2 REGRESSION MODELS FOR A DICHOTOMOUS DEPENDENT VARIABLE

2.1 INTRODUCTION

The purpose of this chapter is to remind the reader of the main building blocks of the logistic regression model and introduce the notations that are used in this volume. Especially those elements needed in the following chapters are emphasized. In this way, the necessary background material for a thorough understanding of the comparison issues in core Chapters 3 through 5 are provided. Having some elementary, basic prior knowledge of logistic regression analysis may be advantageous for the reader's understanding of this chapter. Several introductions with somewhat different emphases can be found in the Sage QASS Series (e.g., Menard, 1995, 2002). For an excellent, rather full coverage of the basic logistic regression model, see Long, 1997 and Long and Freese, 2014.

The two main approaches to logistic/probit regression are discussed in Section 2.2 from the viewpoint of DRM (discrete response model) and in Section 2.3 from the LVM (latent variable model) perspective. In Section 2.4, the important disturbing role of mavericks ("'orthogonal' independent variables") in logistic/probit regression, compared to their role in standard regression analysis, is clarified by means of both LVM and DRM.

The logistic regression model as a DRM is introduced and derived in Section 2.2.1. The dependent variable used here is an observed dichotomous variable Y with $0, 1$ scores. The restriction to a dichotomous dependent variable, here and elsewhere, is only made for reasons of convenience (as mentioned in Section 1.2). The probability of scoring $Y = 1$ is called the *response probability*. In logistic DRM, the response probabilities turn out to have the well-known sigmoid, S-curved relationship with the independent variable(s) X.

One way of measuring the effects of X on Y in logistic DRM is to use some form of *change measures* in which the effects of X on Y are essentially measured by means of the difference between the response probabilities for two different values of X. Discrete and instantaneous change measures as well as the evaluation of the shape of the response profile are discussed in Section 2.2.1. However, as shown, due to the nonlinear, S-shaped character of the relation between X and Y, the use of these change measures for the effects of X on Y may become rather complicated and even arbitrary.

An alternative effect measure is presented in Section 2.2.2, where essentially the same logistic regression model is discussed, but now applied to the response odds instead of to the response probabilities. In this way, the logistic regression coefficients themselves can be used in a straightforward way as effect measures and can be given a nice, simple interpretation in terms of odds ratios. The complications inherent to the use of change measures are thus avoided.

In Section 2.2.2, also the relationship between loglinear/logit models and logistic DRM is mentioned. Because of this relationship, the comparison difficulties and solutions encountered in logistic regression also apply to loglinear and logit models.

The DRM probit regression model is presented in Section 2.2.3. The discussion in this section is mainly meant to show that the comparative interpretation problems in probit DRM are essentially the same as in in logistic DRM.

The last DRM model explained in Section 2.2 is the LPM (linear probability model; Section 2.2.4). LPM is a linear-additive regression model for the response probabilities and is often used and recommended as an attractive alternative for logistic (or probit) DRM. Therefore, it is important to see when and how logistic DRM and LPM differ in their estimated response probabilities and effect estimates.

The other main perspective, the LVM approach, is explained in Section 2.3. The dependent variable in LVM is a latent not directly observed continuous variable Y^* that is connected to the observed (dichotomous) variable Y by means of a specific *threshold model*. The substantive interpretation of this latent variable (like the *propensity* to vote) may be somewhat problematic, as is discussed here.

The effects of the independent variables X on the latent variable Y^* are estimated by means of a standard linear-additive regression equation. The usual assumptions are expected to be valid for this underlying regression equation, such as the assumption that the error terms are independently and identically distributed.

In logistic LVM, dealt with in Section 2.3.1, the assumption is made that the error terms in the underlying regression equation for Y^* have a logistic distribution. In this way, it can be derived that the DRM logistic regression coefficients for the effects of X on the observed variable Y provide scaled estimates of the corresponding unstandardized regression effects in the underlying regression equation for Y^*, where the scaling is accomplished regarding the error variance in the underlying regression equation. This scaling can be seen as a form of standardizing the effects, but not, as usual, by setting the variances of the dependent and/or independent variables to one but by fixing the error variance to a particular constant.

The outcomes of the DRM logistic regression equation for Y can also be used to obtain estimates for the usual standardized regression effects in the underlying regression equation for Y^*, where the standardization is achieved in the customary way regarding the variances of $Y^*$and/or X. Finally, it is possible to estimate the proportion of explained variance in Y^* in the underlying regression equation by means of the outcomes of the DRM logistic regression equation for Y.

The probit LVM, presented in Section 2.3.2, is very much like the logistic LVM, but now starting from the assumption that the error terms in the underlying regression equation for Y^* are normally instead of logistically distributed. Therefore, some differences with the outcomes of the logistic LVM arise and are discussed.

Both in logistic LVM and probit LVM, the assumption is made that the error terms in the underlying regression equation for Y^* are homoscedastically distributed. In standard regression analysis, violation of this assumption does not bias the estimates of the unstandardized regression effects (although it does affect the variance and standard errors estimates) and is therefore often ignored. However, in LVM, the logistic or probit effect estimates are seriously biased as scaled estimates of the underlying unstandardized regression coefficients if there is heteroscedasticity in the underlying regression equation for Y^*. This is illustrated in Section 2.3.3.

Simulated Data Set University

Throughout Chapter 2 (and the first part of Chapter 3), a simulated data set *university* is used that is generated by means of a logistic regression equation. The use of a simulated data set has the didactic advantage that the analyses outcomes and interpretations can be compared with the simulated, "true" state of affairs.

The dependent variable in the data set *university* is referred to as the dichotomous variable Attending University – Y: 1 = *attending*; 0 = *not attending* (which is also labeled as U when this is more convenient). The data are supposed to come from two hypothetical countries A and B represented by the dummy variable C (Country) X_1 $(0 = A\ ; 1 = B)$, labeling this variable either as X_1 or as C for the same convenience reasons. There are three additional independent variables, viz. A (Academic background of parents) X_2 (0 = *no*; 1 = *yes*), S (number of Siblings) X_3 $(0 = 0\,or\,1;\ 1 = \ \geq 2)$ and I (Intelligence) X_4. The latter independent variable is a categorized continuous variable with five categories, scored $-21, -8, 0, +8,$ and $+\ 21$. These scores are the quintile scores of the underlying continuous intelligence variable, which is assumed to be normally distributed with mean 0 and standard deviation 15.

The dichotomous independent variables all have a uniform (.50/.50) distribution.

The key data generating equation for this simulation is a main-effects-only logistic equation: All logistic higher order interaction effects on the response

TABLE 2.1 ■ Logistic, Probit, and LPM Effects on University Attendance (*Y* = 1)

		True, Simulated Logistic effects β	Probit effects γ	LPM effects α
		(1)	(2)	(3)
X_1	Country (1 = B)	ln (9) (= 2.197)	1.256	0.310
X_2	Academic (1 = yes)	ln (3) (= 1.099)	0.627	0.152
X_3	Sibling (1 = 2+)	ln (1/3) (= −1.099)	−0.627	−0.152
X_4	Intelligence	ln (11/10) (= 0.095)	0.054	0.013
	Constant	ln (1/10) (= −2.303)	−1.316	0.173

Source: University Data Set (N = 2,000,000)

variable are assumed to be absent. The precise values of the effect parameters used in the simulation are shown in Table 2.1, Column (1) and Eq. (3.1).

The directions of the simulated effects of the independent variables on the dependent one are in agreement with research results on educational attainment (Blake, 1989; Breen & Jonsson, 2005; Skirbekk, 2008; Teachman, 1987). However, their precise strengths as well as the relationships among the independent variables are chosen here for their practical, illustrative usefulness.

Positive direct effects on university attendance are assumed for academic background and intelligence and a negative direct effect for the number of siblings; the proportion university attendance is supposed to be larger in Country B than in A. The independent variables were made statistically independent of each other, except Academic (X_2) and Sibling (X_3) ($r_{AS} = -0.6$, odds ratio $OR_{AS} = 1/16 = .0625$).

The number of cases was set to two million. This enormously large sample size is chosen to reproduce the simulation parameters as close as possible and to overcome rounding errors that mainly occurred when the simulated probabilities were transferred into discrete frequencies.

2.2 DISCRETE RESPONSE MODEL — DRM

In the discrete response modeling (DRM) approach, the research interest concerns the effects of the independent variables on the observed, discrete outcomes of the dependent variable. With a dichotomous dependent variable, the probability is modeled of observing one of the two possible outcomes of Y: $1 = \textit{outcome of interest occurs}$; $0 = \textit{outcome of interest does not occur}$. The probability of $Y = 1$ occurring is called the *response probability*. In the simulated *university* data set, the response probability $\Pr(Y = 1)$ is the probability to attend university.

A simple, straightforward way of modeling a response probability is to estimate it as a linear-additive function of the $k = 1, \cdots, K$ independent variables X_k, each one of them being weighted by its regression coefficient α_k:

$$\Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right) = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_K x_{Ki} = \alpha_0 + \sum_{k=1}^{K} \alpha_k x_{ki} \quad (2.1)$$

To denote the $i = 1,2, \ldots, N$ independent cases in the data, subscript i is used. The realization, the actually observed value of random variable X_k is denoted as x_{ki}. The conditional response probability on the left-hand side of Eq. (2.1) is assumed to be a linear function of the realizations x_{ki}. The (partial) regression coefficient α_k measures the influence of the independent variable X_k on the response probability, controlling for all the other independent variables in the model. The right-hand side of Eq. (2.1) is called the *linear predictor* (indicated below by μ_i).

The model in Eq. (2.1) is known as the linear probability model (LPM; see Section 2.2.4). There are several problems with LPM, such as the estimation problem caused by the heteroscedasticity of the error terms (Long, 1997, pp. 38–40). (There is no explicit error term in Eq. [2.1] because the model is defined in terms of the response probabilities [i.e., in terms of the expected values of Y, the error terms are actually the differences between the estimated response probabilities and the observed 0/1-realizations of Y].)

Next to strictly statistical estimation issues, a problem is that the left-hand side of Eq. (2.1) contains a bounded variable: Probabilities are numbers between 0 and 1, while the linear-additive function on the right-hand side does not guarantee predictions within this unit interval. Another, and probably even more serious problem, is the functional form in Eq. (2.1), which implies that the left-hand side probabilities in- or decrease linearly with the values of the independent variables. However, given the bounded nature of probabilities, bottom and ceiling effects must be expected and S-curved (sigmoid) rather than linear relationships are often much more plausible (Long, 1997, pp. 39, 40; McKelvey & Zavoina, 1975).

To avoid out-of-bound probability predictions, a mathematical function F must be chosen that maps the right-hand side predicted outcomes in Eq. (2.1) onto the (left-hand side) probabilities scale and provides results within the unit interval:

$$\Pr\left(Y_i = 1|X_{1i}, \cdots, X_{Ki}\right) = F\left(\beta_0 + \sum_{k=1}^{K}\beta_k x_{ki}\right) \tag{2.2}$$

An obvious choice of $F(\cdot)$ in Eq. (2.2) is a *cumulative distribution function*, which links a particular value to a probability. By way of example, think of a standard normally distributed random variable Z with mean 0 and standard deviation 1. The probability $\Pr(Z \leq z)$ for any particular cutoff value $Z \leq z$ can be found from the cumulative distribution function $F(Z)$ of the normal distribution: $\Pr(Z \leq z) = F(Z)$.

