Auxiliary Slides on Logistic Regression for SF2930 Regression analysis

Timo Koski, KTH Royal Institute of Technology

22.02.2017

KTH Matematik

These slides have been edited from a material in another context, and therefore some details of notation e.t.c. do not conform to those used in the guest lecture in SF2930 on the 22nd of Feb..

Your Learning Outcomes

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

Let $\mathbf{x}_i \in \mathcal{X} \subseteq R^p$, $y_i \in \mathcal{Y} = \{+1, -1\}$[1], $\mathbf{X} = (X_1, \ldots, X_p)^T$, a vector. Y is r.v. with two values, -1 and 1.

We consider the problem of modelling the probability

$$P(Y = 1 \mid \mathbf{X}).$$

The model will be called *logistic regression*.

[1] We could use $\mathcal{Y} = \{+1, 0\}$

Modelling of

$$P(Y = 1 \mid \mathbf{X})$$

This is known as the **discriminative approach**.

- The **generative approach**: Model $P(\mathbf{X} \mid Y = 1)$ and $P(Y = 1)$ and use Bayes' formula to find $P(Y = 1 \mid \mathbf{X})$.
- The discriminative approach models $P(\mathbf{X} \mid Y = 1)$ and $P(Y = 1)$ and implicitly.

- Generative: Y $\rightarrow$ Data
- Discriminative: Data $\rightarrow$ Y

KTH Matematik

- *Odds, Odds Ratio*
- *Logit function*
- *Logistic function a.k.a Sigmoid function*

Odds

The **odds** of a statement (event e.t.c) A is calculated as the probability $p(A)$ of observing A divided by the probability of not observing A:

$$\text{odds of A} = \frac{p(A)}{1 - p(A)}$$

E.g., in humans an average of 51 boys are born in every 1000 births, the odds of a randomly chosen delivery being boy are:

$$\text{odds of a boy} = \frac{0.51}{0.49} = 1.04$$

The odds of a certain thing happening are infinite.

Odds ratio

The **odds ratio** ψ is the ratio of odds from two different conditions or populations.

$$\psi = \frac{\text{odds of } A_1}{\text{odds of } A_2} = \frac{\frac{p(A_1)}{1-p(A_1)}}{\frac{p(A_2)}{1-p(A_2)}}$$

Bernoulli Distribution again

	$y = 1$	$y = 0$
$f(y)$	p	$1 - p$

We write this as

$$f(y) = p^y (1 - p)^{1-y}$$

and

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

The logarithmic odds of success is called the logit of p

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

$$f(y) = e^{\ln\left(\frac{p}{1-p}\right)y + \ln(1-p)} = e^{\mathrm{logit}(p)y + \ln(1-p)}$$

KTH Matematik

The function logit(p) and its inverse

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

If $\theta = \mathrm{logit}(p)$, then the inverse function is

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

KTH Matematik

The logit(p) and its inverse

$$\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right), \quad 0 < p < 1$$

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

The function

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

is called the **logistic function** or **sigmoid**.

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

KTH Matematik

In biology the logistic function refers to change in size of a species population. In artifical neural networks it is a network output function (sigmoid). In statistics it is the 'canonical link function' for the Bernoulli distribution (c.f. above).

The logit(p) and the logistic function

Sats

The logit function

$$\theta = \mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right), \quad 0 < p < 1$$

and the logistic function

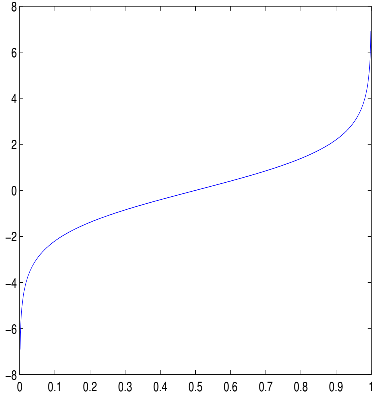
$$p = \sigma(\theta) = \frac{1}{1+e^{-\theta}}, \quad -\infty < \theta < \infty,$$

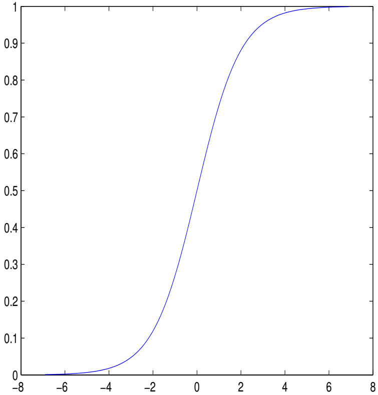
are inverse functions to each other.

