

Applied Logistic Regression

Week 1

1. Introduction to logistic regression I
2. Introduction to logistic regression II
3. Fitting the logistic model (likelihood function)
4. Maximum likelihood estimation
5. Examples using statistical software to fit a logistic regression model
 - SYSTAT
 - STATA
5. Homework

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

GOAL: To find the best fitting, simplest, model possible describing the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables.

← or “covariates”.

What distinguishes a logistic regression model from the linear regression model is that the outcome variable is binary (or dichotomous).

The techniques used in linear regression analysis will provide the motivation for our approach to logistic regression.

Example:

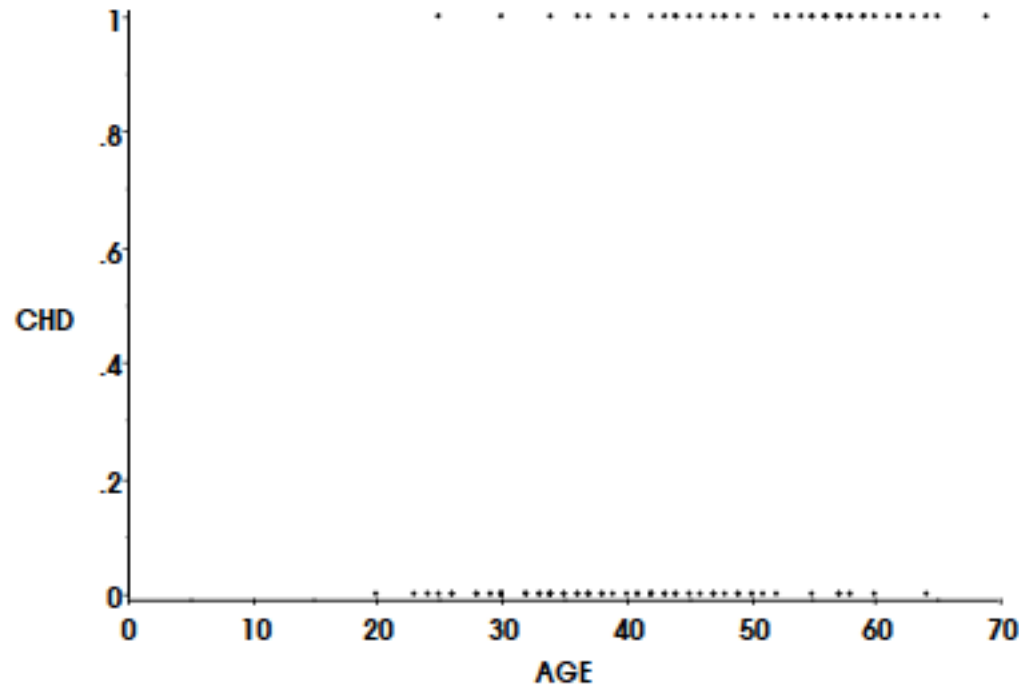
AGE (yrs) and presence or absence of evidence of significant coronary heart disease (CHD) for 100 subjects selected to participate in a study.

ID	AGE	CHD	ID	AGE	CHD	ID	AGE	CHD	ID	AGE	CHD
1	20	0	26	35	0	51	44	1	76	55	1
2	23	0	27	35	0	52	44	1	77	56	1
3	24	0	28	36	0	53	45	0	78	56	1
4	25	0	29	36	1	54	45	1	79	56	1
5	25	1	30	36	0	55	46	0	80	57	0
6	26	0	31	37	0	56	46	1	81	57	0
7	26	0	32	37	1	57	47	0	82	57	1
8	28	0	33	37	0	58	47	0	83	57	1
9	28	0	34	38	0	59	47	1	84	57	1
10	29	0	35	38	0	60	48	0	85	57	1
11	30	0	36	39	0	61	48	1	86	58	0
12	30	0	37	39	1	62	48	1	87	58	1
13	30	0	38	40	0	63	49	0	88	58	1
14	30	0	39	40	1	64	49	0	89	59	1
15	30	0	40	41	0	65	49	1	90	59	1
16	30	1	41	41	0	66	50	0	91	60	0
17	32	0	42	42	0	67	50	1	92	60	1
18	32	0	43	42	0	68	51	0	93	61	1
19	33	0	44	42	0	69	52	0	94	62	1
20	33	0	45	42	1	70	52	1	95	62	1
21	34	0	46	43	0	71	53	1	96	63	1
22	34	0	47	43	0	72	53	1	97	64	0
23	34	1	48	43	1	73	54	1	98	64	1
24	34	0	49	44	0	74	55	0	99	65	1
25	34	0	50	44	0	75	55	1	100	69	1

Let us explore the relationship between AGE and presence or absence of CHD.

