STA3010 Regression Analysis

Feng YIN

The Chinese University of Hong Kong (Shenzhen)

yinfeng@cuhk.edu.cn

April 24, 2020

Overview

Logistic Regression: Definition

Logistic regression is a class of regression models with

- output, y, is qualitative/categorical/discrete-valued;
- input, $\mathbf{x}$, is not constrained.

Binary Logistic Regression

Binary logistic regression models assume that the output has two possible outcomes, generically called "is-the-case" (1) and "not-the-case" (0). For instance, the output can be success vs. failure, alive vs. death, win vs. lose, etc.
In machine learning, this is also known as binary classification.

Multinomial Logistic Regression

Multinomial logistic regression models assume that the output has more than two possible outcomes.

Logistic Regression: Scope and Aim

Logistic regression can be applied, when

1. we want to model the probability of an output as a function of an input/feature vector;
2. we want to predict the probability of a case falling into either category of the binary output as a function of the input/features;
3. we want to classify a case into a suitable category based on the input/features.

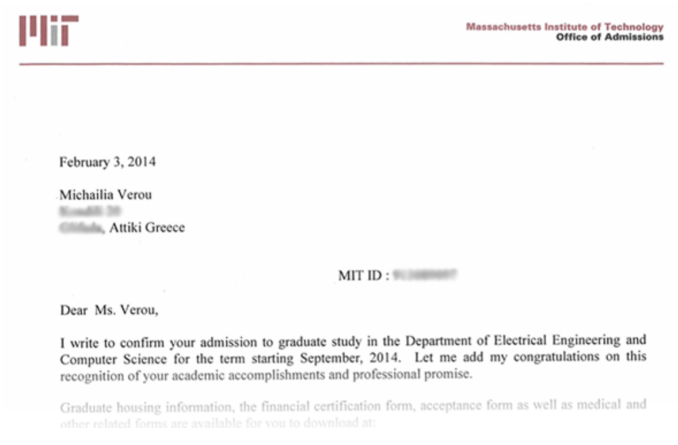
Logistic Regression: Motivating Examples

Example I: academic performance prediction

- The output y is either "pass the exam (1)" or "fail the exam (0)"
- The input $x = [1, x_1, x_2, ..., x_k]^T$ may include "lecture attendance rate", "how many times sleeping in class", "time spent on the lecture materials", "frequency of visiting TAs", etc.

Example II: PhD admission prediction

- The output y is either "accept (1)" or "reject (0)"
- The input $x = [1, x_1, x_2, ..., x_k]^T$ may include "GPA", "nationality", "master or bachelor", "GRE score", "research capability", etc.



Massachusetts Institute of Technology
Office of Admissions

February 3, 2014

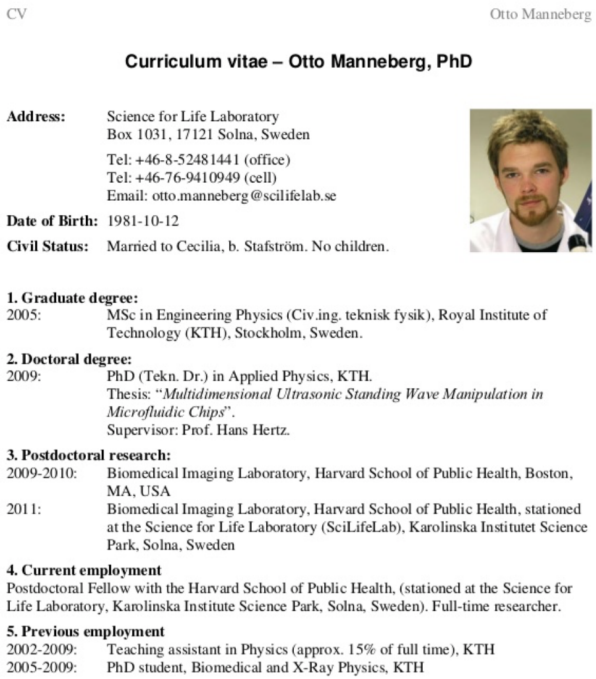
Michailia Verou
[redacted], Attiki Greece

MIT ID : [redacted]

Dear Ms. Verou,

I write to confirm your admission to graduate study in the Department of Electrical Engineering and Computer Science for the term starting September, 2014. Let me add my congratulations on this recognition of your academic accomplishments and professional promise.

