CHAPTER 2

The Impact of Mismeasured Continuous Variables

In the first chapter we gave an intuitive explanation as to how the mismeasurement phenomenon arises, i.e., how measurement error in a continuous explanatory variable typically attenuates the associated regression coefficient. Now we mathematically elucidate this attenuation, first in the simple scenario of Chapter 1 and then in more complex scenarios. There is no attempt yet to adjust estimators to account for measurement error; this is taken up in Chapter 4. Rather, the goal of this chapter is to quantify the degradation of naive inference in various settings. It can be very useful to develop some intuition in this regard, as in some scenarios not enough is known about the mismeasurement process to attempt a formal adjustment.

The majority of the material in this chapter is not new, though arguably it has a wider range and different flavour than other discussions of how well naive estimation performs in the face of measurement error. One notable feature is that most of the present treatment permit a discrepancy between the actual and modelled relationships for the outcome variable in terms of the actual but unobservable explanatory variable. That is, the bias induced by measurement error can be quantified without regard for the extent of model misspecification at hand. In general it is the author's contention that much of the mismeasured variable literature gives short shrift to discussion of how ignoring the mismeasurement affects the quality of inference. In contrast, we take a detailed and leisurely look at when naive estimation does or does not work well before turning our attention to adjustment in Chapter 4.

2.1 The Archetypical Scenario

To return to the scenario described in Chapter 1, say the investigator is interested in the relationship between the continuous response Y and the continuous predictor X, and correctly surmises that $E(Y|X) = \beta_0 + \beta_1 X$ for unknown coefficients β_0 and β_1. However, while Y can be measured correctly, X^* is measured in lieu of X. If the investigator is unaware of this mismeasurement, or chooses to ignore it, he may interpret the estimated coefficients from a linear regression of Y on X^* as describing the relationship of interest between Y and X.

It is straightforward to quantify the damage resulting from this actual or feigned ignorance under some assumptions about the precise predictor X and its imprecise surrogate X^*. Particularly, assume that X has a normal distri-

bution, and say that X^* arises from additive measurement error which is *nondifferential*, *unbiased*, and normally distributed. In general mismeasurement is said to be nondifferential if the distribution of the surrogate variable depends only on the actual explanatory variable and not on the reponse variable. Intuitively this can be regarded as a supposition that the mismeasurement arises in a manner which is blind to the outcome variable, so in some sense the problem is limited to one of difficulty in making measurements. Stated more formally, we say the measurement error is nondifferential when X^* and Y are conditionally independent of Y, so that the conditional distribution of $(X^*|X,Y)$ is identically the conditional distribution of $(X^*|X)$. We further restrict to unbiased measurement error, i.e., $E(X^*|X) = X$, and in fact assume a normal distribution with constant variance for the conditional distribution. All told then we take $X \sim N(\mu, \sigma^2)$ and $(X^*|X,Y) \sim N(X, \tau^2\sigma^2)$ as a simple, prototypical measurement error scenario. Note that with the chosen parameterization $\tau = SD(X^*|X)/SD(X)$ can be interpreted as the magnitude of the measurement error expressed as a fraction of the variability in the predictor X itself. For instance, $\tau = 0.1$ can be viewed as yielding 10% imprecision in the measurement of X.

Under the stated assumptions, standard distributional theory implies that (X, X^*) has a bivariate normal distribution. Using first the nondifferential property and then properties of this bivariate normal distribution, it follows that

$$\begin{aligned} E(Y|X^*) &= E\{E(Y|X)|X^*\} \\ &= \beta_0 + \beta_1 E(X|X^*) \\ &= \beta_0^* + \beta_1^* X_1^*, \end{aligned}$$

where

$$\begin{aligned} \beta_0^* - \beta_0 &= \frac{\mu\beta_1}{1+\tau^2}, \\ \frac{\beta_1^*}{\beta_1} &= \frac{1}{1+\tau^2}. \end{aligned} \tag{2.1}$$

Thus the ordinary slope coefficient from a linear regression applied to sampled values of Y and X^* is an unbiased estimator of β_1^*. However, if this estimator is regarded as describing the relationship between Y and X, then a bias obtains—one is trying to estimate β_1 but actually estimating β_1^*. It is clear from (2.1) that β_1^* lies between zero and β_1. That is, in accord with the plots and simulations in Chapter 1, the effect of measurement error on the slope estimate is one of attenuation. As terminology we refer to (2.1) as the *attenuation factor* (AF), while the magnitude of the *relative bias* is $|(\beta_1^* - \beta_1)/\beta_1| = 1 - AF$. The basic relationship (2.1) between the magnitude of the mismeasurement and the extent of the attenuation is very well known, to the point that it is hard to trace an original reference. Fuller (1987) traces the literature as far back as Adcock (1877, 1878). Other historical references include Durbin (1954), and Cochran (1968).

In Figure 2.1 we plot the attenuation factor (2.1) as a function of τ, refer-

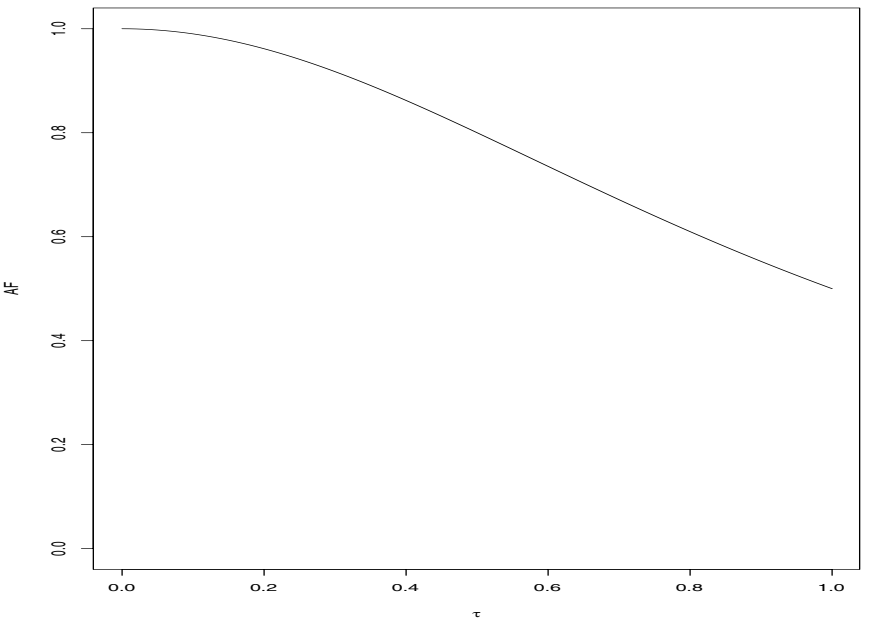

Figure 2.1 *Attenuation curve for linear regression with a single predictor. The curve is the attenuation factor* $(1+\tau^2)^{-1}$ *plotted as a function of* τ.

ring to this as an *attenuation curve*. The first impression from the curve is that a moderate amount of measurement error does *not* cause a substantial attenuation. For instance, $\tau = 0.1$, interpreted as 10% measurement error, yields $AF = 0.99$, interpreted as a negligible 1% relative bias. Further along these lines, a 5% relative bias does not occur until we have 23% measurement error, and 33% measurement error is necessary to produce a 10% relative bias. It is reasonable to surmise that in this scenario one need not worry about the impact of measurement error unless this error is quite large. Of course, as discussed in Chapter 1, large measurement errors are quite common in epidemiological investigations. Moreover, we shall encounter plausible circumstances under which moderate measurement error causes much more severe bias.

2.2 More General Impact

Of course the expression (2.1) for the multiplicative bias induced by measurement error was derived under very specific assumptions about the distributions of X, $X^*|X$, and $Y|X, X^*$. It turns out, however, that a similar expression can be obtained in a much more general framework. In the absence of these distributional assumptions, the large-sample limit of estimated coefficients for

regression of the response on the imprecise predictor can be compared to the corresponding limit for regression of the response on the precise predictor. To be more precise, by large-sample limit we mean the large-n limit of estimated coefficients from n independent and identically distributed realizations of the variables under consideration. We also expand the scope by considering a second predictor Z, which is precisely measured, to reflect the reality that most practical problems involve multiple regressors.

Before presenting the main result of this section we comment that by looking at large-sample limits our findings can be regarded as applying to different methods of parameter estimation. In particular, it is well known that in virtually any circumstance maximum likelihood and Bayesian estimators will tend to the same value as the sample size accumulates, regardless of whether the model leading to the estimators is correct or not. In fact throughout this chapter and the next our discussion of the impact of mismeasurement is not specific to the method of estimation. The Bayesian approach is introduced in Chapter 4 as a route to modelling and inference when the desire is to adjust for mismeasurement.

The impact of measurement error in linear regression with an additional precisely measured explanatory variable can be summarized as follows.

