8. Logistic regression

- overview
- logistic model
- estimating coefficients
- simulation example

Overview

in classification problems, one labels a number to the response variable, Y

$$Y = \begin{cases} 1, & \text{if } \texttt{stroke;} \\ 2, & \text{if } \texttt{drug overdose ;} \\ 3, & \text{if } \texttt{epileptic seizure.} \end{cases}$$

these three conditions can be related to predictors, X

- though least-squares can be used to fit Y, there is no clear reasons to convert the difference between *qualitative* conditions into *quantitative* ordering
- even for binary classification, $Y \in \{0, 1\}$, if we perform least-squares, $\hat{Y}$ could lie outside $[0, 1]$ and it's not clear how to interpret the results
- logistic model is a model that is suitable for *qualitative* response variable

Binary classification

consider the problem of classifying data into two classes: $Y \in \{0, 1\}$

setting:

- we have data (Y, X) where Y is the response variable and X is the predictor
- example: defaults on credit card payment
  - $X = (X_1, X_2, X_3)$ contains `balance`, `income`, `student status`
  - Y is `default` status; $Y = 1$ is 'yes' and $Y = 0$ is 'no'

goal: find a model that provides $P(Y = 1 \mid X = x)$

$$P(\texttt{default = yes} \mid \texttt{balance} = 10,000\text{baht}, \texttt{income} = 200\text{kbaht}, \texttt{student = no})$$

Logistic model

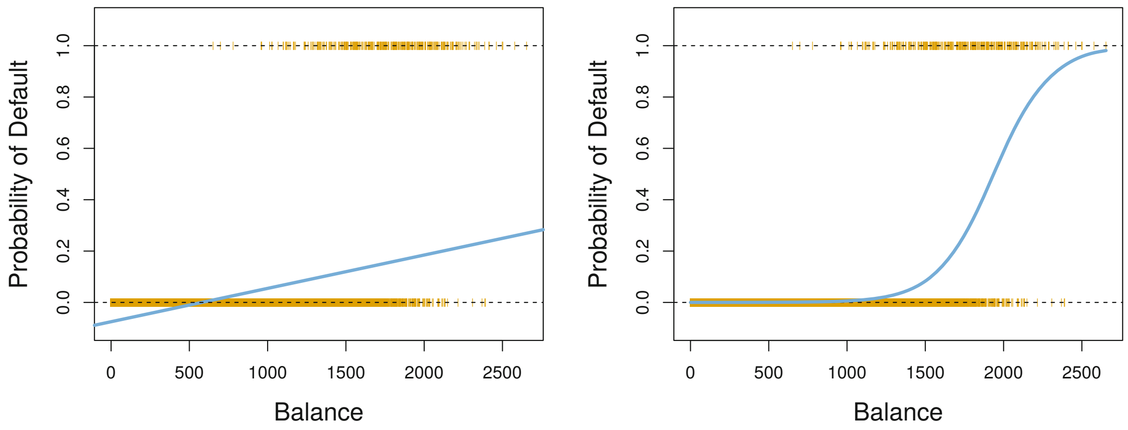
a logistic function is used to gives output between 0 and 1

$$f(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x} \quad \text{has S-shape}$$

(this is a nominal form of logistic, aka. sigmoid function)

a logistic model uses the logistic function to explain Y from predictors thru:

$$P(Y=1|X) = \frac{e^{\beta^T X}}{1+e^{\beta^T X}}, \quad P(Y=0|X) = \frac{1}{1+e^{\beta^T X}}$$



Logistic regression

problem: fitting the logistic model

$$P(Y = 1|X) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}$$

from data set $\{(y_i, x_i)\}_{i=1}^N$ to find parameters β

- the linear predictor term is $\beta^T X = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- if an intercept β_0 is needed, we assume X_k must contain $\mathbf{1}$
- estimation method: maximum likelihood estimation (more on this later)
- for new $X = x$, if $P(Y = 1|X) > 0.5$ we classify that this data belong to class '1', and '0' otherwise (the threshold 0.5 is up to the user)

- the following quantitiy, called odds,

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\beta^T X} \qquad \in (0, \infty)$$

  indicates the ratio of the chance that class '1' occurs to class '0'

