Introduction to Logistic Regression (Ch. 14)

(More details in Generalized Linear Model)

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 $g(\mu)$, is assumed linear with respect to the covariates. This function, $g(\mu)$, 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:

  Response: Y_i = 1 or 0

  (success vs failure, win vs lose, head vs tail, true vs false, survive vs not ....)

  P(Y_i = 1) = probability of success = π_i, π_i changes with the predictors (X's).

- Binomial response

  Response: $Y_i \sim$ Binomial(n_i, π_i), "count of success".

  There are n_i independent Bernoulli outcomes from the exact same x-combination (aka. covariate pattern). Y_i is the number of successes.

  π_i changes with X's.

- Link functions for Binary and Binomial response

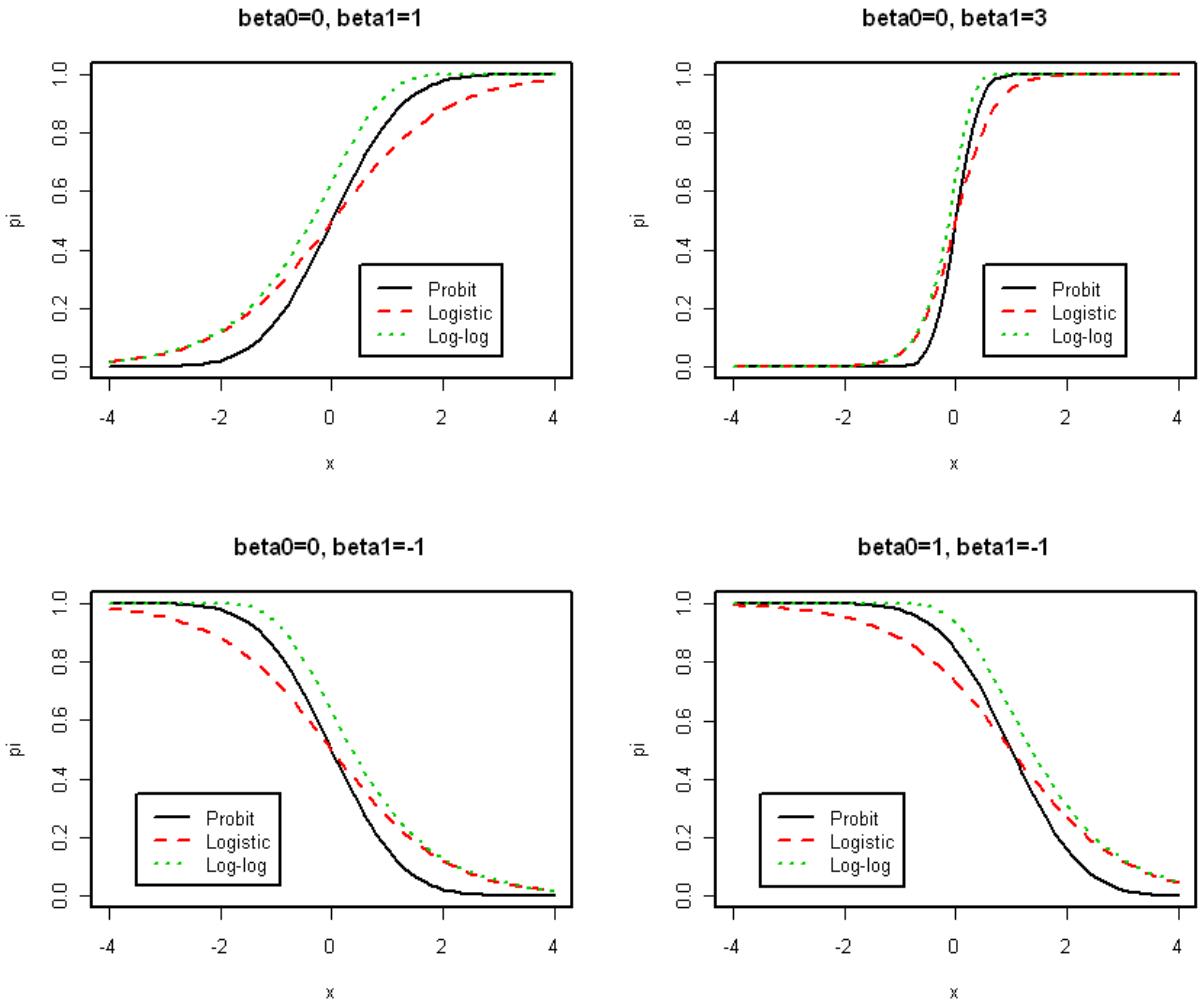
  $0 \le \pi_i \le 1$

  The logit-link (logistic regression):

  $$log\left(\pi_i/(1-\pi_i)\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} \quad (without\ error\ term)$$

  Define Odds = (success probability)/(failure probability) = π_i / (1 - π_i)

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



The above plots illustrate different link functions: (Log-log should be "Complementary log log")

a. All link functions assume a non-linear association between X and π_i.

b. The sign of the slope (beta1) shows the direction of the association (though not linear).

c. Both probit and logit links are symmetrical for small π_i (π_i close to 0) and large π_i (π_i close to 1). But the tails of logistic link are "heavier."

d. The Complementary log log link function has similar behavior to the "Logit" link when π_i is close to 0, but has a thin tail when π_i close to 1. This link is often used to model events with low success probability (rare event).

