead we have the so-called deviance residuals or Pearson or χ^2 residuals.

The corresponding RSS (residual-sum-of-squares) is equal to

– deviance:

$$-2 \log \text{likelihood} = -2 \sum_i \log f(y_i; \hat{\boldsymbol{\beta}}),$$

– or Pearson's χ^2 statistic:

$$\sum_i \frac{(O_i - E_i)^2}{E_i} = \sum_i \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2$$

where $O_i = y_i$ and $E_i = m_i \hat{p}_i$. In both cases, the RSS (approximately) follows a χ^2 distribution with df $=$ (n - num-of-parameters).

Model Comparison

When comparing two nested models, we can use any of the following methods:

- Their RSS difference $\sim \chi^2$ distribution with df equal to the dim difference between the two models;
- Pick the model with smallest AIC/BIC;
- If the two models just differ by one predictor, we can just look at the p-value from the normal test.

[a]

[a]The F-test is used when there is a scale parameter, such as in the ordinary linear regression, or the quasi-Poisson or quasi-logistic regression that has a dispersion parameter.