Usually, the cumulative normal distribution function or the cumulative logistic one is chosen for $F(\cdot)$ in Eq. (2.2). The choice of the cumulative normal distribution leads to the probit model and the choice of the cumulative logistic distribution to the logistic regression model. Both choices imply an S-curved (sigmoidal) rather than a linear relationship between the independent variables and the response probabilities. In this way bottom and ceiling effects are reckoned with.

The logistic distribution looks much like the normal distribution, but is less peaked and with somewhat heavier tails (Long, 1997, p. 43). Hence, the predicted conditional response probabilities from the logistic and probit regression are usually very close, with somewhat larger deviations to be expected for response probabilities close to 1 or 0. The probit regression effects are about a factor 1.7 smaller than the corresponding logistic effects (as further explained in Sections 2.2.3 and 2.3.2). The cumulative logistic distribution is perhaps the more popular choice because it is mathematically easier to handle than the cumulative normal distribution and because the logistic regression coefficients have a simple direct interpretation regarding odds ratios. On the other hand, the probit model is sometimes easier to integrate into models assuming normally distributed variables (see also Liao, 1994, pp. 24, 25; Long, 1997, p. 83). There is also a "disciplinary" flavor to the preference for the probit or the logistic model; for example, economic researchers often automatically apply the probit and sociologists the logistic model.

2.2.1 Logistic Regression, Response Profiles, Discrete (*DC*), and Instantaneous (*IC*) Change Measures

Using the cumulative logistic distribution function $F(\mu_i) = \exp(\mu_i)/\left(1 + \exp(\mu_i)\right)$ to map the linear predictor μ_i onto the probability scale, the LPM in Eq. (2.1) turns into the following nonlinear logistic regression model for the response probabilities:

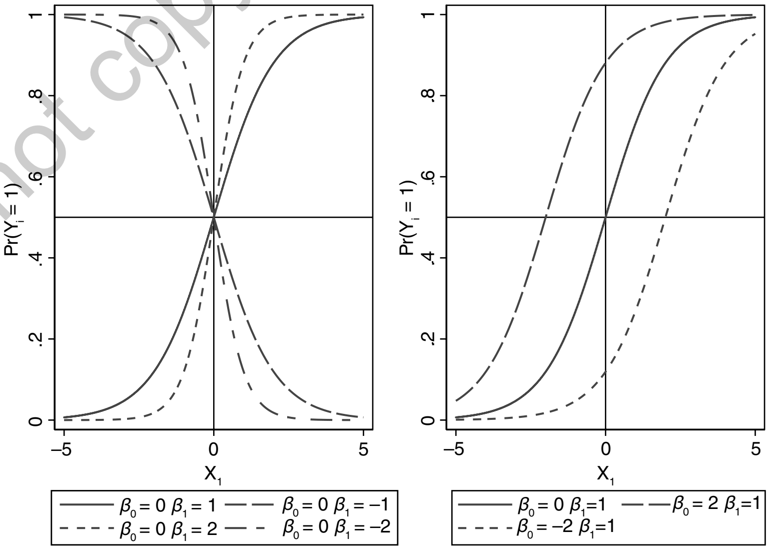
$$\Pr\left(Y_i = 1|X_{1i}, \cdots, X_{Ki}\right) = \frac{\exp\left(\beta_0 + \sum_{k=1}^{K}\beta_k x_{ki}\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{K}\beta_k x_{ki}\right)} \tag{2.3}$$

From here on, the β-coefficients are always used as the symbols for the logistic effects.

The left-hand side probability in Eq. (2.3) can be interpreted as the chance that a randomly chosen subject i scores $Y = 1$ given the scores on the independent variables $X_1, \cdots, X_K$; it can be estimated by means of the corresponding proportion in the sample. Estimates of the logistic regression coefficients β can be found using MLE (maximum likelihood estimation), which also provides the (co)variances estimates of the needed for the estimation of confidence intervals and the application of statistical tests. Moreover, the value of the maximized likelihood can be used to compute overall measures of model fit and to test the significance of nested models against each other (Long, 1997; for more details on MLE see Eliason, 1993).

The S-shaped form of the relation between the conditional response probabilities $\Pr(Y_i = 1|X_{1i}\ldots X_{Ki})$ and the independent variables, implied by the logistic regression model in Eq. (2.3) is illustrated in Figure 2.1 for a simple logistic model with one independent variable X_1 varying between −5 and +5. The response curves or response profiles are presented for a few different values of the intercept β_0 and the effect β_1. (For a more general treatment of response profiles, see Long, 1997, Section 3.7.) It is clearly seen in Figure 2.1 that with de- or increasing values of X_1, away from the middle, the response probability $\Pr(Y_i = 1|X_{1i})$ gets closer and closer to 0 or 1. However, it will never surpass these boundaries. Further, as seen in Figure 2.1, a unit change

FIGURE 2.1 ■ The Logistic Regression Function



in X_1 leads to a larger difference in the response probabilities in the middle range of X_1 values than in the lower or upper range.

In general, when the strength of the effect an independent variable like X_1 has on the response probability is measured by means of the differences in the response probabilities, the model in Eq. (2.3) implies that the strength of the effect varies with the level of $\Pr(Y_i = 1|X_1)$: it is largest around the conditional response probabilities of .50 where the response curve is steepest and it gets smaller when the conditional response probabilities go to 0 or 1 where the response curve becomes flatter.

As also seen in Figure 2.1, the response curve has an approximately linear shape when the (estimated) response probabilities all lie within the .30 − .70 range. Within this range, the steepness of the response curve is more or less constant (Goodman, 1976, p. 92; Long, 1997, p. 64). (Sometimes less conservative ranges are proposed: .25 − .75 or even .20 − .80.)

How much the (expected) response probabilities change for different values of an independent variable, say X_1, can be quantified by means of the *discrete change coefficients DC*. The *DC* coefficients are defined as the difference of the two (expected) conditional response probabilities that correspond with two different values of the independent variable of interest, controlling for other independent variables where appropriate (Long, 1997, pp. 75–79, 137–138, 166–167; Long & Mustillo, 2018):

$$\begin{aligned} DC = {} & \Pr(Y_i = 1|X_{1i} = x_1 + \delta, X_{2i}, \cdots, X_{Ki}) \\ & -\Pr(Y_i = 1|X_{1i} = x_1, X_{2i}, \cdots, X_{Ki}) \end{aligned} \tag{2.4}$$

The symbol δ in Eq. (2.4) indicates the difference between the two chosen values of X_1; the quantity (100)(*DC*) yields the familiar percentage difference as effect measure, often symbolized by *d*% or ε.

According to Eq. (2.4), the *DC* effect of X_1 on the response probability is obtained by subtracting the (expected) response probability for $X_1 = x_1$ from the (expected) response probability for another value $x_1' = x_1 + \delta$, while controlling for the remaining independent variables. This "controlling" is done by conditioning on particular values for the remaining independent variables. Preferably, the values chosen for X_1 and the values to which the remaining independent variables are set will be chosen for their theoretical or practical relevance. However, often a more pragmatic choice is made, for example, by investigating the *DC* effect(s) of X_1 for the means (or medians or modes) of the remaining independent variables.

The obtained value of *DC* as a measure of the direct effect of a particular independent variable X_k, on the response probability may depend strongly on the particular choice of the value x_k (for a given value of δ). A *DC* effect of X_k will generally have a different value when a different value x'_k for X_k is chosen instead of x_k while keeping δ (in Eq. [2.4]) the same. The choice of the values pair $x_k$and $x_k + \delta$will deliver the same sign for the DC effect of X_k on the response probability as the pair $x'_k$and $x'_k + \delta$but their sizes will generally be different. This is a direct consequence of the S-curved response profile.

Moreover, the strength of the DC effect of X_k may also strongly vary with the choices for the selected conditioning values of the other independent variables. This is certainly true when estimated response probabilities are involved with values outside the .30 − .70 range. This also follows from the S-shape of the response profile. Say variable X_c has a direct effect on the response probability and is the independent variable to be controlled. Then, the conditional response probability for a given value $X_k = x_k$ will be generally different when control variable X_c is set to $X_c = x_c$ rather than to another value $X_c = x'_c$. The response probability, given $X_k = x_k$, will be closer to or further away from .50 when $X_c = x_c$ than when $X_c = x'_c$. Therefore, the DC effect of X_k estimated for the same two values of X_k ($X_k = x_k$ and $X_k = x'_k$) will generally be different, larger or smaller, for $X_c = x_c$ compared to $X_c = x'_c$ (although it will have the same sign).

In other words, the effects of X_k on the response probability will interact with X_c when the effect of X_k is quantified in terms of DC and this despite the fact that in the basic logistic equation Eq. (2.3) no explicit interaction effects were inserted.

The simulated *university* data nicely demonstrates the above remarks on DCs. Because this data set was created by means of the main-effects-only logistic model in Eq. (2.3) with the logistic effects given in Column (1) of Table 2.1, the response profile representing the direct effect of a particular independent variable on the response probability follows an S-curve (although, as in many empirical applications, only a segment of the complete S-curve is realized). The estimated (simulated) conditional response probabilities of attending university are presented in Table 2.2 for three categories of X_4-Intelligence: the lowest score (-21), the middle (0) and the highest (21), conditioned on the combined

TABLE 2.2 ■ Response Probabilities of Attending University

X_4-Intelligence		−21		0		21
Favorable conditions: $X_1 = X_2 = 1; X_3 = 0$		0.267		0.730		0.952
	DC		0.462		0.222	
Unfavorable conditions: $X_1 = X_2 = 0; X_3 = 1$		0.004		0.032		0.198
	DC		0.028		0.166	

Source: Simulated data set set *university* scores on the remaining independent variables in terms of favorable versus unfavorable for attending university. The favorable condition is the combination of Country B ($X_1 = 1$), Academic background ($X_2 = 1$), and 0 or 1 Sibling ($X_3 = 0$), and unfavorable, otherwise.

Given a favorable condition, someone with a middle intelligence score (0) has a .462 higher chance of attending university than someone with the lowest intelligence score (-21): $DC = .730 - .267 = .462$. Under the same favorable condition, an equally sized jump δ but now from intelligence 0 to 21 improves the university chances by a smaller amount: $DC = .952 - .730 = .223$. This follows from the S-curved response profile. The effect $DC = .462$ involves the response probabilities .267 and .730, which are symmetrically arranged around the middle value .50, a range in which the response profile is steepest. The effect $DC = .223$ on the other hand is the difference between the response probabilities .730 and .952, which are closer to the endpoint 1 where the response profile is less steep.

Looking at the comparable outcomes but now for the unfavorable condition, belonging to the middle intelligence category 0 instead of the lowest –21 hardly improves one's chances of attending university ($DC = .028$), while belonging to the highest category instead of the middle improves one's chances by $DC = .166$. Again, this is in agreement with the response probabilities for $DC = .028$ being closer to the endpoint 0 than the ones for $DC = .166$.

Looked at from another angle, these same outcomes also show a strong *DC* interaction effect of intelligence and the favorable/unfavorable conditions on the response probabilities. Under the favorable condition, having a middle instead of the lowest intelligence score has an effect on university attendance of $DC = .462$, while under the unfavorable conditions there is hardly any effect: $DC = .028$. Again, this is understandable from the fact that the favorable conditions move the response probabilities upward toward .50, while the unfavorable conditions move them strongly toward 0. The change from middle to highest intelligence has more or less the same effect under the favorable ($DC = .223$) and the unfavorable condition ($DC = .166$), in agreement with the fact that both effects pertain to response probabilities at the (opposite) tails of the (symmetrical) distribution.

