

The Logit Model: Estimation, Testing and Interpretation

Herman J. Bierens

October 25, 2008

1 Introduction to maximum likelihood estimation

1.1 The likelihood function

Consider a random sample Y_1, \dots, Y_n from the Bernoulli distribution:

$$\begin{aligned}\Pr[Y_j = 1] &= p_0 \\ \Pr[Y_j = 0] &= 1 - p_0,\end{aligned}$$

where p_0 is unknown. For example, toss n times a coin for which you suspect that it is unfair: $p_0 \neq 0.5$, and for each tossing j assign $Y_j = 1$ if the outcome is heads and $Y_j = 0$ if the outcome is tails. The question is how to estimate p_0 and how to test the null hypothesis that the coin is fair: $p_0 = 0.5$.

The probability function involved can be written as

$$\begin{aligned}f(y|p_0) &= \Pr[Y_j = y] \\ &= p_0^y (1 - p_0)^{1-y} = \begin{cases} p_0 & \text{if } y = 1, \\ 1 - p_0 & \text{if } y = 0. \end{cases}\end{aligned}$$

Next, let y_1, \dots, y_n be a given sequence of zeros and ones. Thus, each y_j is either 0 or 1. The joint probability function of the random sample Y_1, Y_2, \dots, Y_n is defined as

$$f_n(y_1, \dots, y_n|p_0) = \Pr[Y_1 = y_1 \text{ and } Y_2 = y_2 \dots \text{ and } Y_n = y_n].$$

Because the random variables Y_1, Y_2, \dots, Y_n are independent, we can write

$$\begin{aligned}
 \Pr[Y_1 = y_1 \text{ and } Y_2 = y_2 \dots \text{ and } Y_n = y_n] &= \Pr[Y_1 = y_1] \times \Pr[Y_2 = y_2] \times \dots \times \Pr[Y_n = y_n] \\
 &= f(y_1|p_0) \times f(y_2|p_0) \times \dots \times f(y_n|p_0) \\
 &= \prod_{j=1}^n f(y_j|p_0),
 \end{aligned}$$

hence

$$\begin{aligned}
 f_n(y_1, \dots, y_n|p_0) &= \prod_{j=1}^n p_0^{y_j} (1 - p_0)^{1-y_j} \\
 &= \left(\prod_{j=1}^n p_0^{y_j} \right) \left(\prod_{j=1}^n (1 - p_0)^{1-y_j} \right) \\
 &= p_0^{\sum_{j=1}^n y_j} (1 - p_0)^{n - \sum_{j=1}^n y_j}.
 \end{aligned}$$

Replacing the given non-random sequence y_1, \dots, y_n by the random sample Y_1, Y_2, \dots, Y_n and the unknown probability p_0 by a variable p in the interval $(0, 1)$ yields the likelihood function

$$L_n(p) = f_n(Y_1, \dots, Y_n|p) = p^{\sum_{j=1}^n Y_j} (1 - p)^{n - \sum_{j=1}^n Y_j}$$

For the case $p = p_0$ the likelihood function can be interpreted as the joint probability that we draw a particular sample Y_1, \dots, Y_n .

1.2 Maximum likelihood estimation

The idea of maximum likelihood (ML) estimation is now to choose p such that $L_n(p)$ is maximal. In other words, choose p such that the probability of drawing this particular sample Y_1, \dots, Y_n is maximal.

Note that maximizing $L_n(p)$ is equivalent to maximizing $\ln(L_n(p))$, i.e.,

$$\begin{aligned}
 \ln(L_n(p)) &= \left(\sum_{j=1}^n Y_j \right) \ln(p) + \left(n - \sum_{j=1}^n Y_j \right) \ln(1 - p) \\
 &= n \left(\bar{Y} \ln(p) + (1 - \bar{Y}) \ln(1 - p) \right),
 \end{aligned}$$

where

$$\bar{Y} = \frac{1}{n} \sum_{j=1}^n Y_j$$

is the sample mean. Therefore, the ML estimator \hat{p} in this case can be obtained from the first-order condition for a maximum of $\ln(L_n(p))$ in $p = \hat{p}$:

