SF 2930 Regression Analysis

Lecture 8

Logistic Regresssion, Generalized Linear Regression

Timo Koski

KTH Royal Institute of Technology

2023

Your Learning Outcomes

- Logit function, Logistic function
- Logistic regression
  - definition
  - likelihood function
  - maximum likelihood estimate
  - best prediction & validation

Part 1: Binary Response Y and $E[Y|\mathbf{x}]$

Binary response and Covariates (of any kind)

$$\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}.$$

is a $(p+1) \times 1$ vector of covariates/predictors, which can be binary, ordinal, categorical or continuous. The response Y is binary r.v.. Its values y are coded as $y \in \{0, 1\}$, equivalently recoded as $y \in \{-1, 1\}$, which has its advantages, as seen later.

Example

$Y =$ *Bacterial Meningitis or Acute Viral Meningitis.*
$x =$ *cerebrospinal fluid total protein count,* $p = 1$.

Logistic regression

$y \in \{0, 1\}, \mathbf{x}^T = (1 \quad x_1 \quad x_2 \quad \dots \quad x_{32})$

Example

Diaz, Armando A et. al: Prediction of protein solubility in *Escherichia coli* using logistic regression, *Biotechnology and bioengineering*, 105, 2 pp. 374–383, 2010.
. . . a model for the prediction of the solubility of proteins overexpressed in the bacterium *Escherichia coli*. The model uses the statistical technique of logistic regression. To build this model, 32 covariates x_i that could potentially correlate well with solubility were used.
Logistic regression provides the probability p of a certain protein to belong ($= Y$) to one set or another.

American Journal of Epidemiology

Vol. 161, No. 6
Printed in U.S.A.
DOI: 10.1093/aje/kwi091

Long-Term Mobile Phone Use and Brain Tumor Risk

Stefan Lönn[1], Anders Ahlbom[1], Per Hall[2], Maria Feychting[1], and the Swedish Interphone Study Group

[1] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
[2] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.

Received for publication August 26, 2004; accepted for publication January 12, 2005.

The purpose of a population-based, case-control study was to test the hypothesis that long-term mobile phone use increases the risk of brain tumors. The authors identified all cases aged 20-69 years who were diagnosed with glioma or meningioma during 2000-2002 in certain parts of Sweden. Randomly selected controls were stratified on age, gender, and residential area. Detailed information about mobile phone use was collected from 371 (74%) glioma and 273 (85%) meningioma cases and 674 (71%) controls. For regular mobile phone use, the odds ratio (using logistic regression) was 0.8 (95% confidence interval: 0.6, 1.0) for glioma and 0.7 (95% confidence interval: 0.5, 0.9) for meningioma.

Lönn, Stefan and Ahlbom, Anders and Hall, Per and Feychting, Maria: *Long-term mobile phone use and brain tumor risk. American journal of epidemiology*, 166, pp. 526–535 2005.

Regression Analysis is a Study of Conditional Expectation

Definition

A binary r.v. $Y \sim \text{Be}(p)$, Bernouilli distribution with parameter $p \in (0, 1)$, the probability mass function is

	$y = 1$	$y = 0$
$P(Y = y)$	p	$1 - p$

$$E[Y] = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = p.$$

Regression Analysis is a Study of Conditional Expectation

When Y is a binary response variable with covariate values $\mathbf{x}$, the conditional expectation is by the simplest of simple arguments above

$$E[Y \mid \mathbf{x}] = P(Y = 1 \mid \mathbf{x}). \tag{1}$$

We shall find an expression for $P(Y = 1 \mid \mathbf{x})$, so that Y is said to follow a *logistic regression*.

- One way: Model $P(\mathbf{x} \mid Y = y)$ and $P(Y = y)$ and use Bayes' formula to find $P(Y = 1 \mid \mathbf{x})$.
- A different plan: find an invertible function g and set

$$g(E[Y \mid \mathbf{x}]) = \mathbf{x}^T \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ be $(p+1) \times 1$

Part 2: Search for g

$$g(E[Y \mid \mathbf{x}]) = \mathbf{x}^T \boldsymbol{\beta}, \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ *be* $(p+1) \times 1$.

- *Link function*
  - *Logit function*
  - *Logistic function a.k.a Sigmoid function*

Link Function

Bernoulli Distribution

$$\begin{array}{lll} & y=1 & y=0 \\ P(Y=y) & p & 1-p \end{array}$$

We write this as

$$P(Y=y) = p^y(1-p)^{1-y}$$

and then

$$P(Y=y) = e^{\ln\left(\frac{p}{1-p}\right)y+\ln(1-p)} \tag{3}$$

We shall now endeavour upon a study of $\ln\left(\frac{p}{1-p}\right)$. In statistics this is the 'canonical link function' for the Bernoulli distribution

The function logit(p)

Terminology: $\frac{p}{1-p}$ is known as the odds of success. The logarithmic odds of success is called the logit of p

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$$

We set $\theta = \text{logit}(p)$.

The logit(p) and its inverse

$$\theta := \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

The inverse function is

$$p = \text{logit}^{-1}(\theta) = \frac{e^{\theta}}{1+e^{\theta}}$$

The inverse function is denoted by $\sigma(\theta)$, and is called the **logistic function** or the **sigmoid**.

$$\sigma(\theta) := \frac{1}{1+e^{-\theta}}, \quad -\infty < \theta < \infty,$$

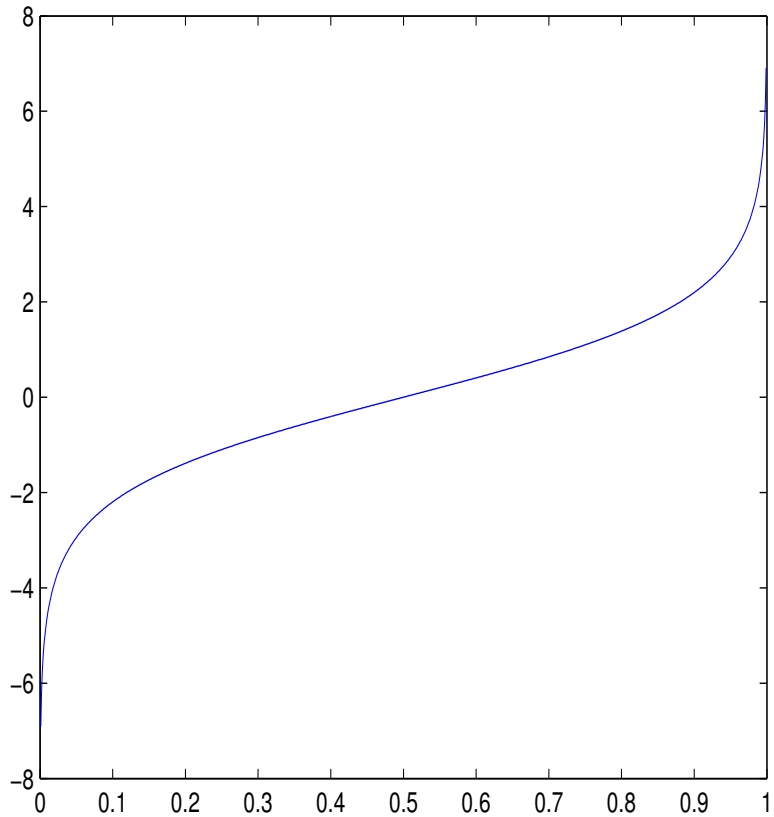
and. Note that $\sigma(0) = \frac{1}{2}$.