- Had our outcome variable been continuous rather than binary we would probably have begun by creating a scatter plot of the dependent vs. the independent variable.

This plot displays the relationship between x and y.



Clearly, in this scatterplot, all points fall on one of two parallel lines representing $\text{CHD} = 0$ and $\text{CHD} = 1$.

We can see that there is some tendency for the individuals with no evidence of CHD ($y = 0$) to be younger than those with CHD ($y = 1$).

While this plot does depict the dichotomous nature of the outcome variable quite clearly, it does not provide a clear picture of the nature of the relationship between CHD and AGE.

To better explore this relationship let us create intervals for the independent variable and compute the mean of the outcome variable within each group.

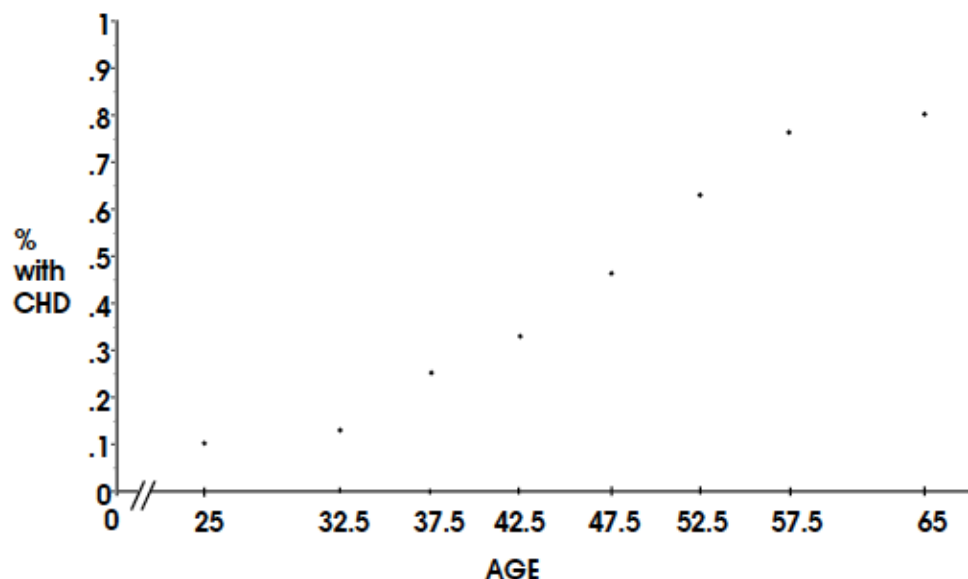
↑
% with CHD = 1

AGE GROUP	n	CHD		MEAN % PRESENT
		ABSENT	PRESENT	
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
	100	57	43	

Here we see that as age increases, the proportion of individuals with evidence of CHD increases.

- However, the *nature* of this relationship is still not clear.

Let us plot the proportion of individuals with CHD vs. the midpoint of each age interval.



Note that with dichotomous response data $0 \leq E(y|x) \leq 1$

- This is certainly not the case when the response variable is continuous as in ordinary linear regression.

Also note that the plot approaches 0 and 1 “gradually”. The change in $E(y|x)$ per unit change in x becomes progressively smaller as the conditional mean gets closer to 0 or 1.

The shape of the curve is said to be “S-shaped”, resembling the plot of a cumulative distribution function of a random var.

Some well known cumulative distributions have been used to provide a model for $E(y|x)$ in the case where y is dichotomous.

The model we will use is the logistic regression model.

- We choose this because
 - (1) from a mathematical point of view, it is an extremely flexible and easily used function and
 - (2) it lends itself to a biologically meaningful interpretation

Let $\pi(x)$ = conditional mean of y given x .

Specifically,

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

A transformation of $\pi(x)$ that will be central to our study of logistic regression is the logit transformation. This is defined as

$$g(x) = \ln \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\}$$

$$\text{but } 1 - \pi(x) = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

Hence,

$$g(x) = \ln \left\{ \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \middle/ \frac{1}{1 + e^{\beta_0 + \beta_1 x}} \right\} = \ln \left\{ e^{\beta_0 + \beta_1 x} \right\} = \beta_0 + \beta_1 x$$

The importance of this transformation is that $g(x)$ has many of the desirable properties of a linear regression model. It may be continuous and is linear in the parameters with the potential for a range between $-\infty$ and $+\infty$ depending on the range of x .