In an actual research project, the obvious next step would be to find substantive explanations for these interaction effects. However, how do these outcomes and possible interpretations then relate to the fact that that the *university* data was constructed using a logistic regression main-effects-only model without any interaction effects? Must this interaction effect not be regarded, in the words of Stinchcombe (1983), as an instance of a "spurious interaction effect" (p. 107). This issue is extensively discussed next in several places.

The need to address this issue is the more urgent if one realizes that the simulated data example *university* is a rather simple one. Often much more variables are involved, many of them with much more categories. Unless all conditional response probabilities range between .30 and .70 where the response profile is more or less linear or unless the conditioning independent variables have only weak effects on the response probabilities, a very large number of (partial, conditional) response profiles and *DC*s are needed to capture adequately the outcomes of a logistic regression equation and the effects of the independent variables on the response probabilities.

These remarks and questions regarding the use of *DC*s apply similarly to the use of *IC*s, the *instantaneous change measures*. Where the discrete change coefficients *DC* can be applied to all kinds of independent variables, whether they are continuous or discrete and measured at nominal, ordinal, or interval level, the instantaneous change measure *IC* can only be used meaningfully if an independent variable X_k is continuous. An *IC* indicates the instantaneous change in the response probability at a particular point on the X_k-axis. *IC* reflects the steepness of the S-shaped response curve at $X_k = x_k$ and, geometrically speaking, equals the slope of the tangent that touches the response curve at $X_k = x_k$. *IC*s can be computed by evaluating the (partial) derivative of the response probability with respect to the pertinent continuous independent variable X_k. Readers not familiar with calculus can think of the derivative as the limiting value when δ in Eq. (2.4) becomes infinitely small. The difference $\Delta X_k = \left[(x_k + \delta) - x_k\right]$ equals ∂X_k when δ becomes infinitely small, in other words when the change in X_k approaches 0.

For the logistic regression model in Eq. (2.3), the partial derivative of the response probability with respect to a continuous independent variable, say X_k, equals (see Long, 1997, Section 3.7.4):

$$\begin{aligned}\frac{\partial \Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right)}{\partial X_k} &= \frac{\exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\right)}{\left(1 + \exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\right)\right)^2} \cdot \beta_k \\ &= \frac{\exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\right)}{1 + \exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\right)} \\ &\quad \cdot \frac{1}{1 + \exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\right)} \cdot \beta_k \\ &= \Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right) \\ &\quad \cdot \left(1 - \Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right)\right) \cdot \beta_k\end{aligned} \tag{2.5}$$

In econometric (and more recently in social science) vernacular, the partial derivative in Eq. (2.5) is called a *marginal effect* (Long, 1997, p. 5). For some scholars who distinguish between partial and marginal effects in terms of controlling or not controlling for other independent variables, this terminology may be confusing. The term marginal effect does not refer here to an effect in a lower order distribution. For example, it does not refer to an effect in marginal table *AB*, obtained by collapsing the joint distribution *ABC* over *C*. Marginal in this context simply means that the interest lies in the expected change of the response probability when the independent variable changes by an infinitely small amount, controlling for all the other independent variables in the model. In that sense, these marginal effects are partial effects. The terms marginal effects, instantaneous change coefficients, and (partial) derivatives, then, all refer here to essentially the same phenomenon. Where the meaning of the term marginal is ambiguous, it is clarified.

In a linear-additive model like the LPM in Eq. (2.1), the partial derivative, analogous to Eq. (2.5), is equal to α_k– the direct (partial) effect coefficient for X_k (see Eq. [2.14] below). For the nonlinear logistic regression model in Eq. (2.3), however, as is obvious from Eq. (2.5), the instantaneous change of the response probability due to X_k is not only a function of the effect β_k, but it also depends on which value of X_k is considered and to which values the other independent variables have been set. The partial derivative in Eq. (2.5) not only includes X_k's effect β_k, but also the effects and values of all independent variables in the model; they are all in the term $\exp\left(\beta_0 + \sum_{k=1}^{K} \beta_k x_{ki}\right)$ in Eq. (2.5) determining the partial derivative. As such, the instantaneous change in the response probability due to X_k not only depends on β_k, but also on the value chosen for X_k, on the chosen values of the remaining independent variables, and on what direct effects these remaining independent variables have on the response probability (as was shown to be true for *DC*).

The last line in Eq. (2.5) shows that the (instantaneous) change of the response probability is larger, the larger the product of the conditional response probability with its conditional converse probability is. This product $\Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right) \cdot \left(1 - \Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right)\right)$ equals the (conditional) variance of Y and so *IC* is largest when $\Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right) = 0.5$ and gets smaller when $\Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right)$ approaches 0 or 1 (as was shown for *DC*). For response probabilities within the range $.30 - .70$, this conditional variance (this product) is more or less constant, for example, $(.50)(.50) = .25$ and $(.70)(.30) = .21$, but not outside this range, for example, $(.95)(.05) = .0475$. Because of the multiplication factor in Eq. (2.5), β_k is multiplied by a factor 5.2 times larger when the conditional response probability equals .50 compared to .95.

Conditional response profiles and *DC*s and *IC*s are especially useful when there are not too many variables, when there are not too many categories for each independent variable, or when there are obvious theoretical or practical reasons to justify the scoring choices to be made (as excellently exemplified by Long & Mustillo, 2018).

Sometimes in their wish to have one simple *DC* or *IC* type of effect measure, researchers apply pragmatic solutions, where the slope of the response curve and the *DC*s and *IC*s are evaluated for a few meaningful values of X_k, but then setting all remaining independent variables at their mean values. Or instead of the mean, other "pragmatic values" are chosen, such as the mean-plus/minus-one-standard-deviation, the median value, or the first or third quartile (see Long, 1997, pp. 74–79, for "pragmatic" solutions for both *DC* and *IC*).

Two often used "one-effect-solutions" for *IC* are *MEM* (*marginal effect at the mean*) and *AME* (*average marginal effect*). For *MEM*, the instantaneous effect of X_k is evaluated at its mean while setting all remaining independent variables also at their mean values.

AME is defined as the instantaneous effect of X_k averaged over all observations, which amounts to:

$$\text{AME}(X_k) = \beta_k \cdot \frac{\sum_{i=1}^{N} \Pr(Y_i = 1 | X_{1i}, \cdots, X_{Ki}) \cdot (1 - \Pr(Y_i = 1 | X_{1i}, \cdots, X_{Ki}))}{N} \quad (2.6)$$

A variant of *AME* is the average partial effect (*APE*). The same Eq. (2.6) is used but then applied to subgroups that have a specific value or a specific range of values on X_k (Karlson et al., 2012, pp. 298–302; Mood, 2010, p. 75).

MEM and *AME* are different measures, having different meanings, and will often deliver different outcomes, and especially, but most importantly, both involve often a rather arbitrary choice from the many possibilities (Long, 1997, pp. 74–75).

A single *DC* and *IC* may present rather distorting pictures of the effects in the data, certainly when more variables and more categories are involved than in our simulated data set and when there are no clear theoretical reasons to choose which one of the many possible (but needed) *DC*s or *IC*s will provide the best answer to the research question.

In the next section, an alternative view on effects in a logistic regression equation is discussed, giving meaning to the β-parameters as effect measures.

2.2.2 Logistic DRM as a Logit Model: Odds Ratios as Effect Measures

So far, the logistic DRM has been mainly described as a "transformed" standard regression model for the response probabilities. Starting point was the additive regression or LPM for the response probabilities in Eq. (2.1). To deal with possible bottom and ceiling effects and with out-of-range predicted response probabilities, the linear predictor μ_i in Eq. (2.1) was mapped onto the probability scale by means of the cumulative logistic distribution function $F(\mu_i) = \exp(\mu_i) / \left(1 + \exp(\mu_i)\right)$, resulting in Eq. (2.3). Analogous to what is common practice in standard regression analysis, the effect of the independent variable was then evaluated by estimating how the dependent variable is expected to change due to a unit change in the independent variable, controlling for the other independent variables in the logistic regression equation.

In the additive regression equation Eq. (2.1), this difference in expected response probabilities for a unit change in X_k is simply and directly rendered by the effect coefficient α_k. However, as seen previously, in logistic regression, all kinds of complications arise when using *DC* and *IC* type effect measures and there is no such simple direct relation between *DC*/*IC* type measures and the logistic effects β_k.

Fortunately, there exists a rather straightforward interpretation of the logistic regression coefficients themselves that is not affected by these complexities. Instead of looking at the consequences of the changes in X_k for the response probabilities, it is investigated what the consequences of the (same) logistic regression model in Eq. (2.3) implies for the response odds. As it turns out, the response odds are a much simpler function of the independent variables with weights β_k where the β_ks can be simply interpreted in terms of odds ratios.

Odds, symbolized here by Ω, are ratios of probabilities. For example, if the variable Voting has three categories: *R(epublican), D(emocrat), I(ndependent)*, the odds of voting *R* rather than *D* equals the ratio $\Pr(R)/Pr(D)$ and the odds of voting *D* rather than *I* are $Pr(D)/Pr(I)$, and so on. In the (dichotomous) *university* example, the response odds of attending university rather than not are defined as the response probability of attending university $Pr(U)$ divided by the converse probability of not attending university $\Pr(U)/(1 - Pr(U))$. The logarithm of the odds is called the logodds or a *logit.* (Note that specifically in STATA, the term odds is not used for a ratio like $Pr(D)/Pr(I)$ but only for $Pr(D)/(1 - Pr(D)$. Odds like $Pr(D)/Pr(I)$ are called the *relative risk*. For the effect measure using the relative risk, the term relative risk ratio is used instead of odds ratio. We do not follow this specific use of the relative risk ratio here).

An effect of an independent variable can be expressed in terms of a ratio between odds, called the *odds ratio*, symbolized by *OR*. For the *university* data, if a researcher wants to know whether there is a relationship between Academic and University, the odds of attending university rather than not are computed among those with an academic home background and among those without an academic home background. Next, the ratio between these two conditional odds is found. This odds ratio expresses the effect of Academic on University by telling how many times larger or smaller the odds of going to university are among Academics compared to the Nonacademics (Agresti, 2002, Chapter 2; Hagenaars, 1990, Chapter 2; Rudas, 1998).

Let the odds $\Omega^{Y|\boldsymbol{X}}_{1/0}$ stand for the conditional response odds that $Y = 1$ rather than $Y = 0$, conditional on a combination of values of the set of independent variables $X_1, \cdots, X_K$ indicated by $\boldsymbol{X}$. The logistic regression model in Eq. (2.3) can be rewritten in such a way that the response odds $\Omega_{1/0}^{Y|\boldsymbol{X}}$ are a multiplicative function of the logistic regression coefficients (where the simplification of the right-hand side of the equation is easily understood from $\frac{a}{1+a}/\left(1-\frac{a}{1+a}\right)=a$):

$$\begin{aligned}\Omega_{1/0}^{Y|\boldsymbol{X}} &= \frac{\Pr(Y_i=1|X_{1i},\cdots,X_{Ki})}{(1-\Pr(Y_i=1|X_{1i},\cdots,X_{Ki}))}\\ &= \frac{\dfrac{\exp\left(\beta_0+\sum_{k=1}^{K}\beta_k x_{ki}\right)}{1+\exp\left(\beta_0+\sum_{k=1}^{K}\beta_k x_{ki}\right)}}{1-\dfrac{\exp\left(\beta_0+\sum_{k=1}^{K}\beta_k x_{ki}\right)}{1+\exp\left(\beta_0+\sum_{k=1}^{K}\beta_k x_{ki}\right)}}\\ &= \exp\left(\beta_0+\textstyle\sum_{k=1}^{K}\beta_k x_{ki}\right)\\ &= \exp(\beta_0)\cdot\exp\left(\beta_1\right)^{x_{1i}}\cdot\ldots\cdot\exp\left(\beta_K\right)^{x_{Ki}}\end{aligned} \tag{2.7}$$

The terms $\exp(\beta_0)$, $\exp(\beta_1)$, and so on are the multiplicative effects on the response odds and are the *anti-logarithms* (the exponential functions) of the logistic regression coefficients β_k.