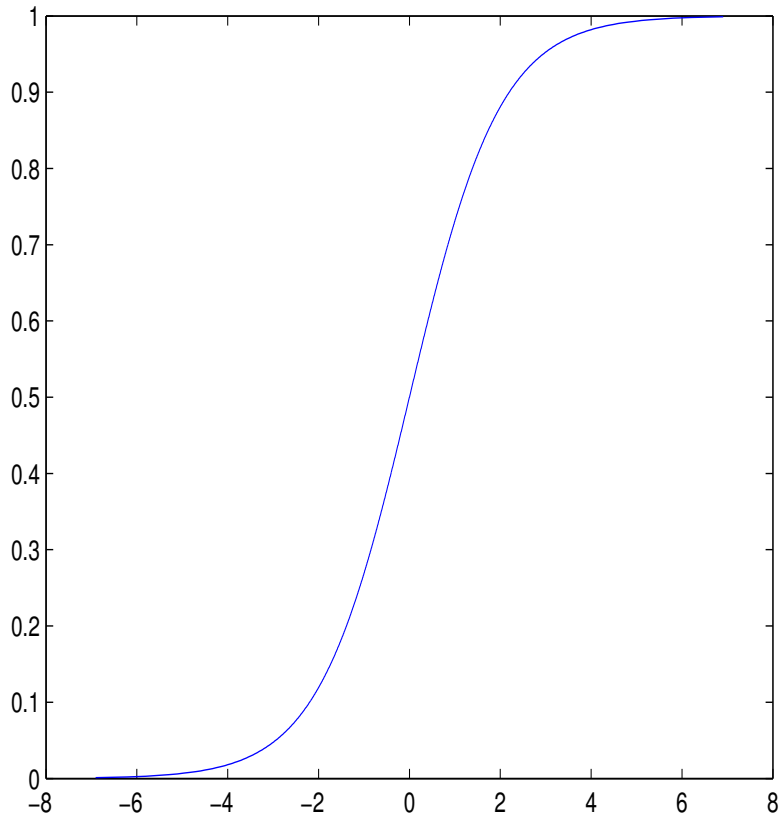
Remark

In biology the logistic function refers to change in size of a species population. In artifical neural networks the sigmoid is a network output function called **sigmoid**.

The logit(odds) and the logistic function

$$\theta = \text{logit}(p) = \log\left(\frac{p}{1-p}\right), \quad 0 < p < 1 \qquad p = \sigma(\theta) = \frac{1}{1+e^{-\theta}}, \quad -\infty < \theta < \infty,$$





Part 3: Logistic Regression

Now let us look at (2), i.e.,

$$g\left(E\left[Y \mid \mathbf{x}\right]\right) = \mathbf{x}^T\boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$. Let us take $g(x) = \operatorname{logit}(x)$ so that we have

$$\operatorname{logit}\left(E\left[Y \mid \mathbf{x}\right]\right) = \mathbf{x}^T\boldsymbol{\beta} \tag{5}$$

Then use the inverse function of $\operatorname{logit}(x)$ in (5), which gives

$$E\left[Y \mid \mathbf{x}\right] = \sigma\left(\mathbf{x}^T\boldsymbol{\beta}\right) = \frac{e^{\mathbf{x}^T\boldsymbol{\beta}}}{1 + e^{\mathbf{x}^T\boldsymbol{\beta}}} = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\beta}}}. \tag{6}$$

Here we have (1), i.e.,

$$E[Y \mid \mathbf{x}] = P(Y = 1 \mid \mathbf{x}).$$

we have now

$$P(Y = 1 \mid \mathbf{x}) = \sigma\left(\mathbf{x}^T\beta\right) = \frac{1}{1 + e^{-\mathbf{x}^T\beta}}.$$

In addition

$$P(Y = 0 \mid \mathbf{x}) = 1 - P(Y = 1 \mid \mathbf{x}) = 1 - \frac{1}{1 + e^{-\mathbf{x}^T\beta}} = \frac{e^{-\mathbf{x}^T\beta}}{1 + e^{-\mathbf{x}^T\beta}}.$$

We have also equivalently

$$P(Y = 1 \mid \mathbf{x}) = \frac{e^{\mathbf{x}^T\beta}}{1 + e^{\mathbf{x}^T\beta}}, P(Y = 0 \mid \mathbf{x}) = \frac{1}{1 + e^{\mathbf{x}^T\beta}}.$$

Logistic regression

Definition

If

$$Y = \begin{cases} 1 & \text{with probability } \sigma\left(\mathbf{x}^T\beta\right) \\ 0 & \text{with probability } 1 - \sigma\left(\mathbf{x}^T\beta\right). \end{cases}$$

then we say that Y follows a logistic regression w.r.t. the predictor variables $x_1, x_2, \ldots, x_p$.

Logistic regression: special case

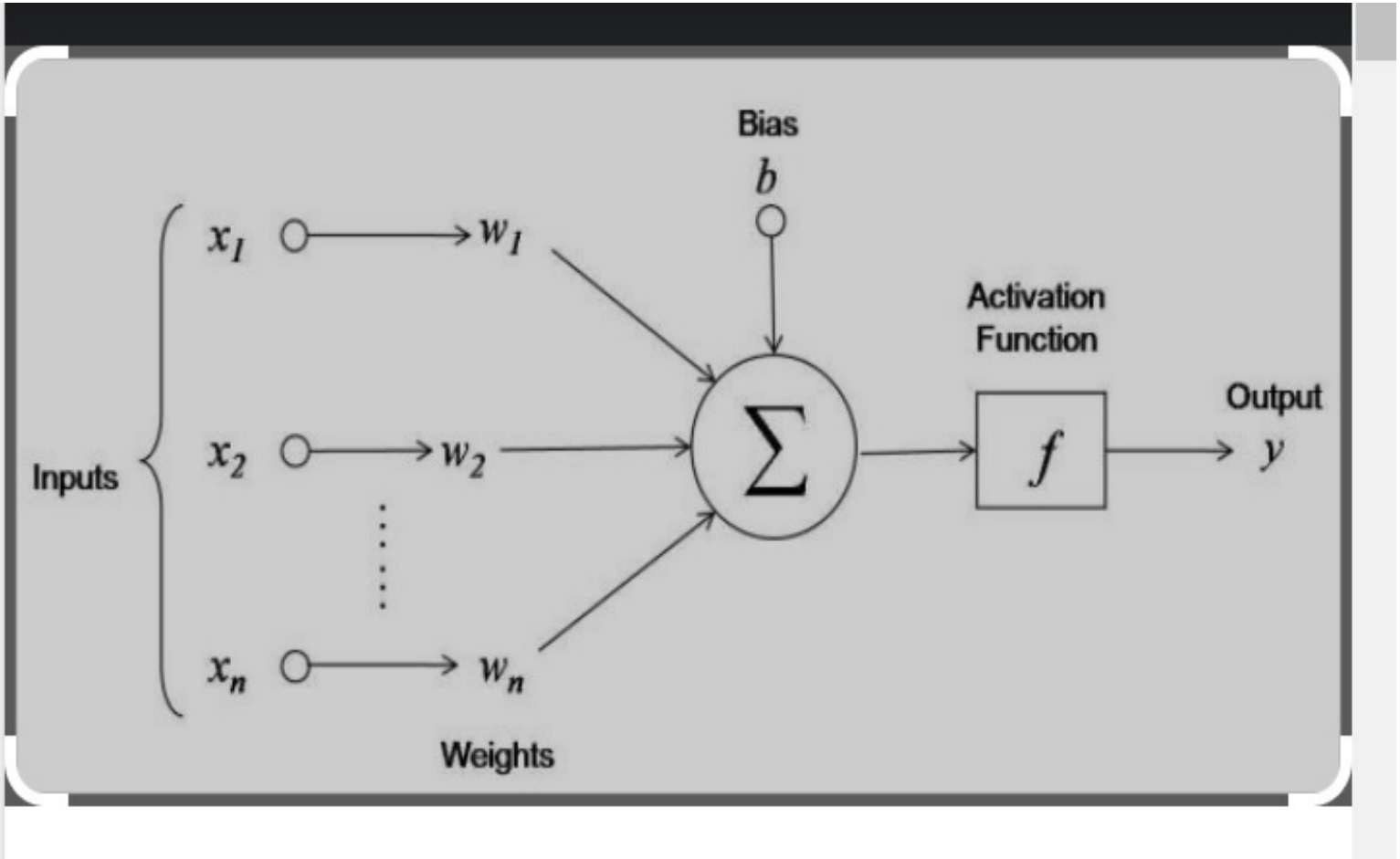
If $\mathbf{x}^T\boldsymbol{\beta} = 0$, then $\sigma(0) = \frac{1}{2}$, and

