Introduction to Logistic Regression (Ch. 14)

1. Generalized Linear Model (GLM)

- The linear (regression) model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \text{ , where } \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

The is equivalent to assuming we have observations (y_i) independently drawn from a Normal distribution with a mean, μ_i, and variance, σ^2:

$$y_i \overset{\text{indep}}{\sim} N(\mu_i, \sigma^2)$$

$$E(y_i) = \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

- ***General linear models*** allow for correlated error terms, and/or non-constant variance on the error. But the response variable and error terms still follow Normal distribution.

- ***Generalized linear models*** extend this model to include certain non-Normally distributed responses. The y_i are assumed to be independently drawn from a probability distribution which is an *exponential family* of distributions (see note at the end of this page). The exponential family of distributions includes the Normal, Bernoulli, Binomial, Poisson, Gamma, and Chi-squared, among others. A function of the mean, denoted as g(μ), is assumed linear with respect to the covariates. This function, g(μ), is called the *link* function.

$$y_i \sim Exponential\ family$$

$$E(y_i) = \mu_i$$

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

- FYI, for a random variable in the Exponential family of distributions, the form of its probability density (or mass) function can be written as $f(y|\theta) = h(y) \exp[\eta(\theta) T(y) - A(\theta)]$ for known functions $h(y)$, $\eta(\theta)$, $T(y)$, and $A(\theta)$ and a parameter of interest, θ. We will focus on the logistic regression model which assumes either a Binomial or Bernoulli distribution for the response.

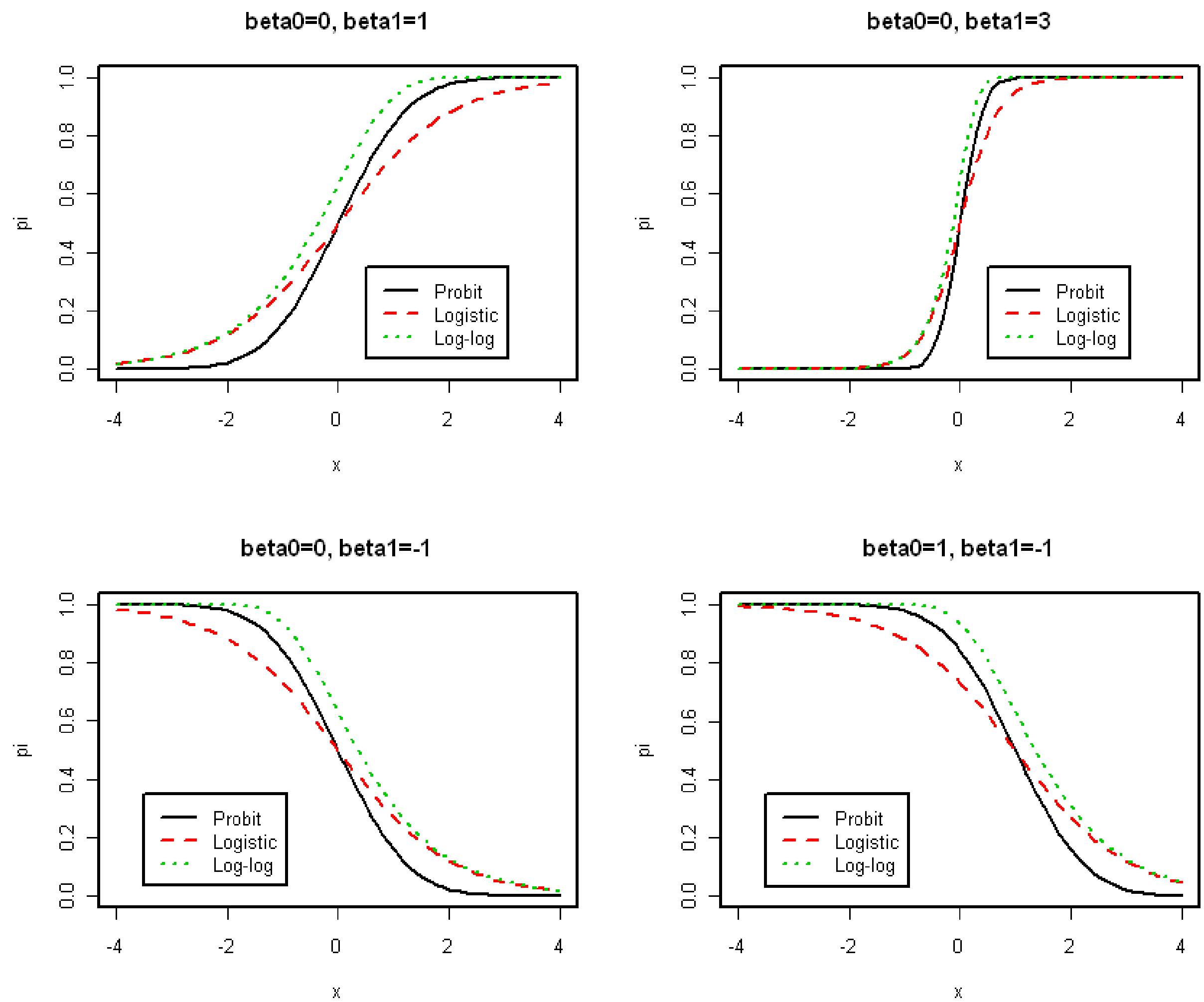
- The choice of the link function depends on the distribution and the data
  - In Normal linear regression models, $g(\mu_i) = \mu_i$ is called the "identity link."
  - If the response variable y_i has a Poisson distribution, use the log( ) link, i.e., $g(\mu_i) = \log(\mu_i)$.
  - Link functions for Binary (0, 1) response and the counts from a Binomial distribution are discussed below.

