When z_t is just a linear combination of the constant term and a single independent variable, the latter is often said to be a **perfect classifier**, because the y_t's can be classified as being 0 or 1 once the value of that variable is known. For example, consider the DGP

$$\begin{aligned} &y_t^* = x_t + u_t, \quad u_t \sim \text{NID}(0,1);\\ &y_t = 1 \text{ if } y_t^* > 0 \quad \text{and} \quad y_t = 0 \text{ if } y_t^* \leq 0. \end{aligned} \tag{15.14}$$

For this DGP, it would seem to be sensible to estimate the probit model

$$E(y_t \,|\, x_t) = \Phi(\beta_0 + \beta_1 x_t). \tag{15.15}$$

But suppose that, in the sample, x_t is always either less than -4 or greater than $+4$. When x_t is less than -4, it is almost certain (the probability is greater than 0.99997) that y_t will be 0, and when x_t is greater than $+4$, it is almost certain that y_t will be 1. Thus, unless the sample size is very large, there are unlikely to be any observations for which $x_t < 0$ and $y_t = 1$ or observations for which $x_t > 0$ and $y_t = 0$. In the absence of such observations, the variable x_t will be a perfect classifier, and it will be impossible to obtain sensible estimates of the parameters of (15.15). Whatever maximization algorithm is being used will simply try to make $\hat\beta_1$ as large as possible.

Although this example is an extreme one, similar problems are likely to occur whenever the model fits very well and the sample size is small. There will be a perfect classifier whenever there exists a separating hyperplane in the space of the explanatory variables such that all the observations with $y_t = 0$ are on one side and all the observations with $y_t = 1$ are on the other. This is likely to happen when the model fits well and when there are only a few observations for which $y_t = 1$ or, alternatively, for which $y_t = 0$. Nevertheless, it may be possible to obtain ML estimates when n is as small as $k + 1$ and when there is only one observation for which $y_t = 1$ or $y_t = 0$.

In regression models, it is common to test the hypothesis that all slopes are zero by using an F test. For binary response models, the same hypothesis can easily be tested by using a likelihood ratio test. A model with a constant term can be written as

$$E(y_t \,|\, \Omega_t) = F\big(\beta_1 + \boldsymbol{X}_{2t}\boldsymbol{\beta}_2\big), \tag{15.16}$$

where $\boldsymbol{X}_{2t}$ consists of $\boldsymbol{X}_t$ without the constant and $\boldsymbol{\beta}_2$ is a $(k-1)$–vector. Under the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$, (15.16) becomes

$$E(y_t \,|\, \Omega_t) = F\big(\beta_1\big) = E(y_t).$$

This just says that the conditional mean of y_t is equal to its unconditional mean, which can be estimated by $\bar y$. Therefore, if we denote the estimate of β_1 by $\bar\beta_1$, $\bar y = F(\bar\beta_1)$. From (15.09), it is easy to work out that the value of