# Logistic Regression Applications

Hu Lunchao

# Contents

# An Example

**Table 1.1 Age and Coronary Heart Disease (CHD) Status of 100 Subjects**

| ID | AGE | AGRP | CHD | ID | AGE | AGRP | CHD |
|---|---|---|---|---|---|---|---|
| 1 | 20 | 1 | 0 | 51 | 44 | 4 | 1 |
| 2 | 23 | 1 | 0 | 52 | 44 | 4 | 1 |
| 3 | 24 | 1 | 0 | 53 | 45 | 5 | 0 |
| 4 | 25 | 1 | 0 | 54 | 45 | 5 | 1 |
| 5 | 25 | 1 | 1 | 55 | 46 | 5 | 0 |
| 6 | 26 | 1 | 0 | 56 | 46 | 5 | 1 |
| 7 | 26 | 1 | 0 | 57 | 47 | 5 | 0 |
| 8 | 28 | 1 | 0 | 58 | 47 | 5 | 0 |
| 9 | 28 | 1 | 0 | 59 | 47 | 5 | 1 |
| 10 | 29 | 1 | 0 | 60 | 48 | 5 | 0 |
| 11 | 30 | 2 | 0 | 61 | 48 | 5 | 1 |
| 12 | 30 | 2 | 0 | 62 | 48 | 5 | 1 |
| 13 | 30 | 2 | 0 | 63 | 49 | 5 | 0 |
| 14 | 30 | 2 | 0 | 64 | 49 | 5 | 0 |
| 15 | 30 | 2 | 0 | 65 | 49 | 5 | 1 |
| 16 | 30 | 2 | 1 | 66 | 50 | 6 | 0 |
| 17 | 32 | 2 | 0 | 67 | 50 | 6 | 1 |
| 18 | 32 | 2 | 0 | 68 | 51 | 6 | 0 |
| 19 | 33 | 2 | 0 | 69 | 52 | 6 | 0 |
| 20 | 33 | 2 | 0 | 70 | 52 | 6 | 1 |
| 21 | 34 | 2 | 0 | 71 | 53 | 6 | 1 |
| 22 | 34 | 2 | 0 | 72 | 53 | 6 | 1 |
| 23 | 34 | 2 | 1 | 73 | 54 | 6 | 1 |
| 24 | 34 | 2 | 0 | 74 | 55 | 7 | 0 |
| 25 | 34 | 2 | 0 | 75 | 55 | 7 | 1 |
| 26 | 35 | 3 | 0 | 76 | 55 | 7 | 1 |
| 27 | 35 | 3 | 0 | 77 | 56 | 7 | 1 |