$$Y = \begin{cases} 1 & \text{with probability } \frac{1}{2} \\ 0 & \text{with probability } \frac{1}{2} \end{cases}$$

or, $Y \sim \mathrm{Ber}(1/2)$. Hence "flipping an honest coin" can be seen as a special case of logistic regression.

Artificial Neural Network with One Layer

Activation function: $\sigma\left(\mathbf{x}^T\mathbf{w}\right)$, $\mathbf{w} \leftrightarrow (\beta_1, \ldots, \beta_p)$, $b \leftrightarrow \beta_0$



WHY?

Why did we take such a tedious or long-winding route to the goal? Why not give the definition of ´Y following a logistic regression´ straightaway?

WHY? A: Generalized Linear Regression

Recall (3)

$$P(Y = y) = e^{\ln\left(\frac{p}{1-p}\right)y + \ln(1-p)}$$

We have $\theta = \mathrm{logit}(p)$ so that

$$P(Y = y) = e^{\theta y - \ln\left((1+e^{\theta}\right)} \tag{7}$$

In statistical theory (8) shows that the Bernouilli distribution belongs to the exponential family of distributions and (8) is its "natural form", θ is called the natural parameter of the Bernouilli distribution. Then we get the logistic regression probabilities as

$$P(Y = y) = e^{y\mathbf{x}^T\boldsymbol{\beta} - \ln\left((1+e^{\mathbf{x}^T\boldsymbol{\beta}}\right)} \tag{8}$$

This is a special case of how Generalized Linear Regression (GLM) are constructed.

Generalized Linear Regression (GLM)

A GLM consists of three elements:

1. A particular distribution from exponential families of probability distributions is modeling Y.
2. A linear predictor $\mathbf{x}^T\beta$, which is made equal the natural parameter $\theta = \mathbf{x}^T\beta$ and
3. A link function g such that $\mathrm{E}(Y \mid \mathbf{x}) = g^{-1}(\theta)$.

The link function provides the relationship between the linear predictor and the mean of the distribution function. There is always one well-defined canonical link function, which is derived from the pdf of the Y in the natural exponential form. Other link functions can be considered.

An Example: Poisson Regression

1. $Y \sim \mathrm{Po}(\lambda)$, $\lambda > 0$, $E[Y] = \lambda$.

$$P(Y = y) = \frac{1}{y!} e^{-\lambda} \lambda^y, \quad y = 0, 1, \ldots$$

$$= \frac{1}{y!} e^{y \ln \lambda - \lambda}$$

Set $\theta = \ln \lambda$. We get the natural form

$$P(Y = y) = \frac{1}{y!} e^{y\theta - e^\theta}$$

2. A linear predictor $\mathbf{x}^T \beta$, which is made equal to the natural parameter $\theta = \mathbf{x}^T \beta$ and
3. The link function $g(x) = \ln x, x > 0$ such that $\ln E(Y \mid \mathbf{x}) = \mathbf{x}^T \beta$

Example: Poisson Regression

Hence the probability mass function of the **Poisson regression** is

$$P(Y = y|\mathbf{x}) = \frac{1}{y!} e^{y\mathbf{x}^T\boldsymbol{\beta} - e^{\mathbf{x}^T\boldsymbol{\beta}}}, \quad y = 0, 1, 2, \ldots$$

Logistic regression: genetic epidemiology

Logistic regression is extensively applied in medical research, where 'success' may mean the occurrence of a disease or death due to a disease, and $x_1, x_2, \ldots, x_p$ are environmental and genetic riskfactors.
Woodward, M. : *Epidemiology: study design and data analysis*, 2013, CRC Press.

Logistic regression: genetic epidemiology

Suppose we have two populations, where $x_1^{(0)}$ in first population and $x_2^{(1)}$ in the second population, all other predictors are equal in the two populations. Then a medical geneticist finds it useful to calculate the logarithm of the odds ratio

$$\ln \psi = \ln \frac{p_1}{1-p_1} - \ln \frac{p_2}{1-p_2}$$

$$= \beta_i \left(x_1^{(0)} - x_2^{(1)} \right)$$

or

$$\psi = e^{\beta_i \left(x_1^{(0)} - x_2^{(1)} \right)}$$

EGAT Study (from Woodward)

Smoker at entry	Cardiovascular death during follow-up		
	Yes	No	Total
Yes	31	1386	1417
No	15	1883	1898
Total	46	3269	3315

EGAT Study (from Woodward)

Logistic regression

$$\widehat{\text{logit}}\left(E\left[Y \mid \mathbf{x}\right]\right) = -4.8326 + 1.0324x$$

was fitted with $x = 1$ for smokers and $x = 0$ for non-smokers. Then the odds ratio is

$$\text{OR} = \frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})} = e^{\widehat{\beta}(x_1 - x_2)} = e^{1.3024(1-0)} = 2.808$$

The log odds for smokers is

$$-4.8326 + 1.0324 \times 1 = -3.8002$$

giving odds= 0.2224. For non-smokers the odds are 0.008.

EGAT Study (from Woodward)

The risk for cardiovascular death for smokers is

$$\frac{1}{1+e^{-4.8326+1.0324\times 1}} = 0.0219$$

For nonsmokers

$$\frac{1}{1+e^{-4.8326+1.0324\times 0}} = 0.0079$$

Logistic regression

$$P(Y = 1 \mid \mathbf{x}) = \sigma(\mathbf{x}^T\beta) = \frac{e^{\mathbf{x}^T\beta}}{1 + e^{\mathbf{x}^T\beta}}$$

$$P(Y = 0 \mid \mathbf{x}) = \frac{1}{1 + e^{\beta^T\mathbf{x}^T}}.$$

Odds ratio (for success)

$$\mathrm{OR} = \mathrm{Odds} \quad \mathrm{Ratio} = \frac{P(Y = 1 \mid \mathbf{x})}{P(Y = 0 \mid \mathbf{x})} = e^{\mathbf{x}^T\beta}$$

Hence a unit change in x_i corresponds to e^{β_i} change in odds and β_i change in logodds.

Mobile Phones and Brain Tumor; Lönn et.al.

