STA 303 / 1002

Note Title 02/02/2012

Generalized Linear Models

Response variable: Y

Explanatory variables: $X_1, \ldots, X_p$

Link function $g(\cdot)$

Model: $$g(E(Y)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Link functions

① Identity $\Rightarrow$ STA 302 Regression model (analysis of variance models)

Distribution: $Y \mid X \sim$ Normal

② Log link $g(E(Y)) = \log(E(Y))$

Model: $\log(E(Y)) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

"Log-linear model"

$E(Y)$ must be positive

Useful for count data, $Y|X \sim$ Poisson
(Poisson regression is coming after logistic)

③ <u>Logit link</u> $\mu = E(Y)$

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

μ must be between 0 and 1

For $Y|X \sim \text{Bernoulli}(\pi)$

$E(Y|X) = \pi$

Logit link: $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$

Log odds in favour of a success

Model: $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$

This is the LOGISTIC REGRESSION MODEL

Note: Link function is a function of the mean of Y

and <u>not</u> a transformation of the data

<u>Logistic Regression</u>

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

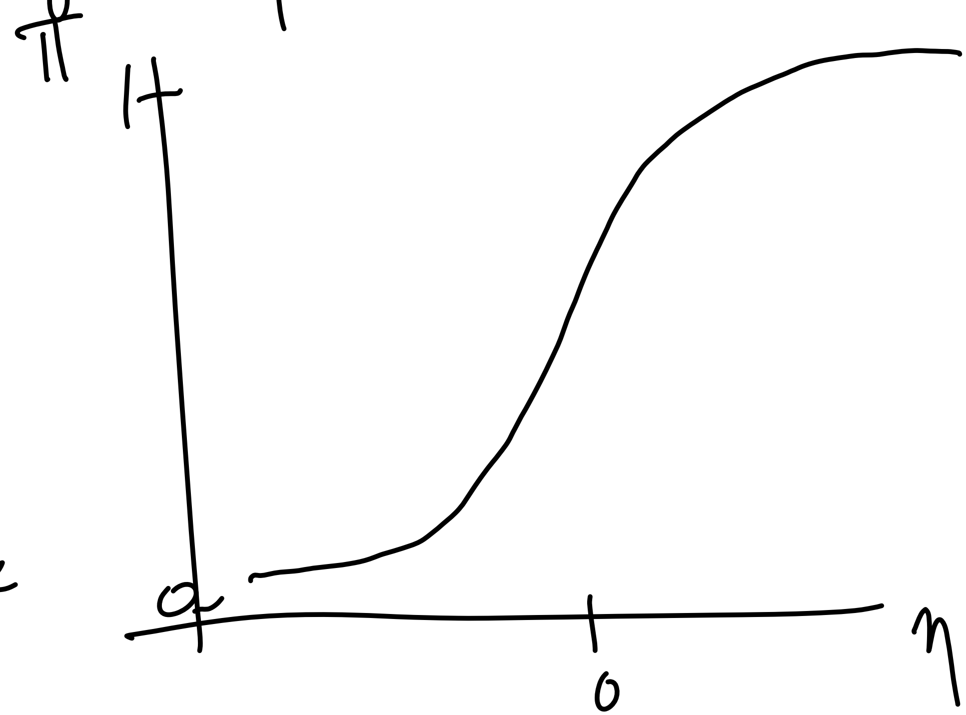
<u>Invert</u>:

$$\pi = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

called the LOGISTIC FUNCTION

What does the logistic function look like?

$$\pi = \frac{e^{\eta}}{1+e^{\eta}}$$



S-shaped

Horizontal asymptote at 0 and 1

Logistic Regression Model

$$E(Y_i \mid X_{i1}, \ldots, X_{ip}) = \pi_i$$

$$Var(Y_i \mid X_{i1}, \ldots, X_{ip}) = \pi_i(1-\pi_i)$$ (variance is not constant)

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}$$

The model does not predict whether y is 0 or 1

Does predict the log-odds that response $= 1$

log-odds are $\in (-\infty, \infty)$ (good characteristic of a 'link' function)

as π_i increases, odds of success increase, and log-odds increase.
(π_i is prob. of success)

Method of estimation of coefficients (β's):
<u>Maximum Likelihood Estimation</u>

Data: $Y_i = \begin{cases} 1 & \text{if } i^{th} \text{ observation is in category of interest} \\ 0 & \text{otherwise} \end{cases}$

$$P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad , i = 1, \dots, n$$

Assume the n observations are independent

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip})}$$

Plug in observed data and think of joint probability as function of β's $\Rightarrow$ Likelihood function

$$L(\beta_0, \ldots, \beta_p) = \prod_{i=1}^{n} \pi_i(\underset{\sim}{\beta})^{y_i} \left(1 - \pi_i(\underset{\sim}{\beta})\right)^{1-y_i}.$$

Maximum Likelihood Estimates: (MLEs) are values of β's that maximize $L(\beta_0, \ldots, \beta_p)$

Log-likelihood <u>function</u>

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \cdots + \beta_p x_{ip})}$$

$$\log L(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} \Big[y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$$

$$- y_i \log(1 + \exp(\beta_0 + \cdots + \beta_p x_{ip}))$$

$$- (1 - y_i) \log(1 + \exp(\beta_0 + \cdots \beta_p x_{ip})) \Big]$$

$$1 - \pi_i = \frac{1}{1 + \exp(\beta_0 + \cdots)} -$$

Find $\hat{\beta}_0, \ldots, \hat{\beta}_p$ by maximizing log-likelihood

Need iterative numerical solution method

such as Newton-Raphson method,

iterativcly re-weighted least squares
(what SAS does)

Resulting $\hat{\beta}_0, \ldots, \hat{\beta}_p$ are MLEs

Large-Sample Properties of MLEs

If correct model, and large enough sample size,
(as $n \to \infty$)

(1) MLEs are unbiased

(2) MLEs have minimum variance

(3) MLEs are Normally distributed.

(4) Formulas for standard errors of MLEs are well-known (estimates of s.e.'s)
(and they are available as by-product of numerical maximization procedure)