Result 2.1 *Assume that (Y, X, Z) jointly have finite second moments, and say X^* and (Z, Y) are conditionally independent given X, with $E(X^*|X) = X$ and $Var(X^*|X) = \tau^2 Var(X)$. Let β and β^* be the large-sample limiting coefficients from least-squares regression of Y on $(1, X, Z)$ and Y on $(1, X^*, Z)$ respectively. Then,*

$$\beta_0^* - \beta_0 = \beta_1 \left(\frac{\tau^2}{1 - \rho^2 + \tau^2} \right) \left\{ \mu_x - \rho \left(\frac{\sigma_x}{\sigma_z} \right) \mu_z \right\}, \tag{2.2}$$

$$\frac{\beta_1^*}{\beta_1} = \frac{1}{1 + \tau^2/(1 - \rho^2)}, \tag{2.3}$$

$$\beta_2^* - \beta_2 = \beta_1 \rho \left(\frac{\sigma_x}{\sigma_z} \right) \left(\frac{\tau^2}{1 - \rho^2 + \tau^2} \right), \tag{2.4}$$

where $\mu_x = E(X)$, $\mu_z = E(Z)$, $\sigma_x^2 = Var(X)$, $\sigma_z^2 = Var(Z)$, and $\rho = Cor(X, Z)$.

We do not offer a proof of Result 2.1 here. Rather, a somewhat more general result is proved later in Section 2.9. To be clear about the generality inherent in Result 2.1, note there is only reference to moments of (X, Z) and $X^*|X$, and in particular no normal distributions or other particular forms are imposed. Also, the expressions for β^* in terms of β make no reference to the actual distribution of the response Y given the precise predictors X and Z. While β describes the large-sample result of fitting a linear model for Y in terms of X and Z, there are no assumptions about the appropriateness of this model. Whereas expressions of the form (2.2) through (2.4) are very well known (see, for instance, Carroll, Ruppert and Stefanski, 1995), they are usually obtained under the assumption that Y really is related to (X, Z) according to the postulated linear model. The more general view of considering the relationship

between the two sets of limiting coefficients is stressed in Gustafson and Le (2002). It is curious that at least asymptotically the bias induced by measurement error is unaffected by how well or poorly the postulated model for $Y|X, Z$ captures the actual relationship. Of course the generality of Result 2.1 does come at the price of being a statement about large-sample limits rather than finite-sample bias as in the previous section. We return to this point in Section 2.5.

Focussing on inference about the effect of X on Y, clearly the attenuation factor (2.3) generalizes (2.1) from the previous section, as the former reduces to the latter in the absence of correlation between X and Z. That is, the measurement error bias is unaffected by the inclusion of an additional precisely measured predictor Z only if this predictor is not correlated with X. On the other hand, the attenuating effect of measurement error worsens if Z is correlated with X. Note in particular that (2.3) goes to zero as ρ goes to one, indicating that very high correlation between X and Z yields extreme attenuation if the estimated effect of X^* is interpreted as estimating the effect of X.

To consider the role of correlation between X and Z more closely, Figure 2.2 displays attenuation curves, that is (2.3) as a function of τ, for various values of ρ. The display underscores that point that for moderate correlation the attenuation (2.3) is not much worse than (2.1) in the single predictor scenario. For large correlations, however, the attenuation is considerably stronger. Thus in practice one should be aware that strong correlations between the imprecisely measured predictor and other precisely measured predictors can warn of moderate measurement errors producing a substantial bias.

The other notable feature of Result 2.1 is that the measurement error in X also biases inference about the effect of Z on Y, despite the fact that Z is measured precisely. Furthermore, the bias $\beta_2^* - \beta_2$ scales with β_1 rather than with β_2. Whereas $\beta_1 = 0$ implies $\beta_1^* = 0$, so that one will not be misled about the absence of an effect of X on Y, $\beta_2 = 0$ will generally yield a nonzero value of β_2^*, provided both ρ and β_1 are nonzero. Thus an additional troubling aspect of measurement error is its ability to induce 'apparent significance' in other precisely measured predictors which actually have no effect on the response.

To further explore the impact of measurement error in X on inference about the effect of Z on Y, consider the situation where in fact Z is the explanatory variable of interest, but X is a suspected confounder. That is, the explanatory variable of interest can be measured precisely, but the confounding variable is subject to measurement error. It is easy to verify that if Y is regressed on $(1, Z)$ only then the large-sample limiting coefficient for Z will be

$$\tilde{\beta}_2 \quad = \quad \beta_2 + \rho \left(\frac{\sigma_x}{\sigma_z} \right) \beta_1,$$

where (β_1, β_2) are limiting coefficients as defined in Result 2.1. That is, if one really wants to estimate the impact of Z on Y given X, but X is omitted from the regression model, then one will be estimating $\tilde{\beta}_2$ rather than β_2 as desired.

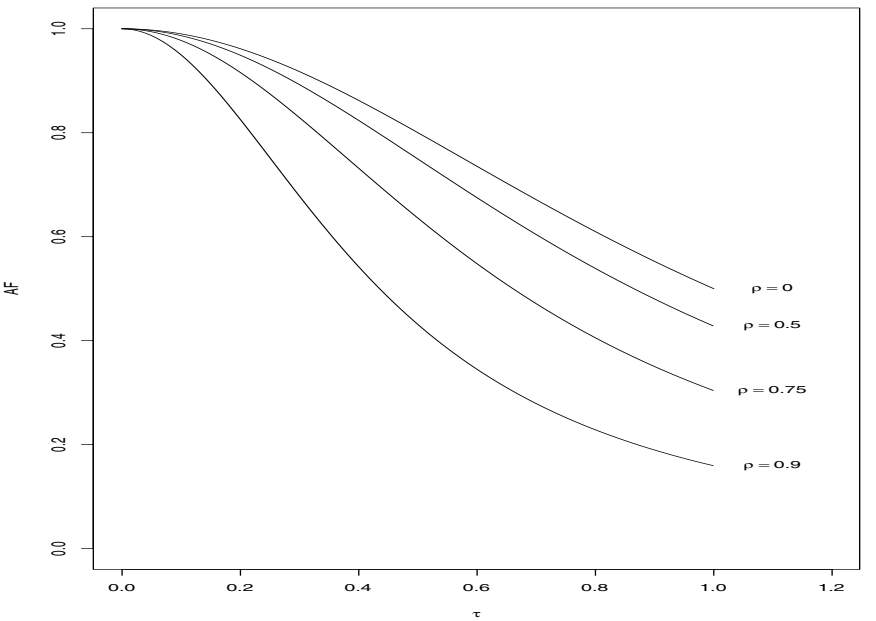

Figure 2.2 *Attenuation curves for linear regression with an additional precisely measured predictor. Each curve is (2.1) as a function of τ, with curves for $\rho = 0$, $\rho = 0.5$, $\rho = 0.75$, and $\rho = 0.9$ appearing.*

In this framework it is interesting to note that (2.4) can be re-expressed as

$$\beta_2^* = \{1 - w(\tau^2, \rho^2)\}\beta_2 + w(\tau^2, \rho^2)\tilde{\beta}_2, \tag{2.5}$$

where

$$w(\tau^2, \rho^2) = \frac{1}{1 + (1 - \rho^2)/\tau^2}.$$

Thus the limiting coefficient for Z is a weighted combination of β_2, which is completely adjusted for confounding, and $\tilde{\beta}_2$, which is completely unadjusted for confounding. When $\tau = 0$ (or $\rho = 0$) all the weight is on the adjusted coefficient, but as τ goes to infinity (or ρ goes to one) all the weight goes to the unadjusted coefficient. In this sense measurement error in a confounder leads to some loss in the ability to control for the confounder in the analysis. This point is made in the case of categorical explanatory variables by Greenland (1980). We also note that correlation between X and Z imparts a 'double-whammy' in that a larger ρ corresponds to a bigger discrepancy between β_2 and $\tilde{\beta}_2$, as well as more weight on $\tilde{\beta}_2$ in (2.5).

Of course Result 2.1 is still limiting in that only one additional precise predictor is considered. Many practical problems will involve numerous such pre-

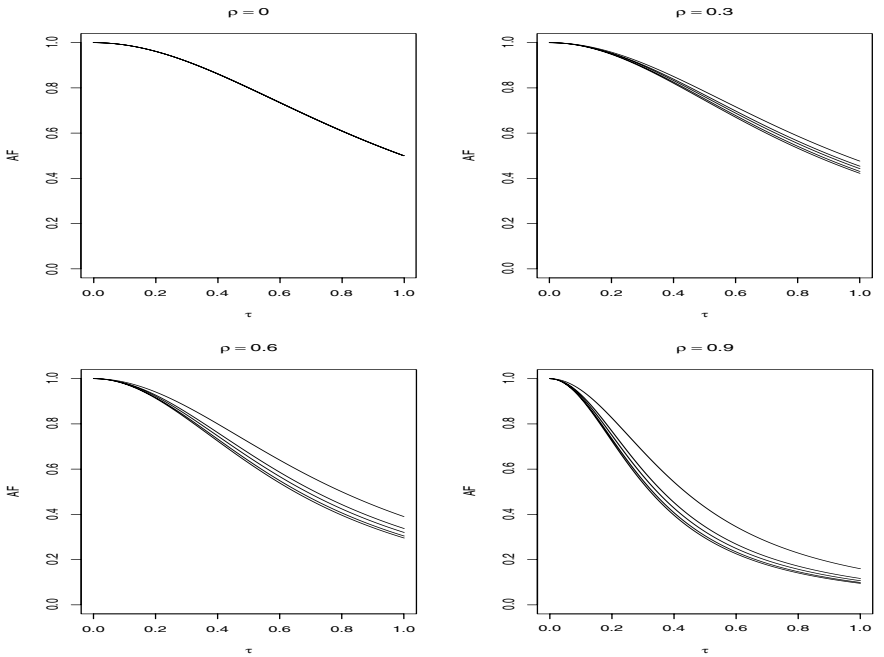

Figure 2.3 *Attenuation curves for linear regression with multiple equi-correlated predictors. Each panel corresponds to a different value of the common correlation ρ. Each curve is the attenuation factor (2.21) as a function of τ. The curves are pointwise decreasing in d, the number of additional precise predictors included in the regression. Curves for $d = 1$, $d = 3$, $d = 5$, $d = 10$, and $d = 20$ appear.*

dictors. Later in Section 2.9 we state and prove Result 2.6 which handles any number of additional precise predictors. The resulting expressions are more unwieldy, but one easily-studied scenario involves equi-correlated predictors. That is, say that X^* and d precise predictors $Z_1, \ldots, Z_d$ are measured, with $\rho = Cor(X, Z_i) = Cor(Z_i, Z_j)$ for all i and j. The expression (2.21) in Section 2.9 gives the attenuation factor in this scenario. Plots of the attenuation curve for various values of d and ρ appear in Figure 2.3. For $\rho > 0$ the attenuation is seen to worsen as the number of additional predictors d increases, but this effect is fairly slight. Thus in practice the presence of many additional precise predictors is only a mild indication of more severe attenuation due to measurement error.