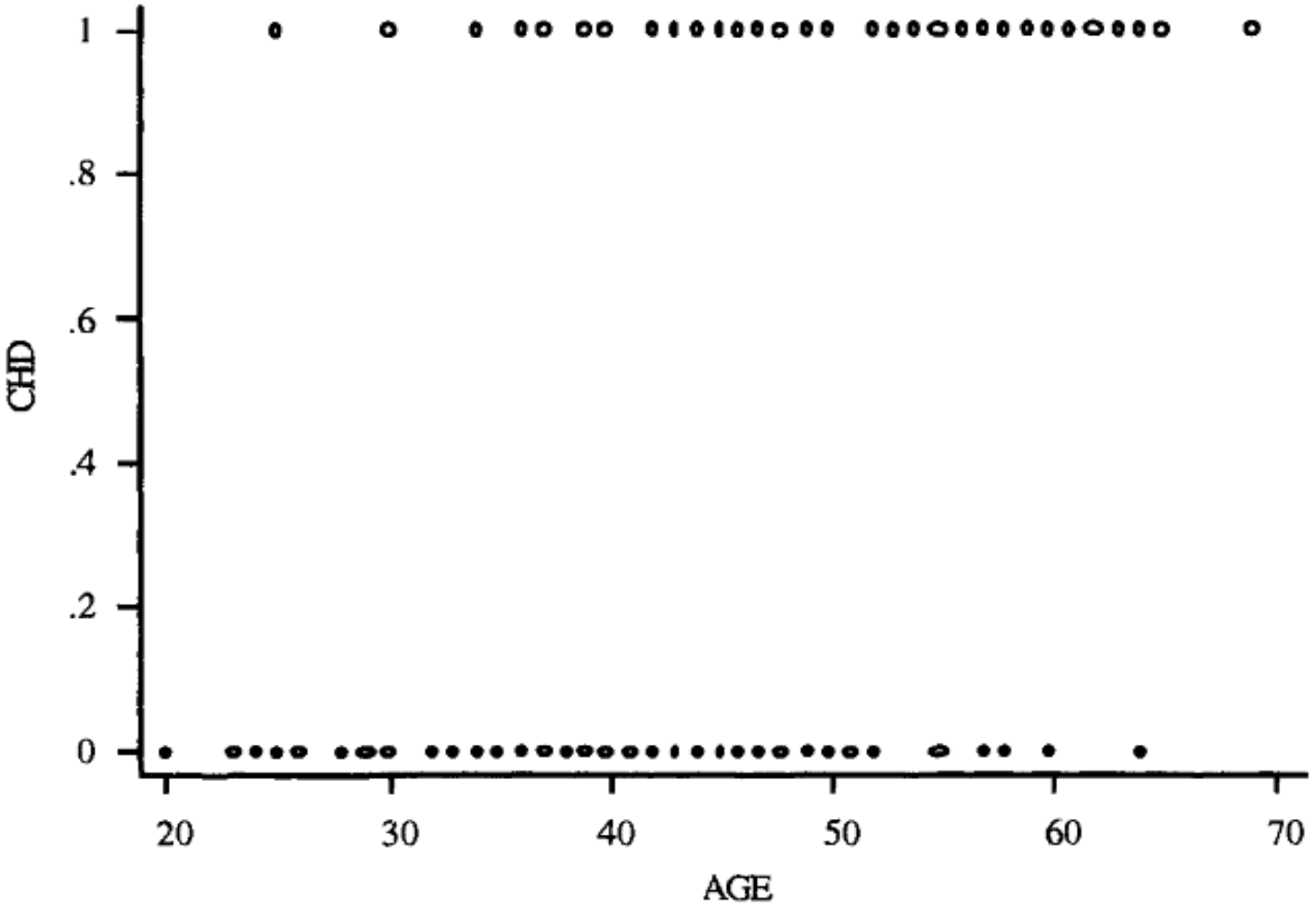



**Figure 1.1** Scatterplot of CHD by AGE for 100 subjects.

# An Example

**Table 1.2 Frequency Table of Age Group by CHD**

| | | CHD | | |
|---|---|---|---|---|
| Age Group | $n$ | Absent | Present | Mean (Proportion) |
| 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 |
| Total | 100 | 57 | 43 | 0.43 |