Graduate housing information, the financial certification form, acceptance form as well as medical and other related forms are available for you to download at:



CV Otto Manneberg

Curriculum vitae – Otto Manneberg, PhD

Address: Science for Life Laboratory
Box 1031, 17121 Solna, Sweden
Tel: +46-8-52481441 (office)
Tel: +46-76-9410949 (cell)
Email: otto.manneberg@scilifelab.se

Date of Birth: 1981-10-12

Civil Status: Married to Cecilia, b. Stafström. No children.

1. Graduate degree:
2005: MSc in Engineering Physics (Civ.ing. teknisk fysik), Royal Institute of Technology (KTH), Stockholm, Sweden.

2. Doctoral degree:
2009: PhD (Tekn. Dr.) in Applied Physics, KTH.
Thesis: "*Multidimensional Ultrasonic Standing Wave Manipulation in Microfluidic Chips*".
Supervisor: Prof. Hans Hertz.

3. Postdoctoral research:
2009-2010: Biomedical Imaging Laboratory, Harvard School of Public Health, Boston, MA, USA
2011: Biomedical Imaging Laboratory, Harvard School of Public Health, stationed at the Science for Life Laboratory (SciLifeLab), Karolinska Institutet Science Park, Solna, Sweden

4. Current employment
Postdoctoral Fellow with the Harvard School of Public Health, (stationed at the Science for Life Laboratory, Karolinska Institute Science Park, Solna, Sweden). Full-time researcher.

5. Previous employment
2002-2009: Teaching assistant in Physics (approx. 15% of full time), KTH
2005-2009: PhD student, Biomedical and X-Ray Physics, KTH

Logistic Regression: Model

Binary logistic regression model in general form:

$$y = f(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon. \tag{1}$$

- The output y takes on two possible outcomes: "is-the-case (1)" and "not-the-case (0)".
- The inputs $\mathbf{x} = [1, x_1, x_2, ..., x_k]^T$ are defined similarly as before in the multiple linear regression model.
- The model parameters $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_k]^T$ are to be estimated.
- The random error is denoted by ε.

Why can't we simply use multiple linear model for $f(\mathbf{x}; \boldsymbol{\beta})$.

We assume that the output y given a specific $\mathbf{x}$ is a Bernoulli random variable with the probability mass function as follows:

y	probability mass function
1	$P(y=1\|\mathbf{x})=p$
0	$P(y=0\|\mathbf{x})=1-p$

$P(y=1|\mathbf{x})=p$ can be read as "the probability that the output y takes outcome "1" given the inputs $\mathbf{x}$ is equal to p.

Due to the model assumptions:

$$y = f(\mathbf{x}; \boldsymbol{\beta}) + \varepsilon \tag{2}$$
$$= E(y|\mathbf{x}, \boldsymbol{\beta}) + \varepsilon \tag{3}$$
$$= P(y = 1|\mathbf{x}) + \varepsilon \tag{4}$$