$$\begin{aligned} 0 &= \frac{d \ln(L_n(\hat{p}))}{d\hat{p}} = n \left(\bar{Y} \frac{d \ln(\hat{p})}{d\hat{p}} + (1 - \bar{Y}) \frac{d \ln(1 - \hat{p})}{d\hat{p}} \right) \\ &= n \left(\bar{Y} \frac{d \ln(\hat{p})}{d\hat{p}} + (1 - \bar{Y}) \frac{d \ln(1 - \hat{p})}{d(1 - \hat{p})} \times \frac{d(1 - \hat{p})}{d\hat{p}} \right) \\ &= n \left(\bar{Y} \frac{1}{\hat{p}} + (1 - \bar{Y}) \frac{1}{1 - \hat{p}} \times (-1) \right) \\ &= n \left(\frac{\bar{Y}}{\hat{p}} - \frac{1 - \bar{Y}}{1 - \hat{p}} \right) = n \left(\frac{\bar{Y}(1 - \hat{p}) - \hat{p}(1 - \bar{Y})}{\hat{p}(1 - \hat{p})} \right) \\ &= n \left(\frac{\bar{Y} - \hat{p}}{\hat{p}(1 - \hat{p})} \right) \end{aligned}$$

where we have used the fact that $d \ln(x)/dx = 1/x$. Thus, in this case the ML estimator \hat{p} of p_0 is the sample mean:

$$\hat{p} = \bar{Y}.$$

Note that this is an unbiased estimator: $E(\hat{p}) = \frac{1}{n} \sum_{j=1}^n E(Y_j) = p_0$.

1.3 Large sample statistical inference

It can be shown (but this requires advanced probability theory) that if the sample size n is large then $\sqrt{n}(\hat{p} - p_0)$ is approximately normally distributed, i.e.,

$$\sqrt{n}(\hat{p} - p_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^n (Y_j - p_0) \sim N[0, \sigma_0^2],$$

where

$$\begin{aligned} \sigma_0^2 &= \text{var}(Y_j) = E[(Y_j - p_0)^2] \\ &= (1 - p_0)^2 p_0 + (-p_0)^2 (1 - p_0) \\ &= p_0(1 - p_0). \end{aligned}$$

Thus, for large sample size n ,

$$\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}} \sim N[0, 1]. \quad (1)$$

This result can be used to test hypotheses about p_0 . In particular, under the null hypothesis that the coin is fair, $p_0 = 0.5$, we have

$$2\sqrt{n}(\hat{p} - 0.5) = \frac{\sqrt{n}(\hat{p} - 0.5)}{\sqrt{0.5 \times 0.5}} \sim N[0, 1],$$

Therefore, $2\sqrt{n}(\hat{p} - 0.5)$ can be used as the test statistic of the standard normal test of the null hypothesis $p_0 = 1/2$, as follows. Recall that for a standard normal random variable U , $\Pr[|U| > 1.96] = 0.05$. Thus, under the null hypothesis $p_0 = 1/2$ one would expect that

$$\begin{aligned} \Pr\left[|2\sqrt{n}(\hat{p} - 0.5)| > 1.96\right] &= 0.05 \\ \Pr\left[|2\sqrt{n}(\hat{p} - 0.5)| \leq 1.96\right] &= 0.95 \end{aligned}$$

If $|2\sqrt{n}(\hat{p} - 0.5)| > 1.96$ then we reject the null hypothesis $p_0 = 1/2$ at the 5% significance level, because this is not what one would expect if the null hypothesis is true, and if $|2\sqrt{n}(\hat{p} - 0.5)| \leq 1.96$ then we accept this null hypothesis, as this result is then in accordance with the null hypothesis $p_0 = 1/2$.

The result (1) can also be used to endow the unknown probability p_0 with a confidence interval, for example the 95% confidence interval, as follows. The result (1) implies

$$\Pr\left[\left|\frac{\sqrt{n}(\hat{p} - p_0)}{\sqrt{p_0(1 - p_0)}}\right| \leq 1.96\right] = 0.95,$$

which, after some straightforward calculations, can be shown to be equivalent to

$$\Pr\left[\underline{p}_n \leq p_0 \leq \bar{p}_n\right] = 0.95$$

where

$$\begin{aligned} \underline{p}_n &= \frac{n\hat{p} + (1.96)^2/2 - 1.96\sqrt{n\hat{p}(1 - \hat{p}) + (1.96)^2/4}}{n + (1.96)^2} \\ \bar{p}_n &= \frac{n\hat{p} + (1.96)^2/2 + 1.96\sqrt{n\hat{p}(1 - \hat{p}) + (1.96)^2/4}}{n + (1.96)^2} \end{aligned}$$

The interval $[\underline{p}_n, \bar{p}_n]$ is now the 95% confidence interval for p_0 .

