STAT 420: Methods of Applied Statistics

Logistic Regression

Ruoqing Zhu, Ph.D. <rqzhu@illinois.edu>

Course website: https://sites.google.com/site/teazrq/teaching/STAT420

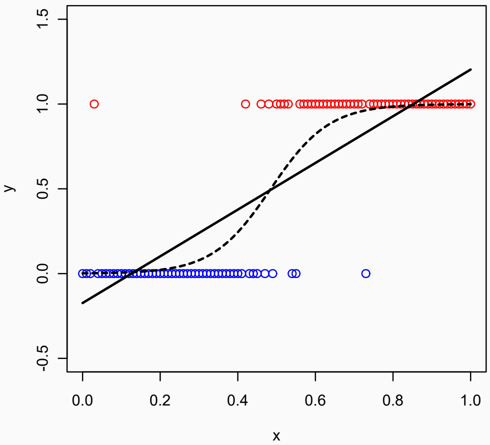
Department of Statistics
University of Illinois at Urbana-Champaign
April 18, 2017

Binary Outcome Variables

- In many applications, the outcome variable is not continuous, e.g.:
  - Whether the patient develops sepsis in hospital
  - Whether a student Receives “A” from STAT 420
- Usually the outcomes are coded as 0 and 1, however, linear regression (treating them as continuous) is not appropriate for this problem.
- Hence we will introduce the logistic regression to deal with this classification problem.

Use linear regression to fit classification problems?

If we treat the binary outcomes 0-1 as continuous and fit linear regression:



Binary Outcome Variables

- What would happen if we use linear regression to fit classification problems?
- In the linear regression, we are modeling the expectation of Y, $E(Y) = X^T\boldsymbol{\beta}$.
- When Y is a binary outcome with 0 or 1, its expectation is just the probably of $\{Y = 1\}$, which will be within $[0, 1]$ regardless of the underlying true model.
- However, this becomes problematic if the fitted value exceeds 1 or falls below 0 (in the previous plot). There is no way to interpret.

Motivation

- Instead of using linear regression, we need to find an appropriate way to describe the relationship between $E(Y)$ (the probability of being 1) and X such that we produce some predictions within $[0, 1]$. A nature target of modeling is

$$\eta(x) = \mathsf{P}(Y = 1 | X = x),$$

  the conditional probably of being 1.
- Interpretation: if we have $\eta(x) > 0.5$, then Y is more likely to be 1.
- Generalized linear model (GLM): use some specific form of $\eta(\cdot)$ such that it is a function of $\boldsymbol{\beta}$ and x, and we solve for $\boldsymbol{\beta}$.
- What specific form to use for binary outcomes?

The logistic link function

- To properly model the probabilities, we need to choose a $\eta(x)$ that is bounded within $[0, 1]$.
- A natural choice is the logistic regression that models:

$$\eta(x) = \text{logit}^{-1}(x^T\boldsymbol{\beta}) = \frac{\exp(x^T\boldsymbol{\beta})}{1 + \exp(x^T\boldsymbol{\beta})},$$

  where the "logit" function is defined as

$$\text{logit}(\eta(x)) = \log\left(\frac{\eta(x)}{1 - \eta(x)}\right) = x^T\boldsymbol{\beta}, \text{ with } \eta(x) \in [0, 1].$$

- The logit function is a way to transform a probably $\eta(x)$ into $(-\infty, +\infty)$, which is the range of $x^T\boldsymbol{\beta}$.

Fitting Logistic Models

- Now we have the logistic regression model by assuming that

$$\mathsf{E}(Y|X=x) = \mathsf{P}(Y=1|X=x) = \eta(x) = \frac{\exp(x^T\boldsymbol{\beta})}{1+\exp(x^T\boldsymbol{\beta})},$$

- As $x^T\boldsymbol{\beta}$ becomes larger ($\to +\infty$), $\eta(x) \to 1$
- As $x^T\boldsymbol{\beta}$ becomes smaller ($\to -\infty$), $\eta(x) \to 0$
- There is no "ϵ" term in the logistic regression. However, that randomness is absorbed into the binomial distribution.