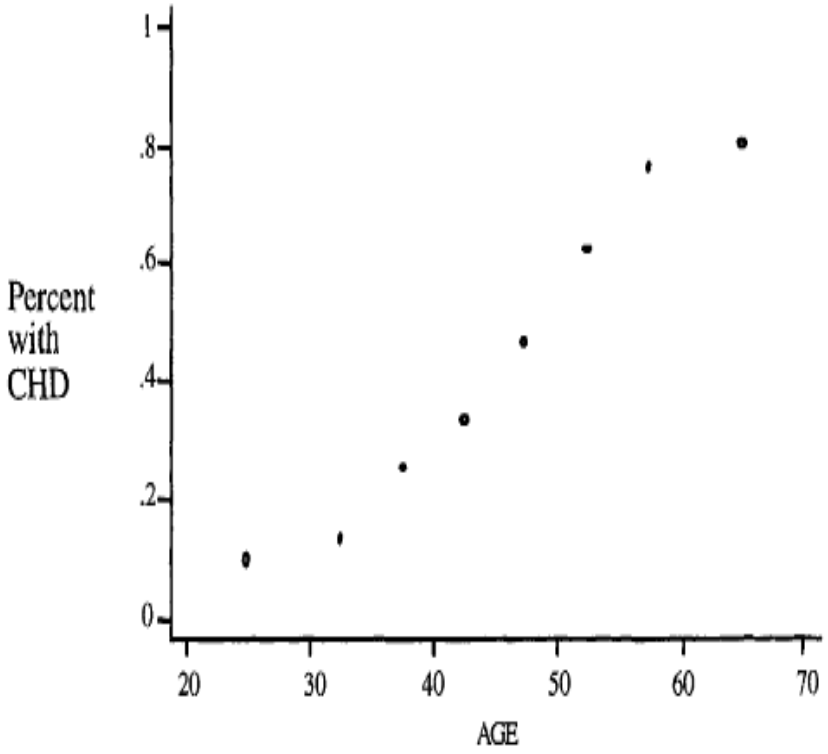



**Figure 1.2** Plot of the percentage of subjects with CHD in each age group.

$$E(Y \mid x) = \beta_0 + \beta_1 x.$$

# What Is Logistic Regression?

Logistic Function:

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

Logit transformation:

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

The variance:

$$y = \pi(x) + \varepsilon$$

$$\varepsilon \sim B(0, \pi(x))$$

# Modeling Categorical Responses

Logistic regression model for binary response variables. For instance, logistic regression can model

- A voter's choice in a presidential election , with predictor variables political ideology, annual income, education level, and religious affiliation .

- Whether a person uses illegal drugs (yes or no), with predictors education level, whether employed, religiosity, marital status, and annual income.

# Modeling Categorical Responses

## Income and Having Travel Credit Cards

Table 15.1: Annual Income (in Thousands of Euros) and Possessing a Travel Credit Card. For example, of the 5 subjects with income 30 thousand euros, 2 possessed a travel credit card.

| Income | Number Cases | Credit Cards | Income | Number Cases | Credit Cards | Income | Number Cases | Credi Cards |
|---|---|---|---|---|---|---|---|---|
| 12 | 1 | 0 | 21 | 2 | 0 | 34 | 3 | 3 |
| 13 | 1 | 0 | 22 | 1 | 1 | 35 | 5 | 3 |
| 14 | 8 | 2 | 24 | 2 | 0 | 39 | 1 | 0 |
| 15 | 14 | 2 | 25 | 10 | 2 | 40 | 1 | 0 |
| 16 | 9 | 0 | 26 | 1 | 0 | 42 | 1 | 0 |
| 17 | 8 | 2 | 29 | 1 | 0 | 47 | 1 | 0 |
| 19 | 5 | 1 | 30 | 5 | 2 | 60 | 6 | 6 |
| 20 | 7 | 0 | 32 | 6 | 6 | 65 | 1 | 1 |

*Source*: Thanks to R. Piccarreta, Bocconi Univ., Milan. The data were originally recorded in Italian lira but have been converted to euros.

# Modeling Categorical Responses

Let
x = annual income and
y = whether have a travel credit card (1 = yes, 0 = no).