> For regular mobile phone use, the odds ratio (using logistic regression) was 0.8 (95% confidence interval: 0.6, 1.0) for glioma and 0.7 (95% confidence interval: 0.5, 0.9) for meningioma.

We see that OR$=1$ is included in one of the intervals and that OR $= 0.7$ has the CI 95% confidence interval $(0.5, 0.9)$.
Lönn et.al. state (loc.cit) that

> *This study includes a large number of long-term mobile phone users, and (we) conclude that the data do not support the hypothesis that mobile phone use is related to an increased risk of glioma or meningioma.*

Part 3: Probability Model for Logistic Regression

Part 3: Probability Model for Logistic Regression

ϵ is a r.v.,

$$\epsilon \sim \text{Logistic}(0, 1)$$

means that the cumulative distribution function (CDF) of the logistic distribution is the logistic function:

$$P(\epsilon \leq x) = \frac{1}{1 + e^{-x}} = \sigma(x)$$

A generative model

We need the following regression model

$$Y^* = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon,$$

where

$$\epsilon \sim \text{Logistic}(0, 1),$$

i.e. the variable Y^* can be written directly in terms of the linear predictor function and an additive random error variable. The logistic distribution (?) is the probability distribution the random error.

Logistic distribution

ϵ is a r.v.,

$$\epsilon \sim \text{Logistic}(0, 1)$$

means that the cumulative distribution function (CDF) of the logistic distribution is the logistic function:

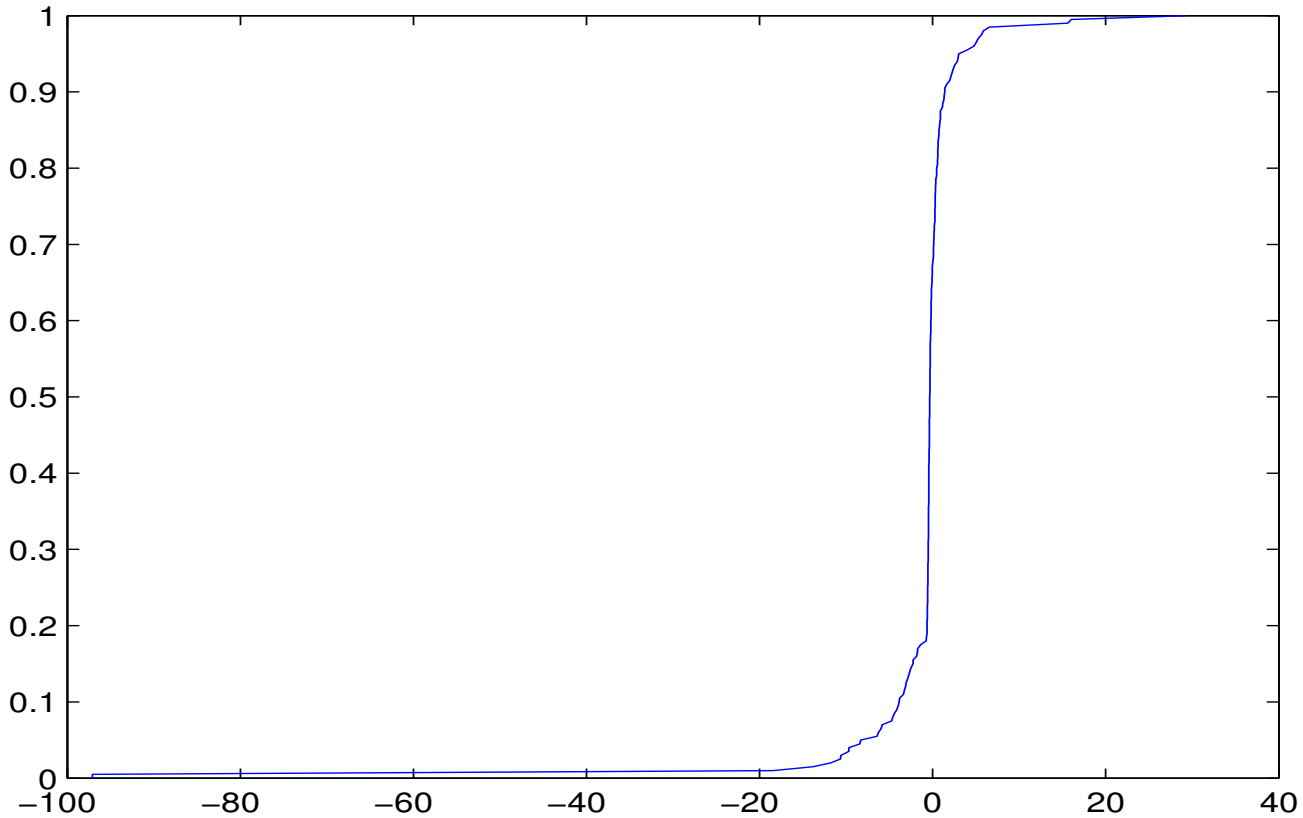
$$P(\epsilon \leq x) = \frac{1}{1 + e^{-x}} = \sigma(x)$$

I.e. $\epsilon \sim \text{Logistic}(0, 1)$*, if the probability density function is*

$$\frac{d}{dx}\sigma(x) = \frac{e^{-x}}{(1 + e^{-x})^2}.$$

Simulating $\epsilon \sim$ Logistic(0, 1)

This is simple: simulate $p_1, \ldots, p_n$ from the uniform distribution on (0, 1) and then do $\varepsilon_i = \mathrm{logit}(p_i)$, $i = 1, \ldots, n$. In the figure we plot the empirical distribution function of ϵ_i for $n = 200$.



A piece of probability

$\epsilon \sim \text{Logistic}(0, 1)$, what is $P(-\varepsilon \leq x)$?

$$P(-\varepsilon \leq x) = P(\varepsilon \geq -x) = 1 - P(\varepsilon \leq -x)$$

$$= 1 - \sigma(-x)$$

$$= 1 - \frac{1}{1+e^x} = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

$$= P(\epsilon \leq x).$$

$\epsilon \sim \text{Logistic}(0, 1) \Leftrightarrow -\epsilon \sim \text{Logistic}(0, 1)$.

Generating model and/or how to simulate

Take a continuous latent variable Y^* (latent= an unobserved random variable) that is given as follows:

$$Y^* = \mathbf{x}^T \boldsymbol{\beta} + \varepsilon$$

and

$$\epsilon \sim \text{Logistic}(0, 1).$$

Define the response Y as the indicator for whether the latent variable is positive:

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \text{ i.e. } -\varepsilon < \mathbf{x}^T\boldsymbol{\beta}, \\ 0 & \text{otherwise.} \end{cases}$$

Then Y follows a logistic regression w.r.t. $\mathbf{x}$. We need only to verify that

$$P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\beta}}}.$$

$$
\begin{aligned}
P(Y = 1 \mid \mathbf{x}) &= P(Y^* > 0 \mid \mathbf{x}) && (9)\\
&= P(\mathbf{x}^T\boldsymbol{\beta} + \varepsilon > 0) && (10)\\
&= P(\varepsilon > -\mathbf{x}^T\boldsymbol{\beta}) && (11)\\
&= P(-\varepsilon < \mathbf{x}^T\boldsymbol{\beta}) && (12)\\
&= P(\varepsilon < \mathbf{x}^T\boldsymbol{\beta}) && (13)\\
&= \sigma(\mathbf{x}^T\boldsymbol{\beta}) && (14)\\
&= \frac{1}{1 + e^{-\mathbf{x}^T\boldsymbol{\beta}}} && (15)
\end{aligned}
$$

where we used in (4) -(5) that the logistic distribution is symmetric (and continuous), as learned above,

$$\Pr(-\varepsilon \leq x) = \Pr(\varepsilon \leq x).$$

Part 4: Logistic Regression Learning and Inference

- *Maximum Likelihood*

Probit analysis

A textbook in biostatistics provides us with the following example: Make and female moths, 20 of both, are administered with various doses of *trans-cypermethrin* in order to examine the lethality of the insecticide. After three days it was registered how many moths were dead or not mobile. We look at only male moths, and model by logistic regression the effect of the dose on the proportion of moths that die or become immobile.

