Binary choice models: an explanation of the standard errors of the Maximum Likelihood estimator

Lennart Hoogerheide

2019

Consider the binary choice model in which observation y_i $(i = 1, 2, \ldots, n)$ has a Bernoulli distribution with $\Pr(y_i = 1|x_i) = p_i = G(x_i'\beta)$ and $\Pr(y_i = 0|x_i) = 1 - p_i = 1 - G(x_i'\beta)$, where $G(.)$ is a cumulative distribution function (CDF) of a continuous distribution with probability density function (pdf) $G'() = g(.)$. We assume that conditionally upon the x_i $(i = 1, 2, \ldots, n)$ the y_i are independent. The loglikelihood is given by

$$\log L(\beta) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \{y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\},$$

where $l_i = y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$ is the contribution of observation i to the loglikelihood.

The large-sample distribution of the maximum likelihood (ML) estimator $\hat{\beta}_{ML}$ of β is (under certain regularity conditions):

$$\hat{\beta}_{ML} \approx N(\beta, \mathcal{I}(\beta)^{-1})$$

with information matrix[1] given by

$$\mathcal{I}(\beta) = E\left(\sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta} \frac{\partial l_i}{\partial \beta'}\right),$$

where the expectation is taken over the y_i (for given values of the x_i). We have

$$\begin{aligned}
\frac{\partial l_i}{\partial \beta} &= y_i \frac{1}{p_i} \frac{\partial p_i}{\partial \beta} + (1 - y_i) \frac{1}{1 - p_i} \left(-\frac{\partial p_i}{\partial \beta}\right) \\
&= \left(\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i}\right) \frac{\partial p_i}{\partial \beta} \\
&= \left(\frac{y_i(1 - p_i) - (1 - y_i)p_i}{p_i(1 - p_i)}\right) \frac{\partial p_i}{\partial \beta} \\
&= \left(\frac{y_i - p_i}{p_i(1 - p_i)}\right) \frac{\partial p_i}{\partial \beta}.
\end{aligned}$$

[1] The information matrix is also given by the equivalent formula $\mathcal{I}(\beta) = -E(\frac{\partial^2 \log L(\beta)}{\partial \beta \partial \beta'})$, but in this case of a binary choice model that formula is harder to work with.

We have

$$\frac{\partial p_i}{\partial \beta} = \frac{\partial G(x_i'\beta)}{\partial \beta} = \frac{\partial G(x_i'\beta)}{\partial (x_i'\beta)} \frac{\partial (x_i'\beta)}{\partial \beta} = g(x_i'\beta) x_i = g_i x_i$$

with $g_i \equiv g(x_i'\beta)$, since the derivative of the CDF $G(.)$ is obviously the pdf $g(.)$.

So, we have information matrix

$$\mathcal{I}(\beta) = E\left(\sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta} \frac{\partial l_i}{\partial \beta'} \right) = E\left(\sum_{i=1}^{n} \frac{(y_i - p_i)^2}{p_i^2 (1-p_i)^2} g_i^2 x_i x_i' \right)$$

where the expectation is taken over the y_i (for given values values of the x_i). We have $E(y_i|x_i) = p_i \times 1 + (1-p_i) \times 0 = p_i$, so that $E((y_i - p_i)^2) = E((y_i - E(y_i))^2)$ is simply the variance of the Bernoulli distribution of y_i (given x_i), which is equal to $p_i(1-p_i)$. So, we have information matrix

$$\begin{aligned} \mathcal{I}(\beta) &= \sum_{i=1}^{n} \frac{E((y_i - p_i)^2)}{p_i^2 (1-p_i)^2} g_i^2 x_i x_i' \\ &= \sum_{i=1}^{n} \frac{p_i(1-p_i)}{p_i^2 (1-p_i)^2} g_i^2 x_i x_i' = \sum_{i=1}^{n} \frac{g_i^2}{p_i(1-p_i)} x_i x_i'. \end{aligned}$$

We consider four binary choice models, which are given in Table 1. The normal, logistic and Cauchy (Student-t with 1 degree of freedom) distributions are symmetric distributions with thin, fat and very fat tails, respectively. The Gumbel distribution (type I extreme value distribution) is asymmetric with a thin left tail and a fat right tail. Figures 1 and 2 show the CDF and pdf of these four distributions.