- the log of odds, called logit

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta^T X$$

  provides a *link function* between the probability and the linear regression expression

- if X_k is one-unit changed
  - in linear regression, the average in Y is changed by β_k
  - in logistic regression, the log odds change by β_k

Estimating regression coefficients

denote the logistic function: $p(x) = e^{\beta^T x}/(1 + e^{\beta^T x})$

β_0, β are chosen to maximize the **likelihood function**

$$
\begin{aligned}
\mathcal{L}(\beta) &= \prod_{i:y_i=1} p(x_i) \prod_{k:y_k=0} (1 - p(x_k)) \\
&= \prod_{i:y_i=1} \frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \prod_{k:y_k=0} \frac{1}{1 + e^{\beta^T x_k}}
\end{aligned}
$$

since $\log(\cdot)$ is increasing, it is the same as maximizing the **log-likelihood**

$$
\log \mathcal{L}(\beta) = \sum_{i:y_i=1} e^{\beta^T x_i} - \sum_{k} \log(1 + e^{\beta^T x_k})
$$

this is a nonlinear unconstrained optimization problem (can be solved by Newton/Quasi-Newton)

Derivation of loglikelihood

suppose $\{(y_i, x_i)\}_{i=1}^n$ are available where $y_i = 0, 1$

- we can write $P(Y = y \mid X = x; \beta) = p(x)^y (1 - p(x))^{1-y}$
- if we have n independent observations, the likelihood function is expressed as

$$\mathcal{L}(y_1, \ldots, y_n \mid x; \beta) = \prod_i P(Y = y_i \mid x_i; \beta) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

$$\begin{aligned} \log \mathcal{L}(y_1, \ldots, y_n \mid x; \beta) &= \sum_{i=1}^n y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)) \\ &= \sum_{i=1}^n y_i \log \left(\frac{e^{\beta^T x_i}}{1 + e^{\beta^T x_i}} \right) + (1 - y_i) \log \left(\frac{1}{1 + e^{\beta^T x_i}} \right) \end{aligned}$$

- substitute $y_i = 1$ for some i and $y_i = 0$ otherwise; this gives $\log \mathcal{L}$ on page 8-7

Default on credit card payment

example of running logistic regression for the default data on page 8-3

	Coefficient	Std. error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	-0.6468	0.2362	-2.74	0.0062

prediction: use $\hat{\beta}$ from the table we can make an estimate of Y

- student/non-student with balance of $1,500$ dollars and income of $40,000$

$$
\begin{array}{lll}
\text{student} & \hat{p}(Y = 1 \mid X = (1500, 40000, 1)) & = 0.068 \\
\text{non-student} & \hat{p}(Y = 1 \mid X = (1500, 40000, 0)) & = 0.105
\end{array}
$$

- with the same balance and income, a non-student is more likely to default

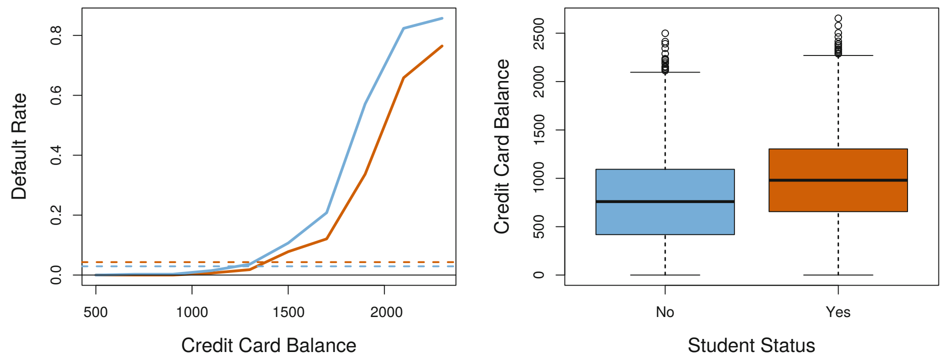
Correlated predictors

compare the results between one predictor (student status) and three predictors

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student [Yes]	−0.6468	0.2362	−2.74	0.0062

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−3.5041	0.0707	−49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- the coefficient of student status is negative (left) and positive (right)
- negative coefficient of student status indicates that students are less likely to default (than non-students) – here we can contradictory results ?



students / non-students

observations:

- in multiple regression (left table), negative coefficient for student indicates that *for a fixed value of balance and income*, a student is less likely to default than a non-student (confirmed by that the orange line is lower than the blue line)
- the horizontal lines show the default rates that are averaged over all values of balance and income – but here the orange line is higher than the blue line
- the box plots suggest that students tend to have higher credit card balance – associated with high default rates

explanations:

- 'student status' and 'balance' are correlated (students tend to have higher debt)
- an *individual* student with a given balance tends to have a lower chance of default, while students *on the whole* tend to have higher credit card balance which further tend to have a higher default rate

conclusions:

- a student is riskier than a non-student if no information about credit card balance is available
- a student is less risky than a non-student with the *same* credit card balance
- a confounding problem: a result obtained from one predictor is different from using multiple predictors when there is correlation among the predictors

K-label classification

the logistic regression can be extended to classify data into K categories

- define the response as indicator variable: $Y = (Y_1, Y_2, \ldots, Y_K)$ where

$$Y_k = 1 \quad \text{if the response fall into } k\text{th category and} \quad Y_j = 0, \quad \forall j \neq k$$

e.g. three medical conditions:

$$Y = \begin{cases} (1,0,0), & \text{if } \texttt{stroke;} \\ (0,1,0), & \text{if } \texttt{drug overdose ;} \\ (0,0,1), & \text{if } \texttt{epileptic seizure.} \end{cases}$$

- the choice of **multinomial** distribution is suitable; π_k is the probability of $Y_k = 1$

$$P(Y = (y_1, \ldots, y_K)) = \frac{1}{y_1! y_2! \cdots y_K!} \; \pi_1^{y_1} \pi_2^{y_2} \cdots \pi_K^{y_K}$$

Multinomial logistic model

denote G the variable indicating the group:

$$Y = (0, 0, \ldots, \underbrace{1}_{k\text{th}}, 0, \ldots, 0) \quad \Longleftrightarrow \quad G = k$$

the response belongs to kth category iff $G = k$

- model: log-odd of each response is linear function of predictors

$$\begin{array}{rcl}
\log \frac{P(G=1 \mid X)}{P(G=K \mid X)} & = & \beta_1^T X \\
\log \frac{P(G=2 \mid X)}{P(G=K \mid X)} & = & \beta_2^T X \\
& \vdots & \\
\log \frac{P(G=K-1 \mid X)}{P(G=K \mid X)} & = & \beta_{K-1}^T X
\end{array}$$

- the choice of last class as the denominator is arbitrary

- the conditional probabilities can be expressed as

$$P(G = k \mid X) = \frac{e^{\beta_k^T X}}{1 + \sum_{l=1}^{K-1} e^{\beta_l^T X}}, \quad k = 1, 2, \ldots, K-1,$$

$$P(G = K \mid X) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_l^T X}} \quad \text{(chosen to be referenced class)}$$

(the sum of K probabilities is one)

- denote $p_k(x; \beta) = P(G = k \,|x)$
- the **log-likelihood** function of (y, x) is expressed from multinomial distribution

$$\log p(y \mid x; \beta) = \sum_{l=1}^{K} y_l \log p_l(x; \beta) - \log(y_1! y_2! \cdots y_K!)$$

entries of $y = (y_1, \ldots, y_K)$ are either 0 or 1 – the last term on RHS is zero

Estimation of multinomial logistic coefficients

suppose data $\{(y^{(i)}, x^{(i)})\}_{i=1}^{n}$ are available (independent samples)

the log-likelihood function to be maximized is

$$\log \mathcal{L}(\beta) = \sum_{i=1}^{n} \log p(y^{(i)} \mid x^{(i)}; \beta)$$
$$= \sum_{i=1}^{n} \sum_{l=1}^{K} y_l^{(i)} \log p_l(x^{(i)}; \beta)$$

note that the term in $\sum_{l=1}^{K}$ reduces to $p_k(x^{(i)}; \beta)$ if $y^{(i)}$ belongs to k class

β can be solved numerically from iterative procedure like Newton-Raphson

`multinom` in R and `mnrfit` in MATLAB

References

All figures and examples are taken from Chapter 4 in

G.James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, 2015

Chapter 4 in

T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd edition, Springer, 2009