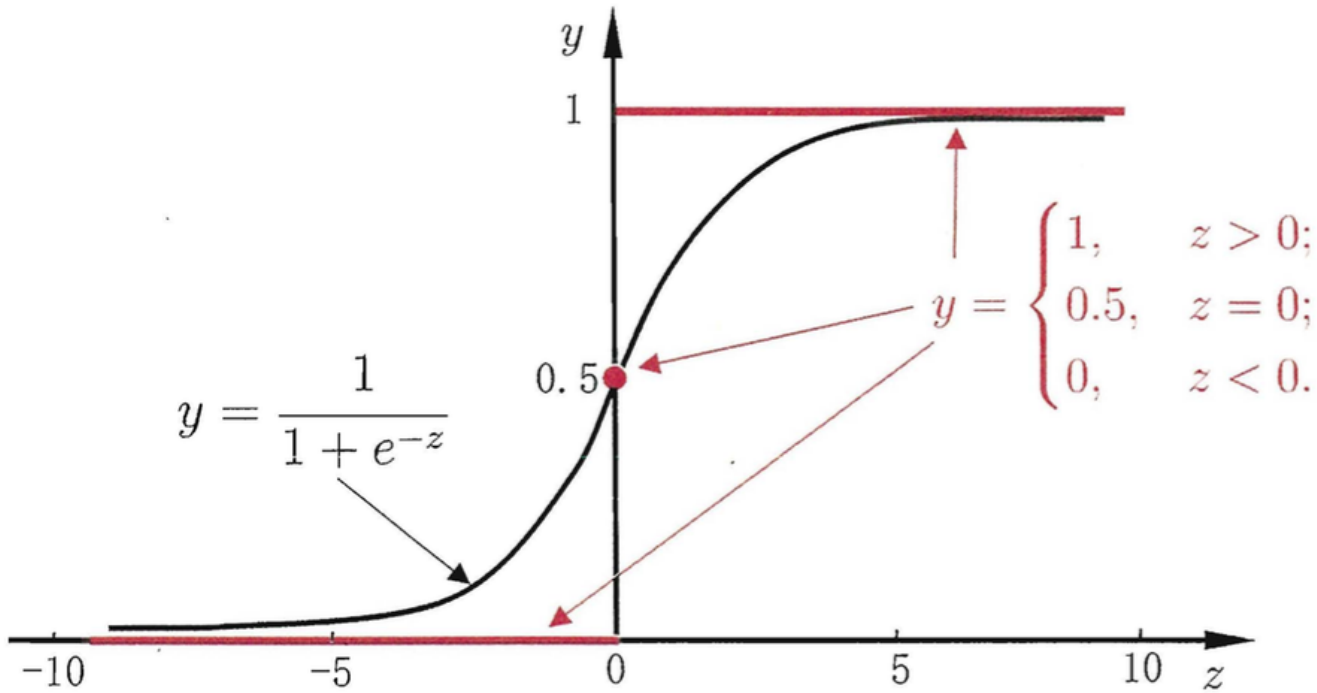
It is most widely used that

$$P(y = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}^T\boldsymbol{\beta})} = \frac{1}{1 + \exp(-\mathbf{x}^T\boldsymbol{\beta})}. \tag{5}$$