1.4 An application election polls

Consider a presidential election with two candidates, candidate A and candidate B , and let p_0 be the fraction of likely voters who favor candidate A , just before the election is held. To predict the outcome of the election, a polling agency draws a random sample of size $n = 3000$, for example, from the population of likely voters.¹ Suppose that 1800 of the respondents express a preference for candidate A . Thus, the fraction of respondents favoring candidate A is $\hat{p} = 0.6$. Substituting $n = 3000$ and $\hat{p} = 0.6$ in the formulas for \underline{p}_n and \bar{p}_n yields

$$\underline{p}_n = 0.58, \bar{p}_n = 0.62$$

Thus, the 95% confidence interval of $100 \times p_0$ is $[58, 62]$. The polling results are therefore stated as: 60% of the likely voters will vote for candidate A , with a margin of error of ± 2 points.

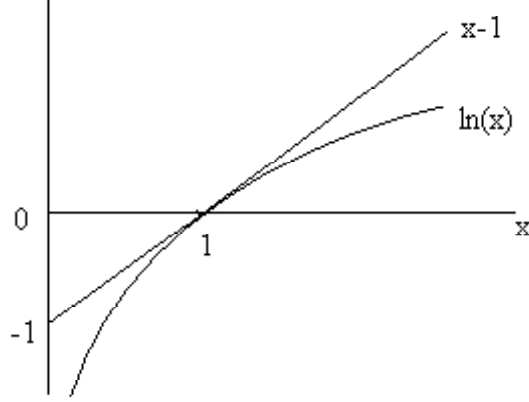
2 Motivation for maximum likelihood estimation

A more formal motivation for ML estimation is based on the fact that for $0 < x < 1$ and $x > 1$,

$$\ln(x) < x - 1.$$

This is illustrated in the following picture:

¹How to draw such a sample is beyond the scope of this lecture note.



$$\ln(x) \leq x - 1.$$

The inequality $\ln(x) < x - 1$ is strict for $x \neq 1$, and $\ln(1) = 0$. Consequently, taking $x = f(Y_j|p)/f(Y_j|p_0)$, we have the inequality

$$\ln \left(\frac{f(Y_j|p)}{f(Y_j|p_0)} \right) \leq \frac{f(Y_j|p)}{f(Y_j|p_0)} - 1.$$

Taking expectations, it follows that

$$\begin{aligned} E \left[\ln \left(\frac{f(Y_j|p)}{f(Y_j|p_0)} \right) \right] &\leq E \left[\frac{f(Y_j|p)}{f(Y_j|p_0)} \right] - 1 \\ &= \frac{f(1|p)}{f(1|p_0)} \Pr[Y_j = 1] + \frac{f(0|p)}{f(0|p_0)} \Pr[Y_j = 0] - 1 \\ &= \frac{p}{p_0} p_0 + \frac{1-p}{1-p_0} (1-p_0) - 1 \\ &= p + 1 - p - 1 = 0, \end{aligned} \tag{2}$$

hence

$$E [\ln (f(Y_j|p))] - E [\ln (f(Y_j|p_0))] = E \left[\ln \left(\frac{f(Y_j|p)}{f(Y_j|p_0)} \right) \right] \leq 0,$$

and therefore,

$$E [\ln (L_n(p))] \leq E [\ln (L_n(p_0))]. \tag{3}$$

Thus, $E [\ln (L_n(p))]$ is maximal for $p = p_0$, and it can be shown that this maximum is unique.

3 Maximum likelihood estimation of the Logit model

3.1 The Logit model with one explanatory variable

Next, let $(Y_1, X_1), \dots, (Y_n, X_n)$ be a random sample from the conditional Logit distribution:

$$\begin{aligned}\Pr[Y_j = 1|X_j] &= \frac{1}{1 + \exp(-\alpha_0 - \beta_0 X_j)}, \\ \Pr[Y_j = 0|X_j] &= 1 - \Pr[Y_j = 1|X_j] \\ &= \frac{\exp(-\alpha_0 - \beta_0 X_j)}{1 + \exp(-\alpha_0 - \beta_0 X_j)}\end{aligned}\tag{4}$$

where the X_j 's are the explanatory variables and α_0 and β_0 are unknown parameters to be estimated. This model is called a Logit model, because

$$\Pr[Y_j = 1|X_j] = F(\alpha_0 + \beta_0 X_j)\tag{5}$$

where

$$F(x) = \frac{1}{1 + \exp(-x)}\tag{6}$$

is the distribution function of the logistic (Logit) distribution.