Table 1: Binary choice models and distributions

model	distribution	CDF $G(z)$	pdf $g(z)$
probit	normal	$G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx$	$g(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$
logit	logistic	$G(z) = \frac{1}{1+\exp(-z)}$	$g(z) = \frac{\exp(-z)}{(1+\exp(-z))^2}$
cauchit	Cauchy	$G(z) = \frac{1}{\pi}\left(\arctan(z) + \frac{\pi}{2}\right)$	$g(z) = \frac{1}{\pi(1+z^2)}$
gompit	Gumbel	$G(z) = \exp(-\exp(-z))$	$g(z) = \exp(-z - \exp(-z))$

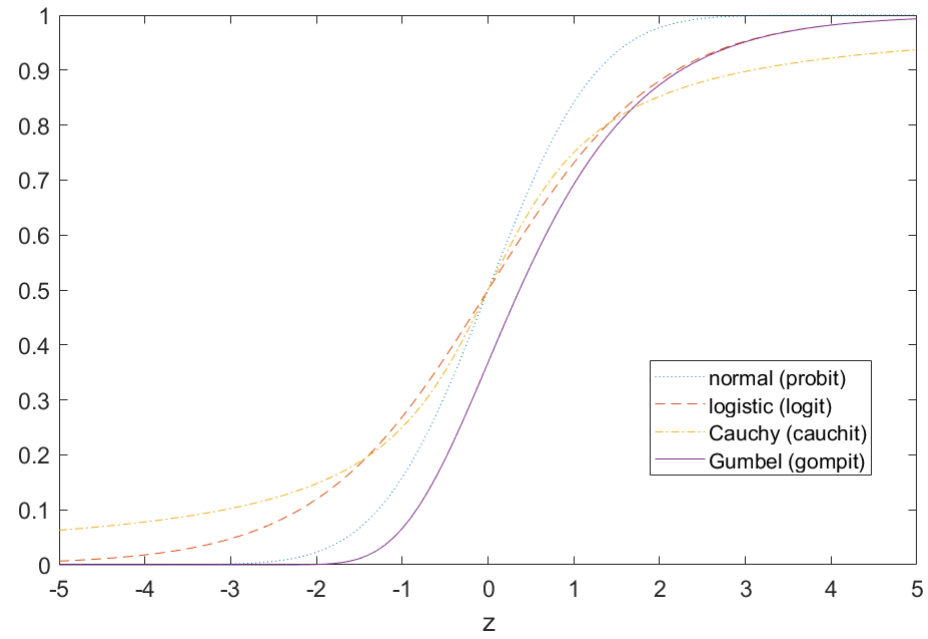

Figure 1: CDF $G(z)$ for the standard normal density (of the probit model), the standard logistic density (of the logit model), the Cauchy density (the Student-t density with 1 degree of freedom of the cauchit model) and the Gumbel density (the type I extreme value density of the Gompit model).

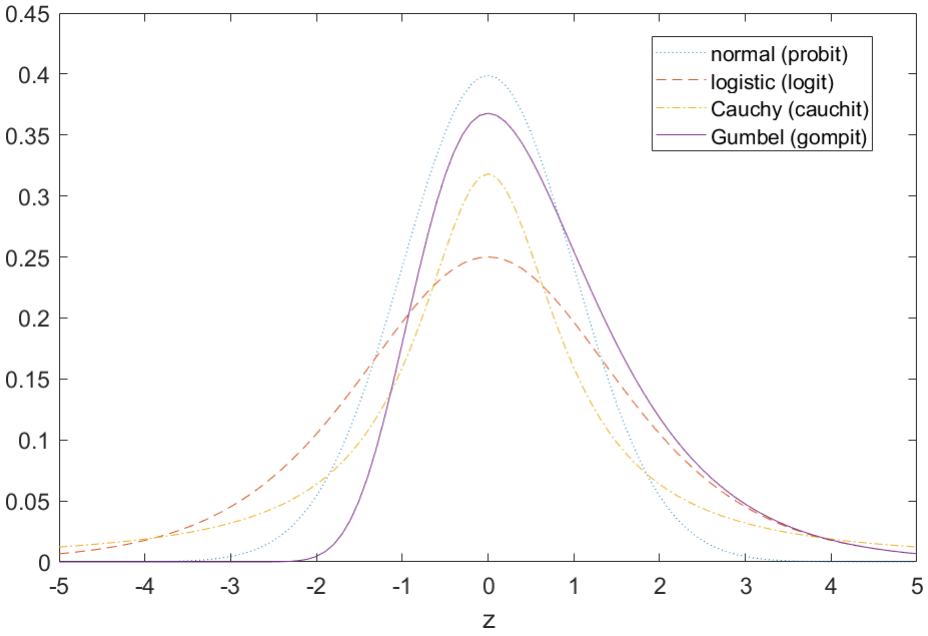

Figure 2: pdf $g(z)$ for the standard normal density (of the probit model), the standard logistic density (of the logit model), the Cauchy density (the Student-t density with 1 degree of freedom of the cauchit model) and the Gumbel density (the type I extreme value density of the Gompit model).