2. Binary (Bernoulli) and Binomial responses

- Binary (Bernoulli) response:

- Binomial response

- Link functions for Binary and Binomial response


beta0=0, beta1=1
beta0=0, beta1=3
beta0=0, beta1=-1
beta0=1, beta1=-1
pi
x
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Probit
Logistic
Log-log

3. Logistic Regression

- Model

- Estimation

- Inference
  - Parameter interpretation
  - Confidence interval and hypotheses test for one parameter
  - Test for several parameters

- Predictions
  - Predict the probability.
  - Predict the group membership (classification) for Binary response.
  - Predict the expected counts of successes (and failures) for Binomial response.

4. Variable selection and model comparison

- Stepwise selection
- Model comparison criteria

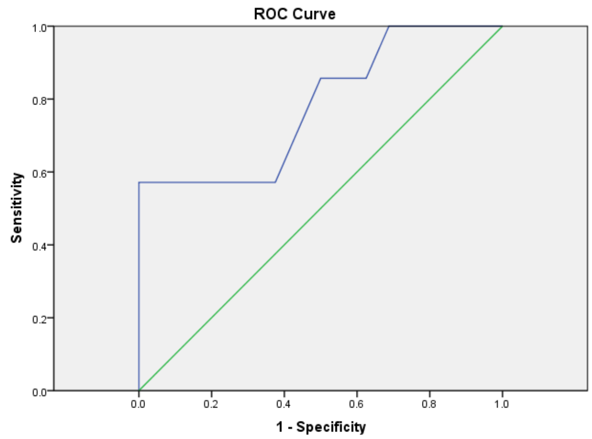
5. Model diagnostics

- Classification table and Receiver Operating Characteristic (ROC) curve (for Binary response)

Classification Table[a]

Observed			Predicted: Failure 0	Predicted: Failure 1	Percentage Correct
Step 1	Failure	0	16	0	100.0
		1	3	4	57.1
	Overall Percentage				87.0

a. The cut value is .500



Diagonal segments are produced by ties.

- Goodness of fit test: Hosmer-Lemeshow test (for Binary response) and Chi-square test (for Binomial counts).

- Residuals and residual plots.

- Leverage, Cook's distance and DFbetas.