The logistic regression model in Eq. (2.7) looks even simpler in its logarithmic form in which the logs of the response odds, the response logits, are a familiar linear-additive function of the independent variables:

$$\begin{aligned}\ln\left(\Omega_{1/0}^{Y|\boldsymbol{X}}\right) &= \ln\left(\frac{\Pr(Y_i=1|X_{1i},\cdots,X_{Ki})}{1-\Pr(Y_i=1|X_{1i},\cdots,X_{Ki})}\right)\\ &= \operatorname{logit}(\Pr(Y_i=1|X_{1i},\cdots,X_{Ki}))=\beta_0+\beta_1x_{1i}+\cdots+\beta_Kx_{Ki}\end{aligned} \tag{2.8}$$

The partial derivative of the response logit for the effect of X_k simply equals β_k:

$$\frac{\ln\left(\Omega_{1/0}^{Y|\boldsymbol{X}}\right)}{\partial X_k}=\frac{\partial\operatorname{logit}\Pr(Y_i=1|X_{1i},\cdots,X_{Ki})}{\partial X_k}=\beta_k \tag{2.9}$$

The logistic regression model in Eqs. (2.7) and (2.8) is a main-effects-only model. There are no interaction effects in the sense that the direct, partial effect β_k of independent variable X_k on the response logit is the same for all values on the remaining independent variables (as also seen in Eq. [2.9]). Further, the response logits increase linearly with the values of X_k: For each additional unit increase of X_k, the response logits increase additionally by a constant β_k. So

Eq. (2.8) represents a linear-additive model for the response logits (and the logistic regression model is an instance of the generalized linear model).

The corresponding multiplicative counterpart in Eq. (2.7) is of course also a main-effects-only model. *Linearity* now implies that the response odds increase by a constant factor: For each additional unit increase of X_k, the odds increase additionally by a constant factor $\exp(\beta_k)$.

As in standard linear regression, nonlinear and interaction effects on the response logits can be introduced in Eq (2.8) by using for the independent variables polynomials, dummy variables and product terms (Jaccard, 2001; Jaccard & Turisi, 2003).

The independent variables in Eq. (2.8) can be continuous or discrete or dummy variables (as in standard regression analysis). If all variables (in Eq. [2.8]) are treated as discrete or categorical, the resulting model is the *(categorical) logit model* (or the loglinear model, given the correspondence between loglinear and logit models; Agresti, 2002; Demaris, 1992; Hagenaars 1990; Knoke & Burke, 1980). The maximum likelihood estimates for the categorical logit model are exactly the same as the ones obtained for the previous logistic regression model. This means that computer programs for and the vast literature on the varieties of categorical loglinear and logit models can also be applied for handling and interpreting DRM logistic regression (and vice versa). It also means that the comparison issues for the DRM logistic regression model discussed next apply similarly to the (categorical) logit and loglinear models.

Regarding terminology, the term logistic regression model usually implies that the independent variables in the logistic regression equation are in principle, but not necessarily seen as, continuous variables. The term logit model is mostly used to imply that the independent variables are treated as categorical variables. However, the terms logistic regression and logit regression are also used in the literature (and here) to refer to the same regression model.

The simulated *university* data set can be used to illustrate the interpretation of the β effects in terms of odds ratios. As said in the beginning of this chapter, this data set has been constructed by means of the main-effects-only logistic regression equation in Eq. (2.8) using the parameter values presented in Column (1) of Table 2.1.

First, the interpretation of the intercept. As easily concluded from Eq. (2.8), the intercept β_0 is the value of the response logit for those research units that score 0 on all independent variables. As in standard linear regression, this may be a *reference set R* of individuals that may be empirically empty or may not really exist because one or more independent variables do not have a score 0. In the *university* example, this reference group R consists of individuals from Country A, with average (middle) intelligence, no academic background, and not more than one sibling. The response odds for this reference group turn out to be $\Omega_{1/0}^{Y|R} = \exp(\beta_0) = \exp(-2.303) = 0.100$ (Table 2.1). The response probability, that is, the probability of attending university, are in the reference group 10 times smaller than the probability of not attending university, viz. .0909 versus .9090. Note that the response probability can be computed from the response odds, for example, in the reference group *R*:

$$\Pr(Y = 1|R) = \Omega_{1/0}^{Y|R} / \left(1 + \Omega_{1/0}^{Y|R}\right) \tag{2.10}$$

The logistic effects β_k show the effects of the independent variables. As follows from Eq. (2.8), the effect β_1 of X_1 (Country) indicates how much the response logit in- or decreases when Country increases by one measurement unit (i.e., when people live in Country B [score 1] instead of Country A [score 0] and all other independent variables are held constant). The logistic effect of Country is $\beta_1 = 2.197$. Due to the direct effect of Country, the response logit is 2.197 larger in Country B than in A. Regarding the response odds, the odds of attending university rather than not are exp(2.197) = 9.000 times larger in Country B than in A. This is the partial odds ratio *OR*, representing the direct partial effect of Country on the response odds. Comparing individuals with the same characteristics on the remaining independent variables, the odds of going to university are much more, viz., nine times more favorable for individuals living in Country B than for those living in A.

This *OR* interpretation of exp(β) applies in general. The multiplicative direct effect $\exp(\beta_k)$ is the partial *OR* that tells by what factor the response odds change if X_k increases by one measurement unit while all remaining independent variables are constant. For example, Intelligence X_4 was given the scores –21, –8, 0, 8, 21, and $\beta_4 = .095$. So for each unit increase in intelligence, wherever on the Intelligence scale, the response logit increases by .095. Given the scores for Intelligence, this means that going from the middle Intelligence score 0 to the next one 8, there is an increase of $(8 - 0)$ Intelligence units and so an expected increase in the response logit of $(8 \cdot .095) = .760$. Regarding the response odds, the multiplicative effect *OR* for one unit increase in Intelligence equals $\exp(.095) = 1.100$. Therefore, given an Intelligence increase from 0 to 8, the response odds increase by a factor: $1.100^8 = 2.144\,(= \exp(.760)$, ignoring rounding errors, holding the remaining independent variables constant.

At first sight, it may seem strange that the same data sets, when analyzed in different ways, leads to such seemingly different conclusions. On the one hand, there is the main-effects-only, additive-linear logistic regression model in Eqs. (2.8) and (2.7) in which there are no interaction effects of the independent variables on the response logits (or odds). On the other hand, when the outcomes of this main-effect logistic regression model are analyzed in terms of response probabilities and *DC*/*IC*s, S-shaped relationships and interaction effects appear. Of course, both analyses and conclusions are mathematically correct: A linear effect on the response logits logically/mathematically implies a particular curvilinear relationship on the response probabilities and also leads to particular interaction effects in terms of *DCs*, *IC*s (and vice versa); different dependent left-hand side elements in the equations are involved, viz. response odds versus response probabilities. Nevertheless, the choice of the effect measure (*DC* or *OR*) may lead to different substantive conclusions from the research outcomes about the effects of the independent variables. More on how to deal with this and on the interpretation of interaction effects is presented in Sections 4.3 and 6.2.

Because the βs are the *natural* effect coefficients in the DRM logistic regression equations and given their parsimonious and direct interpretation in terms of odds and odds ratios, βs and *OR*s are the central effect measures in the next three chapters on the comparative issues for logistic DRM (although some comparisons with *DC*/*IC*–type additive effect measures are made).

2.2.3 Probit Regression

So far, the logistic distribution function has been used for the cumulative distribution function in Eq. (2.2). As an alternative, the cumulative standard normal distribution $F(z) = \Phi(z)$ was mentioned, giving rise to the probit model. Given the similar shape of the logistic and the normal distribution, the estimates of the conditional response probabilities in the probit regression model are rather close to those obtained by the logit regression model, except for estimates very close to 0 or 1. Also the comparative issues involved are very much the same for the probit and logit model.

The main disadvantage of the probit model is that the probit regression coefficients do not have the simple substantive interpretation in terms of odds ratios or something similar, which exists for the logistic regression coefficients. No simple transformation of probits exists that have a simpler interpretation than the probit itself. This means, given the use of the cumulative standard normal distribution, that the probit effects can be interpreted in terms of the expected change in the *z*-scores of the categorical dependent variable due to a unit change in the independent variable. However, for dichotomous dependent variables, such as attending university or not, and for categorical variables in general, the use of *z*-scores usually does not make much theoretical, substantive sense. Hence, researchers tend to use the predicted response probabilities that correspond to the predicted *z*-scores to gauge the effects in probit regression. Mostly, the *DC* and *IC* measure described earlier are being used as effect measures, with the same advantages and complications described for the logistic regression model.

In the probit DRM, the response probability is parameterized as a function of the independent variables in the following way:

$$\begin{aligned}\Pr\left(Y_i = 1|X_{1i}, \cdots, X_{Ki}\right) &= \Phi\left(\gamma_0 + \textstyle\sum_{k=1}^{K} \gamma_k x_{ki}\right) \\ &= \frac{1}{2\pi} \int_{-\infty}^{\gamma_0 + \sum_{k=1}^{K} \gamma_k x_{ki}} \exp\left(-0.5\, t^2\right) \mathrm{d}t\end{aligned} \tag{2.11}$$

Eq. (2.11) yields an S-shaped response curve very similar to the response curve for the logistic model, although less heavy in the tails of the distribution (Long, 1997, p. 43). The linear predictor $\mu_i = \gamma_0 + \sum_{k=1}^{K} \gamma_k x_{ki}$ is a linear-additive function of the independent variables with the probit regression coefficients denoted by the Greek letter γ. The probit effects γ_k can be estimated by means of MLE methods.

Similar to the logit function (Eq. [2.8]) being the inverse of the cumulative standard logistic distribution function with mean 0 and variance $\pi^2/3$, the probit function is the inverse of the cumulative standard normal distribution. (Technically, multiply the right-hand and left-hand sides of Eq. [2.11] by Φ^{-1}.) And as the logit DRM, the probit DRM is a linear-additive model in terms of the probits:

$$\operatorname{probit}\left(\Pr\left(Y_i = 1|X_{1i}, \cdots, X_{Ki}\right)\right) = \gamma_0 + \gamma_1 x_{1i} + \cdots + \gamma_K x_{Ki} \tag{2.12}$$

Column (2) of Table 2.1 shows the estimated probit regression coefficients for the main effects model, applied to the *university* data set. They are systematically smaller in absolute size than the logistic regression coefficients, which is true in general. There are various proposals to make them more comparable, but all those proposals lead to more or less the result: $\beta \approx 1.7\gamma$ (Long, 1997, p. 48). In Section 2.3.2, the logic behind the factor 1.7 is explained. In Table 2.1, the logistic effects β are all close to 1.75 times larger than the corresponding probit effects γ. As can be seen from the STATA output on the chapter's webpage, the estimated response probabilities for the probit model are very much like the response probabilities for the logistic model.