In case of the **binary logit model** we have logistic CDF

$$G(z) = \frac{1}{1+\exp(-z)} = \frac{\exp(z)}{1+\exp(z)}$$

and logistic pdf

$$\begin{aligned}
\frac{\partial G(z)}{\partial z} &= \frac{-1}{(1+\exp(-z))^2}(-\exp(-z)) \\
&= \frac{\exp(-z)}{(1+\exp(-z))^2} \\
&= \frac{1}{1+\exp(-z)}\,\frac{\exp(-z)}{1+\exp(-z)} \\
&= G(z)(1-G(z)),
\end{aligned}$$

since

$$\frac{\exp(-z)}{1+\exp(-z)} = \frac{1}{1+\exp(z)} = 1 - G(z),$$

where the numerator and denominator of $\frac{\exp(-z)}{1+\exp(-z)}$ are multiplied by $\exp(z)$. So, in the binary logit model we have

$$g_i = g(x_i'\beta) = G(x_i'\beta)(1-G(x_i'\beta)) = p_i(1-p_i),$$

so that the information matrix becomes

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} \frac{g_i^2}{p_i(1-p_i)} x_i x_i' = \sum_{i=1}^{n} \frac{p_i^2(1-p_i)^2}{p_i(1-p_i)} x_i x_i' = \sum_{i=1}^{n} p_i(1-p_i) x_i x_i'.$$

Notice that the information matrix depends on three inputs:

- n: the smaller the number of observations, the larger the standard errors.
- x_i: the smaller the variation in the explanatory variables x_i, the larger the standard errors. In the most extreme case, if there is no variation in the explanatory variables x_i, then we can also not measure the effect of x_i on y_i. Then the standard errors are infinite.
- p_i: the closer the p_i are to 0 or 1 (that is, the smaller $p_i(1-p_i)$), the larger the standard errors. The closer the p_i are to 0.5 (that is, the larger $p_i(1-p_i)$), the smaller the standard errors. Figure 3 illustrates that this finding also holds for the other symmetric distributions: also for the logistic and Cauchy distributions of the logit and cauchit models the ratio $\frac{g_i^2}{p_i(1-p_i)}$ is largest for values of p_i closest to 0.5. However, for the Gumbel density the mode is smaller than the median, which causes that $\frac{g_i^2}{p_i(1-p_i)}$ is largest for a smaller value of p_i.

  Note that $\frac{g_i^2}{p_i(1-p_i)}$ is larger for the normal distribution than for the logistic and Cauchy distributions. This does not mean that we can always estimate the probit model better than the logit and cauchit models; it merely reflects that the scale

of the parameters β is smaller in the probit model than in the logit and cauchit models.

Intuitively, it makes sense that if the densities g_i are all small (with many probabilities p_i close to 0 or 1), so that the marginal effects of the inputs $x_i'\beta$ on the probabilities $p_i = G(x_i'\beta)$ are small, then it is difficult to precisely estimate β. On the other hand, if many densities g_i are large, then the probabilities are strongly affected by the input $x_i'\beta$, and then it is easier to precisely estimate β.