The conditional probability function involved is

$$\begin{aligned}f(y|X_j, \alpha_0, \beta_0) &= \Pr[Y_j = y|X_j] \\ &= F(\alpha_0 + \beta_0 X_j)^y (1 - F(\alpha_0 + \beta_0 X_j))^{1-y} \\ &= \begin{cases} F(\alpha_0 + \beta_0 X_j) & \text{if } y = 1, \\ 1 - F(\alpha_0 + \beta_0 X_j) & \text{if } y = 0. \end{cases}\end{aligned}$$

Now the conditional log-likelihood function is

$$\begin{aligned}\ln(L_n(\alpha, \beta)) &= \sum_{j=1}^n \ln(f(Y_j|X_j, \alpha, \beta)) \\ &= \sum_{j=1}^n Y_j \ln(F(\alpha + \beta X_j)) + \sum_{j=1}^n (1 - Y_j) \ln(1 - F(\alpha + \beta X_j)) \\ &= - \sum_{j=1}^n (1 - Y_j) (\alpha + \beta X_j) - \sum_{j=1}^n \ln(1 + \exp(-\alpha - \beta X_j)).\end{aligned}\tag{7}$$

Similar to (3) we have

$$E[\ln(L_n(\alpha, \beta)) | X_1, \dots, X_n] \leq E[\ln(L_n(\alpha_0, \beta_0)) | X_1, \dots, X_n].$$

Again, this result motivates to estimate α_0 and β_0 by maximizing $\ln(L_n(\alpha, \beta))$ to α and β :

$$\ln(L_n(\hat{\alpha}, \hat{\beta})) = \max_{\alpha, \beta} \ln(L_n(\alpha, \beta)).$$

However, there is no longer an explicit solution for $\hat{\alpha}$ and $\hat{\beta}$. These ML estimators have to be solved numerically. Your econometrics software will do that for you.

3.2 Pseudo t-values

It can be shown that if the sample size n is large then

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \sim N(0, \sigma_\alpha^2), \quad \sqrt{n}(\hat{\beta} - \beta_0) \sim N(0, \sigma_\beta^2).$$

Given consistent estimators $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\beta^2$ of the unknown variances σ_α^2 and σ_β^2 , respectively (which are computed by your econometrics software), we then have

$$\frac{\sqrt{n}(\hat{\alpha} - \alpha_0)}{\hat{\sigma}_\alpha} \sim N(0, 1), \quad \frac{\sqrt{n}(\hat{\beta} - \beta_0)}{\hat{\sigma}_\beta} \sim N(0, 1).$$

These results can be used to test whether the coefficients α_0 and β_0 are zero or not. In particular the null hypothesis $\beta_0 = 0$ is of interest, because this hypothesis implies that the conditional probability $\Pr[Y_j = 1 | X_j]$ does not depend on X_j . Under the null hypothesis $\beta_0 = 0$ we have

$$\hat{t}_\beta = \frac{\sqrt{n}\hat{\beta}}{\hat{\sigma}_\beta} \sim N(0, 1).$$

Recall that the 5% critical value of the two-sided standard normal test is 1.96. Thus, for example, the null hypothesis $\beta_0 = 0$ is rejected at the 5% significance level in favor of the alternative hypothesis $\beta_0 \neq 0$ if $|\hat{t}_\beta| > 1.96$, and accepted if $|\hat{t}_\beta| \leq 1.96$.

The statistic \hat{t}_β is called the *pseudo* t-value of $\hat{\beta}$ because it is used in the same way as the t-value in linear regression, and $\hat{\sigma}_\beta$ is called the standard error of $\hat{\beta}$. Your econometric software will report the ML estimators together with their corresponding pseudo t-values and/or standard errors.

