Statistics 191: Introduction to Applied Statistics

Logistic regression

Jonathan Taylor
Department of Statistics
Stanford University

March 3, 2010

Topics

Today's class

- Binary outcomes.
- Logistic regression.
- Generalized linear models.
- Deviance.

Logistic regression

Binary outcomes

- Most models so far have had response Y as continuous.
- Many responses in practice fall into the YES/NO framework.
- Examples:
  1. medical: presence or absence of cancer
  2. financial: bankrupt or solvent
  3. industrial: passes a quality control test or not

Logistic regression

Jonathan Taylor
Department of Statistics
Stanford University

Modelling probabilities

- For $0-1$ responses we need to model

$$\pi(x_1, \dots, x_p) = P(Y = 1 | X_1 = x_1, \dots, X_p = x_p)$$

- That is, Y is Bernoulli with a probability that depends on covariates $\{X_1, \dots, X_p\}$.
- **Note:**

$$\mathrm{Var}(Y) = \pi(1-\pi) = E(Y) \cdot (1 - E(Y))$$

- **Or,** the binary nature forces a relation between mean and variance of Y.

Logistic regression

Jonathan Taylor
Department of Statistics
Stanford University

Flu shot example

- A local health clinic sent fliers to its clients to encourage everyone, but especially older persons at high risk of complications, to get a flu shot in time for protection against an expected flu epidemic.
- In a pilot follow-up study, 50 clients were randomly selected and asked whether they actually received a flu shot. $Y = \texttt{Shot}$
- In addition, data were collected on their age $X_1 = \texttt{Age}$ and their health awareness $X_2 = \texttt{Health.Aware}$

Logistic regression

Jonathan Taylor
Department of Statistics
Stanford University

Model for probabilities

- Simplest model

$$\pi(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- Problems:
  - We must have $0 \leq E(Y) = \pi(X_1, X_2) \leq 1$
  - Ordinary least squares will not work because of relation between mean and variance.

Logistic regression

Logistic model

- Logistic model

$$\pi(X_1, X_2) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}$$

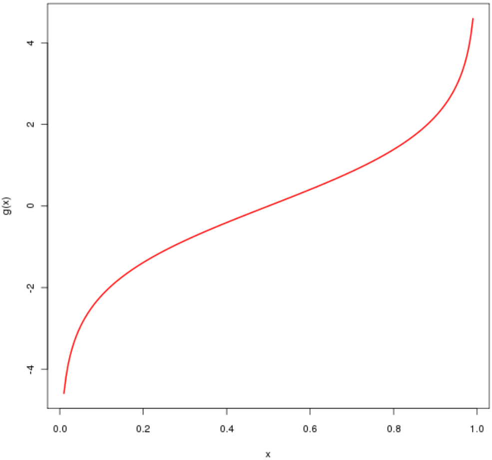
- This automatically fixes $0 \leq E(Y) = \pi(X_1, X_2) \leq 1$.
- **Note:**

$$\text{logit}(\pi(X_1, X_2)) = \log\left(\frac{\pi(X_1, X_2)}{1 - \pi(X_1, X_2)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Logistic curve

Statistics 191: Introduction to Applied Statistics

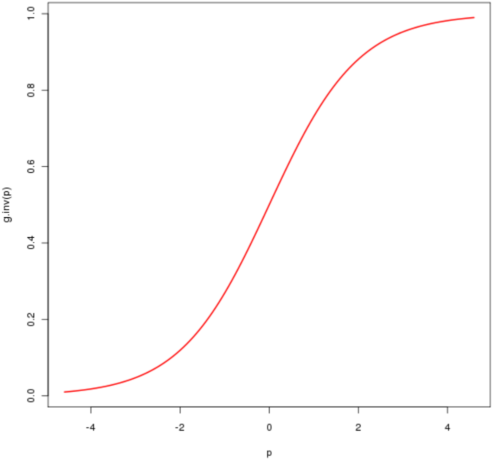
Jonathan Taylor
Department of Statistics
Stanford University