When using the response profiles and *IC* or *DC* as effect measures in probit regression, the (marginal) effect of a particular independent variable X_k is largest in the center of the multivariate distribution and gets smaller toward its tails, as follows from the standard normal density being the largest in the center of the distribution (around $Y_i = 1|X_{1i}, \cdots, X_{Ki} = .50$). Therefore, the sizes of *IC* and *DC* as effect measures for a particular independent variable depend in general, as for the logistic regression model, on the values chosen for that particular variable and on the values and effects of the other independent variables. Nonlinearities and interaction *DC*/*IC* effects will appear, even in the main-effects-only probit regression model in Eqs. (2.11) and (2.12). This is formally shown in Eq. (2.13) for the partial derivative (*IC*, the marginal effect) with $f(z) = \phi(z)$ being the standard normal density:

$$\begin{aligned}\frac{\partial \Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right)}{\partial X_k} &= \phi\left(\gamma_0 + \sum_{k=1}^{K} \gamma_k x_{ki}\right) \cdot \gamma_k \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\left\{\gamma_0 + \sum_{k=1}^{K} \gamma_k x_{ki}\right\}^2}{2}\right) \cdot \gamma_k\end{aligned} \tag{2.13}$$

And of course, from Eq. (2.13), it is easily inferred that the appearance of nonlinearities and interaction effects are also true for the *DC* effect measures.

2.2.4 Linear Probability Model – LPM

The linear probability model was introduced in Eq. (2.1) as a first obvious and attractive choice to model response probabilities. It is often considered as a serious alternative to logit or probit regression. LPM in Eq. (2.1) is a main-effects-only additive-linear regression model, but now for the response probabilities and not for the response logits or probits. The partial derivative for the LPM simply amounts to:

$$\frac{\partial \Pr\left(Y_i = 1 | X_{1i}, \cdots, X_{Ki}\right)}{\partial X_k} = \alpha_k \tag{2.14}$$

The α_k effects in LPM are identical to the familiar unstandardized regression coefficients from a standard regression analysis (with dichotomous dependent variable *Y*). They are also identical to the *DC*s and *IC*s discussed earlier for a unit increase in the independent variable, but without the mentioned complexities. In standard LPM (Eq. [2.1]), there are no curvilinear or S-shaped relations and no interaction effects.

LPM's parameters including their standard errors can be estimated by means of OLS (ordinary least squares). However, the statistical properties of the OLS estimates are based on certain assumptions, among them the homoscedasticity assumption regarding the error terms. This homoscedasticity assumption is not fulfilled when the dependent variable is dichotomous. As a consequence, although estimates of the regression coefficients are still unbiased, estimates of their standard errors are not. In principle, this problem can be tackled by either using robust standard errors or using WLS (weighted least squares) estimation. However, often OLS is used to estimate LPMs, as several authors recommend (Angrist & Pischke, 2008; Long, 1997, pp. 38–39).

Column (3) of Table 2.1 shows the OLS estimates of the main-effects-only LPM model in Eq. (2.1) for the *university* data. The estimate of the intercept α_0 tells that the estimated response probability of attending university equals

.173 in the reference group R (individuals from Country A, with average intelligence, no academic background, and maximum one sibling). Note that according to the true (simulated) data, the response probability in this reference group R is actually .091. The estimate of the effect α_1 of X_1 Country is .310 and means that the response probability is .310 higher in Country B than in A, holding all remaining independent variables constant. The estimate of the effect α_4 of X_4-Intelligence is .013 and implies that an increase in intelligence from 0 to 8 increases the response probability of going to university by $8 \cdot .13 = .104$.

If in a given data set all conditional response probabilities lie within the range .30 and .70, there is a rather close linear connection between the logistic effects β_k and the LPM effect α_k (see Goodman, 1976, p. 92):

$$\beta_k \approx 4\alpha_k \tag{2.15}$$

In the *universit*y data, however, there are a quite a few response probabilities outside the range .30–.70, especially < .30. As the book's website shows, LPM produces several negative estimated conditional response probabilities (for eight of 80 combinations of the independent variables) and several estimated response probabilities that seriously deviate from the probabilities in the data (22 from 80 estimated probabilities deviate more than .10 from the true probabilities where all deviations concern true probabilities that are less than 0.2 or larger than 0.85). In this sense, LPM does not fit as nicely as the logistic regression, which (necessarily) fitted the simulated *university* data perfectly. When the conditional response probabilities cover the whole range 0–1 in S-shaped form, it is easily seen from Figure 2.1 that fitting a (LPM) straight line through the (logistic) S-curve must lead to overestimating the lower tail and underestimating the upper tail response probabilities.

As was true for the logistic model in Eq. (2.8), nonlinearities and interaction effects might be added to the basic main effects LPM in Eq. (2.1) by means of independent variables in the form of higher order polynomials, dummy variables, or product terms.

Finally, a categorical LPM variant exists, in which all independent variables are treated as categorical or discrete (analogous to the relationship between the logistic and the categorical logit model). Grizzle et al. (1969) developed the main framework, using WLS estimation; the model is often called the GSK model (which also includes the WLS logit model; Kuechler & Wides, 1981). Bergsma et al. (2009) developed a similar MLE approach while Davis (1975) explained extensively the logic of additive path models for categorical data, based on LPM notions and *d*% (or *DC*) and loosely applying MLE (see also Aris, 2001).

2.3 LATENT VARIABLE MODEL — LVM

A rather different view on the logistic or probit regression model, different from the DRM perspective, is to understand the logistic or probit regression model as derived from a particular LVM. In logistic and probit LVM, the observed dependent variable Y is conceptualized as the dichotomized outcome of an underlying unobserved continuous dependent variable Y^* and the interest is in the effects of the independent variables X on this latent variable Y^*.

At the observed level, both approaches are equivalent: Logistic LVM and logistic DRM lead to one and the same logistic regression model and are identical to each other in terms of the (expected) response probabilities $Pr(Y_i = 1)$ and the (estimated) logistic effects β on Y. Similarly, probit DRM and probit LVM lead to the same results and effects at the observed level.

To indicate the effect coefficients for the effects the independent variables have on the underlying, latent variable Y^*, Latin instead of Greek letters are used; for logistic LVM: b instead of β and for probit LVM: c instead of γ. (Some other authors do it the other way around and use b for the logistic and β for the underlying effects on Y^* (e.g., Karlson et al., 2012).

The basic idea underlying LVM is that the observed (dichotomized) Y gets the value 1 if a research unit's score on the latent continuous variable Y^* is larger than a particular threshold value τ and that otherwise the observed score on Y is 0:

$$y_i = 1\left(y_i^* > \tau\right) \tag{2.16}$$

where $1(\cdot)$ is an indicator function returning a value of 1 if the logical expression in brackets is true; otherwise, it returns a value of 0.

The latent variable itself is assumed to be a linear-additive function of the $k = 1, \cdots, K$ independent variables X_k:

$$y_i^* = \mu_i + u_i = b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki} + u_i \tag{2.17a}$$

As explained further, it is convenient for showing the connection between b_k in Eq. (2.17a) and the logistic effect β_k (and similarly for c and γ) to rewrite Eq. (2.17a) by replacing u_i by another error term e_i multiplied by a *scaling factor* φ (i.e., $u_i = \varphi e_i$):

$$y_i^* = \mu_i + u_i = b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki} + \varphi e_i \tag{2.17b}$$

The error variance in Eq. (2.17b) equals: $\sigma_u^2 = \varphi^2 \sigma_e^2$ and the scaling factor is $\varphi = \sigma_u / \sigma_e$. (Note that in Section 2.23 on probit regression the symbols φ [and its capital Φ] referred to the [cumulative] normal distribution function, but from here on φ will refer to the scaling factor.)

Except for the fact that Y^* is an unobserved, latent variable with unknown scale, and therefore unknown mean and variance, the model in Eq. (2.17) is otherwise a normal standard regression equation with Y^* as a continuous interval level dependent variable. The partial derivative of Y^* with respect to a particular X_k equals: $\partial Y^* / \partial X_k = b_k$ and the LVM regression coefficients in Eq. (2.17) are the direct effects for each independent variable on the latent variable Y^*.

The usual standard regression assumptions are made for Eq. (2.17), such as homoscedastic error variances and error terms that are not correlated with each other and not with the independent variables. Both assumptions are important of course, but the homoscedasticity assumption is more important than usual. In a normal standard regression equation with an observed dependent variable, if the error variances are not the same within the categories of *X*, heteroscedasticity will not lead to biased OLS estimates of the unstandardized regression coefficients b_k (although the OLS variance estimates and standard errors are biased; as in LPM, Section 2.2.4). However, in logistic (and probit) LVM, heteroscedasticity of the error term in Eq. (2.17) may cause serious bias in the (scaled) estimates of b_k (and c_k), as shown in Section 2.3.3.

Before discussing the precise relationship between the logistic β effects and the underlying effects *b* (and between probit effect γ and *c*), a fundamental question must be raised: What is the conceptual status of the latent variable Y^* and hence the meaning of the LVM approach? Theoretically, the idea of an underlying continuous variable may be a useful and interesting one. It certainly has a certain intuitive appeal. In the words of Long (1997):

> Consider a woman's labor force participation as the observed variable *Y*. The variable *Y* can only be observed in two states: a woman is in the labor force, or she is not. However, not all women in the labor force are there with the same certainty. One woman might be very close to the decision of leaving the labor force, while another woman could be very firm in her decision. In both cases, we observe the same $Y = 1$. The idea of a latent Y^* is that there is an underlying propensity to work that generates the observed state. (p. 40)

Researchers within a discipline like economics in which *utility* as a *primitive notion* and rational choice play a central role find this idea very appealing and make extensive use of it. Moreover, in the statistical literature, it has been probably the mostly used way to derive the properties of the logistic/probit regression model.

Nevertheless, some problematic aspects must be mentioned that have to be taken into account when interpreting the results from LVM. These issues have been brought forward many times. They also echo the notorious dispute from the old days between Yule and Pearson about the nature of categorical data and reverberate the discussions about *revealed* versus *stated preferences* (Berkson, 1951; Hagenaars, 2016; Kuha & Mills, 2017, pp. 13–17).

First of all, it must be asked why such a single crude dichotomous measurement Y is used when the interest is actually in the underlying continuous variable Y^*. It might be true that this is all researchers have at their disposal, but still, then, they should be aware of the possible shortcomings of their "measurement."

Second, as shown next, the response model that links the (unknown) scores on Y^* to the observed scores on Y entails a rather restrictive distributional assumption for the underlying latent variable along with a specific deterministic response mechanism. And these restrictions cannot be empirically validated because they are needed for the identification of the model.

Finally, and perhaps most important, the substantive meaning of the latent variable Y^* may be far from obvious. Take the *university* data as an example. What is the meaning of Y^* here? Within LVM, the focus is not on the probability of attending university or not, but on the *propensity/inclination* to attend university. Outside a definite theoretical framework, this propensity/inclination is a very vague concept. Is propensity, for example, the preference, capacity, ability, willingness, suitability, or encouragement to go to university? And is it often not of interest to know why people who have the propensity to go to university are actually not going rather than interpreting their behavior as realized preference?

All this is not to say that one should not use the latent variable interpretation of logistic regression, but that this choice should be based on careful considerations.