3.3 The general Logit model

The general Logit model takes the form

$$\begin{aligned} \Pr[Y_j = 1|X_{1j}, \dots, X_{kj}] &= \frac{1}{1 + \exp(-\beta_1^0 X_{1j} - \dots - \beta_k^0 X_{kj})} \\ &= \frac{1}{1 + \exp\left(-\sum_{i=1}^k \beta_i^0 X_{ij}\right)}, \end{aligned} \quad (8)$$

where one of the X_{ij} equals 1 for the constant term, for example, let $X_{kj} = 1$, and the β_i^0 's are the true parameter values. This model can be estimated by ML in the same way as before. Thus, the log-likelihood function is

$$\ln(L_n(\beta_1, \dots, \beta_k)) = -\sum_{j=1}^n (1 - Y_j) \sum_{i=1}^k \beta_i X_{ij} - \sum_{j=1}^n \ln\left(1 + \exp\left(-\sum_{i=1}^k \beta_i X_{ij}\right)\right), \quad (9)$$

and the ML estimators $\hat{\beta}_1, \dots, \hat{\beta}_k$ are obtained by maximizing $\ln(L_n(\beta_1, \dots, \beta_k))$:

$$\ln(L_n(\hat{\beta}_1, \dots, \hat{\beta}_k)) = \max_{\beta_1, \dots, \beta_k} \ln(L_n(\beta_1, \dots, \beta_k)).$$

Again, it can be shown that if n is large then for $i = 1, \dots, k$,

$$\sqrt{n}(\hat{\beta}_i - \beta_i^0) \sim N[0, \sigma_i^2].$$

Given consistent estimators $\hat{\sigma}_i^2$ of the variances σ_i^2 , it follows then that

$$\frac{\sqrt{n}(\hat{\beta}_i - \beta_i^0)}{\hat{\sigma}_i} \sim N[0, 1]$$

for $i = 1, \dots, k$. Your econometrics software will report the ML estimators $\hat{\beta}_i$ together with their corresponding pseudo t-values $\hat{t}_i = \sqrt{n}\hat{\beta}_i/\hat{\sigma}_i$ and/or standard errors $\hat{\sigma}_i$.

3.4 Testing joint significance

Now suppose you want to test the joint null hypothesis

$$H_0: \beta_1^0 = 0, \beta_2^0 = 0, \dots, \beta_m^0 = 0, \quad (10)$$

where $m < k$.

There are two ways to do that. One way is akin to the F test in linear regression: Re-estimate the Logit model under the null hypothesis:

$$\ln \left(L_n(0, \dots, 0, \tilde{\beta}_{m+1}, \dots, \tilde{\beta}_k) \right) = \max_{\beta_{m+1}, \dots, \beta_k} \ln \left(L_n(0, \dots, 0, \beta_{m+1}, \dots, \beta_k) \right).$$

and compare the log-likelihoods². It can be shown that under the null hypothesis (10) and for large samples,

$$LR_m = -2 \ln \left(\frac{L_n(0, \dots, 0, \tilde{\beta}_{m+1}, \dots, \tilde{\beta}_k)}{L_n(\hat{\beta}_1, \dots, \hat{\beta}_k)} \right) \sim \chi_m^2,$$

where the degrees of freedom m corresponds to the number of restrictions imposed under the null hypothesis. This is the so-called likelihood ratio test, which is conducted right-sided. For example, choose the 5% significance level, look up in the table of the χ^2 distribution the critical value c such that for a χ_m^2 distributed random variable Z_m , $\Pr[Z_m > c] = 0.05$. Then the null hypothesis (10) is rejected at the 5% significance level if $LR_m > c$ and accepted if $LR_m \leq c$.

An alternative test of the null hypothesis (10) is the Wald test, which is conducted in the same way as for linear regression models.³ Under the null hypothesis (10) the Wald test statistic has also a χ_m^2 distribution.

4 Interpretation of the coefficients of the Logit model

4.1 Marginal effects

Consider the Logit model (5). If $\beta_0 > 0$ then $\Pr[Y_j = 1 | X_j] = F(\alpha_0 + \beta_0 X_j)$ is an increasing function of X_j :

$$\frac{dP[Y_j = 1 | X_j]}{dX_j} = \beta_0 \cdot F'(\alpha_0 + \beta_0 X_j),$$

where F' is the derivative of (6):

²Your econometric software will report the log-likelihood function value.