2.3 Multiplicative Measurement Error

In the previous section the imprecision in the predictor arose from *additive* measurement error. We now turn to the case of *multiplicative* measurement error. That is, we consider situations where the amount of imprecision in the

measurement of a positive predictor is proportional to the actual value of the predictor. Many predictors in biostatistical and epidemiological settings are subject to this sort of measurement error.

A first thought is that multiplicative measurement error can be converted to additive measurement error simply by transforming the predictor in question to the logarithmic scale. Indeed, say that U is the positive explanatory variable and U^* is its imprecise surrogate. We might define the measurement error as being multiplicative if the conditional distribution of U^*/U given U does not depend on U. This is equivalent to saying that the conditional distribution of $X^* - X$ does not depend on X, where $X = \log U$ and $X^* = \log U^*$, which might similarly be thought of as a characterization of additive measurement error. Consequently, separate treatments of additive and multiplicative measurement error are not really required. What is needed, however, is consideration of which function of X (or equivalently which function of U) is used as a regressor in the linear model for the response variable.

Consider the scenario where the positive predictor U and an additional predictor Z are the explanatory variables of interest, but a surrogate U^* and Z are observed. The investigator must postulate a model for Y in term of U and Z. If he chooses $\log U$ and Z as regressors but instead actually uses $\log U^*$ and Z, then Result 2.1 applies. On the other hand, if U itself or perhaps some power of U is used as a regressor then Result 2.1 is not applicable. The next result deals with these possibilities. For consistency with Result 2.1 it is expressed in terms of X^* arising from X via additive measurement error, but as discussed it can be viewed as dealing with $U^* = \exp(X^*)$ arising from $U = \exp(X)$ via multiplicative measurement error. The framework of the result and the attenuation factor determined are somewhat similar to work of Lyles and Kupper (1997), who go on to consider adjustments for mismeasurement in this context. More generally, though, there is little literature concerned with situations where the scale on which the mismeasured predictor is subject to additive measurement error differs from the scale on which the predictor appears in the outcome model.

Result 2.2 *Assume that* (Y, e^{kX}, Z) *jointly have finite second moments. Say* X^* *and* (Z, Y) *are conditionally independent given* X. *Let* $\sigma^2 = Var(X)$ *and assume* $X^*|X \sim N(X, \tau^2\sigma^2)$. *Let* β *be the large-sample limiting coefficients from linear regression of* Y *on* $(1, e^{kX}, Z)$, *and let* β^* *be the large-sample limiting coefficients from linear regression of* Y *on* $(1, e^{kX^*}, Z)$. *Then*

$$\frac{\beta_1^*}{\beta_1} = \frac{1}{\exp(k^2\sigma^2\tau^2/2)\left[1 + c\left\{\exp(k^2\sigma^2\tau^2) - 1\right\}\right]}, \tag{2.6}$$

where

$$c = \frac{E\left(e^{2kX}\right)}{Var\left(e^{kX}\right)\left\{1 - Cor\left(e^{kX}, Z\right)^2\right\}}. \tag{2.7}$$

A proof of this result is deferred to Section 2.9.

In examining (2.6), clearly the denominator is greater than one and in-

creasing in τ, so again an attenuating bias results from the measurement error. Moreover, from (2.7) the attenuation is seen to worsen with increasing correlation between the precise regressors $\exp(kX)$ and Z. Note that the conditions in Result 2.2 are a little more restrictive than those in Result 2.1. In particular, normality of $X^*|X$ is now required to obtain somewhat tractable expressions.

It is somewhat instructive to consider a Taylor series approximation to the denominator of (2.6), expanding in τ about zero. The first derivative term vanishes, giving the small τ approximation

$$\frac{\beta_1^*}{\beta_1} \approx \frac{1}{1+(0.5+c)k^2\sigma^2\tau^2}.$$

Of course this has a similar form to (2.3) in the situation where X itself is considered as the regressor.

To consider a specific case more clearly, say X and Z have a bivariate normal distribution, and let $\rho = Cor(X, Z)$. It is straightforward to verify that (2.7) specializes to

$$c = \frac{e^{k^2\sigma^2}}{\left(e^{k^2\sigma^2}-1\right)\left[1-\left\{k^2\sigma^2/\left(e^{k^2\sigma^2}-1\right)\right\}\rho^2\right]}.$$

In this scenario $U = \exp(X)$ has a log-normal distribution, with $\sigma = SD(X)$ becoming a shape parameter which governs the skewness of this distribution. The top panels in Figure 2.4 display the log-normal density function for U when $\sigma = 0.25$, $\sigma = 1$, and $\sigma = 1.75$. In each case $\mu = E(X) = 0$ so that the median of U is one. Note, however, that this simply fixes the scale of U, and the bias (2.6) will not depend on μ. The panels below each density give multiplicative bias curves when σ takes the value in question, for three different values of $\rho = Cor(X, Z)$. Specifically, three curves appear in each panel. The solid curve is (2.6) with $k = 1$, describing a situation where U itself is used as a regressor. The dashed curve is (2.6) with $k = 0.5$, describing a situation where $U^{1/2}$ is used as a regressor. Finally, the dotted curve is (2.3) from Result 2.1, describing the situation where $\log U$ is used as a regressor.

The figure shows that when $\sigma = 0.25$ the density of U is quite symmetric, and for each value of ρ the three bias curves corresponding to the three choices of regressor are nearly indistinguishable. When $\sigma = 1$, however, the density of U is quite skewed, with a long right tail. In this case, unless ρ is very large, there are substantial differences between the biases for the three choices of regressor. Specifically, the use of U itself leads to the most attenuation, followed by the use of $U^{1/2}$ and then the use of $\log U$. These differences are magnified when $\sigma = 1.75$, corresponding to even more skewness in the distribution of U. Again, however, the behaviour when $\rho = 0.9$ is not in accord with the behaviour for $\rho = 0$ and $\rho = 0.5$.

We surmise that multiplicative measurement error has the potential to yield particularly strong attenuation in the estimated regression coefficient for the regressor subject to measurement error. Specifically, this potential is realized if (i) the positive explanatory variable subject to multiplicative measurement

error has a skewed distribution, and (ii) this variable itself, or possibly a power transformation of this variable, enters the outcome model linearly. On the other hand, if the regressor is obtained via a log transformation, then the induced bias reduces to that obtained in the additive error case as described in Result 2.1. Of course this finding makes good intuitive sense. If U itself is used as a regressor and its distribution has a long right tail, then the multiplicative measurement error yields the largest errors for a few potentially influential data points in the tail of the distribution.

2.4 Multiple Mismeasured Predictors

Of course more than one of the predictors in a given problem may be subject to measurement error. The following result describes what happens when both explanatory variables under consideration are subject to additive measurement error. In addition to allowing for correlation between the two precise predictors, we allow for correlation between the two additive errors. Such dependence would be plausible in a number of circumstances. For instance, if the two regressors are daily levels of two ambient pollutants, it is plausible that the instrumentation used to measure these levels could yield correlated errors in the two measurements.

Result 2.3 *Assume* (Y, X_1, X_2) *jointly have finite second moments, where each* X_i *has been scaled so that* $E(X_i) = 0$ *and* $Var(X_i) = 1$. *Let* $\rho = Cor(X_1, X_2)$. *Assume* $X^* = (X_1^*, X_2^*)$ *is conditionally independent of* Y *given* $X = (X_1, X_2)$, *with* $E(X_i^*|X) = X_i$ *and*

$$Var(X^*|X) = \begin{pmatrix} \tau_1^2 & \lambda\tau_1\tau_2 \\ \lambda\tau_1\tau_2 & \tau_2^2 \end{pmatrix}.$$

Let $\beta = (\beta_0, \beta_1, \beta_2)'$ *be the large-sample limiting coefficients from linear regression of* Y *on* $(1, X_1, X_2)'$, *and let* $\beta^* = (\beta_0^*, \beta_1^*, \beta_2^*)'$ *be the large-sample limiting coefficients from linear regression of* Y *on* $(1, X_1^*, X_2^*)'$. *Then*

$$\beta_1^* = \frac{(1 + \tau_2^2 - \rho^2 - \rho\lambda\tau_1\tau_2)\beta_1 + (\rho\tau_2^2 - \lambda\tau_1\tau_2)\beta_2}{(1 + \tau_1^2)(1 + \tau_2^2) - (\rho + \lambda\tau_1\tau_2)^2}, \tag{2.8}$$

with a symmetric expression obtaining for β_2^*.