Logistic Regression: Sigmoid Function

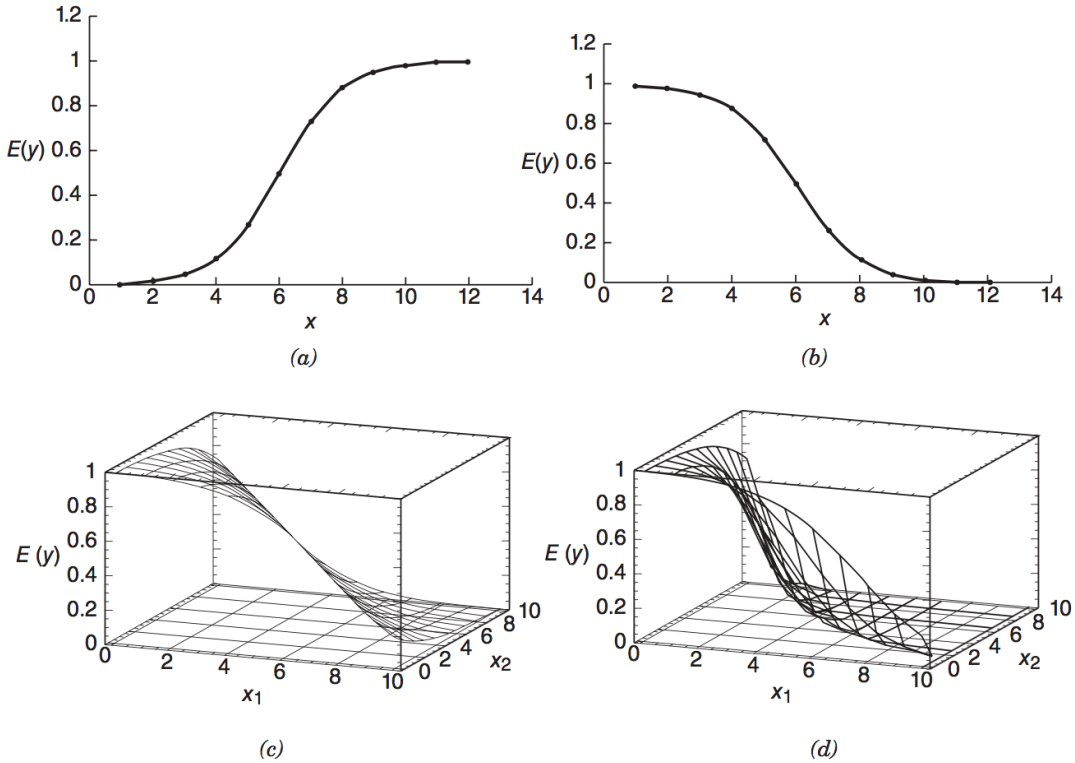
Sigmoid function has the form

$$f(z) = \frac{1}{1 + \exp(-z)} : \mathbb{R} \to \mathbb{R},\ z \in (-\infty, \infty) \tag{6}$$



sigmoid function (in black color) versus the unit step function (in red color).

More examples of sigmoid function with $\mathbf{x}^T\boldsymbol{\beta}$ as argument:



(a) $E(y) = 1/(1 + \exp(-6 + x))$,
(b) $E(y) = 1/(1 + \exp(6 - x))$,
(c) $E(y) = 1/(1 + \exp(-5 + 0.65x_1 + 0.4x_2))$,
(d) $E(y) = 1/(1 + \exp(-5 + 0.65x_1 + 0.4x_2 + 0.15x_1x_2))$

Logistic Regression: Logistic Function

Sigmoid function is a special case of logistic function with

$$f(z) = \frac{a}{1 + \exp(-b(z - c))} : \mathbb{R} \to \mathbb{R},\ z \in (-\infty, \infty) \tag{7}$$

where

- a controls the maximum value of the function;
- b controls the steepness of the function curve;
- c controls the mid-point.

The sigmoid function is a special case with $a = 1, b = 1, c = 0$.

Logistic Regression: Training Phase Vs. Test Phase

1. Training phase: given a set of n data points in $\mathcal{S} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_n, \mathbf{x}_n)\}$, we fit the model parameters $\boldsymbol{\beta}$ and obtain the maximum-likelihood estimate $\hat{\boldsymbol{\beta}}_{ML}$.
2. Test phase: given a novel input $\mathbf{x}_*$, we desire an accurate prediction on the probability of having "0" or "1" based on the trained logistic regression model.

Logistic Regression: Model Fit Using ML Estimation

Given a set of n data points in $\mathcal{S} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), ..., (y_n, \mathbf{x}_n)\}$, the likelihood function is given as follows:

$$L(\mathbf{y}|\mathbf{X}, \beta) = L(y_1, y_2, ..., y_n|\mathbf{X}, \beta) = \prod_{i=1}^{n} P(y_i|\mathbf{x}_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}. \quad (8)$$

It is more convenient to work with the log-likelihood:

$$\ln L(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) = \ln \prod_{i=1}^{n} P(y_i|\mathbf{x}_i) \tag{9}$$

$$= \sum_{i=1}^{n} \ln(1 - p_i) + \sum_{i=1}^{n} y_i \ln\left(\frac{p_i}{1 - p_i}\right) \tag{10}$$

$$= \sum_{i=1}^{n} y_i \mathbf{x}_i^T \boldsymbol{\beta} - \sum_{i=1}^{n} \ln\left(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})\right). \tag{11}$$

Logistic Regression: Model Fit Using ML Estimation

This slide can be skipped!

A modified expression: often in logistic regression models we have repeated observations at each level of the input $\mathbf{x}$.

