Logistic regression

11 Nov 2010

Modeling overview

- Want to capture important features of the *relationship between* a (set of) *variable(s)* and one or more *response(s)*
- Many models are of the form

$$g(Y) = f(\mathbf{x}) + \text{error}$$

- *Differences* in the form of g, f and distributional assumptions about the error term

Examples of models

- Linear: $Y = \beta_0 + \beta_1 x + \epsilon$
- Linear: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
- (Intrinsically) Nonlinear: $Y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} + \epsilon$
- Generalized Linear Model (e.g. Binomial): $\log \frac{p}{1-p} = \beta_0 + \beta_1 x + \beta_2 x_2$
- Proportional Hazards (in Survival Analysis): $h(t) = h_0(t) \exp(\beta x)$

Linear modeling

- A simple linear model: $E(Y) = \beta_0 + \beta_1 x$
- Gaussian measurement model: $Y = \beta_0 + \beta_1 x + \epsilon, \epsilon \ N(0, \sigma^2)$
- More generally: $Y = X\beta + \epsilon$, where Y is $n \times 1$, X is $n \times p$, β is $p \times 1$, ϵ is $n \times 1$, often assumed $N(0, \sigma^2 I_{n \times n})$

Analysis of designed experiments

- An important use of linear models
- Define a (design) matrix X so that for response variable Y:

$$E(Y) = X\beta,$$

  where β is a vector of *parameters* (or contrasts)
- Many ways to define design matrix/contrasts

Model fitting

- For the standard (*fixed effects*) linear model, estimation is usually by *least squares*
- Can be more complicated with *random effects* or when x-variables are subject to measurement error as well

Model checking

- Examination of *residuals*
  - Normality
  - Time effects
  - Nonconstant variance
  - Curvature
- Detection of *influential observations*

Linear regression model (again)

- Linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Another way to write this:

$$Y \sim N(\mu, \sigma^2), \quad \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- Suitable for a *continuous* response
- **NOT** suitable for a *binary* response

Modified model

- Instead of modeling the response directly, could instead model the *probability* of '1'
- Problems:
  - could lead to fitted values outside of $[0, 1]$
  - normality assumption on errors is wrong
- Instead of modeling the expected response *directly* as a linear function of the predictors, model a *suitable transformation*
- For binary data, this is generally taken to be the *logit* (or *logistic*) transformation

Logit transformation

- $\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$
- Therefore,

$$p(x_1, \ldots x_k) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}$$

- The parameter β_k is such that $\exp(\beta_k)$ is the *odds* that the response takes value 1 when x_k increases by one, when the remaining variables are constant
- Estimate parameters by *maximum likelihood*

Binary response in a linear model

- In a standard linear model, the *response variable* is modeled as a *normally distributed*
- However, if the response variable is *binary*, it does not make sense to model the outcome as normal
- Generalized linear models (GLMs) are an extension of linear models to model non-normal response variables
- We are using *logistic regression* for a binary response

Generalized linear models: some theory

- Allows unified treatment of statistical methods for several important classes of models
- Response Y assumed to have *exponential family distribution*:

$$f(y) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

- For a standard linear model

  $$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon, \text{ with } \epsilon \sim N(0, \sigma^2)$$

- The *expected response* is $E[Y \mid x] = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- Let η denote the *linear predictor* $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- For a standard linear model, $E[Y \mid x] = \eta$
- In a *generalized linear model*, there is a *link function* g between η and the expected response:

  $$g(E[Y \mid x]) = \eta$$

- For a standard linear model, $g(y) = y$ (*identity link*)

Link function

- When the response variable is binary (with values coded as 0 or 1), then $E[Y \mid x] = P(Y = 1 \mid x)$
- A convenient function in this case is
  $$E[Y \mid x] = P(Y = 1 \mid x) = \frac{e^{\eta}}{1+e^{\eta}}$$
- The corresponding link function (inverse of this function) is called the *logit*
- $\text{logit}(x) = \log(x/(1-x))$
- Regression using this model is called *logistic regression*

Link function: examples

	Family Name				
Link	binomial	Gamma	gaussian	inverse.gaussian	poisson
logit	D				
probit	•				
cloglog	•				
identity		•	D		•
inverse		D			
log		•			D
1/mu^2				D	
sqrt					•

Analogous to linear regression

- The logit function g has many of the desirable properties of a linear regression model
- Mathematically convenient and flexible
- Can meaningfully interpret parameters
- Linear in the parameters
- A difference: Error distribution is binomial (not normal)

Fitting the model

- For linear regression, typically use *least squares*
- When outcome dichotomous, the ‘nice’ statistical properties of least squares estimators no longer hold
- The general estimation method that leads to least squares (for normally distributed errors) is *maximum likelihood*

Maximum likelihood estimation

- Likelihood: $f(x_i) = p(x_i)^{y_i}[1 - p(x_i)]^{1-y_i}$
- Assuming independent observations, the likelihood $l(\beta) = \prod_{i=1}^{n} f(x_i)$
- log likelihood
  $L(\beta) = \log[l(\beta)] = \sum_{i=1}^{n}(y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i)))$
- To find β that maximize the log likelihood, differentiate wrt each β_i and set the derivative equal to 0
- In linear regression these equations are easily solved
- In logistic regression, these are nonlinear in β and are solved iteratively

Assessing model fit

- In linear regression, an anova table partitions SST, the total sum of squared deviations of observations about their mean, into two parts:
  - SSE, or residual (observed - predicted) sum of squares
  - SSR, or regression sum of squares
- Large SSR suggests the explanatory variable(s) is(are) important
- Use same guiding principle in logistic regression: compare observed response to predicted response obtained from models with/without the variable(s)
- Comparison based on log likelihood function

Deviance

- In standard linear models, estimate parameters by minimizing residual sum of squares
- (Equivalent to ML for normal model)
- In GLM, estimate parameters by ML
- The *deviance* is (proportional to) $2 \times l$
- (Analogous to SSE)
- Obtaining 'absolute' measure of goodness of fit depends on some assumptions that may not be satisfied in practice
- Usually focus on comparing competing models
- When the models are *nested*, can carry out likelihood ratio test

Comparing models

- In linear regression, consider coefficient significant if (squared) standardized value $\hat{\beta}/SE(\hat{\beta})$ is 'large'
- Can also do this for logistic regression (Wald test), but there are some problems with it
- Preferred approach: likelihood ratio test
- Deviance $D = -2\sum_{i=1}^{n} y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1-y_i)\log\left(\frac{1-\hat{p}_i}{1-y_i}\right)$
- To compare models, compute $G = D(\text{submodel}) - D(\text{bigger model})$
- Under the null (*i.e.* the submodel), $G \sim \chi^2$ with df $=$ difference in the number of estimated parameters