³In *EasyReg International* the Wald test can be conducted simply by point-and-click.

$$\begin{aligned}
F'(x) &= \frac{\exp(-x)}{(1 + \exp(-x))^2} = \frac{1 + \exp(-x)}{(1 + \exp(-x))^2} - \frac{1}{(1 + \exp(-x))^2} \\
&= \frac{1}{1 + \exp(-x)} - \frac{1}{(1 + \exp(-x))^2} = F(x) - F(x)^2 \\
&= F(x)(1 - F(x)).
\end{aligned}$$

Therefore, the marginal effect of X_j on $\Pr[Y_j = 1|X_j]$ depends on X_j :

$$\frac{dP[Y_j = 1|X_j]}{dX_j} = \beta_0 \cdot F(\alpha_0 + \beta_0 X_j) (1 - F(\alpha_0 + \beta_0 X_j)),$$

which renders the interpretation of β_0 difficult.

However, the coefficient β_0 can be interpreted in terms of relative changes in odds.

4.2 Odds and odds ratios

The odds is the ratio of the probability that something is true divided by the probability that it is not true. Thus, in the Logit case (4),

$$\text{Odds}(X) = \frac{\Pr[Y_j = 1|X_j]}{\Pr[Y_j = 0|X_j]} = \frac{F(\alpha_0 + \beta_0 X_j)}{1 - F(\alpha_0 + \beta_0 X_j)} = \exp(\alpha_0 + \beta_0 X_j). \quad (11)$$

The odds ratio is the ratio of two odds for different values of X_j , say $X_j = x$ and $X_j = x + \Delta x$:

$$\frac{\text{Odds}(x + \Delta x)}{\text{Odds}(x)} = \frac{\exp(\alpha + \beta x + \beta \Delta x)}{\exp(\alpha + \beta x)} = \exp(\beta \Delta x),$$

where Δx is a small change in x . Then

$$\begin{aligned}
&\lim_{\Delta x \rightarrow 0} \frac{1}{\Delta x} \left(\frac{\text{Odds}(x + \Delta x) - \text{Odds}(x)}{\text{Odds}(x)} \right) = \lim_{\Delta x \rightarrow 0} \frac{\exp(\beta_0 \Delta x) - 1}{\Delta x} \\
&= \beta_0 \lim_{\beta_0 \Delta x \rightarrow 0} \frac{\exp(\beta_0 \Delta x) - 1}{\beta_0 \Delta x} = \beta_0 \times \left. \frac{d \exp(u)}{du} \right|_{u=0} = \beta_0 \exp(0) = \beta_0.
\end{aligned}$$

Thus, β_0 may be interpreted as the *relative* change in the odds due to a small change Δx in X_j :

$$\frac{\text{Odds}(x + \Delta x) - \text{Odds}(x)}{\text{Odds}(x)} = \frac{\text{Odds}(x + \Delta x)}{\text{Odds}(x)} - 1 \approx \beta_0 \Delta x \quad (12)$$

If X_j is a binary variable itself, $X_j = 0$ or $X_j = 1$, then the only reasonable choices for $x + \Delta x$ and x are 1 and 0, respectively, so that then

$$\frac{\text{Odds}(1)}{\text{Odds}(0)} - 1 = \frac{\text{Odds}(1) - \text{Odds}(0)}{\text{Odds}(0)} = \exp(\beta_0) - 1.$$

Only if β_0 is small we may then use the approximation $\exp(\beta_0) - 1 \approx \beta_0$. If not, one has to interpret β_0 in terms of the log of the odds ratio involved:

$$\ln\left(\frac{\text{Odds}(1)}{\text{Odds}(0)}\right) = \beta_0.$$

The interpretation of the coefficients $\beta_i^0, i = 1, \dots, k - 1$ in the general Logit model (8) is similar as in the case (12):

$$\frac{\text{Odds}(X_{1j}, \dots, X_{i-1,j}, X_{i,j} + \Delta X_{i,j}, X_{i+1,j}, \dots, X_{k,j})}{\text{Odds}(X_{1j}, \dots, X_{i-1,j}, X_{i,j}, X_{i+1,j}, \dots, X_{k,j})} - 1 \approx \beta_i^0 \Delta X_{i,j}$$

if $\Delta X_{i,j}$ is small. For example, β_i^0 may be interpreted as the percentage change in $\text{Odds}(X_{1j}, \dots, X_{k,j})$ due to a small percentage change $100 \times \Delta X_{i,j} = 1$ in $X_{i,j}$.