KTH Matematik

The logit(p) and the logistic function

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





Part II: Logistic Regression

- *A regression function*
- *log odds of* $Y \leftarrow$ *a regression function*
- *How to generate* Y*, logistic noise*

KTH Matematik

Let $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ be $(p+1) \times 1$ vector and $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)$ be a $(p+1) \times 1$ -vector of (predictor) variables. We set, as in multiple regression,

$$\boldsymbol{\beta}^T \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Then

$$G(\mathbf{X}) = \sigma(\boldsymbol{\beta}^T \mathbf{X}) = \frac{1}{1 + e^{-\beta^T \mathbf{X}}} = \frac{e^{\beta^T \mathbf{X}}}{1 + e^{\beta^T \mathbf{X}}}$$

KTH Matematik

The predictor variables $(X_1, X_2, \ldots, X_p)$ can be binary, ordinal, categorical or continuous.

$$G(\mathbf{X}) = \sigma(\boldsymbol{\beta}^T \mathbf{X}) = \frac{e^{\boldsymbol{\beta}^T \mathbf{X}}}{1 + e^{\boldsymbol{\beta}^T \mathbf{X}}}$$

By construction $0 < G(\mathbf{X}) < 1$. Then logit is well defined and

$$\mathrm{logit}(G(\mathbf{X})) = \ln \frac{G(\mathbf{X})}{1 - G(\mathbf{X})} = \boldsymbol{\beta}^T \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

Let now Y be a binary random variable such that

$$Y = \begin{cases} 1 & \text{with probability } G(\mathbf{X}) \\ -1 & \text{with probability } 1 - G(\mathbf{X}) \end{cases}$$

Definition

If the logit of $G(\mathbf{X})$ (or log odds of Y) is

$$\operatorname{logit}(G(\mathbf{X})) = \ln \frac{G(\mathbf{X})}{1 - G(\mathbf{X})} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p,$$

then we say that Y follows a logistic regression w.r.t. the predictor variables $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)$.

Logistic regression

For fixed $\boldsymbol{\beta}$ we have the hyperplane

$$D(\boldsymbol{\beta}) = \{\mathbf{X} \mid \boldsymbol{\beta}^T \mathbf{X} = 0\}$$

For any $\mathbf{X} \in D(\boldsymbol{\beta})$, $G(\mathbf{X}) = \sigma(0) = \frac{1}{2}$ and

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

An Aside: Deciban

We might use (suggestion by Alan Turing for logarithm of posterior odds)

$$\mathrm{e}(Y|\mathbf{X}) = 10 \log_{10} \frac{G(\mathbf{X})}{1 - G(\mathbf{X})}$$

and call it the evidence for Y given $\mathbf{X}$. Turing called this 'deciban'. The unit of evidence is then *decibel* (db). 1 db change in evidence is the smallest increment in of plausibility that is perceptible for our intuition.

KTH Matematik

Logistic regression

$$Y = \begin{cases} \text{success} & \text{with probability } G(\mathbf{X}) \\ \text{failure} & \text{with probability } 1 - G(\mathbf{X}) \end{cases}$$

$$\mathrm{logit}(G(\mathbf{X})) = \ln \frac{G(\mathbf{X})}{1 - G(\mathbf{X})} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p.$$

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 $X_i = x_1$ in first population and $X_i = x_2$ 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

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

$$= \beta_i \left(x_1 - x_2 \right)$$

or

$$\psi = e^{\beta_i (x_1 - x_2)}$$

KTH Matematik

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

$$\widehat{\text{logit}} = -4.8326 + 1.0324x$$

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

$$\psi = e^{\beta(x_1 - x_2)} = e^{1.3024(1-0)} = 2.808$$

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

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

KTH Matematik

Logistic regression

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

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

$$= \frac{1}{1 + e^{\beta^T \mathbf{X}}}.$$

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

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:

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

KTH Matematik

A generative model

We need the following regression model

$$Y^* = \boldsymbol{\beta}^T \mathbf{X} + \epsilon$$

where $\boldsymbol{\beta}^T \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ and

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

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.

KTH VETENSKAP OCH KONST

KTH Matematik

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:

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

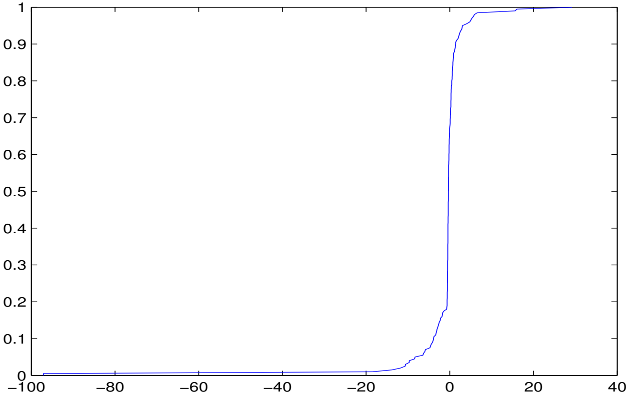
I.e. $\epsilon \sim \text{Logistic}(0, 1)$, *if the probability density function is*

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

KTH Matematik

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

This is simple: simulate $p_1, \ldots, p_n$ from the uniform distribution on $(0, 1)$ and then do $\epsilon_i = \text{logit}(p_i)$, $i = 1, \ldots, n$. 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 $P(-\epsilon \leq x)$?

