Unobserved heterogeneity in logistic regression

Maarten L. Buis

November 18, 2005

1 Introduction

Regression is used to find the relation between various explanatory variables and an explained variable. For instance, we might think that there is a relation between income and the respondents age and his intelligence. Often we do not have all variables that influence the explained variable (income), in this case we might not know the intelligence the respondents. This is not a problem in a linear regression model, as long as the expected value or mean of the unmeasured variables is zero and does not change for different values of the measured variables (age)[1]. The idea behind this is simple: The estimated regression may sometimes overestimate (when someone is dumber than average) and sometimes underestimate (when someone is smarter than average) the explained variable, but these positive and negative errors cancel each other out.

Sometimes, there is reason to believe that the relationship between variables is not linear. This is for instance the case when the explained variable can have only two values, e.g. pass or fail, marry or not marry, live or die, etc. In that case we want to explain how the proportion of passed, married, or alive persons depend on the explanatory variables. The explained variable (say, proportion of successes) must lie between 0 and 1, since it is impossible to have less than 0% or more than 100% successes. Thus, explanatory variables can not have an effect that leads to predicted proportions below zero or above one. So the explanatory variables cannot have a linear effect. In this case the positive and negative errors do not cancel each other out. If the known variable leads to a high predicted proportion of success, the positive effect of the unknown variable will be 'squashed' against the maximum proportion of successes while negative effects are not limited, even though the unknown variable itself has an expected value of zero. Thus at values of the known variables that lead to high predicted proportion of success, the proportion of successes will be too low, because the negative effects of the unknown variables are not fully compensated by the positive effects. Similarly, if the known variable leads to a low predicted proportion of success, the negative effect of the unknown variable will be 'squashed' against the minimum proportion of successes while positive effects are not limited. Thus at values of the known variables that lead to low predicted proportion of success, the proportion of successes will be too high, because the positive effects of the unknown variables are not fully compensated by the negative effects.

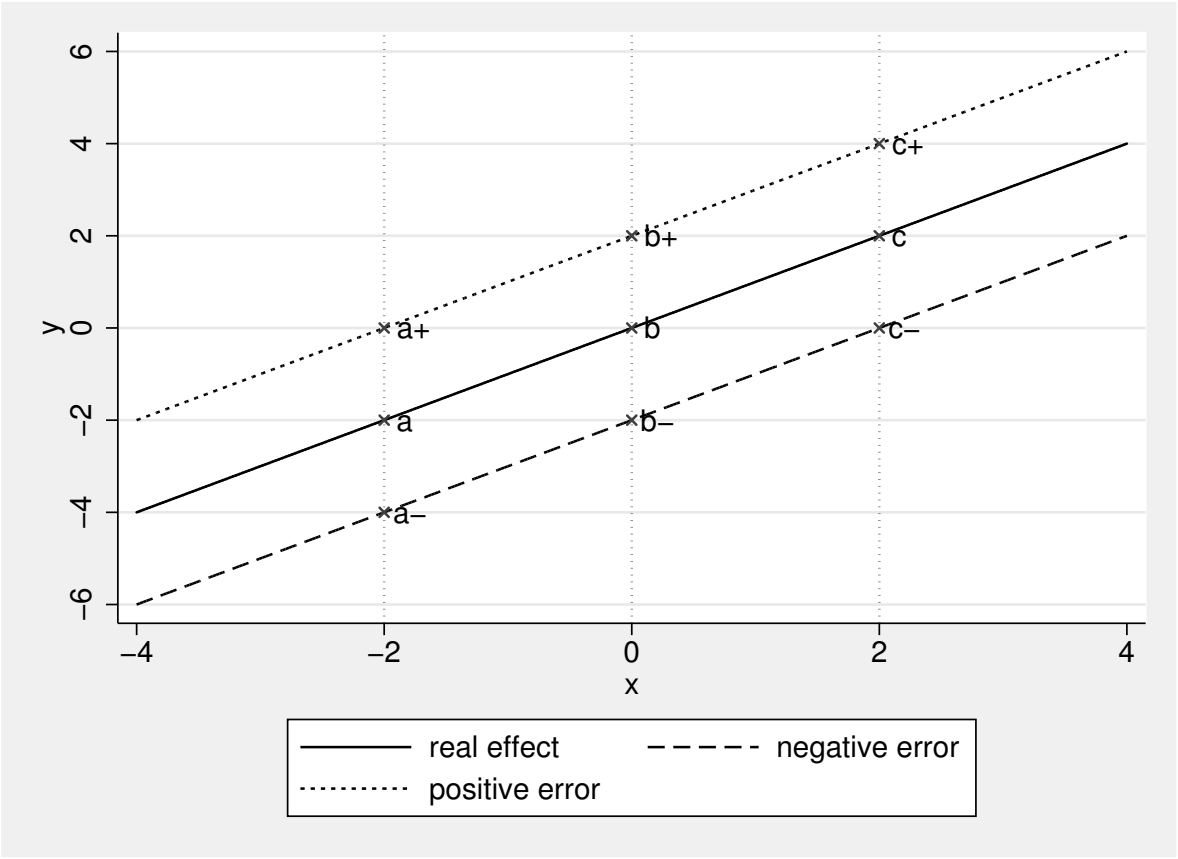
These ideas will be discussed in more detail in two stages: First, the reason that unobserved variables in a linear model do not lead to biased estimates will be discussed. Second,