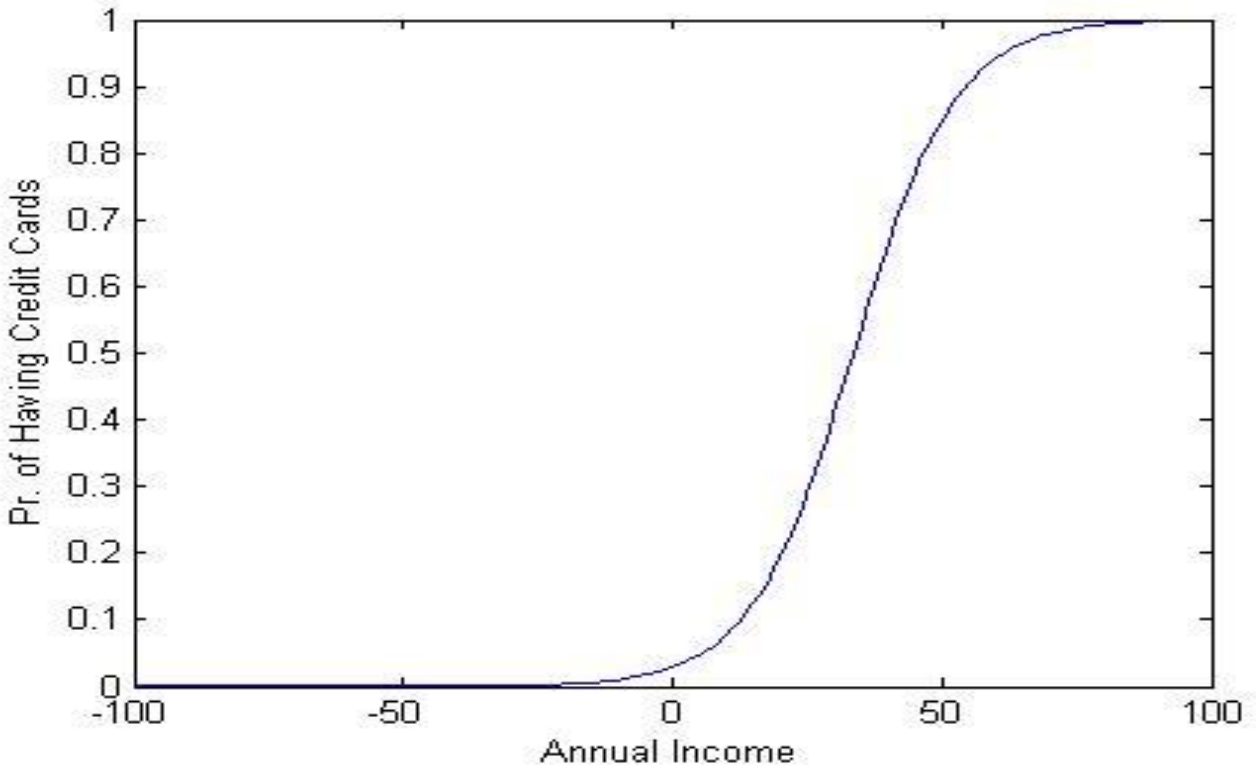
Software provides results shown in Table

| Fit of Logistic Regression Model for Italian Credit Card Data | | | |
|---|---|---|---|
| | beta | S.E. | Exp(beta) |
| income | .1054 | .0262 | 1.111 |
| Constant | -3.5179 | .7103 | |

# Modeling Categorical Responses

The logistic prediction equation is

$$\log it[\hat{P}(y=1)] = -3.518 + 0.105x$$



# Modeling Categorical Responses

$$P(y=1)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad P(y=1)=\frac{e^{-3.52+0.105x}}{1+e^{-3.52+0.105x}}$$

Interpretation for beta:

- The sign of beta tells us whether it is increasing or decreasing as x increases ;

- The curve has slope beta*P(y = 1)[1-P(y = 1)], where P(y = 1) is the probability at that point.

- Interpretation Using the Odds and Odds Ratio

$$\frac{P(y=1)}{1-P(y=1)} = e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x}$$

This exponential relationship implies that every unit increase in x has a multiplicative effect of $e^{beta}$ on the odds.

# Modeling for Ordinal Variables

Many applications have a categorical response variable with more than two categories.
For instance, the General Social Survey recently asked subjects whether government spending on the environment should increase, remain the same, or decrease