A proof of Result 2.3 is not given, but can be established along the same lines as the proof Result 2.6 in Section 2.9. Such expressions are well established in the literature (see, for example, Carroll, Ruppert and Stefanski 1995, Sec. 2.2.4), but again they are typically derived under the assumption that the postulated linear model for Y given (X_1, X_2) is correct. Thus we emphasize that again in this scenario the bias induced by measurement error does not depend on the actual distribution of the response given the precise predictors.

With variances τ_1^2 and τ_2^2 and correlations ρ and λ at play, it is difficult to glean much understanding by examining (2.8) directly. Rather, we fix $(\beta_1, \beta_2) = (0.5, 0.25)$, and compute (β_1^*, β_2^*) for various values of τ_1, τ_2, ρ, and λ. The results appear in Table 2.1.

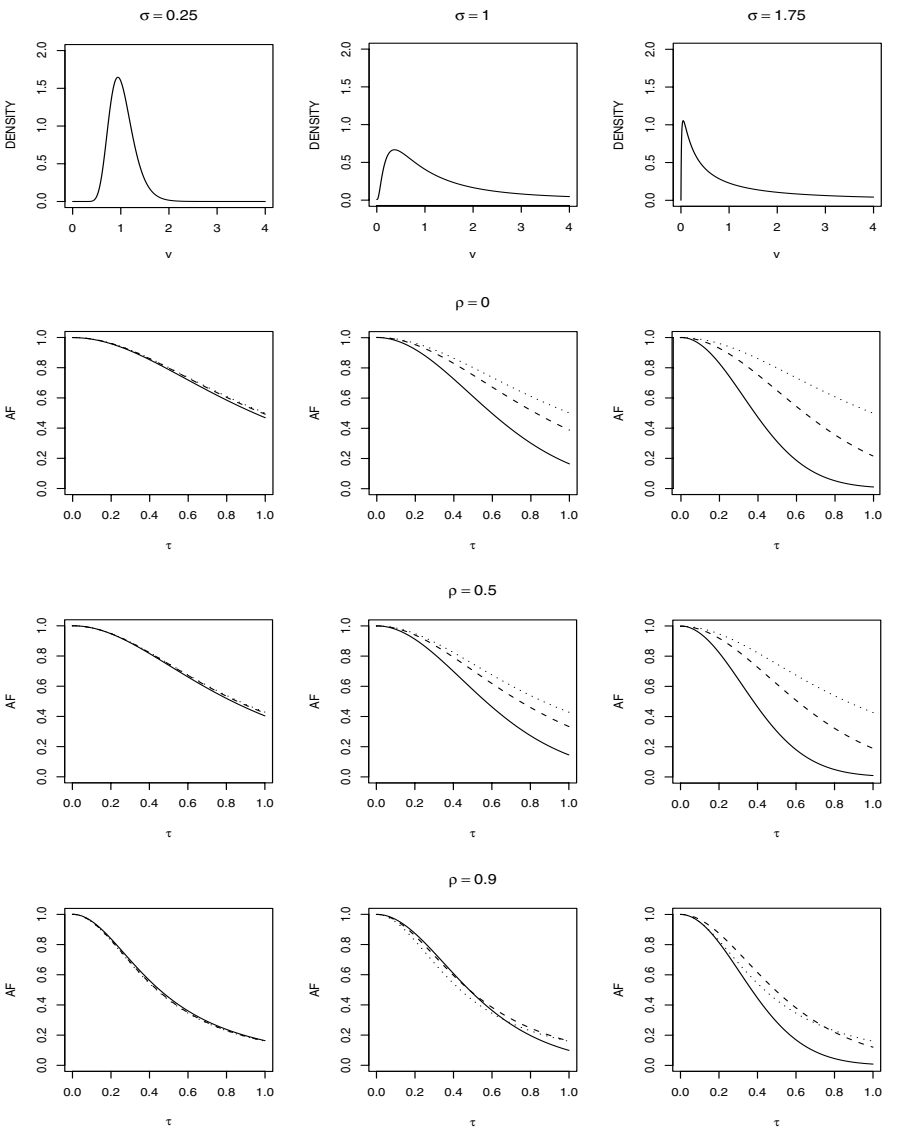

Figure 2.4 *Log-normal densities and corresponding attenuation curves. The columns of plots correspond to log-normal shape parameters* $\sigma = 0.25$, $\sigma = 1$, *and* $\sigma = 1.75$. *The top panels give the log-normal density functions scaled to have median zero. The panels below give the resulting attenuation curves, with the three rows corresponding to different values of* $\rho = Cor(\log U, Z)$. *Each panel contains three curves. The solid curve corresponds to the use of* U *as a regressor, the dashed curve corresponds to the use of* $U^{1/2}$ *as a regressor, and the dotted curve corresponds to the use of* $\log U$ *as a regressor.*

λ	τ_1	τ_2	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.9$
0.0	0.2	0.2	(0.48 , 0.24)	(0.48 , 0.25)	(0.46 , 0.28)
	0.2	0.5	(0.48 , 0.20)	(0.51 , 0.20)	(0.56 , 0.15)
	0.5	0.2	(0.40 , 0.24)	(0.38 , 0.30)	(0.25 , 0.45)
	0.5	0.5	(0.40 , 0.20)	(0.40 , 0.24)	(0.37 , 0.30)
0.5	0.2	0.2	(0.48 , 0.23)	(0.48 , 0.24)	(0.47 , 0.26)
	0.2	0.5	(0.47 , 0.18)	(0.51 , 0.18)	(0.61 , 0.10)
	0.5	0.2	(0.39 , 0.22)	(0.38 , 0.28)	(0.22 , 0.47)
	0.5	0.5	(0.38 , 0.16)	(0.40 , 0.20)	(0.37 , 0.26)
0.9	0.2	0.2	(0.47 , 0.22)	(0.48 , 0.23)	(0.48 , 0.24)
	0.2	0.5	(0.47 , 0.17)	(0.51 , 0.16)	(0.67 , 0.03)
	0.5	0.2	(0.39 , 0.21)	(0.37 , 0.27)	(0.19 , 0.49)
	0.5	0.5	(0.38 , 0.13)	(0.40 , 0.17)	(0.40 , 0.20)

Table 2.1 *Values of* (β_1^*, β_2^*) *when both predictors are subject to additive measurement error. The limiting coefficients in the absence of measurement error are* $(\beta_1, \beta_2) = (0.5, 0.25)$. *Various values of* $\tau_i = Var(X_i^*|X)$, $\rho = Cor(X_1, X_2)$, *and* $\lambda = Cor(X_1^*, X_2^*|X)$ *are considered.*

Perhaps the most notable observation from Table 2.1 is that the effect of imprecision in two explanatory variables is not always attenuation. That is, some scenarios yield $|\beta_1^*| > |\beta_1|$, or $|\beta_2^*| > |\beta_2|$. Also, the relative magnitudes of the estimated effects can be distorted by the measurement error, as $\beta_1^* < \beta_2^*$ in some scenarios even though $\beta_1 = 2\beta_2$. In short, with more than one imprecise covariate there is more danger of drawing substantively erroneous conclusions, and the nature of the induced bias is not easily predicted without knowledge of the magnitudes of the measurement errors, the correlation between the underlying precise predictors, and the correlation between the measurement errors.

Given that Result 2.3 applies after each regressor has been standardized, we can regard the predictor with the larger $|\beta_i|$ as being the more important of the two in explaining variation in Y. Of course if the data are analyzed without considering the measurement error then the predictor with the larger $|\beta_i^*|$ will be *apparently* more important. We have already noted from Table 2.1 that the predictor which is apparently more important may not be actually more important. Obviously this sort of *reversal* is very damaging, as it leads to a qualitatively wrong assessment of the relationship between the response and the predictors.

Looking more closely at Table 2.1, we see reversals when the two underlying precise predictors are highly correlated, but the one that is actually more important is measured with more error than the other. To study the potential for reversal more carefully, consider the case of uncorrelated measurement errors, that is $\lambda = 0$. Say $\beta_1 > 0$ and $\beta_2 > 0$, with $\beta_1/\beta_2 = \gamma > 1$. From (2.8) one would mistakenly conclude that X_2 is more important if

$$(1 + \rho\gamma)\tau_1^2 + (1 - \rho^2) \quad > \quad (\gamma + \rho)\tau_2^2 + \gamma(1 - \rho^2). \tag{2.9}$$

If the precise predictors have little or no correlation, then reversal is not likely to be a problem. For instance, say $\gamma = 1.25$, $\tau_2 = 0.1$, and $\rho = 0$. Then (2.9) shows that reversal would require a substantial measurement error of $\tau_1 > 0.51$. The situation changes, however, when the precise predictors are highly correlated. In fact, as ρ tends to one, the inequality (2.9) tends to $\tau_1 > \tau_2$, for any value of γ. That is, in the limit of perfect correlation the more precisely measured predictor will be implicated as the more important predictor, regardless of the actual underlying relationship between Y and X. As examples with strong but not perfect correlation, again say that $\gamma = 1.25$ and $\tau_2 = 0.1$. When $\rho = 0.75$, (2.9) shows that $\tau_1 > 0.26$ is required for reversal, and when $\rho = 0.9$, $\tau_1 > 0.18$ is required. Thus reversal is clearly more of a danger if the two predictors in question are highly correlated. Related questions of whether the right explanatory variable is implicated in an environmetrics scenario are considered by Zidek, Wong, Le and Burnett (1996).