Model validation: the χ^2-test

Let us return to the moth data. We can write the data for males as

	Dose (μg)					
	1	2	4	8	16	32
Die	1	4	9	13	18	20
Survive	19	16	11	7	8	0

A difficulty

We look at only male moths, and model by logistic regression the effect of the dose on the proportion of moths that die or become immobile. This looks straightforward. But:

A difficulty

All twenty male moths were dead or immobile in three days after a dose of 32 μg.

$$\text{logit}(p_i) = \alpha + \beta \cdot \text{dose}_i$$

How do we handle the infinite odds at 32 μg dose ?

A difficulty & a solution due to Laplace

We have infinite odds at 32 μg dose, if we use the estimate

$$\ln\left(\frac{s_i/n_i}{(n_i - s_i)/n_i}\right)$$

where s_is are the frequencies in the table and $n_i = (20)$ is the total number of units. But we can use the adjusted values

$$\ln\left(\frac{s_i + \frac{1}{2}}{n_i - s_i + \frac{1}{2}}\right)$$

Solution with Laplacian adjustment

Maximum likelihood

We have the training set

$$\mathcal{D}_{tr} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

The likelihood function is

$$L(\boldsymbol{\beta}) \stackrel{\text{def}}{=} \prod_{i=1}^{n} \sigma(\mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{x}_i))^{1-y_i}.$$

Maximum likelihood

Some simple manipulation gives that

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

There is no closed form solution to the minimization of $-\ln L(\boldsymbol{\beta})$. The function is twice continuously differentiable, convex and even strictly convex if the data is not linearly separable. There are standard optimization algorithms for minimization of functions with these properties.

Model validation: the χ^2-test

Using the MLE-estimates $\widehat{\alpha} = -1.9277$ and $\widehat{\beta} = 0.2972$ we can calculate the probability of death for the dose $x = 1$ as

$$\frac{1}{1 + e^{1.9277 - 0.2972}} = 0.1638$$

and then the expected frequency of death at $x = 1$ is

$$20 \cdot 0.1638 = 3.275$$

In the same way we can calculate the probabilities of death and survival for the other doses x.

Model validation: the χ^2-test

We use the chi-square goodness-of-fit test statistic Q

$$Q = \sum_{i=1}^{r} \frac{(\text{observed freq}_i - \text{expected freq}_i)^2}{\text{expected freq}_i} = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i} .$$

where r is the number of groups in the grouped data. It can be shown that Q is approximatively $\chi^2(r/2-2)$- distributed (chi square with $r/2-2$ degrees of freedom) under the (null) hypothesis that the probabilities of death and survival are as given by the estimated model. The reduction with two degrees of freedom is for the fact that we have estimated two parameters.

Model validation: the χ^2-test

$$Q = \sum_{i=1}^{r} \frac{(\text{observed freq}_i - \text{expected freq}_i)^2}{\text{expected freq}_i} = \sum_{i=1}^{r} \frac{(x_i - np_i)^2}{np_i}.$$

E.g., $n_2 = 4$ and

$$n \cdot p_2 = 20 \cdot \widehat{P}(Y = 1 \mid x = 2) = \frac{20}{1 + e^{-\widehat{\alpha} - 2\widehat{\beta}}}$$

Model validation: the χ^2-test

We get

$$Q = \sum_{i=1}^{12} \frac{(\text{observed freq}_i - \text{expected freq}_i)^2}{\text{expected freq}_i}$$

$$= \frac{(1 - 3.275)^2}{3.275} + \ldots + \frac{(20 - 19.99)^2}{0.010} = 4.2479$$

The **p-value** is

$$P(Q \geq 4.24) = 0.3755$$

where Q is $\chi^2(6-2)$- distributed. Hence we do not reject the logistic regression model[1].

[1] Here the expected frequency of 0 taken as 0.01 in the textbook cited.

Part III: More on Maximum Likelihood

- *Likelihood function rewritten*
- *Training: an algorithm for computing the Maximum Likelihood Estimate*
- *Linear Separability and Regularization*

The trick applied to rewriting the logistic probability

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Let us recode $y \in \{-1, +1\}$. *Then we get*

$$P\left(Y = y \mid \mathbf{x}^T\right) = \sigma\left(y\mathbf{x}^T\beta\right)$$

Logistic Regression: a check of recoding

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

$$P(Y = 1 \mid \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{x}^T\beta}} = \sigma\left(+1\mathbf{x}^T\beta\right).$$

$$P(Y = 0 \mid \mathbf{x}) = 1 - P(+1 \mid \mathbf{x}) = 1 - \frac{1}{1 + e^{-\mathbf{x}^T\beta}}$$

$$= \frac{1 - 1 + e^{-\mathbf{x}^T\beta}}{1 + e^{-\mathbf{x}^T\beta}}$$

$$= \frac{e^{-\mathbf{x}^T\beta}}{1 + e^{-\mathbf{x}^T\beta}} = \frac{1}{1 + e^{\mathbf{x}^T\beta}} = \sigma\left(-1\mathbf{x}^T\beta\right).$$

Logistic Regression: Likelihood of β

$$P(y \mid \mathbf{x}; \beta) = \sigma\left(y\mathbf{x}^T\beta\right)$$

A training set

$$\mathcal{D}_{tr} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

where now $y_i \in \{-1, 1\}$. The likelihood function of β is

$$L(\beta) \stackrel{\text{def}}{=} \prod_{i=1}^{n} P(y_i \mid \mathbf{x}_i; \beta)$$

Logistic Regression: -log likelihood of β

The negative log likelihood

$$-l(\beta) \stackrel{\text{def}}{=} -\ln L(\beta) =$$

$$= \sum_{i=1}^{n} -\ln P(y_i \mid \mathbf{x}_i; \beta)$$

$$= \sum_{i=1}^{n} -\ln \sigma\left(y_i \mathbf{x}_i^T \beta\right)$$

$$= \sum_{i=1}^{n} -\ln \left[\frac{1}{1 + e^{-y_i \mathbf{x}_i^T \beta}}\right]$$

$$= \sum_{i=1}^{n} \ln \left[1 + e^{-y_i \mathbf{x}_i^T \beta}\right]$$

Logistic Regression: MLE (1)

$$-l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln\left[1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\beta}}\right]$$

Let us recall that

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)\,.$$

Then

$$\frac{\partial}{\partial \beta_0} \ln\left[1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\beta}}\right] = -y_i \frac{e^{-y_i(\mathbf{x}_i^T \boldsymbol{\beta})}}{1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\beta}}}$$

$$= -y_i \sigma(-y_i \mathbf{x}_i^T \boldsymbol{\beta}) = -y_i \left(1 - P\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right)\right)$$