# Cumulative Probabilities

Let y denote an ordinal response variable. Let P(y <=j) denote the probability that the response falls in category j or below.
With four categories, for example, the cumulative probabilities are

$$P(y=1), P(y \le 2) = P(y=1) + P(y=2),$$

$$P(y \le 3) = P(y=1) + P(y=2) + P(y=3)$$

and $P(y \le 4)=1$

# Cumulative Logits

A c-category response has c cumulative probabilities

$$P(y \le 1) \le P(y \le 2) \le \cdots \le P(y \le c)$$

With c = 4, for example, the logits are

$$\log it[P(y \le 1)] = \log[\frac{P(y=1)}{P(y>1)}] = \log[\frac{P(y=1)}{P(y=2)+P(y=3)+P(y=4)}]$$

$$\log it[P(\le 2)] = \log[\frac{P(y \le 2)}{P(y>2)}] = \log[\frac{P(y=1)+P(y=2)}{P(y=3)+P(y=4)}]$$

$$\log it[P(\le 3)] = \log[\frac{P(y \le 3)}{P(y>3)}] = \log[\frac{P(y=1)+P(y=2)+P(y=3)}{P(y=4)}]$$

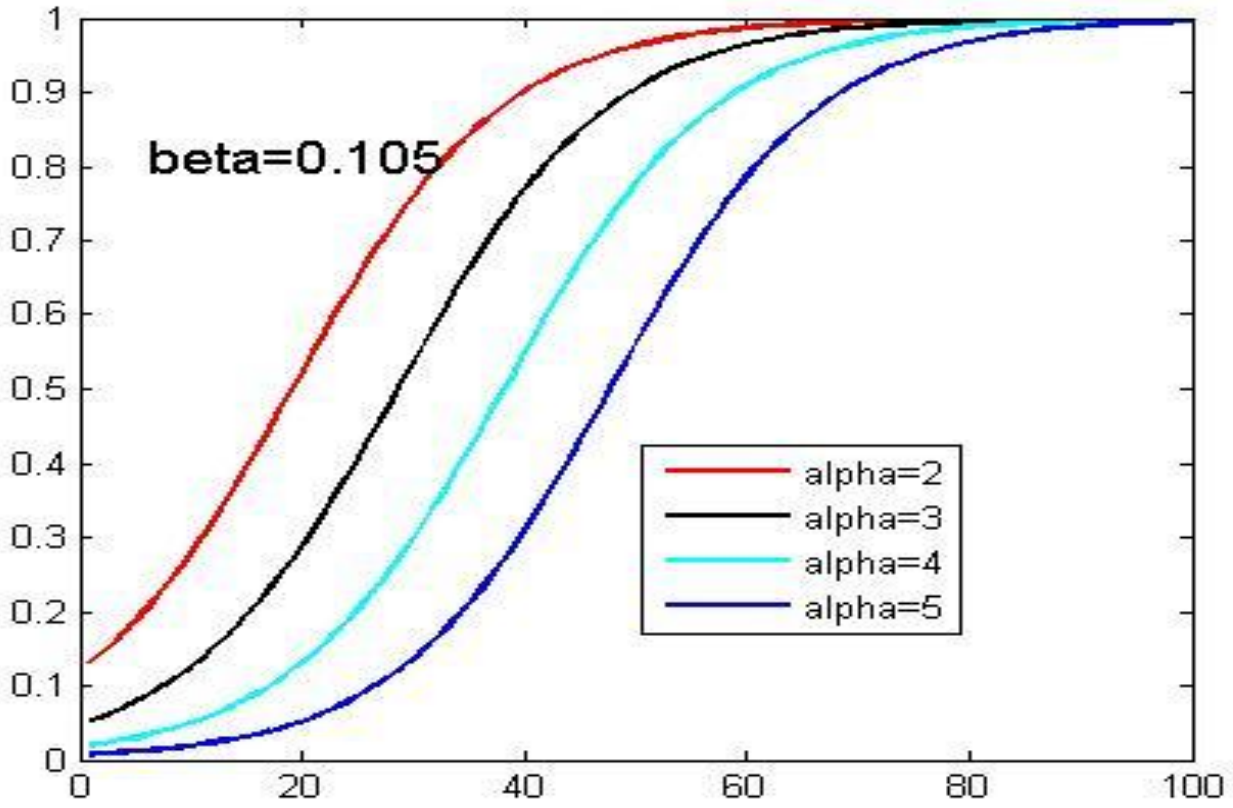
$$\log it[P(\le 4)] = 0$$

# Cumulative Logit Models for an Ordinal Response

$$\log it[P(y \le j)] = \alpha_j - \beta x, \quad j = 1, 2, \cdots, c-1$$

- The model requires a separate intercept parameter j for each cumulative probability. Since the cumulative probabilities increase as j increases, so do $\{\alpha_j\}$.
- The effect of an explanatory variable on all the cumulative probabilities for y should be the same, so beta keeps unchanged.