Now, recall that in linear regression our model is

$$y = E(y|x) + \varepsilon$$

constant over x

We assume $\varepsilon \sim N(0, \sigma^2)$. It follows that $y|x \sim N(E(y|x), \sigma_{y|x}^2)$

This is not true with a dichotomous outcome variable.

Here the model is

$y = \pi(x) + \varepsilon$ but ε may assume one of two possible values:

If $y = 1$ then $\varepsilon = 1 - \pi(x)$ with probability $\pi(x)$

and if $y = 0$ then $\varepsilon = -\pi(x)$ with probability $1 - \pi(x)$

$$\therefore E(\varepsilon) = 0$$

$$\text{var}(\varepsilon) = \pi(x)(1 - \pi(x))$$

i.e., $y|x \sim \text{binomial with parameter } \pi(x)$

In linear regression we used the method of least squares to estimate the parameters β_0 and β_1 . These minimize $\sum (y_i - \hat{y}_i)^2$. Under the usual assumptions for linear regression the method of least squares yields estimators with a number of desirable properties.

Unfortunately, when the method of least squares is applied to a model with a dichotomous outcome, the estimators no longer have these same properties.

The general method of estimation that leads to the least squares function under the linear regression model (when the error terms are normally distributed) is called MAXIMUM LIKELIHOOD.

The method of maximum likelihood yields values for the unknown parameters that maximize the probability of obtaining the observed set of data.

With this method we first construct the likelihood function. This gives the probability (or likelihood) of the data for some arbitrary values of the parameters.

We then determine specific values for the parameters that maximize the likelihood function.

Let $y = \begin{cases} 0 \\ 1 \end{cases}$. Then $\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \Pr(y = 1 | x)$.

It follows that $1 - \pi(x) = \frac{1}{1 + e^{\beta_0 + \beta_1 x}} = \Pr(y = 0 | x)$.

For an arbitrary value of $\underline{\beta} = (\beta_0, \beta_1)$, $\Pr(y = 1 | x) = \pi(x)$
 $\Pr(y = 0 | x) = 1 - \pi(x)$

Thus, for those pairs (x_i, y_i) where $y_i = 1$, the contribution to the likelihood function is $\pi(x_i)$, and for those pairs where $y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$.

A convenient way to express the contribution to the likelihood function for the pair (x_i, y_i) is through the term

$$\xi(x_i) = \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i}$$

xi
(pronounced: ksai)

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given above.

i.e.,

$$l(\underline{\beta}) = \prod_{i=1}^n \xi(x_i)$$

In maximum likelihood we use as our estimate of β the value that maximizes $l(\beta)$. However it is easier, mathematically, to maximize the log of $l(\beta)$. This is called the log likelihood, denoted $L(\beta)$.

$$L(\underline{\beta}) = \ln\{l(\underline{\beta})\} = \sum_{i=1}^n \left[y_i \ln\{\pi(x_i)\} + (1 - y_i) \ln\{1 - \pi(x_i)\} \right]$$

To find the value of $\hat{\underline{\beta}}$ that maximizes $L(\underline{\beta})$ we differentiate $L(\underline{\beta})$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations are as follows:

$$\begin{aligned} & \sum_{i=1}^n (y_i - \pi(x_i)) = 0 \\ \text{and} \quad & \sum_{i=1}^n x_i (y_i - \pi(x_i)) = 0 \end{aligned} \quad \left. \vphantom{\sum_{i=1}^n} \right\} \text{likelihood equations}$$

In linear regression the likelihood equations are obtained by differentiating the sum of squared deviations function wrt β_0 and β_1 . These equations are linear in the unknown parameters and thus are easily solved.

For logistic regression the likelihood equations are non-linear in β_0 and β_1 and require special methods for their solution.

These methods are iterative in nature and have been programmed into available logistic regression software.

We let $\hat{\underline{\beta}}$ denote the maximum likelihood estimate of $\underline{\beta}$.

$\hat{\pi}(x_i)$ denotes the maximum likelihood estimate of $\pi(x_i)$.

← estimates the conditional probability that $y = 1$ given $x = x_i$.