Assuming we have n different input levels $\mathbf{x}_i$, $i = 1, 2, ..., n$, for each individual $\mathbf{x}_i$, we collect n_i observations y_{ij}, $j = 1, 2, ..., n_i$ with

$$P(y_{ij} = 1|\mathbf{x}_i) = p_i \tag{12}$$

$$P(y_{ij} = 0|\mathbf{x}_i) = 1 - p_i \tag{13}$$

The dataset is then:

$$\begin{aligned}\mathcal{S} =& \{(y_{1,1}, \mathbf{x}_1), (y_{1,2}, \mathbf{x}_1), ..., (y_{1,n_1}, \mathbf{x}_1),\\ & (y_{2,1}, \mathbf{x}_2), (y_{2,2}, \mathbf{x}_2), ..., (y_{2,n_2}, \mathbf{x}_{2,n_2}),\\ & \vdots\\ & (y_{n,1}, \mathbf{x}_n), (y_{n,2}, \mathbf{x}_n), ..., (y_{n,n_n}, \mathbf{x}_{n,n})\}\end{aligned} \tag{14}$$

Logistic Regression: Model Fit Using ML Estimation

This slide can be skipped!

The likelihood function can be written, for the new dataset in eq.(13), as follows:

$$L(\mathbf{y}|\mathbf{X},\beta) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} p_i^{y_{ij}}(1-p_i)^{1-y_{ij}} \tag{15}$$

Similarly, the log-likelihood function is derived as

$$\ln L(\mathbf{y}|\mathbf{X},\beta) = \sum_{i=1}^{n} t_i \ln(p_i) + \sum_{i=1}^{n}(n_i - t_i)\ln(1-p_i). \tag{16}$$

Herein, t_i denotes the total number of 1's observed for the same input $\mathbf{x}_i$ in n_i independent Bernoulli trials. For large n_i, the ratio $\frac{t_i}{n_i} \approx p_i$.

Logistic Regression: Newton's Method

- We define $l(\beta) = -\ln L(\mathbf{y}, \beta|\mathbf{X})$ and find the maximum-likelihood estimate via

$$\hat{\beta}_{ML} = \arg\min_{\beta} l(\beta) \tag{17}$$

- Newton's method requires both the gradient and the Hessian matrix of the cost function $l(\beta)$, details are given in the manuscript.

The final results are:

$$\mathbf{g} = \nabla_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) = \mathbf{X}^T(\mathbf{p} - \mathbf{y}), \tag{18}$$

$$\mathbf{H} = \nabla_{\boldsymbol{\beta}} \nabla_{\boldsymbol{\beta}}^T l(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{V} \mathbf{X}, \tag{19}$$

where

- $\mathbf{p} = [p_1, p_2, ..., p_n]^T$;
- $\mathbf{V} = diag(p_1(1-p_1), p_2(1-p_2), ..., p_n(1-p_n))$.

Newton Method

- Find an initial estimate $\hat{\boldsymbol{\beta}}^0$
- Select a threshold δ_T
- For $\eta = 0, 1, 2, ...,$ do
  1. Select a proper step-length α^η (can be $\alpha^\eta = 1$ to have the vanilla version)
  2. Compute $\hat{\boldsymbol{\beta}}^{\eta+1} = \hat{\boldsymbol{\beta}}^\eta - \alpha^\eta \left(\mathbf{H}^\eta\right)^{-1} \mathbf{g}^\eta$.
  3. Terminate if the change in two consequent costs and/or parameter estimates is smaller than the threshold δ_T, or if the maximal number of iterations has been reached.
  4. Otherwise $\eta := \eta + 1$ and repeat the above iterative process.

Prediction with the Trained Model

After the ML based training,

- given a novel input $\mathbf{x}_*$, we can perform prediction in light of $P(y_* = 1|\mathbf{x}_*, \hat{\boldsymbol{\beta}}_{ML}) = \frac{\exp(\mathbf{x}_*^T \hat{\boldsymbol{\beta}}_{ML})}{1+\exp(\mathbf{x}_*^T \hat{\boldsymbol{\beta}}_{ML})}$.
- based on the above probability, we can further make a binary decision, i.e., choosing either 1 or 0, if necessary.