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

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

KTH Matematik

Generating model and/or how to simulate

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

$$Y^* = \boldsymbol{\beta}^T \mathbf{X} + \epsilon$$

where $\boldsymbol{\beta}^T \mathbf{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ and

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

KTH Matematik

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

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

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

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

KTH Matematik

$$
\begin{aligned}
P(Y = 1 \mid \mathbf{X}) &= P(Y^* > 0 \mid \mathbf{X}) && (1)\\
&= P(\boldsymbol{\beta}^T \mathbf{X} + \varepsilon > 0) && (2)\\
&= P(\varepsilon > -\boldsymbol{\beta}^T \mathbf{X}) && (3)\\
&= P(-\varepsilon < \boldsymbol{\beta}^T \mathbf{X}) && (4)\\
&= P(\varepsilon < \boldsymbol{\beta}^T \ \mathbf{X}) && (5)\\
&= \sigma(\boldsymbol{\beta}^T \mathbf{X}) && (6)\\
&= \frac{1}{1 + e^{-\boldsymbol{\beta}^T \mathbf{X}}} && (7)
\end{aligned}
$$

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

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

KTH VETENSKAP OCH KONST

KTH Matematik

- *Some Probit Analysis*
- *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.

various doses of *trans-cypermethrin*

amine the lethality of the insecticide. After three days it was

v many moths were dead or immobilized. Data are shown in

v:

Sex	Dose (μg) 1	2	4	8	16	32
Males	1	4	9	13	18	20
Females	0	2	6	10	12	16

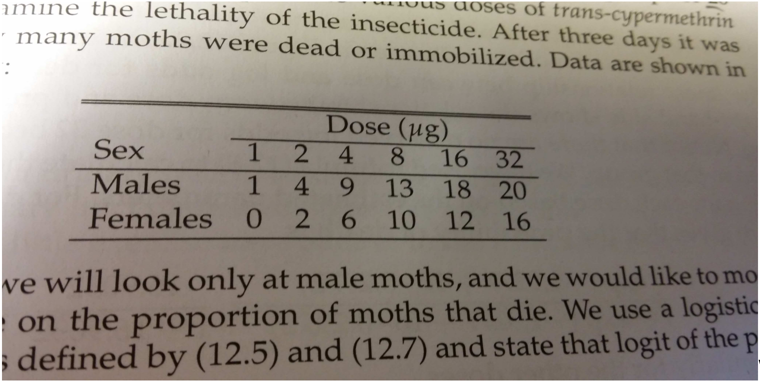
we will look only at male moths, and we would like to mo

e on the proportion of moths that die. We use a logisti

s defined by (12.5) and (12.7) and state that logit of the p

We

A difficulty



amine the lethality of the insecticide. After three days it was many moths were dead or immobilized. Data are shown in :

Sex	Dose (μg)					
	1	2	4	8	16	32
Males	1	4	9	13	18	20
Females	0	2	6	10	12	16

we will look only at male moths, and we would like to mo on the proportion of moths that die. We use a logistic defined by (12.5) and (12.7) and state that logit of the p

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.

amine the lethality of the insecticide. ...various doses of *trans-cypermethrin* many moths were dead or immobilized. After three days it was : Data are shown in

Sex	Dose (μg)					
	1	2	4	8	16	32
Males	1	4	9	13	18	20
Females	0	2	6	10	12	16

we will look only at male moths, and we would like to mo on the proportion of moths that die. We use a logistic s defined by (12.5) and (12.7) and state that logit of the p

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

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

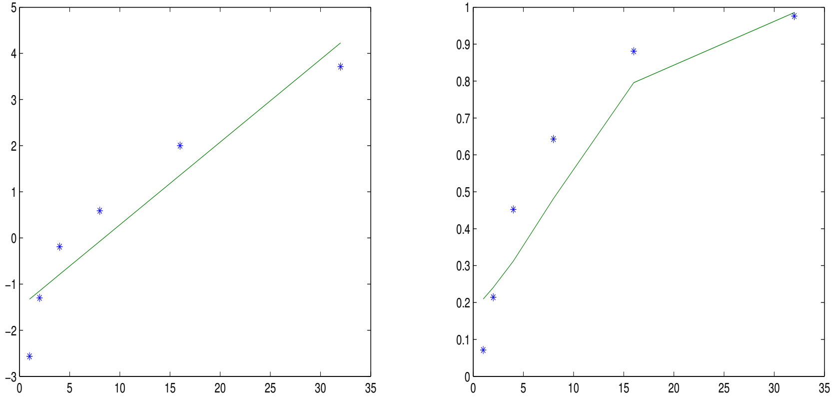
$$\ln\left(\frac{s_i\,/\,n_i}{(n_i - s_i)\,/\,n_i}\right)$$

