Finansiell Statistik, GN, 15 hp, VT2008
Lecture 17-1: Regression with dichotomous outcome variable - Logistic Regression

Gebrenegus Ghilagaber, PhD, Associate Professor

May 7, 2008
1 Introduction

Let $Y$ be a binary (0, 1) variable defined as

$$Y = \begin{cases} 
1 & \text{if an individual has experienced the event of interest} \\
& \text{(has the characteristic of interest)} \\
0, & \text{otherwise}
\end{cases}$$

Let $p = P(Y = 1)$, be the probability that a randomly selected individual has the characteristic of interest. Our goal is to model this probability $p$, as a function of one or more explanatory variables such as Education, Residence, etc. These variables are usually denoted as $Z_1, Z_2, \ldots$ etc.
Investigators may be tempted to consider using an ordinary linear regression model:

\[ p = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 \]  

(1)

where \( \beta_0 \) is the constant (intercept), \( \beta_1 \) is the effect of \( Z_1 \), \( \beta_2 \) is the effect of \( Z_2 \), and \( \beta_3 \) is the effect of \( Z_3 \). The parameters to be estimated (\( \beta_0, \beta_1, \beta_2 \) and \( \beta_3 \)) may then be obtained by simple regression procedures, such as Least Squares Method. If \( Z_1, Z_2, \) and \( Z_3 \) are also dichotomous variables taking values 0 and 1 (for instance for Uneducated and Educated women), then the above equation reduces to

\[ p = \beta_0 + \beta_1 + \beta_2 + \beta_3 \]  

(2)
Do we see any problem in the equation

\[ p = \beta_0 + \beta_1 + \beta_2 + \beta_3 \quad ???
\]

However, the right-hand side of equations (1) or (2) may give largely positive or largely negative values depending on the estimates. In other words it may be less than 0 or greater than 1. However, we know that our dependent variable, \( p \), is a probability and as such it cannot assume a value outside the range (0, 1).
2 Solution: Logistic Regression

To solve, the above paradox we model, instead, the logit of the probability as a function of the explanatory variables. The logit of $p$ is defined as

$$\log it(p) = \ln \left( \frac{p}{1-p} \right), \quad (4)$$

where $\ln(p)$ is the natural logarithm of $p$. The logistic regression model is then given by

$$\log it(p) = \ln \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 Z_1 + \ldots + \beta_k Z_k = \sum_{i=0}^{k} \beta_i Z_i \quad (5)$$
If we simplify the above equation, we get

\[
\frac{p}{1 - p} = \exp \left[ \sum_{i=0}^{k} \beta_i Z_i \right]
\]  

(6)

\[\implies p = (1 - p) \exp \left[ \sum_{i=0}^{k} \beta_i Z_i \right]
\]  

(7)

\[= \exp \left[ \sum_{i=0}^{k} \beta_i Z_i \right] - p \exp \left[ \sum_{i=0}^{k} \beta_i Z_i \right]
\]  

(8)

\[\implies (1 + p) \left[ \sum_{i=0}^{k} \beta_i Z_i \right] = \left[ \sum_{i=0}^{k} \beta_i Z_i \right]
\]  

(9)
or that

\[
p = \frac{\exp \left( \sum_{i=0}^{k} \beta_i Z_i \right)}{1 + \exp \left( \sum_{i=0}^{k} \beta_i Z_i \right)} = \frac{1}{1 + \exp \left( - \sum_{i=0}^{k} \beta_i Z_i \right)}
\]  

From equation (10) we are guaranteed that \( p \) lies in the interval \( (0, 1) \), as it should, because the denominator is always greater than the numerator (since the exponential function cannot be negative).

- Looking at (5) we note that the estimated coefficients (the \( \beta_i \)) are the effects on the log-odds \( \ln \left( \frac{p}{1-p} \right) \).

- Accordingly, \( \exp(\beta_i) \) are interpreted as the odds of having the event of interest \( \left[ \frac{p_1}{1-p_1} \right] \) for individuals with certain characteristics indexed by 1.
(say, the Educated) relative to the odds of having the event of interest \[
\frac{p_0}{1-p_0}
\] of a reference (baseline) group of individuals (indexed by 0, say the uneducated):

\[
\exp(\beta_i) = \text{Relative Odds Ratio} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \frac{p_1}{p_0} \frac{1-p_0}{1-p_1}
\] (11)