2.3.1 Logistic Latent Variable Model

To be able to estimate the effects of the independent variables on the latent variable Y^* in Eq. (2.17), a precise formal link must be established between the unobserved scores on Y^* and the response probabilities of $Y = 1$. Starting point is the threshold equation Eq. (2.16) in which regression equation Eq. (2.17b) is inserted:

$$y_i = 1\left(b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki} + \varphi\, e_i > \tau\right) \tag{2.18}$$

From Eq. (2.18), it follows that the probability of observing $Y_i = 1$ equals the following expression:

$$\begin{aligned}\Pr(Y_i = 1 | X_{1i}, \cdots, X_{Ki}) &= \Pr\left(\varphi e_i > \tau - \left(b_0 + b_1 x_{1i} + \cdots + b_K x_{Ki}\right)\right) \\ &= \Pr\left(e_i > \frac{\tau}{\varphi} - \left(\frac{b_0}{\varphi} + \frac{b_1}{\varphi} x_{1i} + \cdots + \frac{b_K}{\varphi} x_{Ki}\right)\right)\end{aligned} \tag{2.19}$$

Now assume that the error e_i has a certain distribution that allows one to compute the probability that e_i exceeds or falls below a certain value. If one assumes this distribution to be symmetric (such as the normal or the logistic distribution), Eq. (2.19) can also be written as:

$$\Pr(Y_i = 1 | X_{1i}, \cdots, X_{Ki}) = \Pr\left(e_i \leq \left(\frac{b_0}{\varphi} + \frac{b_1}{\varphi} x_{1i} + \cdots + \frac{b_K}{\varphi} x_{Ki}\right) - \frac{\tau}{\varphi}\right) \tag{2.20}$$

The right-hand side of Eq. (2.20) equals the definition of a cumulative distribution function (cdf): $F(z) = \Pr(Z \leq z)$. The random variable for this cdf is error variable E and the right-hand side of Eq. (2.20) provides the probability of that random variable being smaller than a certain value, namely the expression to the right of the inequality sign ≤.

In logistic LVM, the standard logistic distribution is assumed for the error variable E with a mean of 0 and variance of $\sigma_e^2 = \pi^2/3$; in probit LVM, the standard normal distribution is used for E with a mean of 0 and $\sigma_e^2 = 1$ (Long, 1997, p. 43).

Application of the standard logistic distribution to the error variable E turns Eq. (2.20) into:

$$\Pr(Y_i = 1 | X_{1i}, \cdots, X_{Ki}) = \frac{\exp\left(\frac{b_0 - \tau}{\varphi} + \frac{b_1}{\varphi} x_{1i} + \cdots + \frac{b_K}{\varphi} x_{Ki}\right)}{1 + \exp\left(\frac{b_0 - \tau}{\varphi} + \frac{b_1}{\varphi} x_{1i} + \cdots + \frac{b_K}{\varphi} x_{Ki}\right)} \tag{2.21}$$

which is equivalent to Eq. (2.3). Comparison of the coefficients of Eq. (2.3) and Eq. (2.21) shows the following relationships between the logistic DRM coefficients β_k and the underlying LVM coefficients b_k:

$$\beta_0 = \frac{b_0 - \tau}{\varphi} \text{ and } \beta_k = \frac{b_k}{\varphi};\ k = 1, \cdots, K \tag{2.22a}$$

$$\text{with } \varphi = \frac{\sigma_u}{\sigma_e} = \frac{\sigma_u}{\pi/\sqrt{3}} \tag{2.22b}$$

The logistic DRM coefficients β_k in Eq. (2.3) are scaled versions of the corresponding underlying LVM coefficients b_k the researcher is actually interested in. The scaling factor φ as such is unknown, being the ratio of the unknown true standard deviation of U and the fixed standard deviation of E. This scaling (or error-standardizing) factor φ is a key factor in the comparative interpretation of the logistic regression coefficients, as the next chapters show.

From Eq. (2.22a), it is further seen that the threshold value τ only plays a role in intercept β_0. This role is related to the fact that the mean of latent variable Y^* is unknown and therefore the intercept β_0 is a function, not only of b_0 and φ, but also of τ. Usually, it is arbitrarily assumed that $\tau = 0$ (sometimes $b_0 = 0$). As long as the latent dependent variable Y^* is regarded as an interval, rather than a ratio level variable, this is not a serious restriction: assigning an arbitrary value to τ amounts to switching the origin (0) of the interval scale, which in case of an interval level variable is arbitrary anyhow.

Often, due to the scaling problem in LVM, researchers will not try to evaluate the estimated b effects on the latent variable Y^* but instead just focus on the consequences the LVM model has for the response probabilities. The estimates of the response probabilities are not affected by the size of the scaling factor, that is, they are not scale dependent. If the logistic distribution had been assigned a standard deviation different from $\pi/\sqrt{3}$, the same estimated response probabilities would have been found (they are estimable functions; see Long, 1997, pp. 49–50).

So, as for the logistic DRM discussed earlier, also in LVM the effects of the independent variables might be evaluated by using response profiles or one more of variants of the *IC* or *DC* effect measures. However, given that the estimated response probabilities are the same in both DRM and LVM, these "effect measures" will of course be the same in both approaches and the same complexities will arise in LVM as in DRM due to the non-linear, S-shaped relations and the many interaction effects. As in DRM, the use of odds ratios might solve these complexity problems. But in this way, there is essentially no difference between the DRM and the LVM approach: The purpose of finding the linear-additive effects b of the X variables on the unobserved Y^* has disappeared. Therefore, despite the scaling problems, the focus in the next chapters regarding LVM is on what can be inferred about the underlying effects b on Y^* in the three main comparative situations.

Underlying Standardized Effects and Explained Variance of Y^*

So the main source of the difficulties involved when using the logistic regression coefficients β to draw conclusions about the LVM regression coefficients b is the unknown size of the underlying error variance σ_u^2 and hence the unknown variance $\sigma^2_{Y^*}$ of Y^*. One general way of solving this problem is to assign the value 1 to the variance of Y^* and transform the underlying latent variable Y^* into the standardized variable Z_{Y^*} with variance $\sigma^2_{Z_{Y^*}} = 1$ (and mean 0; Long, 1997, pp. 70–74; McKelvey & Zavoina, 1975; Winship & Mare, 1983, 1984). The focus of the LVM analysis then switches from the underlying unstandardized regression coefficients *b* to the underlying standardized effects. As shown next, these standardized effects can be estimated from the observed data.

For the *fully standardized regression coefficients*, all variables, the dependent and the independent ones, are standardized with mean equal to 0 and standard deviation equal to 1, yielding the fully standardized coefficients b_k^{fs}:

$$b_k^{fs} = \frac{\sigma_k}{\sigma_{Y^*}} b_k \tag{2.23}$$

A *semi-* or *half-standardized regression coefficient* in the form of only *Y^*-standardized*, $b_k^{s_{Y^*}}$ equals:

$$b_k^{s_{Y^*}} = \frac{b_k}{\sigma_{Y^*}} \tag{2.24}$$

Half-standardized effects in the form of only *X*-standardized are not considered in this volume, because they do not solve the scaling problem. Half-standardized effects always refer to the $b_k^{s_{Y^*}}$ coefficients. Also the fully standardized effects are mostly ignored. Certainly with independent variables in the form of dummy variables, the Y^*-standardized effects make more sense than the fully standardized (see also Section 3.1). The half-standardized coefficients in Eq. (2.24) leave the initial scoring of the independent variables as they are but standardize Y^*, whose scaling is not known anyhow. The Y^*-standardized effect indicates how many standard deviations the latent variable in- or decreases by an increase of one measurement unit on the independent variable. Although the standard deviation of Y^* remains unknown, this is the type of interpretations that is regularly used also in standard regression analysis.

Although σ_{Y^*} and b_k in Eq. (2.24) are both unknown, latent, the (half-) standardized effect $b_k^{s_{Y^*}}$ can be obtained from the data. Starting point is the equation $\sigma^2_{Y^*} = \sigma^2_{\hat{Y}^*} + \sigma^2_u$: The total variance $\sigma^2_{Y^*}$ equals the sum of the explained variance $\sigma^2_{\hat{Y}^*}$ plus the error variance σ^2_u, where $\sigma^2_u = \varphi^2 \sigma^2_e = \varphi^2 (\pi^2/3)$. The explained variance $\sigma^2_{\hat{Y}^*}$ can be written as a function of the (co)variances of the independent variables and their effects *b*:

$$\mathrm{Var}(\mu_i) = \sigma^2_{\hat{Y}^*} = \sum_{k=1}^{K} \sum_{l=1}^{K} b_k \cdot b_l \cdot \sigma_{kl} \tag{2.25}$$

in which σ_{kl} (for $k \neq l$) is the covariance $\mathrm{Cov}(X_k, X_l)$ and σ_{kk} (for $k = l$) the variance σ_k^2. As it turns out, when using the outcomes of the logistic regression equation Eq. (2.3) for computing the total variance $\sigma_{Y^*}^2$ and the standardized effects $b_k^{s_{Y^*}}$, the scaling factor φ cancels out. This can be simply illustrated for a regression equation with one independent variable X_1 (and straightforwardly generalized to more independent variables):

$$\begin{aligned} b_1^{s_{Y^*}} \quad = \quad & \frac{b_1}{\sigma_{Y^*}} = \frac{b_1}{\sqrt{b_1^2\sigma_1^2 + \sigma_u^2}} = \frac{\varphi\beta_1}{\sqrt{\varphi^2\beta_1^2\sigma_1^2 + \varphi^2\frac{\pi^2}{3}}} \\ & = \frac{\varphi\beta_1}{\varphi\sqrt{\beta_1^2\sigma_1^2 + \frac{\pi^2}{3}}} = \frac{\beta_1}{\sqrt{\beta_1^2\sigma_1^2 + \frac{\pi^2}{3}}} \end{aligned} \tag{2.26}$$

Because all elements in the last right-hand side component of Eq. (2.26) can be estimated from the observed data, the underlying half-standardized effects $b_1^{s_{Y^*}}$ can be computed and they can be interpreted in the usual way (Long, 1997, pp. 70–71). The advantages and disadvantages of using standardized versus unstandardized coefficients are essentially the same in LVM as for standard regression models. The issue is well discussed in general treatments of standard regression analysis; it returns in later chapters.

The notions underlying Eqs. (2.25) and (2.26) can also be used to estimate how much of the variance in latent variable Y^* is explained by the independent variables. Several *pseudo* $- R^2$ coefficients both within the DRM and the LVM framework have been proposed in the literature and implemented in the main statistical packages (Long, 1997, Section 4.3). For LVM, a rather straightforward and intuitively appealing coefficient $R_{Y^*}^2$, suggested by McKelvey and Zavoina (1975), is based on an estimate of $\sigma_{\hat{Y}^*}^2/\sigma_{Y^*}^2$, the proportion explained variance in Y^* (pp. 111–112; Long, 1997, p. 105). Coefficient $R_{Y^*}^2$ can be computed from the data as illustrated in Eq. (2.27) for one independent variable X_1 (and easily generalized to more independent variables using Eq. [2.25]; see Long, 1997, p. 105):

$$\begin{aligned} R_{Y^*.1}^2 \quad & = \frac{\sigma_{\hat{Y}^*}^2}{\sigma_{Y^*}^2} = \frac{b_1^2\sigma_1^2}{b_1^2\sigma_1^2 + \sigma_u^2} = \frac{b_1^2\sigma_1^2}{b_1^2\sigma_1^2 + \varphi^2\sigma_e^2} \\ & = \frac{\varphi^2\beta_1^2\sigma_1^2}{\varphi^2\beta_1^2\sigma_1^2 + \varphi^2\frac{\pi^2}{3}} = \frac{\beta_1^2\sigma_1^2}{\beta_1^2\sigma_1^2 + \frac{\pi^2}{3}} \end{aligned} \tag{2.27}$$

2.3.2 Probit Latent Variable Model

If the error variable E in Eq. (2.17b) is assumed to be standard normally distributed with standard deviation equal to 1, the result is the probit LVM. The underlying LVM probit effects are indicated by c_k, replacing b_k in Eq. (2.17).