Logistic Regression: MLE (2)

$$-l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln\left[1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\beta}}\right]$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)\,.$$

$$\frac{\partial}{\partial \beta_k} \ln\left[1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\beta}}\right] = -y_i \mathbf{x}_i^T \frac{e^{-y_i\left(\mathbf{x}_i^T \boldsymbol{\beta}\right)}}{1 + e^{-y_i \mathbf{x}_i^T \boldsymbol{\beta}}}$$

$$= -y_i \mathbf{x}_i^T \sigma(-y_i \mathbf{x}_i^T \boldsymbol{\beta}) = -y_i \mathbf{x}_i^T \left(1 - P\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right)\right)$$

Logistic Regression: MLE (3)

$$-l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln\left[1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i^T}\right]$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$.

$$\frac{\partial}{\partial \beta_0} \ln\left[1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i^T}\right] = -y_i \left(1 - P(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})\right)$$

$$\frac{\partial}{\partial \beta_k} \ln\left[1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i^T}\right] = -y_i \mathbf{x}_i^T \left(1 - P(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})\right)$$

Logistic Regression: MLE (3) Update

$$\frac{\partial}{\partial \beta_0} \ln\left[1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i^T}\right] = -y_i \left(1 - P\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right)\right)$$

$$\frac{\partial}{\partial \beta_k} \ln\left[1 + e^{-y_i \boldsymbol{\beta}^T \mathbf{x}_i^T}\right] = -y_i \mathbf{x}_i^T \left(1 - P\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right)\right)$$

Parameters can then be updated by selecting training samples at random and moving the parameters in the opposite direction of the partial derivatives (stochastic gradient descent algorithm).

Logistic Regression: MLE (4) Update

Parameters can then be updated by selecting training samples at random and moving the parameters in the opposite direction of the partial derivatives

$$\beta_0 \leftarrow \beta_0 + \eta y_i \left(1 - P\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right)\right)$$

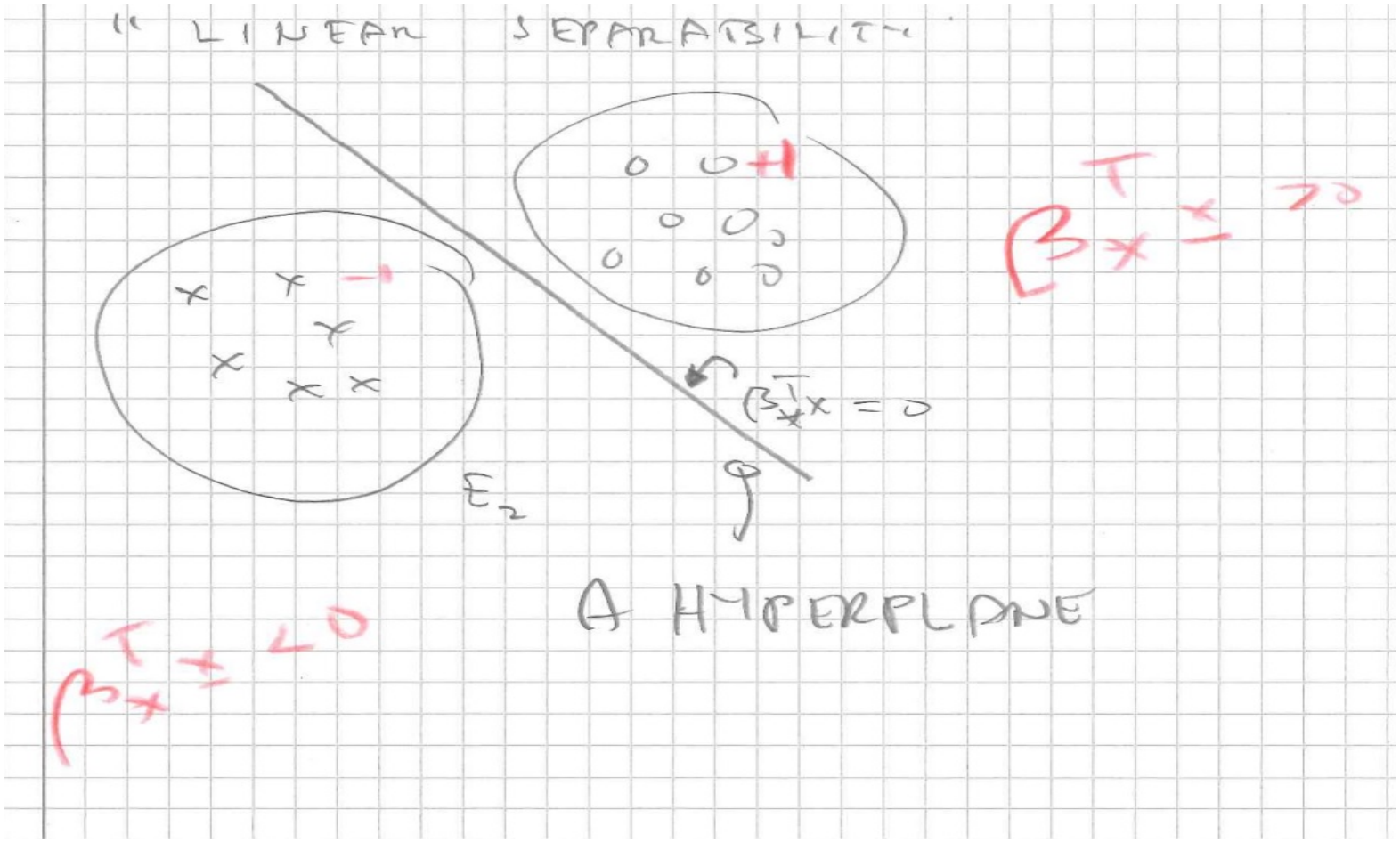
$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} + \eta y_i \mathbf{x}_i \left(1 - P\left(y_i \mid \mathbf{x}_i; \boldsymbol{\beta}\right)\right)$$

Logistic Regression: MLE (5) Update

$$\beta_0 \leftarrow \beta_0 + \eta y_i \left(1 - P\left(y_i \mid \mathbf{x}_i; \beta\right)\right)$$

$$\beta \leftarrow \beta + \eta y_i \mathbf{X}_i \left(1 - P\left(y_i \mid \mathbf{x}_i; \beta\right)\right).$$

A Difficulty of MLE



Separability

For a training set $\mathcal{D}_{tr} = (\mathbf{x}_i, y_i)_{i=1}^n$ set

$$E_+ := \{i | y_i = +1\}, E_- := \{i | y_i = -1\}.$$

Suppose there exists a β_s such that

$$\begin{aligned} \mathbf{x}_i^T \beta_s > 0 \qquad & \text{if } i \in E_+ \\ \mathbf{x}_i^T \beta_s < 0 \qquad & \text{if } i \in E_- \end{aligned} \tag{16}$$

We say that the hyperplane $\mathbf{x}_i^T \beta_s = 0$ separates linearly the training set.