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

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

KTH Matematik

Solution with Laplacian adjustment



We have data

$$\mathcal{S} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$$

Next the labels are coded as $+1 \mapsto +1, -1 \mapsto 0$. The likelihood function is

$$L(\beta) \stackrel{\text{def}}{=} \prod_{i=1}^{l} G(\mathbf{X}_i)^{y_i} (1 - G(\mathbf{X}_i))^{1-y_i}.$$

KTH Matematik

Maximum likelihood

Some simple manipulation gives that

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

There is no closed form solution to the minimization of $- \ln L(\beta)$. 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

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

Model validation: the χ^2-test

Using the ML-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.

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

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

KTH Matematik

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

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

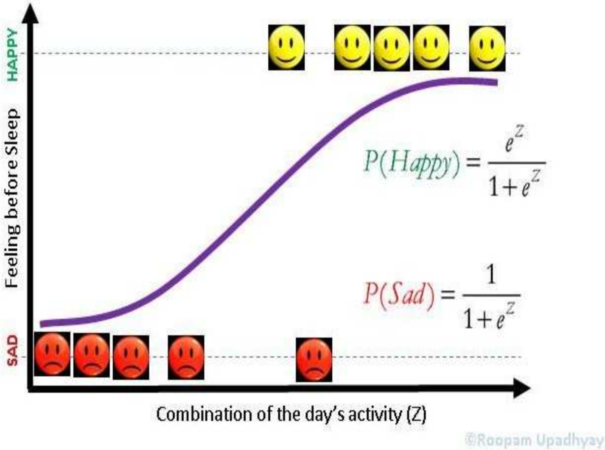
where Q is $\chi^2(6-2)$- distributed. Hence we do not reject the logistic regression model[2].

[2] 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*

KTH VETENSKAP OCH KONST
KTH Matematik


HAPPY
Feeling before Sleep
SAD
$P(Happy) = \frac{e^Z}{1+e^Z}$
$P(Sad) = \frac{1}{1+e^Z}$
Combination of the day's activity (Z)
©Roopam Upadhyay

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

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

KTH Matematik

Logistic Regression: a check

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

$y \in \{+1, -1\}$.

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

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

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

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

Logistic Regression: Likelihood of β

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

A training set

$$\mathcal{S} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$$

The likelihood function of β is

$$L(\beta) \stackrel{\text{def}}{=} \prod_{i=1}^{n} P(y_l \mid \mathbf{X}_l; \beta)$$

KTH Matematik

Logistic Regression: -log likelihood of $\boldsymbol{\beta}$

The negative log likelihood

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

$$= \sum_{i=1}^{n} -\ln P\left(y_l \mid \mathbf{X}_l; \boldsymbol{\beta}\right)$$

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

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

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

KTH VETENSKAP OCH KONST

KTH Matematik

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

Let us recall that

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

Then

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

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

Logistic Regression: ML (2)

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

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

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

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

Logistic Regression: ML (3)

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

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

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

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

KTH Matematik

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

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

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 algorithm).

Logistic Regression: ML (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

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

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

Logistic Regression: ML (5) Update

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

ML & Regularizer

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

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

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

More on ML

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

Then we recall that $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)$. Thus

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

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

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

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

$$\sum_{i=1}^{n} y_i \mathbf{X}_i \sigma(-y_i \boldsymbol{\beta}^T \mathbf{X}_i) = \mathbf{0}$$

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

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

$\Leftrightarrow$

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

KTH Matematik

Part IV:

- *Prediction*
- *Crossvalidation*

When we insert $\widehat{\boldsymbol{\beta}}$ back to $P(y \mid \mathbf{X})$ we have

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

or

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

KTH VETENSKAP OCH KONST
KTH Matematik

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

1) $t = \boldsymbol{\beta}^T\mathbf{X} > 0$. 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.

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

KTH Matematik

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

$$yt < y^* t$$

and it follows in the same way as above that

$$P\left(y^* \mid \mathbf{X}\right) > P\left(y \mid \mathbf{X}\right)$$

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

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

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

KTH Matematik

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.

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

$$\mathcal{S} = \{(\mathbf{X}_1, y_1), \ldots, (\mathbf{X}_n, y_n)\}$$

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

$$\mathcal{T} = \left\{ \left(\mathbf{X}_1^t, y_1^t\right), \ldots, \left(\mathbf{X}_m^t, y_m^t\right) \right\}$$

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

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)

KTH Matematik

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