Note that

$$\sum_{i=1}^n y_i = \sum_{i=1}^n \hat{\pi}(x_i)$$

Let us now consider the AGE/CHD data. Use of a logistic regression program produces the following output:


Term	$\hat{\beta}_i$	$\widehat{SE}(\hat{\beta}_i)$	$\hat{\beta}_i / \widehat{SE}(\hat{\beta}_i)$
AGE	0.11092	0.02046	4.610
CONSTANT	-5.30950	1.13400	-4.683

log likelihood = -53.677

From the output we also know that:

	CHD	
	<u>YES</u>	<u>NO</u>
$n =$	43	57
\bar{x}_{age}	51.28	39.18

Hence $\hat{\beta}_0 = -5.30950$ and $\hat{\beta}_1 = 0.11092$

fitted values are given by $\hat{\pi}(x) = \frac{e^{-5.31+0.11x}}{1+e^{-5.31+0.11x}}$ 

SYSTAT

```
>format=5
>model chd=constant+age
>loptions means
>estimate
```

```
=====
BINARY LOGIT ANALYSIS
=====
```

DEPENDENT VARIABLE: CHD

INPUT RECORDS: 100

SAMPLE SPLIT
=====

CATEGORY CHOICES

RESP		43
REF		57
-----+		
		100

INDEPENDENT VARIABLE MEANS
=====

PARAMETER		1	0	OVERALL
-----		-----		
1	CONSTANT	1.00000	1.00000	1.00000
2	AGE	51.27907	39.17544	44.38000
-----		-----		

```

L-L AT ITER      1 IS      -69.31472
L-L AT ITER      2 IS      -54.24651
L-L AT ITER      3 IS      -53.68311
L-L AT ITER      4 IS      -53.67655

```

CONVERGENCE ACHIEVED

RESULTS OF ESTIMATION

=====

LOG LIKELIHOOD: -53.67655

PARAMETER	ESTIMATE	S.E.	T-RATIO	P-VALUE
1 CONSTANT	-5.30945	1.13344	-4.68439	0.00000
2 AGE	0.11092	0.02406	4.61104	0.00000

PARAMETER	ODDS RATIO	95.0% BOUNDS	
		UPPER	LOWER
2 AGE	1.11731	1.17125	1.06585

```

LOG LIKELIHOOD OF CONSTANTS ONLY MODEL = LL(0) = -68.33149
2*[LL(N)-LL(0)] = 29.30989 WITH 1 DOF, CHI-SQ P-VALUE = 0.00000
MCFADDEN'S RHO-SQUARED = 0.21447

```

STATA

```
. use chdage.dta
```

```
. sum AGE CHD
```

Variable	Obs	Mean	Std. Dev.	Min	Max
AGE	100	44.38	11.72133	20	69
CHD	100	.43	.4975699	0	1

```
. tabulate CHD, summarize( AGE)
```

CHD	Summary of AGE		
	Mean	Std. Dev.	Freq.
0	39.175439	10.201755	57
1	51.27907	9.9793253	43
Total	44.38	11.721327	100

. logit CHD AGE

Iteration 0: log likelihood = -68.331491
Iteration 1: log likelihood = -54.170558
Iteration 2: log likelihood = -53.681645
Iteration 3: log likelihood = -53.676547
Iteration 4: log likelihood = -53.676546

Logit estimates

Number of obs = 100
LR chi2(1) = 29.31
Prob > chi2 = 0.0000
Pseudo R2 = 0.2145

Log likelihood = -53.676546

CHD	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
AGE	.1109211	.0240598	4.610	0.000	.0637647	.1580776
_cons	-5.309453	1.133655	-4.683	0.000	-7.531376	-3.087531

. vce

	AGE	_cons
AGE	.000579	
_cons	-.026677	1.28517

```
. logistic CHD AGE
```

Logit estimates

```
Number of obs   =      100
LR chi2(1)       =      29.31
Prob > chi2      =      0.0000
Pseudo R2       =      0.2145
```

Log likelihood = -53.676546

CHD	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
AGE	1.117307	.0268822	4.610	0.000	1.065842	1.171257

The estimated logit, $\hat{g}(x)$, is given by

$$\hat{g}(x) = -5.31 + 0.11 \times AGE$$

The log-likelihood = -53.677 is the value of $L(\underline{\beta})$ computed using $\hat{\beta}_0$ and $\hat{\beta}_1$.