A Difficulty of MLE

Then the loglikelihood from the above

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln \sigma\left(y_i \mathbf{x}_i^T \boldsymbol{\beta}\right) = \sum_{i \in E_+} \ln \sigma\left(y_i \mathbf{x}_i^T \boldsymbol{\beta}\right) + \sum_{i \in E_-} \ln \sigma\left(y_i \mathbf{x}_i^T \boldsymbol{\beta}\right).$$

$$= \sum_{i \in E_+} \ln \left[\frac{1}{1 + e^{-\mathbf{x}_i^T \boldsymbol{\beta}_s}}\right] + \sum_{i \in E_-} \ln \left[\frac{1}{1 + e^{\mathbf{x}_i^T \boldsymbol{\beta}}}\right]$$

A Difficulty of MLE

Take $\lambda > 0$ and consider $l(\lambda\beta_s)$. The terms with $i \in E_+$ are

$$\ln\left[\frac{1}{1+e^{-\lambda \mathbf{x}_i^T \beta_s}}\right].$$

. Let now $\lambda \to +\infty$. Then

$$\ln\left[\frac{1}{1+e^{-\lambda \mathbf{x}_i^T \beta_s}}\right] \to 0,$$

since $\mathbf{x}_i^T \beta_s > 0$ for $i \in E_+$.

A Difficulty of MLE

$l(\lambda\beta_s)$. When $\lambda \to +\infty$ the terms with $i \in E_-$ converge to zero

$$\ln\left[\frac{1}{1+e^{\lambda \mathbf{x}_i^T \beta_s}}\right] \to 0$$

since $\mathbf{x}_i^T \beta_s < 0$ for $i \in E_-$. Thus the likelihood function

$$L(\lambda\beta_s) = e^{l(\lambda\beta_s)} \to 1,$$

as $\lambda \to +\infty$ But for every β

$$L(\beta) < 1$$

since $L(\beta)$ is a product of probabilities. Hence the MLE does not exist for a linearly separable training set.

MLE & Regularizer

To avoid linear separability due to small training sets we minimize *the regularizer + the negative loglikelihood function* or

$$\frac{\lambda}{2}\beta^T\beta + \sum_{i=1}^{n} \ln\left[1 + e^{-y_i \mathbf{x}_i^T \beta}\right]$$

where λ is a parameter that measures the strength of regularization.

More on MLE

$$-l(\boldsymbol{\beta}) = \sum_{i=1}^{n} -\ln \sigma\left(y_i \mathbf{x}_i^T \boldsymbol{\beta}\right)$$

Then we recall that $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)$. Thus

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{F}(\boldsymbol{\beta}) = \sum_{i=1}^{I} y_i \mathbf{x}_i \sigma(-y_i \mathbf{x}_i^T \boldsymbol{\beta}).$$

This follows by the preceding, or expressing the preceding in vector notation

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{x}^T \boldsymbol{\beta} = \mathbf{x}^T$$

Thus if we set the gradient $\frac{\partial}{\partial \boldsymbol{\beta}} \mathbf{F}(\boldsymbol{\beta}) = \mathbf{0}$ (= a column vector of $p+1$ zeros) we get

$$\sum_{i}^{n} y_i \mathbf{x}_i^T \sigma\left(-y_i \mathbf{x}_i^T \boldsymbol{\beta}\right) = \mathbf{0}$$

The MLE estimate $\widehat{\boldsymbol{\beta}}$ will satisfy

$$\mathbf{0}_{p+1} = \sum_{i=1}^{n} y_i \sigma(-y_i \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}) \mathbf{x}_i$$

$\Leftrightarrow$

$$\mathbf{0}_{p+1} = \sum_{i=1}^{n} y_i (1 - P(y_i \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}^T) \mathbf{x}_i$$

CI

$$\widehat{\beta}_{MLE,i} \pm \lambda_{\alpha/2} \cdot \mathit{stderr}(\widehat{\beta}_{MLE,i})$$

Part : Prediction and Crossvalidation

- *Prediction*
- *Crossvalidation*

When we insert $\widehat{\boldsymbol{\beta}}$ back to $P\left(y \mid \mathbf{x}^T\right)$ we have

$$\widehat{P}\left(y \mid \mathbf{x}\right) = \sigma\left(y\mathbf{x}^T\widehat{\boldsymbol{\beta}}\right)$$

or

$$\widehat{P}\left(Y = 1 \mid \mathbf{x}^T\right) = \sigma\left(\mathbf{x}^T\widehat{\boldsymbol{\beta}}\right)$$

Logistic Regression

We can drop the notations $\widehat{P}$ and $\widehat{\beta}$ for ease of writing. For given $\mathbf{X}$ the task is to maximize $P(y \mid \mathbf{x}) = \sigma\left(y\mathbf{x}^T\widehat{\beta}\right)$. There are only two values $y = \pm 1$ to choose among. There are two cases to consider.

1) $t = \mathbf{x}^T\widehat{\beta} > 0$. Then if $y = +1$, and $y^* = -1$

$$y^*t < 0 < yt \Rightarrow e^{y^*t} < e^{yt} \Rightarrow e^{-yt} < e^{-y^*t}$$

$$\Rightarrow 1 + e^{-yt} < 1 + e^{-y^*t} \Rightarrow \frac{1}{1+e^{-y^*t}} < \frac{1}{1+e^{-yt}}$$

i.e.

$$P(y \mid \mathbf{x}) = \sigma(yt) > \sigma(y^*t) = P(y^* \mid \mathbf{x})$$

2) $t = \mathbf{x}^T\widehat{\boldsymbol{\beta}} < 0$. If $y = +1$, and $y^* = -1$, then

$$yt < y^*t$$

and it follows in the same way as above that

$$P(y^* \mid \mathbf{x}) > P(y \mid \mathbf{x})$$

Hence: the maximum probability is assumed by y that has the same sign as $\mathbf{x}^T\widehat{\boldsymbol{\beta}}$.

Given $\widehat{\boldsymbol{\beta}}$, the best probability predictor of Y, denoted by $\widehat{Y}$, for given $\mathbf{x}$ is

$$\widehat{Y} = \mathbf{sign}\left(\mathbf{x}^T \widehat{\boldsymbol{\beta}}\right)$$

Model validation: Cross-validation

A way to check a model's suitability is to assess the model against a set of data (testing set) that was not used to create the model: this is called **cross-validation**. This is a **holdout** model assessment method.

Cross validation

We have a training set of l pairs $y_i \in \{0, 1\}$ and the corresponding values of the predictors.

$$\mathcal{D}_{tr} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$$

and use this to estimate β by $\widehat{\beta} = \widehat{\beta}(\mathcal{S})$, e.g., by MLE.
We must have another set of data, testing set, of holdout samples

$$\mathcal{D}_{test} = \left\{ \left(\mathbf{x}_1^t, y_1^t \right), \ldots, \left(\mathbf{x}_m^t, y_m^t \right) \right\}$$

Having found $\widehat{\beta}$ we should apply the optimal predictor $\widehat{P}\left(y \mid \mathbf{x}_j^t\right)$ on $\mathcal{D}_{test}$, and compare the prediction to y_j^t for all j. Note that in this $\widehat{\beta} = \widehat{\beta}(\mathcal{S})$

Cross-validation: categories of error

- prediction of -1 when the holdout sample has a -1 (True Negatives, the number of which is TN)
- prediction of -1 when the holdout sample has a 1 (False Negatives, the number of which is FN)
- prediction of 1 when the holdout sample has a -1 (False Positives, the number of which is FP)
- prediction of 1 when the holdout sample has a 1 (True Positives, the number of which is TP)

Evaluation of logistic regression (and other) models

False Positives = FP , True Positives = TP
False Negatives = FN , True Negatives= TN

	$Y = +1$	$Y = -1$
$\widehat{Y} = +1$	TP	FP
$\widehat{Y} = -1$	FN	TN

Cross-validation

One often encounters one or several of the following criteria of evaluation:

- Accuracy $= \frac{TP+TN}{TP+FP+FN+TN}$ =fraction of observations with correct predicted classification
- Precision $=$ PositivePredictiveValue $= \frac{TP}{TP+FP}$ =Fraction of predicted positives that are correct
- Recall $=$ Sensitivity $= \frac{TP}{TP+FN}$ =fraction of observations that are actually 1 with a correct predicted classification
- Specificity $= \frac{TN}{TN+FP}$ =fraction of observations that are actually -1 with a correct predicted classification

Appendix: MLE for Simple Logistic Regression

- *Special case:* $\mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x$.
- *Likelihood*
- *Maximum Likelihood*
- `logisticmle.m`

We consider the model:

$$\mathbf{x}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x.$$

$$P(Y = 1 \mid x) = \sigma(\mathbf{x}^T\boldsymbol{\beta}) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}$$

Notation:

$$P(Y = y \mid x) = \sigma(\mathbf{x}^T\beta)^y(1 - \sigma(\mathbf{x}^T\beta)^{1-y}$$

$$= \begin{cases} \sigma(\mathbf{x}^T\beta) & \text{if } y = 1 \\ 1 - \sigma(\mathbf{x}^T\beta) & \text{if } y = 0 \end{cases}$$

Data $(x_i, y_i)_{i=1}^n$, likelihood function with the notation above

$$L(\beta_0, \beta_1) = P(Y = y_1 \mid x_1) \cdot P(Y = y_2 \mid x_2) \cdots P(Y = y_n \mid x_n)$$

$$= \sigma(\mathbf{x}^T\beta_1)^{y_1}(1 - \sigma(\mathbf{x}^T\beta_1)^{1-y_1} \cdots \sigma(\mathbf{x}^T\beta_n)^{y_n}(1 - \sigma(\mathbf{x}^T\beta_n)^{1-y_n}$$

$$= A \cdot B$$

$$A = \sigma(\mathbf{x}^T\beta_1)^{y_1} \cdot \sigma(\mathbf{x}^T\beta_2)^{y_2} \cdots \sigma(\mathbf{x}^T\beta_n)^{y_n}$$

$$B = (1 - \sigma(\mathbf{x}^T\beta_1)^{1-y_1}(1 - \sigma(\mathbf{x}^T\beta_2)^{1-y_2} \cdots (1 - \sigma(\mathbf{x}^T\beta_n)^{1-y_n}$$

$$\ln L(\beta_0, \beta_1) = \ln A + \ln B$$

$$= \sum_{i=1}^{n} y_i \ln \sigma(\mathbf{x}^T \boldsymbol{\beta}_i) + \sum_{i=1}^{n} (1 - y_i) \ln(1 - \sigma(\mathbf{x}^T \boldsymbol{\beta}_i))$$

$$= \sum_{i=1}^{n} \underbrace{\ln(1 - \sigma(\mathbf{x}^T \boldsymbol{\beta}_i))}_{= -\ln\left(1 + e^{\beta_0 + \beta_1 x_i}\right)} + \sum_{i=1}^{n} y_i \underbrace{\ln \frac{\sigma(\mathbf{x}^T \boldsymbol{\beta}_i)}{1 - \sigma(\mathbf{x}^T \boldsymbol{\beta}_i)}}_{= \beta_0 + \beta_1 x_i}$$

In summary:

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln \left(1 + e^{\beta_0 + \beta_1 x_i}\right)$$

$$\frac{\partial}{\partial \beta_1} \ln L(\beta_0, \beta_1)$$

$$= \sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} \ln\left(1 + e^{\beta_0 + \beta_1 x_i}\right)$$

$$\frac{\partial}{\partial \beta_1} \ln\left(e^{\beta_0 + \beta_1 x_i}\right) = \frac{e^{\beta_0 + \beta_1 x_i} x_i}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$= P(Y = 1 \mid x_i) \cdot x_i.$$

$$\frac{\partial}{\partial \beta_1} \ln L(\beta_0, \beta_1) = 0$$

$$\Leftrightarrow$$

$$\sum_{i=1}^{n} (y_i x_i - P(Y = 1 \mid x_i) \cdot x_i) = 0.$$

In the same manner we can also find

$$\frac{\partial}{\partial \beta_0} \ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - P(Y = 1 \mid x_i)) = 0$$

These two equations have no closed form solution w.r.t. β_0 and β_1.

Newton-Raphson for MLE: a one-parameter case

For one parameter θ, set $f(\theta) = \ln L(\theta)$. We are searching for the solution of

$$f'(\theta) = \frac{d}{d\theta} f(\theta) = 0.$$

Newton-Raphson method

$$\theta^{new} = \theta^{old} + \frac{f'\left(\theta^{old}\right)}{f''\left(\theta^{old}\right)},$$

where a good initial value is desired.

Newton-Raphson for logistic MLE

$$\begin{pmatrix} \beta_0^{new} \\ \beta_1^{new} \end{pmatrix} = \begin{pmatrix} \beta_0^{old} \\ \beta_1^{old} \end{pmatrix} + H^{-1}(\beta_0^{old}, \beta_1^{old}) \begin{pmatrix} \frac{\partial}{\partial \beta_0} \ln L\left(\beta_0^{old}, \beta_1^{old}\right) \\ \frac{\partial}{\partial \beta_1} \ln L\left(\beta_0^{old}, \beta_1^{old}\right) \end{pmatrix}.$$

where $H^{-1}(\beta_0^{old}, \beta_1^{old})$ is the matrix inverse of the 2×2 matrix (next slide)

Newton-Raphson for logistic MLE

$$H(\beta_0^{old}, \beta_1^{old}) = \begin{pmatrix} \frac{\partial^2}{\partial \beta_0^2} \ln L\left(\beta_0^{old}, \beta_1^{old}\right) & \frac{\partial^2}{\partial \beta_0 \beta_1} \ln L\left(\beta_0^{old}, \beta_1^{old}\right) \\ \frac{\partial^2}{\partial \beta_1 \beta_0} \ln L\left(\beta_0^{old}, \beta_1^{old}\right) & \frac{\partial^2}{\partial \beta_1^2} \ln L\left(\beta_0^{old}, \beta_1^{old}\right) \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^n P\left(Y = 1 \mid x_i\right)\left(1 - P(Y = 1 \mid x_i\right) & \sum_{i=1}^n P\left(Y = 1 \mid x_i\right)\left(1 - P\left(Y = 1 \mid x_i\right) \cdot x_i\right) \\ \sum_{i=1}^n P\left(Y = 1 \mid x_i\right)\left(1 - P\left(Y = 1 \mid x_i\right)\right) \cdot x_i & \sum_{i=1}^n P\left(Y = 1 \mid x_i\right)\left(1 - P(Y = 1 \mid x_i)\right) x_i^2 \end{pmatrix}$$

```
[bs,stderr,phat,deviance] =
logisticmle(y,x)
```

Input:
y - responses, a binary vector, values 0 and 1
x - the covariate, as a vector
Output:
bs - estimators of beta0 and beta1
stderr - standard error of the estimate $=$ square roots of the diagonal elements of H^{-1}.
phat - estimator of p= P(Y=1)
deviance - deviance

Confidence Interval with degree of confidence $1-\alpha$

$$\widehat{\beta}_{MLE,i} \pm \lambda_{\alpha/2} \cdot \mathit{stderr}(\widehat{\beta}_{MLE,i})$$