Fitting Logistic Models

- To fit the logistic regression and solve for the parameters $\boldsymbol{\beta}$, we need to use the maximum likelihood approach again.
- If Y following a binomial distribution with mean $\eta(x)$, the likelihood for each y_i (0 or 1) is

$$\eta(x_i)^{y_i} \left(1 - \eta(x_i)\right)^{(1-y_i)}$$

- Then the joint likelihood for all y_i's is

$$\prod_i^n \eta(x_i)^{y_i} \left(1 - \eta(x_i)\right)^{(1-y_i)}$$

- And the log-likelihood is given by

$$\sum_i^n -\log\left(1 + \exp(x_i^T \boldsymbol{\beta})\right) + \sum_i^n y_i (x_i^T \boldsymbol{\beta})$$

Fitting Logistic Models

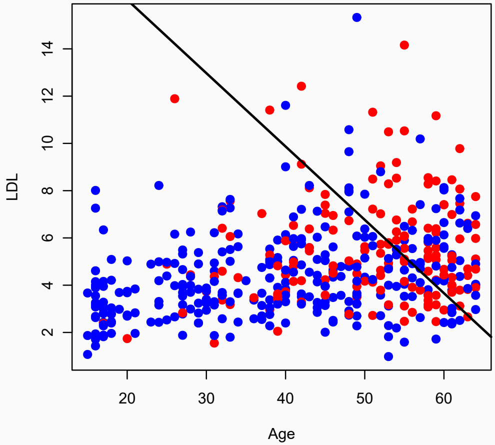
- There is no close form solution to this objective function. We need to solve it through numerical optimization (we will come back to this topic later on).
- An example: the South Africa Heart Disease Data `SAheart` in the `ElemStatLearn` package.
- Lets start with modeling the event of coronary heart disease (`chd`) using `age` and `ldl`.

South Africa Heart Disease Data

```
> library(ElemStatLearn)
> data(SAheart)
> heart.fit = glm(chd~ ldl + age, data=SAheart, family=binomial)
> summary(heart.fit)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.201040   0.477629  -8.796  < 2e-16 ***
ldl          0.188541   0.053462   3.527 0.000421 ***
age          0.058510   0.008831   6.626 3.46e-11 ***
```

South Africa Heart Disease Data



Interpreting the Logistic Model Fit

- We have two parameters age (0.058510) and ldl (0.188541), both are positive.
- Interpreting: higher age and cholesterol are associated with higher risk of developing heart disease.
- Recall the interpretation of linear regression parameters, each unit increase of X results in β increase of the mean value of Y.
- Logistic regression cannot be interpreted this way. β does not represent a “linear” increase of the probability.
- Instead, each unit increase of X results in β increase of the logit of the probably of $\{Y = 1\}$.
- Another commonly used interpretation is the odds ratio.

Example

- From the logistic regression model, calculate the fitted probably of developing heart disease at $\mathrm{age} = 50$ and $\mathrm{ldl} = 6$.
  - A). -0.1443034
  - B). 0.4639866
  - C). 0.9829890
  - D). 0.9962306

Odds ratio

- The odds ratio is nothing but a math trick to interpreted the β parameters.
- If we have two persons, with the same age = 50, and the ldl measures are 6 and 7 respectively.
- Hence we can let $x_1 = c(1, 7, 50)$, $x_2 = c(1, 6, 50)$.
- Then we can calculate their predicted probabilities, say η_1 and η_2.
- We know that logit$(\eta) = x^\intercal \boldsymbol{\beta}$. If we define the odds as $\eta/(1-\eta)$, then the log of odds ratio is

$$\log\left(\frac{\eta_1/(1-\eta_1)}{\eta_2/(1-\eta_2)}\right) = \log\left(\frac{\eta_1}{1-\eta_1}\right) - \log\left(\frac{\eta_2}{1-\eta_2}\right) = x_1^\intercal\boldsymbol{\beta} - x_2^\intercal\boldsymbol{\beta},$$

which is just the parameter estimate of ldl.