[1] The expected value of the unmeasured variables may even be some constant that does not change for different values of the measured variables. In that case the constant in the regression equation (β_0 in equation (1)) takes care of the 'constant' part of the error, so that the remaining 'variable' part of the error has an expected value of zero.

the reason that this process does not work in non-linear regressions will be discussed. This is discussed in more generality and as a consequence more complexity in Neuhaus and Jewell (1993) and Neuhaus et al. (1991).

2 Unobserved heterogeneity in a linear model

Figure 1: Unobserved heterogeneity in linear regression



The effect of unobserved heterogeneity (error) in a linear model is discussed here to be able to contrast it with the effect of unobserved heterogeneity in a logistic or other non-linear model, even though it is well known and documented, e.g. (Gujarati, 1995). Assume that there are two variables, x and z, that both influence a variable y according to equation (1). The strengths of the effects of x and z are represented by β_1 and γ respectively. It will be convenient to represent the effect of the unobserved variable (γz) by a singel character, ε, since neither z nor γ are observed. This is done in the second part of equation (1). Assume that half the population has for ε a value of -2 and the other half a value of 2[2], ε and x are uncorrelated and ε is not observed. If we assume that β_0 is zero and β_1 is one, than equation (1) turns into equation (2), which is graphed in figure 1. The solid line represents the relationship between

[2]This assumption about the distribution of ε is for exposition purposes only. The following line of reasoning will work with any distribution whose expected value is zero. Indeed, the standard representation of the effect of error in a linear model assumes that ε is normally distributed with a mean of zero.

x and y if one accurately controls for ε, the 'short-dashed' line represents the relationship between y and x when ε is two, and the 'long-dashed' line represents the relationship between y and x when ε is minus two. If x is zero y should be zero, however for half the population the y is actually minus two (for those whose ε is minus 2), while for the other half the y is two. Still the expected value of y is correct ($.5 \times 2 + .5 \times -2 = 0$). The positive error (the distance between 'b' and 'b+') is as large as the negative error (the distance between 'b' and 'b-'), so they cancel each other out. The same is true for x is two (the distance between 'c' and 'c+' is as large as the distance between 'c' and 'c-'), minus two (the distance between 'a' and 'a+' is as large as the distance between 'a' and 'a-'), or any other value of x

$$y = \beta_0 + \beta_1 x + \gamma z = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

$$y = x + \varepsilon = \begin{cases} x + 2 & \text{if } \varepsilon = 2 \\ x - 2 & \text{if } \varepsilon = -2 \end{cases} \tag{2}$$

3 Unobserved heterogeneity in a logistic model

Now assume that x and z, influence a proportion of success according to the logistic function (3), again replacing γz by ε. Estimating a logistic regression while just ignoring ε will lead to a correct estimate of β_0, but β_1 will be underestimated. This conclusion can be explained with the help of figure 2. Again, the real β_0 is assumed to be zero and the real β_1 is assumed to be one. This means that equation (3) turns into equation (4).

$$\Pr = \frac{e^{\beta_0 + \beta_1 x + \gamma z}}{1 + e^{\beta_0 + \beta_1 x + \gamma z}} = \frac{e^{\beta_0 + \beta_1 x + \varepsilon}}{1 + e^{\beta_0 + \beta_1 x + \varepsilon}} \tag{3}$$

$$\Pr = \frac{e^{x+\varepsilon}}{1 + e^{x+\varepsilon}} = \begin{cases} \frac{e^{x+2}}{1+e^{x+2}} & \text{if } \varepsilon = 2 \\ \frac{e^{x-2}}{1+e^{x-2}} & \text{if } \varepsilon = -2 \end{cases} \tag{4}$$