2.5 What about Variability and Small Samples?

So far in this chapter we have used large-sample limits to compare regression coefficients based on imprecise predictors to those based on precise predictors. An obvious question is how well do the resulting asymptotic biases reflect what actually happens for a particular fixed sample size.

To investigate we return to the context of Section 2.2 and perform a simulation study with three different distributions for the precise predictors (X, Z) and two different distributions for the response Y given these predictors. Thus six scenarios are considered in all. The three different distributions for the predictors all have $E(X) = E(Z) = 0$, $Var(X) = Var(Z) = 1$, and $\rho = Cor(X, Z) = 2/3$. Thus by Result 2.1 the large-sample multiplicative bias in estimating the effect of X is the same in all six scenarios.

To be more specific, the first distribution of predictors is 'short-tailed.' Both X and Z are taken to have uniform distributions on $(-\sqrt{3}, \sqrt{3})$. But rather than having a jointly uniform distribution, some probability is moved from the quadrants where X and Z have differing signs to the quadrants where X and Z have the same signs, to achieve $Cor(X, Z) = 2/3$. The second distribution of predictors is simply bivariate normal. The third distribution is bivariate Student's t, with four degrees of freedom for each component and the desired covariance matrix.

Both distributions for the response involve a normal distribution for $Y|X, Z$ with $Var(Y|X, Z) = 1$. In the first case,

$$E(Y|X, Z) = X + (0.5)Z, \tag{2.10}$$

giving scenarios in which the postulated linear model for the response given the precise predictors is correct. In the second case,

$$E(Y|X, Z) = e^X + (0.5)I\{Z > 0\}, \tag{2.11}$$

under which the postulated linear model is clearly not correct. Result 2.1 indicates that asymptotically the measurement error bias is the same for all

distributions of the response given the precise predictors, but by trying both (2.10) and (2.11) we will see how well this translates to fixed sample sizes.

We simulate 2000 samples of (Y, X, Z, X^*) for each scenario and for each of three sample sizes: $n = 50$, $n = 200$, and $n = 400$. In each instance the measurement error is induced via $X^*|X, Z, Y \sim N(X, \tau^2)$ with $\tau = 0.5$. For each sample we compute $\hat{\beta}$, the estimated coefficients from regression of Y on $(1, X, Z)$, and $\hat{\beta}^*$, the estimated coefficients from regression of Y on $(1, X^*, Z)$. Figure 2.5 uses boxplots to illustrate the sampling distribution of the finite sample *realized attenuation*, $\hat{\beta}_1^*/\hat{\beta}_1$, for each scenario and sample size. The large-sample attenuation factor (2.3), which equals 20/29 in the present scenario, is also marked on the plots.

The Figure indicates considerable variability in the sampling distribution of the realized attenuation, especially when n is small. Even when $n = 50$, however, the large-sample attenuation factor is a very good descriptor of the centre of the sampling distribution. In particular, the centre exhibits very little variation across scenarios. The same is not true of the spread in the sampling distributions. Somewhat predictably, the realized attenuation is more variable when (X, Z) have the long-tailed distribution, especially when the response variable is not linear in the predictors. For the other (X, Z) distributions, however, the distribution of the realized attenuation is almost the same for the two response distributions. In all, the simulations lend credence to the large-sample attenuation factor as a good descriptor of the measurement error impact in smaller samples.

Another point concerning the variability of naive estimators which don't correct for measurement error concerns the reported uncertainty associated with such estimators. To make the point simply, consider the scenario above where the distribution of X is normal and (2.10) holds, so that $E(Y|X, Z)$ and $E(Y|X^*, Z)$ are both linear in their arguments, with coefficient vectors β and β^* respectively. Naive estimation ignoring measurement error yields $\hat{\beta}_1^* \pm 1.96 \times SE[\hat{\beta}_1^*]$ as a reported 95% confidence interval for β_1. Of course the primary problem here is that the interval is centred in the wrong place, as has been discussed. However, there is also the secondary problem that the interval is too narrow. Intuitively, it does not reflect the loss of information associated with measuring only X^* and not X. To elaborate, say the attenuation factor is known to be $AF \in (0, 1)$, i.e., $\beta_1^* = AF\beta_1$. Then a simple corrected estimate for β_1 would be $AF^{-1}\hat{\beta}_1^*$. Clearly an approximate 95% confidence interval centred at this estimate must take the form $AF^{-1}\hat{\beta}_1^* \pm 1.96 \times AF^{-1}SE[\hat{\beta}_1^*]$, in order to achieve the nominal coverage. This arises easily by transforming the approximate confidence interval for β_1^*. Thus to construct an appropriate interval estimate from the naive interval we must both move the centre of the interval *and* make the interval wider by a factor of AF^{-1}. While this chapter and the next emphasize the bias inherent in naive estimation, it should be remembered that underestimation of apparent variability is also manifested.

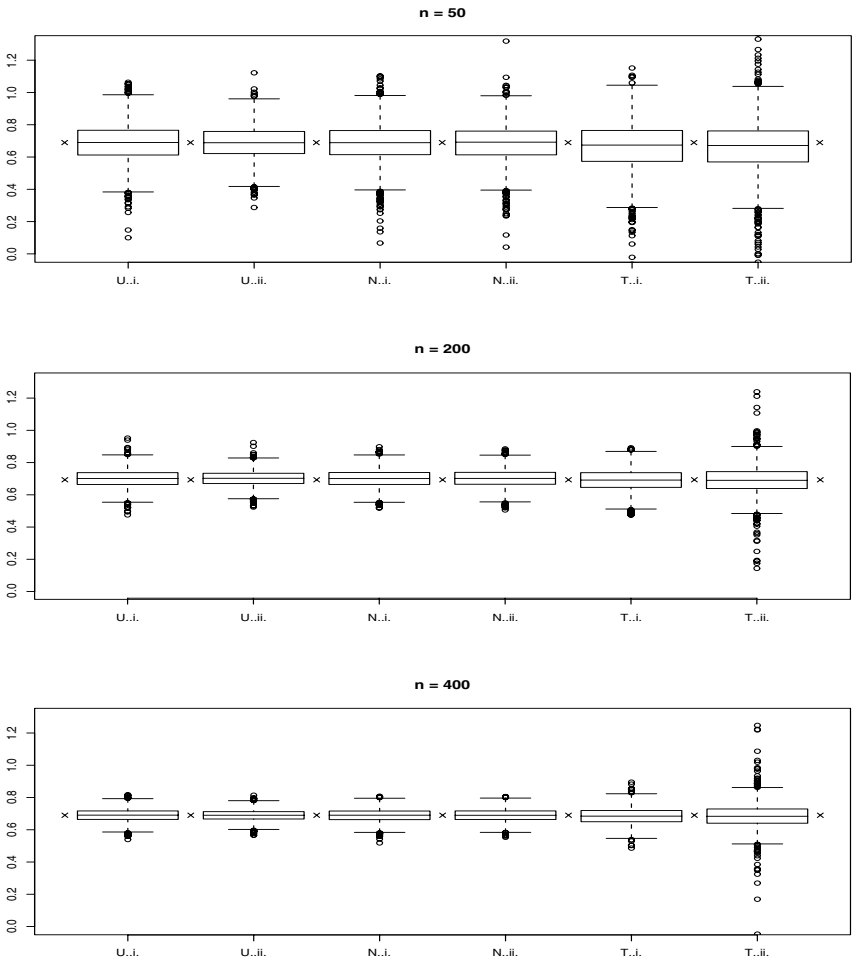

Figure 2.5 *Sampling distribution of the realized attenuation for six scenarios and three sample sizes. At each sample size the first two boxplots correspond to the modified uniform distribution of predictors, the second two correspond to the bivariate normal distribution, and the last two correspond to the bivariate t distribution with four degrees of freedom. In each pair, the first boxplot corresponds to the linear response (2.10) and the second to the nonlinear response (2.11). The crosses between the boxplots mark the asymptotic bias which is the same for all scenarios. Each boxplot is based on* 2000 *simulated samples.*

2.6 Logistic Regression

Up to this point we have considered the bias induced by measurement error when linear regression is used to relate a continuous response variable to explanatory variables. Of course many analyses that arise in biostatistics and epidemiology in particular involve a binary response variable. In such cases the most common inferential procedure is logistic regression, which is one particular case of a generalized linear model. Roughly stated, a 'folk theorem' pervading statistics is that generalized linear models behave very similarly to linear models. We put this to the test by examining the impact of measurement error when logistic regression is applied to data with a binary outcome variable. There is a considerable literature on measurement error in binary regression models, though some of this focusses on probit regression rather than logistic regression for the sake of numerical tractability (Carroll, Spiegelman, Lan, Bailey and Abbott, 1984).