Similar to logistic LVM, the probit effect coefficients γ_k in Eq. (2.12) are scaled versions of the underlying LVM coefficients c_k:

$$\gamma_0 = \frac{c_0 - \tau}{\varphi} \text{ and } \gamma_k = \frac{c_k}{\varphi};\ \ k = 1, \cdots, K \tag{2.28a}$$

$$\text{with } \varphi = \frac{\sigma_u}{\sigma_e} = \frac{\sigma_u}{1} = \sigma_u \tag{2.28b}$$

The consequences of this scaling problem are the same for the probit LVM as for the logistic LVM. They can be circumvented for the probit LVM analogously to the logistic LVM by defining (half) standardized effects or using the estimated response probabilities, with the same advantages and problems. The effect coefficients γ_k of the probit DRM did not have the same clear and nice (odds ratio) interpretation as was possible for the logistic DRM. However, the underlying effects c_k (probit) and b_k (logit) on the latent dependent Y^* have the same kind of standard regression interpretation. The only difference between logistic and probit LVM is that in probit LVM the error term is assumed to be standard normally distributed instead of standard logistically, which does not affect the basic interpretation of the underlying effects.

It can now be explained why (in DRM) the logistic effect β_k is about 1.7 times larger in absolute value than the corresponding probit effect γ_k. First of all, the standard deviation of the error terms in the assumed standard logistic distribution equals $\pi/\sqrt{3} = 1.814$, while the standard deviation of the error terms in the assumed standard normal distribution equals 1. Taking this difference into account, it can be expected from comparing Eqs. (2.28) and (2.22) that the logistic effects are about $\left(\pi/\sqrt{3}\right)/1$ times stronger than the probit effect: $\beta_k = 1.814\gamma_k$, just due to the different scaling of the error terms.

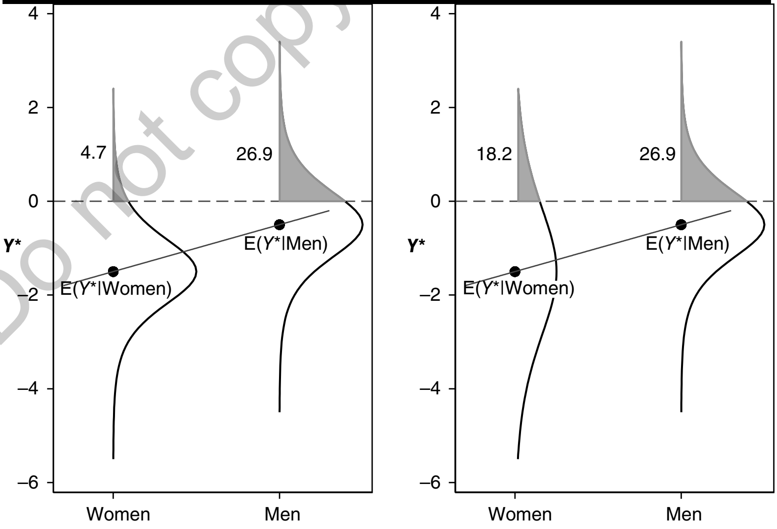
However, although the standard logistic and the standard normal distribution have approximately the same form, the shapes are not completely the same and differ not only with regard to their respective standard deviations. Taking this into account, somewhat different results might be obtained. Andreß et al. (2013) suggested to restrict the marginal effect for Pr $\left(Y_i = 1\right) = 0.5$ to be the same for probit and logit LVM, yielding: $\beta_k = 1.6\gamma_k$ (p. 226). The same result, as cited by Long (1997), is obtained by Amemiya who suggested making the cdfs of the logistic and the normal distributions as close as possible. Long's own calculations along similar lines resulted in $\beta_k = 1.7\gamma_k$ (p. 48).

2.3.3 Heteroscedastic Errors, Unequal Thresholds, and Biased Effects

Homoscedasticity of the error terms u_i in the underlying LVM equation (Eq. (2.17) is an essential assumption in logistic/probit LVM. It is strictly needed for the identification of the underlying *b*s in terms of $b_k = \varphi\beta_k$ and cannot be empirically tested due to the latent nature of Y^*. If the error terms u_i in the underlying LVM equation (Eqs. (2.17a, b) are not homoscedastically distributed, the crucial relationship $b_k = \varphi\beta_k$ (Eq. 2.22) are no longer valid and inferences about the underlying *b*s on the basis of the "observed" *β*s are biased. A thorough understanding of the consequences of heteroscedasticity is therefore important. The simulated data example used for this purpose also provides further insight into the role of the thresholds in LVM.

The observed dichotomous variable *Y* in the simulated data example is (again) denoted as Attending university or Not (1 = Attending; 0 = Not). The underlying variable Y^* represents the propensity to attend university. There is only one independent variable Gender (1 = Men; 0 = Women). The underlying equation equals: $Y^* = b_0 + b_{Y^*G}G + u_i$. In the simulation, the "true" underlying effects are known. The intercept is given the value $b_0 = -1.5$ and the effect of Gender is set to one: $b_{Y^*G} = 1$. Hence, the mean latent score of Y^* for Women equals $E\left(Y_W^*\right) = -1.5$ and for Men $E\left(Y_M^*\right) = -.5$. For the distribution of the errors u_i around these expected values, the logistic distribution is used

FIGURE 2.2 ■ Homoscedastic (a) and Heteroscedastic (b) Error Variance in LVM



and shown in Figure 2.2. The threshold value τ is set at $Y^* = 0$ for both Men and Women.

As can be readily seen from Figure 2.2, the conditional response probability $Pr(Y = 1|G)$ will in principle vary with and depend on the threshold value τ and on the expected value of Y^* but also on the error terms (i.e., on the spread around the expected values of Y^*).

First, the homoscedastic scenario is simulated and presented in Figure 2.2a. In this scenario, the errors u_i are logistically distributed with a standard deviation that is identical for Men and Women and made .5 times as small as the error standard deviation σ_e in the standard logistic distribution: $\sigma_{uM} = \sigma_{uW} = .5\sigma_e = (.5)(\pi/\sqrt{3}) = (.5)(1.8138) = .9069$ and the scaling factor equals $\varphi = \sigma_u/\sigma_e = .5$. Also the threshold is the same for Men and Women. In this way, the difference in the response probabilities for Men and Women is only a function of the difference in expected values (i.e., of the underlying effect b_{Y^*G}).

The response probability for Men resulting from this simulation is $Pr(Y_M = 1) = \left(e^{(-1.5+1)/0.5}\right)/\left(1 + e^{(-1.5+1)/0.5}\right) = .2689$ and the response odds are $\Omega_{1/0}^{Y|M} = .2689/(1 - .2689) = .3679$. For Women, the simulated result for the response probability is $Pr(Y_W = 1) = \left(e^{-1.5/0.5}\right)/\left(1 + e^{-1.5/0.5}\right) = .0472$ and for the response odds: $\Omega_{1/0}^{Y|W} = .0472/(1 - .0472) = .0495$. The odds ratio for the effect of Gender on the observed variable Y equals $OR = \Omega_{1/0}^{Y|M}/\Omega_{1/0}^{Y|W} = .2689/.0472 = 7.4246$ and $\ln OR = \beta_{YG} = 2.00$. The logistic effect β_{YG} is twice as large as the underlying effect b_{Y^*G}; this corresponds with the simulated value of φ.

In the homoscadestic scenario, it should be true that $b_k = \varphi\beta_k$. For these simulated data, the scale factor φ is known and was set to: $\varphi = 0.5$. Applying the scale factor to find the underlying effect from the logistic effect yields: $b_{Y^*G} = \varphi\ \beta_{YG} = .5 \cdot 2 = 1$, which is the true, simulated underlying effect of Gender. In the homoscedastic scenario, the logistic effect $\beta_{YG}\,(= 2)$ is an unbiased, albeit φ scaled estimate of $b_{Y^*G}\,(= 1)$. If the standard deviation σ_e had been made equal to the true eror standard deviation σ_u, logistic effect β_{YG} would have been the same as the underlying effect b_{Y^*G}.

In the heteroscedastic scenario in Figure 2.2b, everything stays the same for Men and Women as in the homoscedastic scenario, except that for Women the spread around the expected value has been doubled: $\sigma_{uW} = 2 \cdot .9069 = 1.8138\,(= \pi/\sqrt{3})$. Just due to the larger error variance for Women, a much larger percentage of the Women will now obtain the score $Y = 1$: $Pr\left(Y_W = 1\right) = \left(e^{-1.5}\right)/\left(1 + e^{-1.5}\right) = .1824$ (instead of .0474). For Men, the response probability and response odds remain the same as in the homoscedastic scenario. In the heteroscedastic scenario, the difference between the observed response probabilities for Men (.2689) and Women (.1824) is not only a function of the difference in their expected values (i.e., of b_{Y^*G}), but also a function of the difference in their error variances: in this heteroscedastic scenario, $OR = (.2689/(1 - .2689))/(.1824/(1 - .1824)) = .3679/.2231 = 1.649$ and $\beta_{YG} = ln\,OR = \ln(1.649) = .50$ $\left(= -1.5 + 2\right)$. While the underlying regression effect b_{Y^*G} on Y^* is the same in the two scenarios, the logistic effect β_{YG} is very much smaller in the heteroscedastic than in the homoscedastic scenario (.5 vs. 2).

For real data sets, the researcher would not know the heteroscedasticity in the underlying equation for Y^*. The question is, then, what would happen when the simulated data under the heteroscedastic scenario were analyzed as if the error terms were homoscedastically distributed?

The value of the single pooled estimate of the two error variances for Men and Women would depend on the distribution of Gender. (But note that the relative conditional distributions of Y^* for Men and Women do not change when only the marginal distribution of Gender is changed and so b, OR and β will keep the same values.) For a uniform (.50/.50) distribution of Gender, the pooled error variance would be equal to the unweighted average of the two error variances and the square root of this average variance equals $\sigma_{u-pooled} = 1.4340$. The scaling factor would be equal to $\varphi = 1.4340/1.8138 = .7906$ and so b_{Y^*G} would be estimated as $b_{Y^*G} = \varphi\beta_{YG} = .7906 \cdot .5 = .3953$, a far cry from the true value 1. Similar calculations for a .75/.25 distribution Women/Men would lead to $\sigma_{u-pooled} = 1.6349$ and $b_{Y^*G} = \varphi\beta_{YG} = .4507$ and a .25/.75 distribution to $\sigma_{u-pooled} = 1.1997$ and $b_G = \varphi\beta_{YG} = .3307$.

Obviously, assuming homoscedasticity of the underlying error terms when they are actually heteroscedastic may seriously bias the inferences about the underlying b coefficients in LVM by means of the βs. How heteroscedasticity precisely affects the outcomes, certainly when there are several independent variables involved is hard to tell. It depends on the amount of the heteroscedasticity and on how the heteroscedasticity is related to one or more of the independent variables. But if there is reason to believe that there is heteroscedasticity, this simple example warns the researcher to be careful.

In the remainder of this volume, it will just be assumed that there is homoscedasticity of the underlying errors, except for the subgroup comparisons dealt with in Chapter 4.

From Figure 2.2, it is also easily seen that different thresholds τ_M and τ_W for Men and Women would lead to serious distortions of the estimates of the underlying effects on Y^* if $\varphi\beta_{Y^*G}$ is used to make inferences about b_{Y^*G}. If one moves the horizontal threshold line in Figure 2.2a up- or downward only for Men and not for Women, while keeping the expected values for Men $E(Y_M^*)$ and Woman $E(Y_W^*)$ and therefore the value of b_{Y^*G} the same, the response probability for Men changes and the key equation $b_k = \varphi\beta_k$ no longer applies.