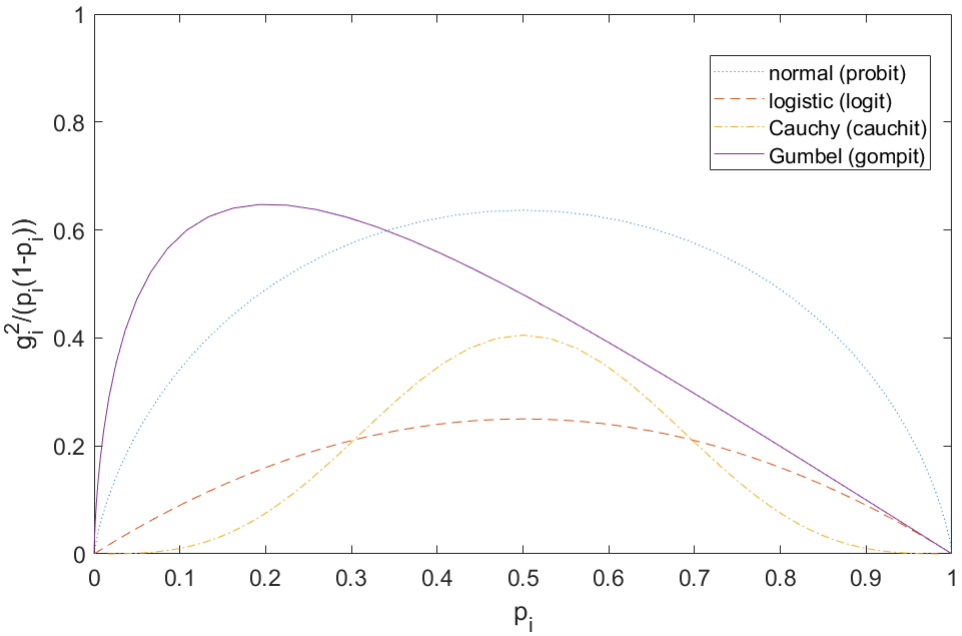

Figure 3: $\frac{g_i^2}{p_i(1-p_i)}$ (the ratio between the squared pdf and CDF (1-CDF)) against p_i for the standard normal density (of the probit model), the standard logistic density (of the logit model), the Cauchy density (the Student-t density with 1 degree of freedom of the cauchit model) and the Gumbel density (the type I extreme value density of the Gompit model).

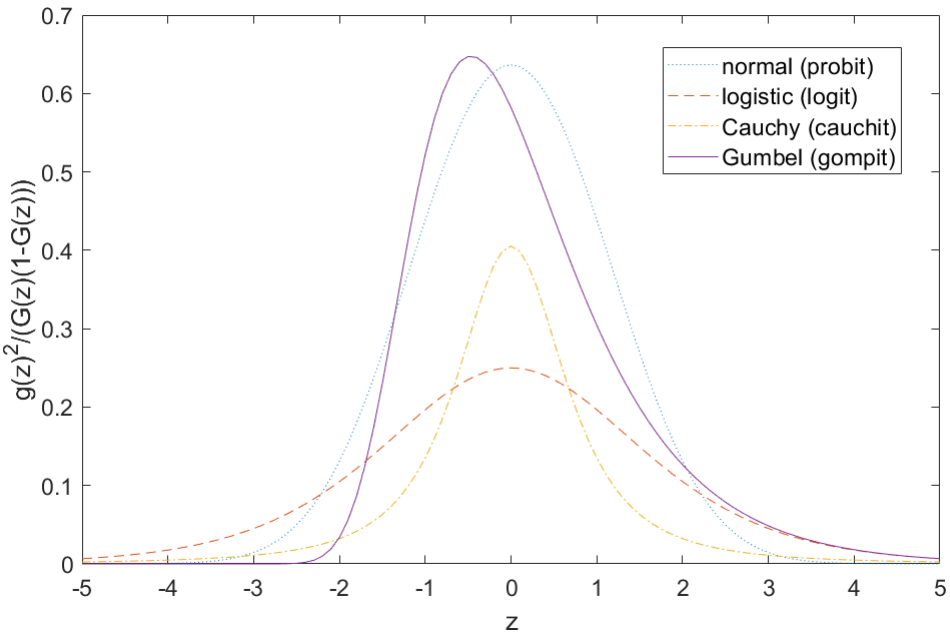

Figure 4: $\frac{g(z)^2}{G(z)(1-G(z))}$ (the ratio between the squared pdf and CDF (1-CDF)) for the standard normal density (of the probit model), the standard logistic density (of the logit model), the Cauchy density (the Student-t density with 1 degree of freedom of the cauchit model) and the Gumbel density (the type I extreme value density of the Gompit model).