Unfortunately, closed-form expressions for the bias induced by measurement error in logistic regression do not exist. Thus some numerical work is required to mimic analytic findings such as Results 2.1, 2.2, and 2.3 for linear regression. For instance, consider the following scenario which parallels that of Result 2.1. An investigator is interested in a logistic regression of a binary outcome variable Y on continuous regressors X and Z. Due to measurement error, however, he actually performs a logistic regression of Y on X^* and Z, where X^* and (Y, Z) are conditionally independent given X, $E(X^*|X) = X$, and $Var(X^*|X) = \tau^2 Var(X)$. If we define $T = (1, X, Z)'$ and let β be the large-sample limiting coefficients from logistic regression of Y on the components of T, then standard large-sample theory (e.g., White 1980) for logistic regression gives β as the solution to a system of three equations, compactly denoted in vector form as

$$E\left[T\left\{Y - \frac{1}{1+\exp(-\beta' T)}\right\}\right] \quad = \quad 0. \tag{2.12}$$

Similarly, if $T^* = (1, X^*, Z)'$ and β^* is the large-sample limiting coefficients from logistic regression of Y on T^*, then

$$E\left[T^*\left\{Y - \frac{1}{1+\exp(-\beta^{*\prime} T^*)}\right\}\right] \quad = \quad 0. \tag{2.13}$$

Stefanski and Carroll (1985) use a variant of (2.13) to provide large-sample approximations to the bias incurred in estimating β_1 when using T^* rather than T as regressors.

We proceed by noting that under the stated conditions $E(X^*Y) = E(XY)$, and hence $E(T^*Y) = E(TY)$. This leads to β^* as a function of β being determined as the solution to

$$E\left[T^*\left\{\frac{1}{1+\exp(-\beta^{*\prime} T^*)}\right\}\right] \quad = \quad E\left[T\left\{\frac{1}{1+\exp(-\beta' T)}\right\}\right]. \tag{2.14}$$

While the system of equations (2.14) does not have an analytic solution, its form implies that the relationship between β^* and β is not affected by

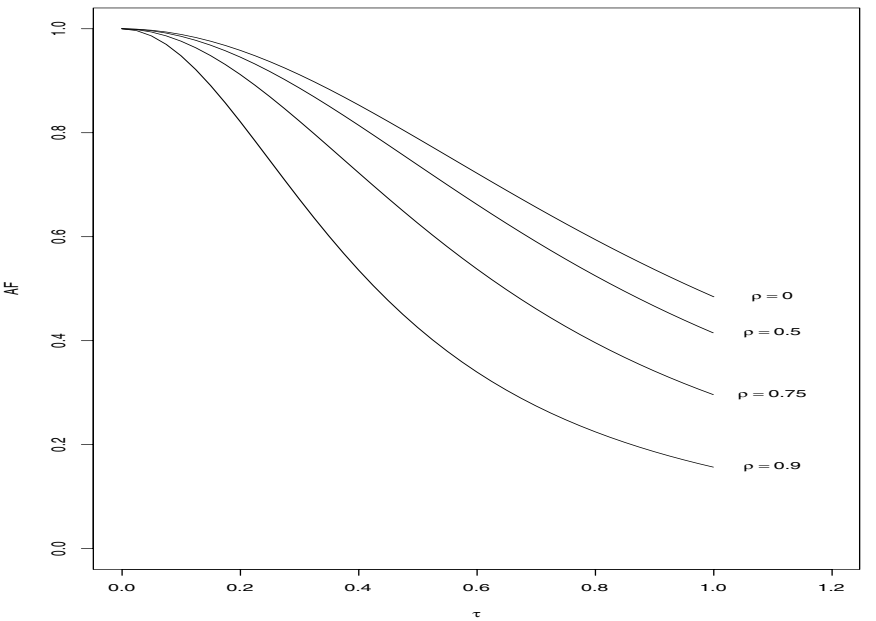

Figure 2.6 *Attenuation curves for logistic regression in the case of one additional precise predictor. Each curve is* β_1^*/β_1 *as a function of* τ*, with curves for* $\rho = 0$*,* $\rho = 0.5$*,* $\rho = 0.75$*, and* $\rho = 0.9$ *appearing. The large-sample limiting coefficients for logistic regression of* Y *on* X *and* Z *are taken to be* $\beta = (0, 0.5, 0.25)'$*.*

the actual distribution of $Y|X, Z$. This is in accord with the linear regression case as per Result 2.1. On the other hand, some nice features of the earlier result do not carry over to the present setting. For instance, in Result 2.1 the attenuation factor β_1^*/β_1 is the same for all values of β. The solution to (2.14) does not have this property, although our numerical experience suggests that the dependence of β_1^*/β_1 on β is quite mild.

In Section 2.9 we state and prove Result 2.7 which yields a variant of (2.14), and describe a computational scheme to compute β^* for a given β approximately. In particular, this scheme uses a combination of Monte Carlo simulation and the solution of 'pseudo' logistic regression likelihood equations. This scheme is applied to solve (2.14) in a scenario where X and Z have a bivariate normal distribution with standardized marginals, $X^*|X$ is normally distributed, and $\beta = (0, 0.5, 0.25)'$. Attenuation curves, that is β_1^*/β_1 as a function of τ, appear in Figure 2.6, for various values of $\rho = Cor(X, Z)$. These curves bear a very strong resemblance to those for linear regression in Figure 2.2. This supports the notion that the impact of measurement error in logistic regression is very similar to the impact in linear regression.

We also investigate the effects of multiplicative measurement error in logistic

regression, along the lines of Result 2.2 and the ensuing discussion in Section 2.3. Say (X, Z) have a bivariate normal distribution with $E(X) = E(Z) = 0$, $Var(X) = \sigma^2$, and $Var(Z) = 1$, while $X^*|X, Z, Y \sim N(X, \tau^2\sigma^2)$. In analogy to Result 2.2 we consider logistic regression of Y on regressors $T^* = (1, e^{kX^*}, Z)'$ in lieu of $T = (1, e^{kX}, Z)'$. Again Result 2.7 and the associated computational scheme permit approximation of the large-sample coefficients β^* for T^* given large-sample coefficients β for T.

To illustrate we fix $\beta = (0, 0.5, 0.25)'$, and then compute attenuation curves which appear in Figure 2.7. As before, the additive measurement error for X can be regarded as multiplicative measurement error for $V = \exp(X)$. Three different values of $\sigma = Var(X)$ and three different values of $\rho = Cor(X, Z)$ are considered. In each case the solid curve corresponds to using $V^* = \exp(X^*)$ as a regressor in lieu of $V = \exp(X)$, the dashed curve corresponds to the use of $(V^*)^{1/2} = \exp(X^*/2)$ in lieu of $V^{1/2} = \exp(X/2)$, and the dotted curve corresponds to the use of $\log V^* = X^*$ in lieu of $\log V = X$. Thus Figure 2.7 describes the same scenarios as Figure 2.4, except that linear regression for a continuous response variable has been replaced with logistic regression for a binary response variable. The curves in the two figures are strikingly similar, again supporting the notion that the impact of measurement error in logistic regression mimics the impact in linear regression.

Finally we consider logistic regression with two regressors that are both subject to measurement error. Again the mathematical details are subsumed by Result 2.7 and the associated computational scheme. The joint distribution of $X = (X_1, X_2)$ and $X^* = (X_1^*, X_2^*)$ is taken to be multivariate normal, with moment structure exactly as in Result 2.3 for linear regression. In particular, X_1 and X_2 are both standardized, while $\tau_i^2 = Var(X_i^*|X)$, $\lambda = Cor(X_1^*, X_2^*|X)$, and $\rho = Cor(X_1, X_2)$. The large-sample limiting coefficients for logistic regression of Y on $(1, X_1, X_2)$ are taken to be $\beta = (0, 0.5, 0.25)'$. Table 2.2 gives (β_1^*, β_2^*) (the intercept β_0^* is not shown) under the same combinations of τ_i, ρ, and λ considered in Table 2.1 for linear regression. The agreement of the tables is very striking; entries differ by no more than 0.01 between the linear regression and logistic regression scenarios. Once again we have strong support for the notion that measurement error impacts logistic regression in very much the same manner that it impacts linear regression.

2.7 Beyond Nondifferential and Unbiased Measurement Error

To this point we have considered the impact of 'fair' measurement error, meaning (i) on some scale the imprecise predictor is an unbiased estimate of the precise predictor, and (ii) the distribution of the imprecise predictor given the precise predictor, other predictors and the response depends only on the precise predictor. However there are realistic situations where measurement error may be biased or differential. We briefly consider two such scenarios. Proofs of the results are omitted, as they are much in the spirit of the proof of Result 2.6 given in Section 2.9.