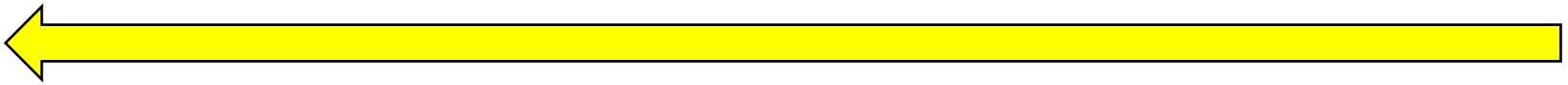
Software estimates the parameters using all the cumulative probabilities at once.



# Do Republicans tend to be more conservative than Democrats?

Table 15.7: Political Ideology by Party Affiliation

| Party Affiliation | Political Ideology | | | | |
|---|---|---|---|---|---|
| | Very Liberal | Slightly Liberal | Moderate | Slightly Conservative | Very Conservative |
| Democratic | 29 | 17 | 36 | 4 | 5 |
| Republican | 2 | 7 | 23 | 23 | 19 |



Let x be a dummy variable for party affiliation, with x = 1 for Democrats and x = 0 for Republicans.

$$\log it[P(y \le j)] = \alpha_j - \beta x, \quad j = 1, 2, \cdots, c-1$$

Table 15.8: Printout for Cumulative Logit Model Fitted to Table 15.7

| | Estimate | Std. Error | Wald | df | Sig. |
|---|---|---|---|---|---|
| INTERCP1 [ideology=1] | -3.051 | .336 | 82.30 | 1 | .000 |
| INTERCP2 [ideology=2] | -2.150 | .300 | 51.28 | 1 | .000 |
| INTERCP3 [ideology=3] | -.183 | .226 | .66 | 1 | .418 |
| INTERCP4 [ideology=4] | .983 | .250 | 15.41 | 1 | .000 |
| PARTY | -2.241 | .338 | 44.05 | 1 | .000 |

Democrats tending to be more liberal than Republicans

# Matched Case-Control Studies

- Case-control studies are used to identify factors that may contribute to a medical condition by comparing subjects who have that condition (the 'cases') with patients who do not have the condition but are otherwise similar.
- Within each stratum, samples of cases(y=1) and controls(y=0) are chosen. The number of cases and controls need not be constant across strata, but the most common matched designs include one case and form one to five controls per stratum and are thus referred to as 1-M matched studies.

In a case–control study, the logistic regression outcome variable *Y is binary; here Y = 1* for cases and *Y = 0 for controls. Then, it is written as:*

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

# ML Approaches to Estimate beta: unconditional and conditional

- The unconditional likelihood is

$$L_U = \prod_{j=1}^{n} \frac{e^{(\beta_0+\beta_1 x_j)y_j}}{[1+e^{(\beta_0+\beta_1 x_j)}]^{y_j}} [1+e^{(\beta_0+\beta_1 x_j)}]^{y_j-1}$$

where *n is the total number of individuals.*

# ML Approaches to Estimate beta: unconditional and conditional

- The conditional likelihood is

$$L_C = \frac{\prod_{j=1}^{m} e^{\beta_1 x_j}}{\sum \prod_{j=1}^{m} e^{\beta_1 x_j}}$$

where *m is the number of cases, and the sum in the denominator is over the* $C_m^n$ possibilities of dividing the numbers from 1 to *n into one group* $\{1, 2, \ldots, m\}$ *of* size *m and its complement* $\{l_{m+1}, \ldots, l_n\}$.