The assumption that the measurement model linking the underlying latent variable Y^* to the observed variable Y is the same for all subjects and, as part of this, that the thresholds are the same for all subjects is a specific instance of the general requirement that the measurement instruments should have the same meaning for all subjects. Without such an assumption the variable scores of the subjects could never be compared. Nevertheless, its validity might be doubted, as may especially occur in cross-cultural research settings. The consequences of unequal thresholds are discussed again in the chapter on subgroup comparisons (Section 4.1).

2.4 INSERTING MAVERICKS, "ORTHOGONAL" INDEPENDENT VARIABLES, INTO EQUATIONS

As the next chapters show, the controversies surrounding the comparative interpretations of logistic or probit effects are often discussed under the headings *collapsing and confounding* and *differences in unobserved heterogeneity* (or, equivalently, instead of unobserved heterogeneity: differences in error variances/in scaling factors). For a clear understanding of these discussions and the link between the notions unobserved heterogeneity and collapsing effect, it is very helpful to have a good insight into the consequences of adding to or omitting from a particular regression equation an independent variable that has a direct effect on the dependent variable but is statistically independent of the other independent variables.

In standard multiple regression equations, the addition of such an independent variable will increase the explained variance, but will not affect the values of the (un)standardized regression coefficients for the other independent variables. In logit/probit regression on the other hand, this is not true: The other logistic (probit) effects β (γ) will be affected; to be more precise: They keep their sign, but their strengths will increase. As Mood (2010, p. 69) noted this difference is often forgotten. Even in the widely used introductory text on logistic regression by Menard, this difference with standard multiple regression is overlooked (Menard, 1995, p. 59; also Menard, 2002, p. 69).

In the context of logistic regression, Hauck et al. (1991) aptly call such a variable that is statistically independent of the other independent variables but directly influences the dependent variable, a maverick. One might say, a *loner* that disturbs the established, expected order of things. A maverick is defined here in a strict sense (i.e., in terms of statistical independence of the other independent variables). In standard linear models, orthogonality,—a less stringent requirement, viz. not being correlated ($r = 0$) with the other independent variables—is sufficient to arrive at similar results, but here the stricter condition of statistical independence is used.

What precisely happens to the effect parameters of a logistic regression equation when a maverick is added to the equation can be more easily shown by means of the LVM than by the DRM framework, but of course with identical results for the logistic regression equation.

Starting point is the LVM regression equation Eq. (2.17a) for the effects on latent variable Y^*. To such an equation with independent variables X_k, maverick X_Q is added. The regression equation including X_Q is called the *full* equation, with underlying regression effects b_k^f (and b_Q^f). The corresponding regression equation but without X_Q is called the *reduced* equation, with underlying regression coefficients b_k^r.

Because Eq. (2.17) is a normal standard regression equation, b_k^r in the reduced equation will have the same value as b_k^f in the full equation: $b_k^r = b_k^f$. However, the logistic effects β_k^f and β_k^r will be different from each other. They will have the same sign but different strengths, β_k^f being stronger. That β_k^f and β_k^r have the same sign, viz. the sign of b_k^f (or, for that matter, b_k^r) follows directly from $b_k = \varphi\beta_k$ and the fact that the scaling factor φ is always positive ($\varphi > 0$), being a ratio of two standard deviations.

The inequality $|\beta_k^f| > |\beta_k^r|$ follows from the unequal error variances and scale factors in the underlying reduced and full equation. The unexplained variance $\left(\sigma_u^f\right)^2$ in the full equation is smaller than $\left(\sigma_u^r\right)^2$ in the reduced equation due to the extra nonzero direct effect b_Q^f of X_Q: $\sigma_u^f < \sigma_u^r$. However, in logistic LVM, both equations, the reduced and the full equation, are estimated as logistic regression equations with the restriction that the error variance is fixed to σ_e^2: $\sigma_e^f = \sigma_e^r = \pi/\sqrt{3}$. Therefore, the scaling factor φ_f for the full logistic equation equals $\varphi_f = \sigma_u^f/\sigma_e$ and the scaling factor φ_r for the reduced equation: $\varphi_r = \sigma_u^r/\sigma_e$. Because $\sigma_u^f < \sigma_u^r$, it follows that $\varphi_f < \varphi_r$ and therefore $|\beta_k^f| > |\beta_k^r|$. After adding a maverick, the strength of the effect of X_k on the response probability will be larger in the full than in the reduced equation, in formula:

$$\frac{\beta_k^f}{\beta_k^r} = \frac{b_k^f/\varphi_f}{b_k^r/\varphi_r} = \frac{b_k/\varphi_f}{b_k/\varphi_r} = \frac{\varphi_r}{\varphi_f} = \frac{\sigma_u^r/(\pi/\sqrt{3})}{\sigma_u^f/(\pi/\sqrt{3})} = \frac{\sigma_u^r}{\sigma_u^f} > 1 \tag{2.29}$$

From the other perspective, omitting a maverick from the equation (collapsing the joint distribution over X_Q) decreases the strengths of the effects β_k without changing their signs.

Although deleting or adding mavericks in an equation changes the strengths of the effects, it usually hardly changes the relative effects in the form of their ratios. For example, the ratio $\beta_k/\beta_{k'}$ of the effects of the independent variables X_k and $X_{k'}$ will remain practically the same after adding or deleting a maverick X_Q: $\beta^r_k/\beta^r_{k'} \approx \beta^f_k/\beta^f_{k'}$. This is easily seen by noting that the underlying *b*s do not change by adding or deleting a maverick and that within each equation, within the full and within the reduced one, the scaling factor is the same for all effects. (However, the incompatibility issue involving some possible incompatibilities between the reduced equation without the maverick and the full equation including the maverick, further discussed in Section 5.1.2, may have a small distorting effect on the ratio.) Relative effects are further discussed in Chapter 3.

How much the addition or deletion of a maverick changes the other logistic effects β_k in the equation depends on the distribution of X_Q and on the absolute strength of its direct effect b^f_Q on the dependent variable Y^*. The larger the variance of X_Q and the larger its direct effect b^f_Q, the more variance of Y^* is explained in the full compared to the reduced underlying equation and the larger the ratios σ^r_u/σ^f_u, φ_r/φ_f and β^f_k/β^r_k are (see also, Mood, 2010). Moreover, because ratios and relative changes are important here, the relative effect of the other independent variables compared to the effects of the maverick also plays a role. If the maverick reduces the error variance by an amount *x*, this generally has much more impact on the ratio σ^r_u/σ^f_u (and hence on β^f_k/β^r_k) for smaller than for larger values of σ^r_u, that is, for larger than for smaller amounts of explained variance. And of course, the larger the effects of the other independent in the reduced equation, the smaller the error variance is.

Analogous result can be obtained for the probit LVM and the ratio γ^f_k/γ^r_k, using $\sigma_e = 1$ instead of $\sigma_e = \pi/\sqrt{3}$.

Because application of LVM leads in the end to the same logistic (or probit) equation for the observed *Y* as DRM, the same consequences of adding or omitting a maverick X_Q are true within the DRM framework for the logistic (probit) effects. These consequences can also be investigated without recourse to a latent dependent variable, strictly within a DRM framework. The formulas are obtained by deriving the multiplicative effect parameters in the collapsed (reduced) table or distribution as sums of the multiplicative effects in the complete (full) table or distribution. Ultimately, the consequences of adding or omitting a maverick have to do with elementary properties of logarithms, for example, that generally the sum of logs is not equal to the log of the sum: $[\ln(a) + \ln(b)] \neq \ln(a + b)$ but $[\ln(a) + \ln(b)] = \ln(ab)$ (in contrast to: $\sum(X_i + W_i) = (\sum X_i + \sum W_i)$).

For a rather simple distribution in the form of a 2 × 2 × 2 table, Hauck et al. (1991) provide a formula for the influence of omitting X_Q from the full logistic equation (collapsing the full distribution over X_Q) on the effects of the other independent variables. In this way, a proportionality factor can be obtained to investigate the effects on the logistic effects of introducing a maverick. But even for this simple data set in 2 × 2 × 2 table, this proportionality factor looks more complicated than Eq. (2.29) (Hauck et al, 1991, p. 81). Aris presents a similar proportionality formula for the same situation, carries out many simulations and also investigates the consequences for the variances of the estimates and the power of significance tests (Aris, 2001, Section 6.3.2; Aris et al., 2000). Hagenaars and Andreß (2020) propose a (possibly less cumbersome) simulation procedure SIMMAV to approximately evaluate the consequences of introducing mavericks in logistic regression (see Chapter 5). For still more, somewhat different DRM approaches for investigating the consequences of introducing mavericks, see Agresti, (2002, pp. 498–500), Karlson et al. (2012, p. 305), and Zeger et al. (1988, p. 1054).

How big the changes in the strength of β_k might be and to what extent the substantive conclusions about the effects of X_k might change after introduction of a maverick is further discussed and illustrated in the next three core chapters. But to give a first impression, Aris (2001, Section 6.3.2) simulated the effects of deleting a maverick from the full logistic regression equation. He found that if the strength of the logistic effect β_Q^f of maverick X_Q on Y is less than half the strength of the logistic effect β_k^f of X_k, the decrease in the strength of β_k after deletion of the maverick from the equation is less than 2% (and so, using $1/.98 = 1.0204$: $1 < (\beta_k^f/\beta_k^r) \leq 1.02$). If β_Q^f and β_k^f are about the same size, the decrease is about 6% (and using $1/.94 = 1.0638 : (\beta_k^f/\beta_k^r) \approx 1.06$). However, for very large β_Q^f, decreases in β_k^f of 25% or even somewhat more can be found, although β_k^f keeps at least 50% of its original value (and so, $1.33 < (\beta_k^f/\beta_k^r) \leq 2.00$). Of course, these outcomes, as is always true in simulations, depend on the precise nature of the chosen simulation parameters (for details, see Aris, 2001), but they are in line with what was to be expected from the above and do not contradict what was found elsewhere and is found in the following chapters. In many cases, the substantive conclusions will not differ whether a maverick is introduced or not. However, in particular circumstances, the numerical consequences of introducing mavericks on the existing effects may be rather large and have to be reckoned with, certainly when rather precise effect estimates matter.

Finally, contrary to what is sometimes suggested, also the value of *DC* as effect measure is affected by the introduction of a maverick, despite the fact that *DC* measures are essentially unstandardized regression coefficients. This is because the *DC*s are not computed on the basis of response probabilities estimated by means of standard linear-additive regression equations (or such as LPM) but by means of a logistic regression equation.

As discussed in Section 2.2.1, the value of *DC* for the effect of X_k on the response probabilities not only depends on the chosen values of X_k but also on the values chosen for the other independent variables in the equation, including maverick X_Q. Depending on the chosen value of maverick X_Q, the *DC* effect of X_k will be evaluated at different points of the S-curve with different slopes, yielding different *DC* values. This is also true for the *IC* measures (although *AME* [Eq. (2.6)] can be expected to be differently, often less, affected by the introduction of mavericks [Mood, 2010; Woolridge, 2002, p. 471]).

Given the often rather arbitrary nature of the *DC* and *IC* measures due to the nonlinearity of the response profiles and given the emphasis in this volume on the interpretation of the effect coefficients in the pertinent equations, *DC* and *IC* effect measures are mostly ignored in the remainder of this volume (but see Breen et al., 2018; Cramer, 2003; Mood, 2010; Woolridge, 2002, Chapter 15).