Logistic transform

Statistics 191:
Introduction
to Applied
Statistics

Jonathan
Taylor
Department of
Statistics
Stanford
University



Logistic regression

Jonathan Taylor
Department of Statistics
Stanford University

Binary regression models

- Models $E(Y)$ as some increasing function of $\beta_0 + \beta_1 X_1 + \beta_2 X_2$.
- The logistic model uses the function $f(x) = e^x/(1+e^x)$.
- Can be fit using Maximum Likelihood / Iteratively Reweighted Least Squares.
- Coefficients have nice interpretation in terms of **odds ratios**
- Inference (?)

Odds Ratios

- One reason logistic models are popular is that the parameters have simple interpretations in terms of **odds**

$$ODDS(A) = \frac{P(A)}{1 - P(A)}.$$

- Logistic model:

$$OR_{X_j} = \frac{ODDS(\dots, X_j = x_j + 1, \dots)}{ODDS(\dots, X_j = x_j, \dots)} = e^{\beta_j}$$

- If $X_j \in 0, 1$ is dichotomous, then odds for group with $X_j = 1$ are e^{β_j} higher, other parameters being equal.

Logistic regression

Rare disease hypothesis

- When incidence is rare, $P(Y=0) \approxeq 1$ no matter what the covariates X_j's are.
- In this case, odds ratios are almost ratios of probabilities:

$$OR_{X_j} \approxeq \frac{\mathbb{P}(Y=1|\dots,X_j=x_j+1,\dots)}{\mathbb{P}(Y=1|\dots,X_j=x_j,\dots)}$$

- Hypothetical example: in a lung cancer study, an OR of 5 means for smoking vs. non-smoking means smokers are $e^5 \approx 150$ times more likely to develop lung cancer
- In flu example, the odds for a 45 year old with health awareness 50 compared to a 35 year old with the same health awareness are $e^{2.2178} = 9.18$, but ratio of probs is $0.1932/0.0254 \approx 7.61$.

Logistic regression

Fitting the model ($g(\pi) = \text{logit}(\pi)$)

1. Initialize $\widehat{\pi}_i = \bar{Y}, 1 \leq i \leq n$
2. Define $Z_i = g(\widehat{\pi}_i) + g'(\widehat{\pi}_i)(Y_i - \widehat{\pi}_i)$
3. Fit weighted least squares model

$$Z_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}, \qquad w_i = \widehat{\pi}_i(1 - \widehat{\pi}_i)$$

4. Set $\widehat{\pi}_i = \text{logit}^{-1}\left(\widehat{\beta}_0 + \sum_{j=1}^{p} \widehat{\beta}_j X_{ij}\right)$.
5. Repeat steps 2-4 until convergence.

This is *basically* Newton-Raphson.

Logistic regression

Statistics 191: Introduction to Applied Statistics

Jonathan Taylor
Department of Statistics
Stanford University

Inference

- The IRLS procedure suggests using approximation

$$\widehat{\beta} \approx N(\beta, (X'WX)^{-1})$$

- This allows us to construct CIs, test linear hypotheses, etc.
- What about comparing $\mathcal{M}_F$ and $\mathcal{M}_R$?

Logistic regression

Jonathan Taylor
Department of Statistics
Stanford University

Deviance

- For a model $\mathcal{M}$, $DEV(\mathcal{M})$ replaces $SSE(\mathcal{M})$.
- Model is fit to minimize

$$DEV(\mathcal{M}) = -2 \log L(\mathcal{M}|Y, X) + \dots$$

- In least squares regression, we use

$$SSE(\mathcal{M}_R) - SSE(\mathcal{M}_F) \sim \chi^2_{df_R - df_F}$$

- This is replaced with

$$DEV(\mathcal{M}_R) - DEV(\mathcal{M}_F) \overset{n \to \infty}{\sim} \chi^2_{df_R - df_F}$$

- Resulting tests **do not** agree with those coming from IRLS (Wald tests). Both are often used.

