1

Logistic Regression

H. EL GHAZI

2

Outline

- Simple Example Of Logistic Regression
- Maximum Likelihood Estimation
- Interpreting Logistic Regression Output
- Inference: Are The Predictors Significant?
- Odds Ratio And Relative Risk

3

Simple Example Of Logistic Regression

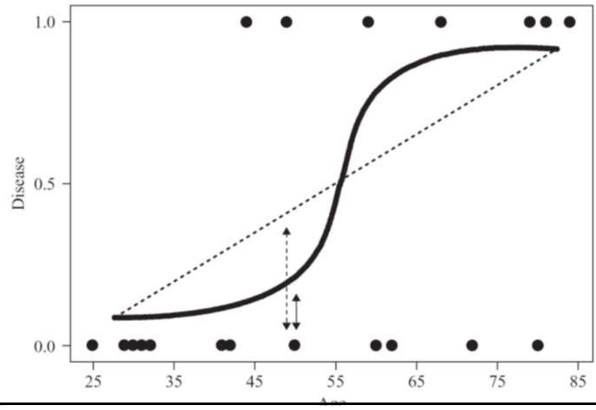
- Age of 20 Patients, with Indicator of Disease

Patient ID	Age, x	Disease, y	Patient, ID	Age, x	Disease, y
1	25	0	**11**	50	0
2	29	0	**12**	59	1
3	30	0	**13**	60	0
4	31	0	**14**	62	0
5	32	0	**15**	68	1
6	41	0	**16**	72	0
7	41	0	**17**	79	1
8	42	0	**18**	80	0
9	44	1	**19**	81	1
10	49	1	**20**	84	1

4

Logistic Regression Model

- Plot of *disease* versus age, with least-squares and logistic regression lines.



S-shaped form is called Sigmoid function and take the form:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

5

Logistic Regression Model

- Linear regression models assume that $Y = \beta 0 + \beta 1x + \varepsilon$, where the error term ε is normally distributed with mean zero and constant variance.
- The response variable in logistic regression $Y = \pi(x) + \varepsilon$ is assumed to follow a binomial distribution with probability of success $\pi(x)$.

$$\pi(x) = P(Y = 1|x) \qquad \pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

- Since the response in logistic regression is dichotomous, the error ε can take two form:
  - If $Y = 1$ then $\varepsilon = 1 - \pi(x)$,
  - if $Y = 0$ then $\varepsilon = 0 - \pi(x) = -\pi(x)$,
- The variance of ε is $\pi(x)\ [1 - \pi(x)]$,

6

Logistic Regression Model

- A useful transformation for logistic regression is the *logit transformation*, as follows:

$$g(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x$$

Thus,

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

7

Maximum Likelihood Estimation

- The *likelihood function* $l(\boldsymbol{\beta} \mid x)$ is a function of the parameters $\boldsymbol{\beta}=\beta 0,\beta 1,\ldots,\beta m$ which expresses the probability of the observed data, x.
- By finding the values of $\boldsymbol{\beta} = \beta 0, \beta 1, \ldots, \beta m$ that maximize $l(\boldsymbol{\beta} \mid x)$, we thereby uncover the *maximum likelihood estimators*, the parameter values most favored by the observed data.
  - Since $Yi = 0$ or 1, the contribution to the likelihood of the *i*th observation may be expressed as
    $$[\pi(x_i)]^{y_i}\,[1-\pi(x_i)]^{1-y_i}$$
  - Thus,
    $$l(\boldsymbol{\beta}|x) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i}\,[1-\pi(x_i)]^{1-y_i}$$
  - The log likelihood
    $$L(\boldsymbol{\beta}|x) = \ln\,[l(\boldsymbol{\beta}|x)] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1-y_i)\ln[1-\pi(x_i)]\}$$

8

Maximum Likelihood Estimation

- The maximum-likelihood estimators may be found by differentiating $L(\boldsymbol{\beta} \mid x)$ with respect to each parameter, and setting the resulting forms equal to zero.
- Unfortunately, unlike linear regression, closed-form solutions for these differentiations are not available.
- Therefore, other methods must be applied, such as iterative weighted least squares, Gradient Descent,..

9

Interpreting Logistic Regression Output

- Results of Logistic Regression of *Disease* on *Age*

Logistic Regression Table

Predictor	Coef	StDev	Z	P	Odds Ratio	95% CI Lower	Upper
Constant	-4.372	1.966	-2.22	0.026			
Age	0.06696	0.03223	2.08	0.038	1.07	1.00	1.14

Log-Likelihood = -10.101
Test that all slopes are zero: G = 5.696, DF = 1, P-Value = 0.017

$$\hat{g}(x) = -4.372 + 0.06696(age)$$

- For example, for a 50-year-old patient, we have $\hat{g}(x) = -4.372 + 0.06696(50) = -1.024$

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} = \frac{e^{-1.024}}{1 + e^{-1.024}} = 0.26$$

Thus, the estimated probability that a 50-year-old patient has the disease is 26%

10

Inference: Are The Predictors Significant?

Wald test

- Another hypothesis test used to determine whether a particular predictor is significant is the *Wald test*
- Under the null hypothesis that β1 = 0, the $Z_{\text{Wald}} = \dfrac{b_1}{\text{SE}(b_1)}$ follows a standard normal distribution, where SE refers to the standard error of the coefficient

Logistic Regression Table

Predictor	Coef	StDev
Constant	-4.372	1.966
Age	0.06696	0.03223

$$Z_{\text{Wald}} = \frac{0.06696}{0.03223} = 2.08$$

The *p*-value is then reported as *P*(|z| > 2.08) = 0.038

11

Inference: CI for Coefficient

CI for Coefficient

- We may construct 100(1 − *a*)% confidence intervals for the logistic regression coefficients as follows:

$$\boxed{\begin{array}{l} b_0 \pm z \cdot \mathrm{SE}(b_0) \\ b_1 \pm z \cdot \mathrm{SE}(b_1) \end{array}}$$

- In our example, a 95% confidence interval for the slope $\beta 1$ could be found thus:

$$\begin{aligned} b_1 \pm z \cdot \mathrm{SE}(b_1) &= 0.06696 \pm (1.96)(0.03223) \\ &= 0.06696 \pm 0.06317 \\ &= (0.00379,\ \ 0.13013) \end{aligned}$$

12

END