Logistic regression can be seen as predicting the proportion of successes at different values of x. The solid line in figure 2 represents the relationship between the proportion of successes and x if one adequately controls for ε. The 'short-dashed' line represents the relationship between the proportion of successes and x when ε is two, and the 'long-dashed' line represents the relationship between the proportion of successes and x when ε is minus two (Respectively, the upper and lower equation in equation (4).). When x is zero the proportion of successes should be 'b' (.5), however the the proportion is 'b-' ($\approx .12$) for those individuals whose ε is -2 and 'b+' ($\approx .88$) for those individuals whose ε is +2. The observed proportion of successes is thus correct at this value of x ($.5 \times .12 + .5 \times .88 = .5$). However when x is two the real proportion getting a success should be 'c' ($\approx .88$), while actually half of them have a probability of 'c-' (.5) and the other half has a probability of 'c+' ($\approx .98$). The 'observed' proportion of success at $x = 2$ ('c obs') is $.5 \times .5 + .5 \times .98 = .74$. This is because the distance between c and c- is larger than the distance between c and c+, which is 'squashed' against the maximum proportion of 100%. Consequently, the positive error does completely cancel the negative error. Exactly the opposite happens at x is minus two. The distance between 'a' and 'a+' is larger than the distance between 'a' and 'a-', so the negative error does not completely cancel out the positive error. More generally, the expected proportion of successes at each value of x is shown in equation (5), which is represented by the 'dash-dot' line in

Figure 2: Unobserved heterogeneity in logistic regression

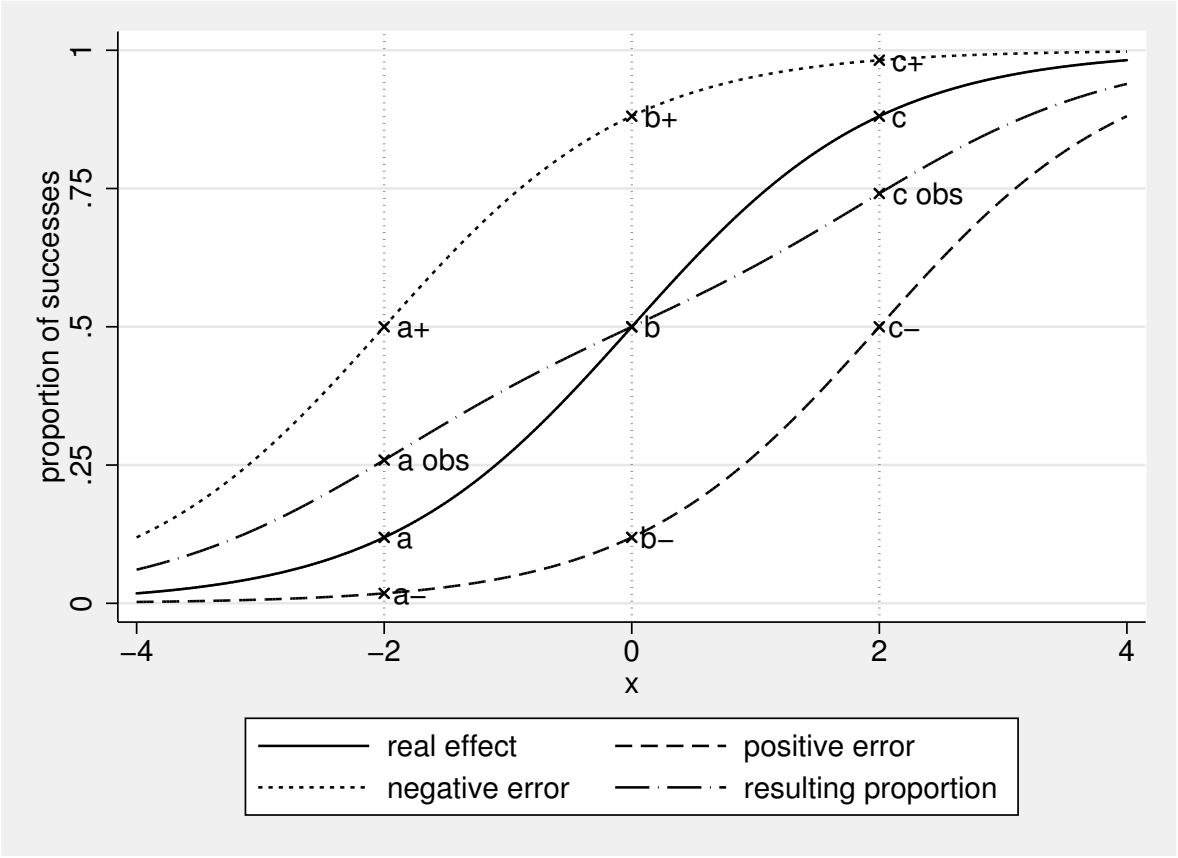


figure 2. When one draws a sample than the 'dash-dot' line represents the distribution from which the individual observations are drawn. Trying to fit a logistic regression without taking into account the unobserved variables is thus the same as trying to fit a logistic line through the 'dash-dot' line. The result is shown in figure 3: It leads to a curve that is too shallow, i.e. an estimated effect of x that is too small. In short, ε causes there to be not enough successes when x is positive and to much successes when x is negative (if x has a positive effect).