Binary regression

Deviance

- $$DEV(\mu|Y) = -2\log L(\mu|Y) + -2\log L(Y|Y)$$

  where μ is a location estimator for Y
- If Y is Gaussian with independent $N(\mu_i, \sigma^2)$ entries

$$DEV(\mu|Y) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu_i)^2$$

- If Y is binary, with mean π

$$DEV(\pi|Y) = -2\sum_{i=1}^{n} Y_i\log(\pi_i) + (1 - Y_i)\log(1 - \pi_i)$$

Binary regression

Jonathan Taylor
Department of Statistics
Stanford University

Deviance

- In the logistic model,

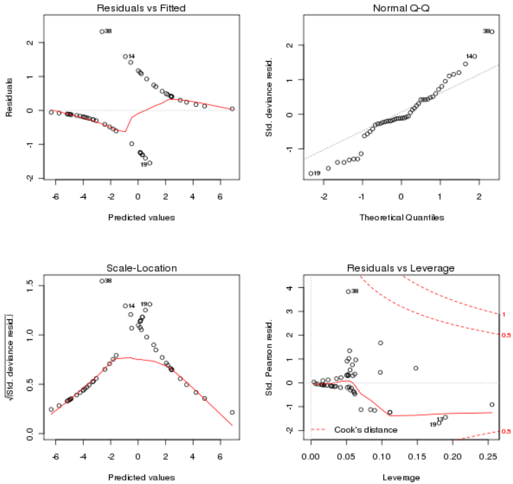
$$
\begin{aligned}
DEV(\beta|Y) &= -2\sum_{i=1}^{n} Y_i \text{logit}(\pi_i(\beta)) + \log(1-\pi_i(\beta)) \\
&= -2\sum_{i=1}^{n} Y_i \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}\right) + \log(1-\pi_i(\beta))
\end{aligned}
$$

Logistic regression

Other points

- Diagnostics: similar to least square regression, only residuals used are *deviance residuals* . . .
- Model selection: because it is fit based on likelihood, stepwise selection can be used easily . . .

Diagnostics for logistic model



Binary regression

Probit transform

- Probit regression model:

$$\Phi^{-1}(\mathbb{E}(Y_i)) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}$$

  where Φ is CDF of $N(0,1)$, i.e. $\Phi(t) = \texttt{pnorm(t)}$.

- Complementary log-log model (cloglog):

$$-log(-log(\mathbb{E}(Y_i)) = \beta_0 + \sum_{j=1}^{p-1} \beta_j X_{ij}.$$

- In logit, probit and cloglog $\mathrm{Var}(Y_i) = \pi_i(1-\pi_i)$ but the model for the mean is different.

Generalized linear models

Link & variance functions

Given a dataset $(Y_i, X_{i1}, \dots, X_{ip}), 1 \leq i \leq n$ we consider a model for the distribution of $Y|X_1, \dots, X_p$.

- If
$$\eta_i = g(\mathbb{E}(Y_i)) = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$
then g is called the *link* function for the model.
- If
$$\mathrm{Var}(Y_i) = \phi \cdot V(\mathbb{E}(Y_i)) = \phi \cdot V(\mu_i)$$
for $\phi > 0$ and some function V, then V is the called *variance* function for the model.
- Canonical reference: *Generalized Linear Models*, McCullagh and Nelder.

Binary regression as GLM

Binary (again)

- For a logistic model,

$$g(\mu) = \mathrm{logit}(\mu), \qquad V(\mu) = \mu(1-\mu).$$

- For a probit model,

$$g(\mu) = \Phi^{-1}(\mu), \qquad V(\mu) = \mu(1-\mu).$$

- For a cloglog model,

$$g(\mu) = -\log(-\log(\mu)), \qquad V(\mu) = \mu(1-\mu).$$