First, in the context of Result 2.1 we replace $E(X|X^*) = X$ with $E(X^*|X) =$

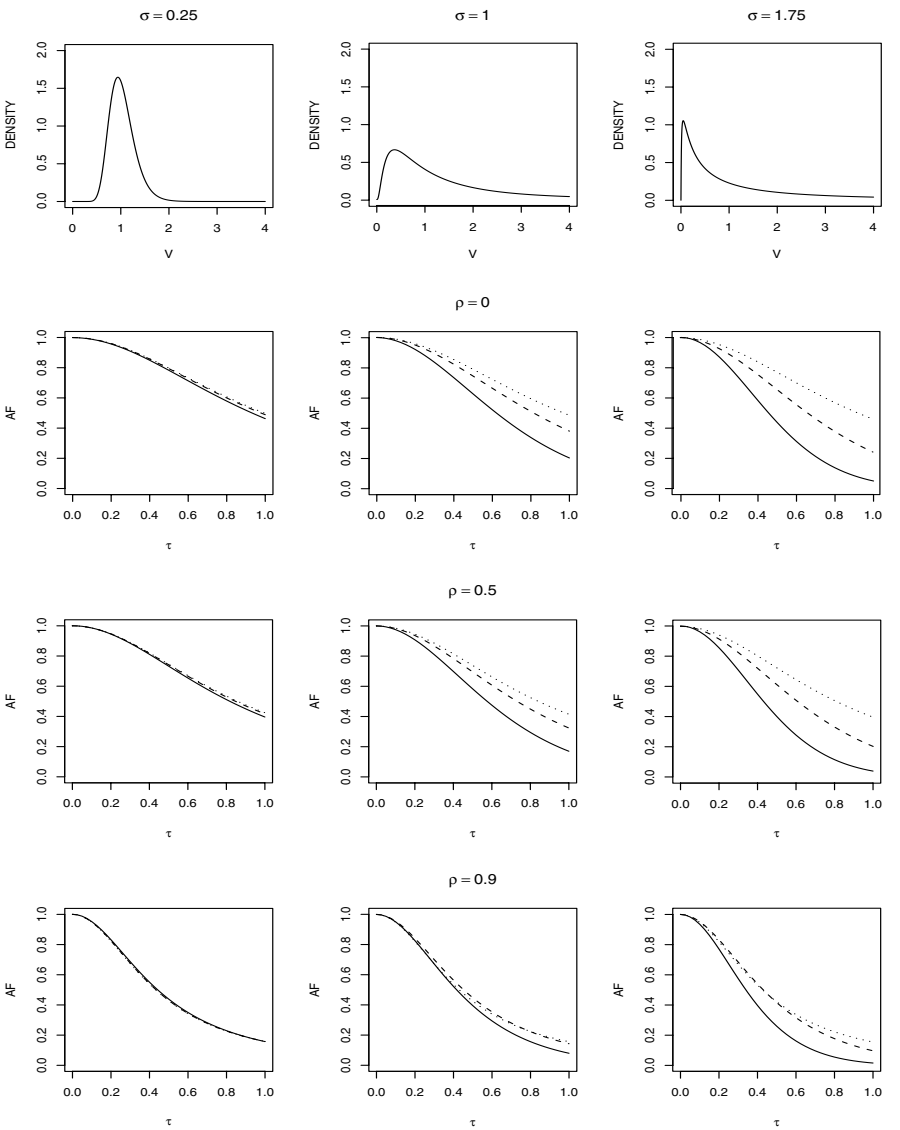

Figure 2.7 *Log normal densities and corresponding multiplicative bias curves for logistic regression with multiplicative measurement error. The top panels give density functions for* $V = \exp(X)$*, where* $X \sim N(0, \sigma^2)$*, for* $\sigma = 0.25$*,* $\sigma = 1$*, and* $\sigma = 1.75$*. The panels below give multiplicative bias curves for different values of* $\rho = Cor(X, Z)$*. Each panel contains three curves. The solid curve corresponds to the use of* V *as a regressor, the dashed curve corresponds to the use of* $V^{1/2}$ *as a regressor, and the dotted curve corresponds to the use of* $\log V$ *as a regressor. The large-sample limit for logistic regression of* Y *on* $(1, V, Z)$ *is taken to be* $\beta = (0, 0.5, 0.25)$ *in each case.*

λ	τ_1	τ_2	$\rho = 0.0$	$\rho = 0.5$	$\rho = 0.9$
0.0	0.2	0.2	(0.48 , 0.24)	(0.48 , 0.25)	(0.45 , 0.28)
	0.2	0.5	(0.48 , 0.20)	(0.50 , 0.20)	(0.56 , 0.15)
	0.5	0.2	(0.39 , 0.24)	(0.38 , 0.30)	(0.25 , 0.45)
	0.5	0.5	(0.39 , 0.20)	(0.40 , 0.24)	(0.36 , 0.29)
0.5	0.2	0.2	(0.47 , 0.23)	(0.48 , 0.24)	(0.46 , 0.26)
	0.2	0.5	(0.47 , 0.18)	(0.50 , 0.18)	(0.60 , 0.10)
	0.5	0.2	(0.38 , 0.22)	(0.37 , 0.28)	(0.22 , 0.47)
	0.5	0.5	(0.38 , 0.16)	(0.39 , 0.20)	(0.36 , 0.26)
0.9	0.2	0.2	(0.47 , 0.22)	(0.48 , 0.23)	(0.48 , 0.24)
	0.2	0.5	(0.46 , 0.16)	(0.51 , 0.16)	(0.66 , 0.03)
	0.5	0.2	(0.38 , 0.20)	(0.37 , 0.27)	(0.19 , 0.49)
	0.5	0.5	(0.37 , 0.13)	(0.39 , 0.16)	(0.39 , 0.20)

Table 2.2 *Values of (β_1^*, β_2^*) when both regressors in a logistic regression are subject to additive measurement error. The limiting coefficients in the absence of measurement error are $(\beta_0, \beta_1, \beta_2) = (0, 0.5, 0.25)$. Various values of $\tau_i = Var(X_i^*|X_i)$, $\rho = Cor(X_1, X_2)$, and $\lambda = Cor(X_1^*, X_2^*|X)$ are considered.*

$a + bX$. Thus we are considering a systematic bias in the measurement of the predictor. The modified result is as follows.

Result 2.4 *Assume that (Y, X, Z) jointly have finite second moments, where X and Z have been scaled so that $E(X) = E(Z) = 0$ and $Var(X) = Var(Z) = 1$. Say X^* and (Z, Y) are conditionally independent given X, with $E(X^*|X) = a + bX$ and $Var(X^*|X) = \tau^2$. Let β and β^* be the large-sample limiting coefficients from least-squares regression of Y on $(1, X, Z)$ and Y on $(1, X^*, Z)$ respectively. Then,*

$$\frac{\beta_1^*}{\beta_1} = \frac{1}{b + \tau^2/\{b(1-\rho^2)\}}. \tag{2.15}$$

It is not surprising that the attenuation factor does not depend on a. Invariance considerations imply that translation error in the form of adding the same constant to each predictor will not have any effect on the estimate of a slope parameter. On the other hand, a multiplicative bias in the measurement process, as manifested when $b \neq 1$, does alter the multiplicative bias in the estimated coefficient. For moderate measurement error ($\tau^2 < 1 - \rho^2$), (2.15) is a decreasing function of b in a neighbourhood of $b = 1$. Thus compared to the $b = 1$ case, attenuation is worse when the measurement error involves an upward bias ($b > 1$) which tends to overstate the magnitude of the predictor.

There are also plausible situations under which the measurement error is differential. Say Y is an indicator of health status coded such that larger values correspond to poorer health, while X is the level of a possible risk factor. If X^* is ascertained by some sort of 'self-report' mechanism, there may be a possibility of differential measurement error. Depending on society's view of Y and X, individuals with high levels of Y might have a tendency to blame X

for their condition. Or perhaps individuals with high levels of Y might want to deny the possibility that X has caused their condition. In either scenario, differential measurement error can result. A modification of Result 2.1 which considers this possibility is as follows.

Result 2.5 *Assume that* (Y, X, Z, X^*) *jointly have finite second moments, where* X *and* Z *have been scaled so that* $E(X) = E(Z) = 0$ *and* $Var(X) = Var(Z) = 1$. *Say* $E(X^*|X, Z) = X$ *and* $Var(X^*|X, Z) = \tau^2$, *but* X^* *and* Y *are not necessarily conditionally independent given* (X, Z). *Let* β *and* β^* *be the large-sample limiting coefficients from least-squares regression of* Y *on* $(1, X, Z)$ *and* Y *on* $(1, X^*, Z)$ *respectively. Then,*

$$\beta_1^* = \frac{\beta_1}{1 + \tau^2/(1 - \rho^2)} + \frac{\Delta}{1 - \rho^2 + \tau^2}, \tag{2.16}$$

where $\Delta = E\{Cov(X^*, Y|X, Z)\}$.

Note, in particular, that the two terms in (2.16) can push β_1^* in opposite directions. Say β_1 is positive. The first term in (2.16) corresponds to the usual attenuation factor. However, if X^* and Y are positively dependent given (X, Z), then $\Delta > 0$ and the second term in (2.16) has an accentuating effect on β_1^*. Thus the combined effect could be one of either attenuation or accentuation. The practical warning is to look out for situations where the measurement error may be differential. In particular, the impact of differential error is harder to predict in advance than the impact of nondifferential error.

2.8 Summary

Having now determined the bias due to measurement error in a number of scenarios, we can give some qualitative guidance on when this bias is likely to be substantial. Of course the first consideration is the magnitude of the measurement error, though this alone does not usually suffice to determine the bias in the corresponding regression coefficient. The next consideration is whether the predictor subject to measurement error is highly correlated with other precisely measured predictors. If so, a strong attenuation in the regression coefficient can result. Another consideration is whether the measurement error is additive or multiplicative. There is more potential for substantive bias in the second case, though whether this potential is realized depends on the shape of the predictor's distribution, and on how the predictor is included in the regression model. We also examined imprecision in several predictors, and 'unfair' measurement error that is either biased or differential. In these situations the nature of the bias in the regression coefficients is more complicated, and attenuation will not necessarily result. Overall, this chapter aims to gives some intuition about when adjustment for measurement error will be sorely needed, and when it will make only a slight difference.