Logistic Regression: Other Existing Models

1. Cumulative normal distribution function:

$$P(y = 1|\mathbf{x}) = \Phi\left(\mathbf{x}^T\beta\right). \tag{20}$$

The inverse function is called probit function.

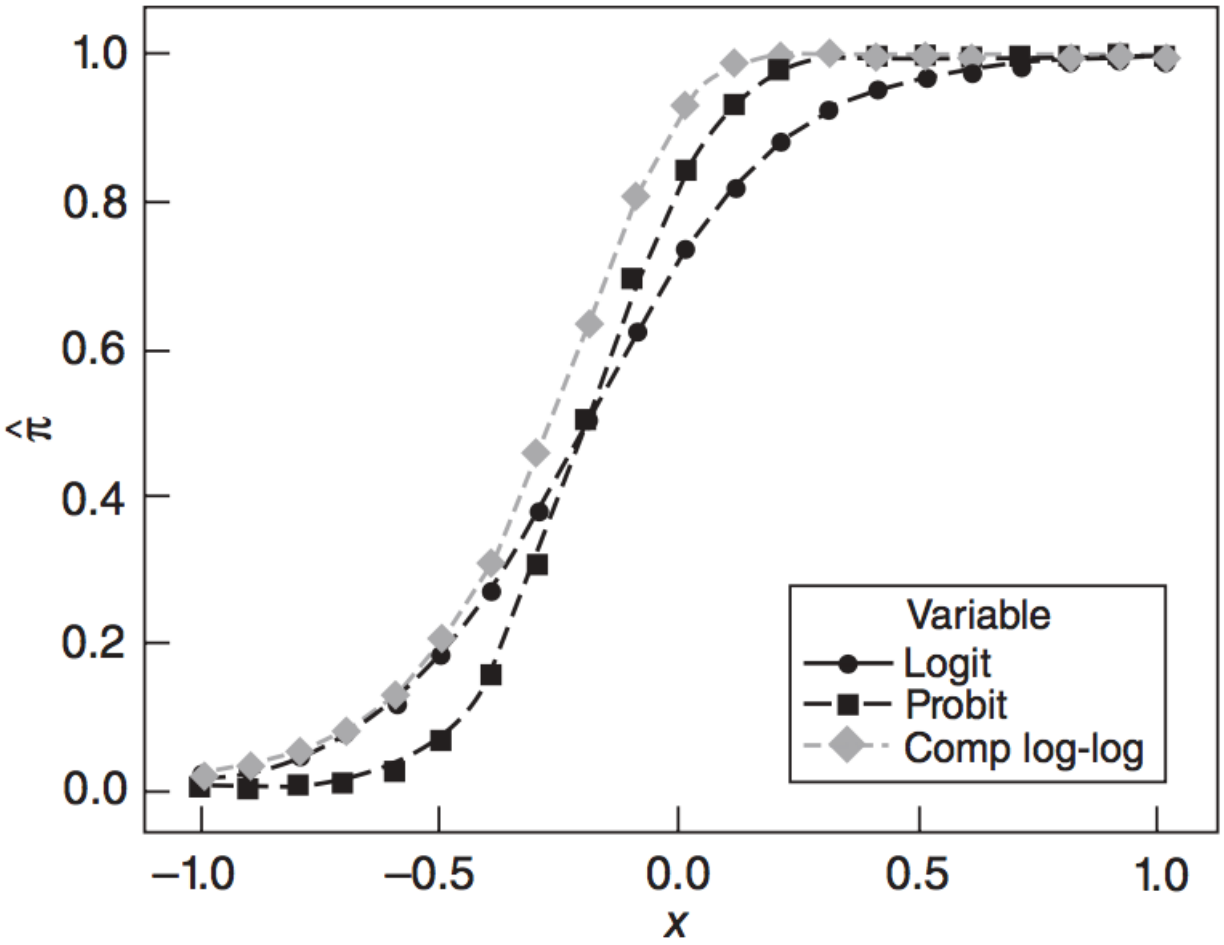
2. Log-Weibull distribution function:

$$P(y = 1|\mathbf{x}) = 1 - \exp\left(-\exp\left(\mathbf{x}^T\beta\right)\right). \tag{21}$$

The inverse function is called complimentary log-log function.