- From these coefficients, it is possible to compute the corresponding probabilities and, hence, ratio of the probabilities corresponding to various categories. The corresponding ratios are called the Odds-Ratio. Standard Programs like MINITAB, SPSS, yield the estimates \((\beta_i)\), the Relative Odds Ratio \(\frac{p_1}{p_0} \frac{1-p_0}{1-p_1}\), and the 95% confidence intervals for such ratio.
2.1 Logistic Regression Model When the Independent Variable is Dichotomous

<table>
<thead>
<tr>
<th>Outcome Variable y</th>
<th>Independent Variable x</th>
<th>$x = 1$</th>
<th>$x = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = 1$</td>
<td>$p_1 = \frac{\exp[\beta_0 + \beta_1]}{1+\exp[\beta_0 + \beta_1]}$</td>
<td>$p_0 = \frac{\exp[\beta_0]}{1+\exp[\beta_0]}$</td>
<td></td>
</tr>
<tr>
<td>$y = 0$</td>
<td>$1 - p_1 = \frac{1}{1+\exp[\beta_0 + \beta_1]}$</td>
<td>$1 - p_0 = \frac{1}{1+\exp[\beta_0]}$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- The Odds of the outcome being present among those individuals with $x = 1$ is defined as
The Odds of the outcome being present among those individuals with $x = 0$ is defined as

$$\frac{p_1}{1 - p_1}$$  \hspace{1cm} (12)

- The Odds of the outcome being present among those individuals with $x = 0$ is defined as

$$\frac{p_0}{1 - p_0}$$  \hspace{1cm} (13)

- The log of the Odds is known as the logit and is defined as
The Odds Ratio, denoted by \( \psi \), is defined as the ratio of the odds for \( x = 1 \) to that of the odds for \( x = 0 \):

\[
\psi = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}
\]

The log of the Odds Ratio, termed as log-odds ratio or simply log-odds is
\[ \ln \psi = \ln \left( \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} \right) = \ln \left( \frac{p_1}{1-p_1} \right) - \ln \left( \frac{p_0}{1-p_0} \right) = \log it(p|x = 1) - \log it(p|x = 0) \]

(17)

which is the logit-difference.

- Now using the expressions in above Table, the Odds Ratio, is

\[ \psi = \left\{ \left[ \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)} \right] \right\} \left\{ \left[ \frac{1}{1 + \exp(\beta_0 + \beta_1)} \right] \right\}^{-1} \left\{ \left[ \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \right] \right\} = \frac{\exp [\beta_0 + \beta_1]}{\exp [\beta_0]} = \exp [\beta_1] \]

(18)
• Thus, for logistic regression with a dichotomous independent variable the odds ratio is $\psi = \exp[\beta_1]$ and the logit-difference, or log-odds is $\ln(\psi) = \ln \{\exp[\beta_1]\} = \beta_1$.

• This fact concerning the interpretability of the coefficients is the fundamental reason why logistic regression has proven such a powerful analytic tool.

• The Odds Ratio $\psi = \exp[\beta_1]$ is a measure of the association which has found wide use as it approximates how much more likely (or unlikely) it is for the outcome to be present among those with $x = 1$ than among those with $x = 0$. 

13
• For example, if \( y \) denotes the presence (\( y = 1 \)) or absence (\( y = 0 \)) of own firm and \( x \) denotes whether a woman is educated (\( x = 1 \)) or not (\( x = 0 \)), then an estimated \( \psi = \exp[\beta_1] = 2 \) indicates that ownership of firm is twice as frequent among the educated than among the educated in the study population.

• Similarly, a value of \( \psi = \exp[\beta_1] = 0.5 \) would mean that the frequency of the event of interest (firm ownership) among the educated is just half as that among the uneducated.
3 Example

MINITAB: Stat—Regression—Binary Logistic Regression

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Coef</th>
<th>SE</th>
<th>Z</th>
<th>P</th>
<th>Odds Ratio</th>
<th>95% CI Lower</th>
<th>95% CI Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.117601</td>
<td>1.02277</td>
<td>-0.11</td>
<td>0.908</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>-0.204174</td>
<td>0.420078</td>
<td>-0.49</td>
<td>0.627</td>
<td>0.82</td>
<td>0.36</td>
<td>1.86</td>
</tr>
<tr>
<td>C3</td>
<td>-0.0101876</td>
<td>0.0323412</td>
<td>-0.32</td>
<td>0.753</td>
<td>0.99</td>
<td>0.93</td>
<td>1.05</td>
</tr>
</tbody>
</table>
4 Polytomous Logistic Regression (outcome variable with more than 2 levels)