3. Logistic Regression

➢ Model

$Y_i \sim$ Binomial(n_i, π_i), (It is Bernoulli when $n_i = 1$)

$$log\left(\pi_i/(1-\pi_i)\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

Equivalently, $\pi_i = \frac{\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{p-1}x_{i,p-1})}{1+\exp(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_{p-1}x_{i,p-1})}$

➢ Estimation

- OLS will NOT work.
- Use MLE (Maximum Likelihood Estimation) method
- There is NOT a "closed-form" solution for the MLE estimate. The computation will be done in numeric iterations.

➢ Inference

- Parameter interpretation

  If X_1 increases by 1 unit, after adjusting the effects of other predictors,

  $\boldsymbol{\beta_1}$: log(odds) will increase β_1,

  $\boldsymbol{e^{\beta_1}}$: the odds will change by a factor (or "multiplier") of e^{β_1}.

  The odds will increase by $(e^{\beta_1} - 1) \times 100\%$, or decrease by $(1 - e^{\beta_1}) \times 100\%$.

  The **Odds ratio** is e^{β_1}. (most commonly used)

- Confidence interval and hypotheses test for one parameter

  Confidence Interval: $\widehat{\boldsymbol{\beta_1}} \pm (\boldsymbol{z_{crit}}) \times \boldsymbol{se}(\widehat{\boldsymbol{\beta_1}})$

  Use standard Normal distribution, N(0, 1), for the critical values and for testing 1 parameter. No longer using t-distribution.

- Test for several parameters

  In MLR, we use partial-F test (aka. general linear test) with Full vs Reduced models

  In logistic regression (similar idea, but use Deviance of the model)

  - Full model vs Reduced model (plugging Ho into the Full to get the Reduced)
  - Compute the "Deviance" (in R: Residual Deviance) for both full model and reduced model.
  - Change in Deviance (test statistic): G^2 = Deviance(R) - Deviance(F)

- P-value = P (Chi-square(df) > G^2), where df = the difference in the number of parameters between Full and Reduced models.
- This testing method may be referred as:
  - Likelihood ratio test (name it by the methods)
  - Deviance test (name it by the statistic used in the test)
  - Test for multiple parameters in logistic regression (name it by the purpose)

- Predictions
  - Predict the probability.

    First, get the MLE: $\widehat{\beta_0},\ \widehat{\beta_1},\ \widehat{\beta_2}, \ldots,\ \hat{\beta}_{p-1}$

    Given a set of x-value ($x_1, x_2, \ldots, x_{p-1}$)

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_{p-1} x_{p-1})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_{p-1} x_{p-1})}$$

  - Predict the group membership (classification) for Binary response (y = 0 or 1).

    First, decide a threshold c. Often c = 0.5. But other values can be used as well.

    Given a set of x-value ($x_1, x_2, \ldots, x_{p-1}$)

    - If $\hat{\pi} \le c$, then $\hat{y} = 0$
    - If $\hat{\pi} \ge c$, then $\hat{y} = 1$

  - Predict the expected counts of successes (and failures) for Binomial response.

    Given a set of x-value ($x_1, x_2, \ldots, x_{p-1}$) and number of trials n

$$\hat{y} = n\hat{\pi}$$

4. Variable selection and model comparison

- Stepwise selection

  Similar to MLR.

- Model comparison criteria

  AIC and BIC are commonly used criteria.

  Caution: Some software will compute a "psudo-R^2". It does NOT have the same interpretation as the R^2 in linear regression (SLR or MLR).

5. Model diagnostics

- Classification table (aka. Confusion Matrix) for Binary response.

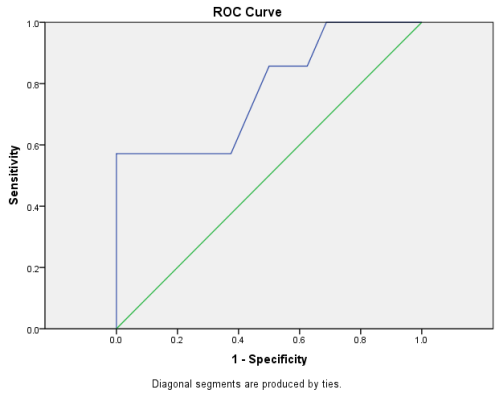
Classification Table[a]

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

a. The cut value is .500

A "bad" classification table (aka. confusion matrix) does not necessarily mean a bad model. If the success probability is close to 0.5, we can expect high mis-classification rate due to how we "predict" the individual outcome.

- Receiver Operating Characteristic (ROC) curve for Binary response.



Good model will have larger area under the ROC curve.

If the area under the curve is close to 0.5, the model is as good (or bad) as guessing at random.

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

For binary response, group (aka. bin) the cases by $\hat{\pi}$, then implement the Pearson's test.

- Residuals and residual plots.

Different residuals. More difficult to read.

- Leverage, Cook's distance and DFbetas.

For outliers and influential cases.