Logistic Regression: Other Models

Comparison of the three models:



Logistic Regression: More than Two Outcomes

- Consider the case where there are $m+1$ ($m>1$) possible categorical outcomes but the outcomes are nominal (i.e., no ranking).
- Let the outcomes be represented by $0, 1, 2, ..., m$.

The probabilities that the responses on observation i take on one of the $m+1$ possible outcomes can be modeled as:

$$P(y_i = 0|\mathbf{x}_i) = \frac{1}{1 + \sum_{j=1}^{m} \exp(\mathbf{x}_i^T \boldsymbol{\beta}^j)}, \tag{22}$$

$$P(y_i = 1|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}^1)}{1 + \sum_{j=1}^{m} \exp(\mathbf{x}_i^T \boldsymbol{\beta}^j)}, \tag{23}$$

$$\vdots = \vdots$$

$$P(y_i = m|\mathbf{x}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}^m)}{1 + \sum_{j=1}^{m} \exp(\mathbf{x}_i^T \boldsymbol{\beta}^j)}. \tag{24}$$

Note that there are m parameter vectors $\boldsymbol{\beta}^j$, $j = 1, 2, ..., m$ to be tuned.

Comparing each response category to a "baseline" produces logits

$$\ln\left(\frac{P(y_i=1|\mathbf{x}_i)}{P(y_i=0|\mathbf{x}_i)}\right)=\mathbf{x}_i^T\boldsymbol{\beta}^1, \tag{25}$$

$$\ln\left(\frac{P(y_i=2|\mathbf{x}_i)}{P(y_i=0|\mathbf{x}_i)}\right)=\mathbf{x}_i^T\boldsymbol{\beta}^2, \tag{26}$$

$$\vdots = \vdots \tag{27}$$

$$\ln\left(\frac{P(y_i=m|\mathbf{x}_i)}{P(y_i=0|\mathbf{x}_i)}\right)=\mathbf{x}_i^T\boldsymbol{\beta}^m, \tag{28}$$

where our choice of zero as the baseline category is arbitrary.

Summary

- Regression with categorical output
- Binary logistic regression
- Sigmoid function versus logistic function
- Maximum-likelihood parameter estimation using Newton's method
- Asymptotic property of the ML parameter estimator
- Extension to more than two categorical outcomes

Appendix: Bernoulli Trials

Bernoulli trials occurs when a Bernoulli experiment is performed several independent times and the probability of "is-the-case", denoted by p, remains the same from trial to trial. In addition, we shall frequently let $1 - p$ denote the probability of "not-the-case".

Example: Suppose that the probability of germination of a beet seed is 0.8 and the germination of a seed is called "is-the-case". If we plant 10 seeds and can assume that the germination of one seed is independent of the germination of another seed, this would correspond to 10 Bernoulli trials with $p = 0.8$.

Appendix: Bernoulli Distribution

Let y be a random variable associated with a Bernoulli trial with $y = 1$ representing "is-the-case" while $y = 0$ representing "not-the-case".

When y is Bernoulli distributed, the probability mass function (pmf) is as follows:

$$P(y = 1) = p, \tag{29}$$
$$P(y = 0) = 1 - p. \tag{30}$$

Alternatively, $P(y) = p^y(1 - p)^{(1-y)}$, $y = 0$ or 1.

The expectation of a Bernoulli distributed random variable, y, is

$$\mu = E(y) = \sum_{y \in \{0,1\}} y\, p^y (1-p)^{(1-y)} = p. \tag{31}$$

The variance of a Bernoulli distributed random variable, y, is

$$\sigma^2 = var(y) = \sum_{y \in \{0,1\}} (y-p)^2 p^y (1-p)^{(1-y)} = p(1-p). \tag{32}$$

In a sequence of n Bernoulli trials, we shall let y_i denote the observation associated with the ith Bernoulli trial.