Keeping only a fraction c of the observations with $y_i = 0$

Suppose that we have a dataset with relatively very few observations with $y_i = 1$. That is, most observations have $y_i = 0$. In that case one may consider to only keep a randomly selected fraction c ($0 < c < 1$) of the observations with $y_i = 0$ (while keeping all observations with $y_i = 1$). To analyze the effect of this, we introduce the variable s_i with

$$s_i = \begin{cases} 1 & \text{if observation } i \text{ is kept,} \\ 0 & \text{if observation } i \text{ is not kept.} \end{cases}$$

Then we have the conditional probabilities $\Pr(s_i = 1|y_i = 1) = 1$ (keeping all observations with $y_i = 1$) and $\Pr(s_i = 1|y_i = 0) = c$ (keeping fraction c of the observations with $y_i = 0$). So for the remaining data we have

$$\begin{aligned} \Pr(y_i = 1|s_i = 1) &= \frac{\Pr(y_i = 1, s_i = 1)}{\Pr(s_i = 1)} \\ &= \frac{\Pr(s_i = 1|y_i = 1)\Pr(y_i = 1)}{\Pr(s_i = 1|y_i = 1)\Pr(y_i = 1) + \Pr(s_i = 1|y_i = 0)\Pr(y_i = 0)} \\ &= \frac{p_i}{p_i + c \cdot (1 - p_i)} \\ &= \frac{G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))} \end{aligned}$$

where we used $\Pr(A|B) = \frac{\Pr(A,B)}{\Pr(B)}$, $\Pr(s_i = 1|y_i = 1) = 1$, $\Pr(y_i = 1) = p_i = G(x_i'\beta)$, $\Pr(s_i = 1|y_i = 0) = c$ and $\Pr(y_i = 0) = 1 - p_i = 1 - G(x_i'\beta)$. Note that all probabilities are conditional on x_i. So, we can still correctly (consistently) estimate the binary choice model, using ML for the reduced dataset (of kept observations with $s_i = 1$) if we use the adapted probabilities

$$\begin{aligned} \Pr(y_i = 1|s_i = 1) &= \frac{G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))} \\ \Pr(y_i = 0|s_i = 1) &= 1 - \frac{G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))} = \frac{c \cdot G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))}. \end{aligned}$$

For the logit model we have

$$\begin{aligned} \Pr(y_i = 1|s_i = 1) &= \frac{G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))} = \frac{\frac{1}{1+\exp(-x_i'\beta)}}{\frac{1}{1+\exp(-x_i'\beta)} + c \cdot \frac{\exp(-x_i'\beta)}{1+\exp(-x_i'\beta)}} \\ &= \frac{1}{1 + c \cdot \exp(-x_i'\beta)} = \frac{1}{1 + \exp(\ln(c)) \cdot \exp(-x_i'\beta)} \\ &= \frac{1}{1 + \exp(\ln(c) - x_i'\beta)} = \frac{1}{1 + \exp(-(x_i'\beta - \ln(c)))}, \end{aligned}$$

where we used that $1 - p_i = 1 - \frac{1}{1+\exp(-x_i'\beta)} = \frac{\exp(-x_i'\beta)}{1+\exp(-x_i'\beta)}$, and where we multiplied numerator and denominator with $1 + \exp(-x_i'\beta)$. That is, we again obtain a logit model where only the constant term in $x_i'\beta$ has increased by $-\ln(c)$. We can simply estimate a logit

model for the reduced dataset and in the end add the negative value $\ln(c)$ to the constant term to obtain estimates of β in the logit model for the full dataset.

Obviously, deleting observations implies a loss of information, and an increase in the standard errors. However, the increase in the standard errors may be relatively limited. The information matrix

$$\mathcal{I}(\beta) = \sum_{i=1}^{n} p_i(1-p_i)x_i x_i'$$

will decrease due to the smaller number of observations n, but the increase from p_i to $\Pr(y_i = 1|s_i = 1)$ (and $p_i(1-p_i)$ to $\Pr(y_i = 1|s_i = 1)(1 - \Pr(y_i = 1|s_i = 1))$) compensates part of this decrease.

Note that we can also estimate probit, cauchit and gompit models for the reduced dataset. There we only need to adapt the formulas for the probabilities

$$\begin{aligned}
\Pr(y_i = 1|s_i = 1) &= \frac{G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))} \\
\Pr(y_i = 0|s_i = 1) &= 1 - \frac{G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))} = \frac{c \cdot G(x_i'\beta)}{G(x_i'\beta) + c \cdot (1 - G(x_i'\beta))}
\end{aligned}$$

which have a different form than in the original probit, cauchit and gompit models. Also here the increase in the standard errors may be relatively limited.