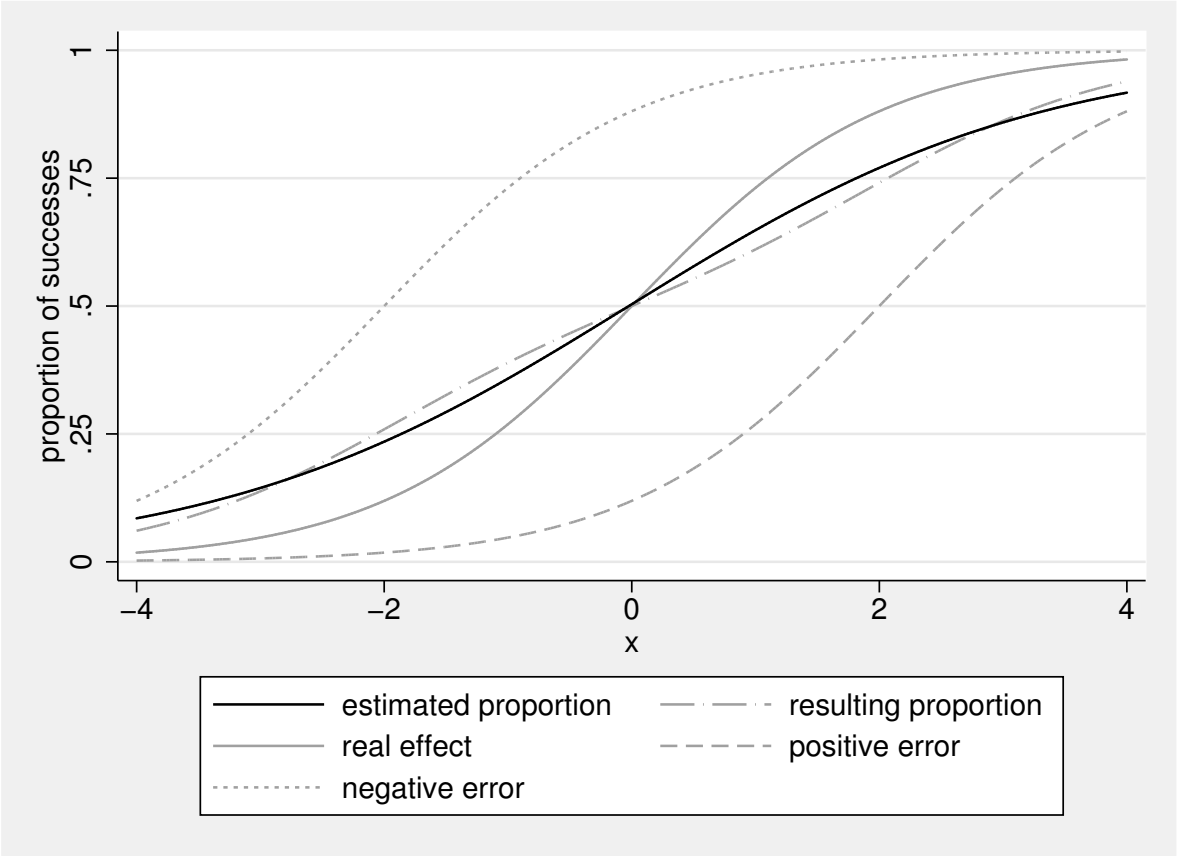
$$\mathrm{E}(\Pr) = .5 \times \frac{e^{x+2}}{1+e^{x+2}} + .5 \times \frac{e^{x-2}}{1+e^{x-2}} \tag{5}$$

Basically, I claim that there is a double stochastic process: there is a stochastic error term (i.e. unobserved variable) in determining the probability of success *and* the way the probability leads to an outcome is stochastic. Normal logistic regression just ignores the first stochastic process, and assumes that all random variation is due to the second stochastic process.

4 Conclusion

In the case logistic regression, the *effect* of an unobserved variable depends on the values of the observed variables even if the the unobserved and observed variables themselves are uncorrelated. Consequently, the proportion of successes and failures at different values of x are

Figure 3: estimating a logistic regression with unobserved heterogeneity



wrong. More specifically, there are too many successes when the observed variables predict a low probability of success, and there are too few successes when the observed variables predict a high probability of success. As a result the effects of the observed variables are underestimated if one does not take the unobserved variables into account.

References

Gujarati, Damodar N., *Basic Econometrics*, third ed., McGraw-Hill, Inc., 1995.

Neuhaus, John M. and Nicholas P. Jewell, "A Geometric Approach to Assess Bias Due to Omited Covariates in Generalized Linear Models," *Biometrika*, December 1993, *80* (4), 807–815.

_ , J.D. Kalbfleisch, and W.W. Hauck, "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, 1991, *59* (1), 25–35.