# What Factors Attribute to Low-birth-weight

Dataset：189 women, of whom 59 had low-birth -weight babies and 130 had normal-weight babies

Risk factors:

weight in pounds at the last menstrual period (LWT),

presence of hypertension (HT),

smoking status during pregnancy (Smoke),

presence of uterine irritability (UI).

The woman’s age (Age),used as the matching variable.

# Data Input

```
data LBW;
   input id Age Low LWT Smoke HT UI @@;
   Time=2-Low;
   datalines;
 25  16   1   130   0  0  0    143  16   0   110   0  0  0
166  16   0   112   0  0  0    167  16   0   135   1  0  0
189  16   0   135   1  0  0    206  16   0   170   0  0  0
216  16   0    95   0  0  0     37  17   1   130   1  0  1
 45  17   1   110   1  0  0     68  17   1   120   1  0  0
```

The variable Low is used to determine whether the subject is a case (Low=1, low-birth-weight baby) or a control (Low=0, normal-weight baby). The dummy time variable Time takes the value 1 for cases and 2 for controls

# Conditional logistic regression Result

The variable Time is the response, and Low is the censoring variable. The matching variable Age is used in the STRATA statement so that each unique age value defines a stratum.

| Summary of the Number of Event and Censored Values | | | | | |
|---|---|---|---|---|---|
| Stratum | Age | Total | Event | Censored | Percent Censored |
| 1 | 18 | 7 | 1 | 6 | 85.71 |
| 2 | 17 | 12 | 5 | 7 | 58.33 |
| 3 | 18 | 10 | 2 | 8 | 80.00 |
| 4 | 19 | 16 | 3 | 13 | 81.25 |
| 5 | 20 | 18 | 8 | 10 | 55.56 |
| 6 | 21 | 12 | 5 | 7 | 58.33 |
| 7 | 22 | 13 | 2 | 11 | 84.62 |
| 8 | 23 | 13 | 5 | 8 | 61.54 |
| 9 | 24 | 13 | 5 | 8 | 61.54 |
| 10 | 25 | 15 | 6 | 9 | 60.00 |
| 11 | 26 | 8 | 4 | 4 | 50.00 |
| 12 | 27 | 3 | 2 | 1 | 33.33 |
| 13 | 28 | 9 | 2 | 7 | 77.78 |
| 14 | 29 | 7 | 1 | 6 | 85.71 |
| 15 | 30 | 7 | 1 | 6 | 85.71 |
| 16 | 31 | 5 | 1 | 4 | 80.00 |
| 17 | 32 | 6 | 1 | 5 | 83.33 |
| Total | | 174 | 54 | 120 | 68.97 |

| Convergence Status |
|---|
| Convergence criterion (GCONV=1E-8) satisfied. |

| Model Fit Statistics | | |
|---|---|---|
| Criterion | Without Covariates | With Covariates |
| -2 LOG L | 159.069 | 141.108 |
| AIC | 159.069 | 149.108 |
| SBC | 159.069 | 157.084 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 17.9613 | 4 | 0.0013 |
| Score | 17.3152 | 4 | 0.0017 |
| Wald | 15.5577 | 4 | 0.0037 |

| Analysis of Maximum Likelihood Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Parameter | DF | Parameter Estimate | Standard Error | Chi-Square | Pr > ChiSq | Hazard Ratio |
| LWT | 1 | -0.01498 | 0.00706 | 4.5001 | 0.0339 | 0.985 |
| Smoke | 1 | 0.80805 | 0.36797 | 4.8221 | 0.0281 | 2.244 |
| HT | 1 | 1.75143 | 0.73932 | 5.6120 | 0.0178 | 5.763 |
| UI | 1 | 0.88341 | 0.48032 | 3.3827 | 0.0659 | 2.419 |

Results indicate that women were more likely to have low-birth-weight babies if they were underweight in the last menstrual cycle, were hypertensive, smoked during pregnancy, or suffered uterine irritability

# THANKS FOR YOUR ATTENTION!