Some of the expressions for bias presented in this chapter are either identical or very similar to expressions in the literature. In particular the books of

Fuller (1987) and Carroll, Ruppert and Stefanski (1995) cover much of the same ground. Typically, however, the context surrounding the expressions is somewhat different. In particular, the approach of comparing the large-sample limiting coefficients with imprecise predictors to the corresponding limit with precise predictors does not seem to have been stressed previously, except recently in Gustafson and Le (2002).

2.9 Mathematical Details

2.9.1 A Generalization of Result 2.1

Result 2.6 *Suppose that (Y, X, Z) jointly have finite second moments, where Y and X are scalar, and Z is a vector with d components. Without loss of generality assume X and the components of Z have been scaled so that $E(X) = E(Z_i) = 0$ and $Var(X) = Var(Z_i) = 1$. Furthermore, say X^* and (Z, Y) are conditionally independent given X, with $E(X^*|X) = X$ and $Var(X^*|X) = \tau^2$. Let β and β^* be the large-sample limiting coefficients from least-squares regression of Y on $(1, X, Z)$ and Y on $(1, X^*, Z)$ respectively. Then*

$$\begin{aligned} \beta_0^* &= \beta_0, \\ \frac{\beta_1^*}{\beta_1} &= \left(\frac{1}{1+\tau^2}\right)\left\{1 - \left(\frac{\tau^2}{1+\tau^2}\right) r' \left(R - \frac{rr'}{1+\tau^2}\right)^{-1} r\right\}, \quad (2.17) \\ \beta_2^* - \beta_2 &= \left(\frac{\beta_1 \tau^2}{1+\tau^2}\right)\left(R - \frac{rr'}{1+\tau^2}\right)^{-1} r, \end{aligned}$$

where $R = E(ZZ')$ and $r = E(ZX)$. The scaling of X and Z implies that R is the $d \times d$ correlation matrix of Z, while r is the $d \times 1$ vector of correlations between X and the components of Z.

Note that we have assumed X and Z are standardized in order to simplify the foregoing expressions. However, it is a simple matter to transform the expressions to describe the nonstandardized situation, as in Result 2.1 for instance.

Proof of Result 2.6. Define $T = (1, X, Z)'$, $T^* = (1, X^*, Z)'$, and $M = E(TT')$, $N = E(T^*T^{*\prime})$. Standard results for wrong-model asymptotics imply that β solves $M\beta = E(TY)$ while β^* solves $N\beta^* = E(T^*Y)$. But from the assumed properties of X^* we have $E(X^*Y) = E\{YE(X^*|X, Z, Y)\} = E(YX)$, so that $E(T^*Y) = E(TY)$, and hence $\beta^* = N^{-1}M\beta$. Partition M and N so that the on-diagonal blocks are 2×2 and $d \times d$, and note that $M_{22} = N_{22} = E(ZZ') = R$. Furthermore,

$$M_{12} = \begin{pmatrix} 0 & \dots & 0 \\ r_1 & \dots & r_d \end{pmatrix}, \quad (2.18)$$

and since $E(ZX^*) = E\{ZE(X^*|X, Z)\} = E(ZX)$, $N_{12} = M_{12}$. Finally, note that $M_{11} = I$, while

$$N_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1+\tau^2 \end{pmatrix}. \quad (2.19)$$

Using the standard expression for the inverse of a partitioned matrix (see, for instance, Gentle 1998, p. 61) gives

$$N^{-1}M = \begin{pmatrix} N_{11}^{-1}(I + M_{12}Z^{-1}M_{21}N_{11}^{-1}) & -N_{11}^{-1}M_{12}Z^{-1} \\ -Z^{-1}M_{21}N_{11}^{-1} & Z^{-1} \end{pmatrix} \times \begin{pmatrix} I & M_{12} \\ M_{21} & M_{22} \end{pmatrix},$$

where $Z = M_{22} - M_{21}N_{11}^{-1}M_{12}$. Carrying out the blocked multiplication gives

$$N^{-1}M = \begin{pmatrix} N_{11}^{-1}\{I + M_{12}Z^{-1}M_{21}(N_{11}^{-1} - I)\} & 0 \\ Z^{-1}M_{21}(I - N_{11}^{-1}) & I \end{pmatrix}, \quad (2.20)$$

and of course $\beta^* = N^{-1}M\beta$. Noting that $Z = R - (1+\tau^2)^{-1}rr'$ and substituting the specific forms (2.18) and (2.19) into (2.20) yields the desired result.□

It is easy to check that if $d = 1$, that is only a single precise predictor is considered, then Result 2.6 specializes to Result 2.1. The other specialization of note is the equi-correlated case described in Section 2.2. Here we have $r = (\rho, \ldots, \rho)'$ and $R = (1-\rho)I + \rho J$, where J is the $d \times d$ matrix with every entry being one. Some manipulation reduces (2.17) to

$$\frac{\beta_1^*}{\beta_1} = \left(\frac{1}{1+\tau^2}\right)\left[1 - \frac{\rho^2\tau^2}{1+\tau^2}\left\{\frac{d}{1-\rho} - \frac{d^2 c_{\rho,\tau}}{(1-\rho)(1-\rho+dc_{\rho,\tau})}\right\}\right], \quad (2.21)$$

where $c_{\rho,\tau} = \rho - \{\rho^2/(1+\tau^2)\}$. Evaluation of (2.21) yields the multiplicative bias curves in Figure 2.3. Of course when $d = 1$ the unwieldy expression (2.21) reduces to (2.3) from Result 2.1, as must be the case.

2.9.2 Proof of Result 2.2 from Section 2.3

Note that fitted regression coefficients for $\exp(kX)$ or $\exp(kX^*)$ are not changed by linear rescaling of Z, and let $\tilde{Z} = \{Z - E(Z)\}/SD(Z)$. Let β and β^* be the large-sample limiting coefficients from linear regression of Y on $T = (1, e^{kX}, \tilde{Z})'$ and Y on $T^* = (1, e^{kX^*}, \tilde{Z})'$ respectively. Then mimicking the proof of Result 2.6, $M\beta = E(TY)$ and $N\beta^* = E(T^*Y)$, where $M = E(TT')$ and $N = E(T^*T^{*\prime})$. In the former case,

$$\begin{pmatrix} 1 & E\left(e^{kX}\right) & 0 \\ E\left(e^{kX}\right) & E\left(e^{2kX}\right) & E\left(\tilde{Z}e^{kX}\right) \\ 0 & E\left(\tilde{Z}e^{kX}\right) & 1 \end{pmatrix}\beta = \begin{pmatrix} E(Y) \\ E\left(Ye^{kX}\right) \\ E\left(Y\tilde{Z}\right) \end{pmatrix}. \quad (2.22)$$

Now the normality of $X^*|X$ and the assumption of nondifferential error yields $E\left(e^{kX^*}\right) = \gamma E\left(e^{kX}\right)$, $E\left(e^{2kX^*}\right) = \gamma^4 E\left(e^{2kX}\right)$, $E\left(\tilde{Z}e^{kX^*}\right) = \gamma E\left(\tilde{Z}e^{kX}\right)$,

and $E\left(Ye^{kX^*}\right) = \gamma E\left(Ye^{kX}\right)$, where $\gamma = \exp(k^2\sigma^2\tau^2/2)$. Thus β^* satisfies

$$\begin{pmatrix} 1 & \gamma E\left(e^{kX}\right) & 0 \\ \gamma E\left(e^{kX}\right) & \gamma^4 E\left(e^{2kX}\right) & \gamma E\left(\tilde{Z}e^{kX}\right) \\ 0 & \gamma E\left(\tilde{Z}e^{kX}\right) & 1 \end{pmatrix} \beta^* = \begin{pmatrix} E(Y) \\ \gamma E\left(Ye^{kX}\right) \\ E\left(Y\tilde{Z}\right) \end{pmatrix},$$

or equivalently

$$\begin{pmatrix} 1 & \gamma E\left(e^{kX}\right) & 0 \\ E\left(e^{kX}\right) & \gamma^3 E\left(e^{2kX}\right) & E\left(\tilde{Z}e^{kX}\right) \\ 0 & \gamma E\left(\tilde{Z}e^{kX}\right) & 1 \end{pmatrix} \beta^* = \begin{pmatrix} E(Y) \\ E\left(Ye^{kX}\right) \\ E\left(Y\tilde{Z}\right) \end{pmatrix}. \tag{2.23}$$

Thus the left-hand sides of (2.22) and (2.23) are equal, giving an explicit relationship between β^* and β. Solving for β_1^* in terms of β using Cramer's rule gives

$$\frac{\beta_1^*}{\beta_1} = \frac{Var\left(e^{kX}\right) - E\left(\tilde{Z}e^{kX}\right)^2}{\gamma\left\{Var\left(e^{kX}\right) - E\left(\tilde{Z}e^{kX}\right)^2\right\} + (\gamma^3 - \gamma)E\left(e^{2kX}\right)}.$$

Upon noting that

$$Var\left(e^{kX}\right) - E\left(e^{kX}\tilde{Z}\right)^2 = Var\left(e^{kX}\right)\left[1 - Cor\left(e^{kX}, Z\right)^2\right],$$

the desired result follows immediately.□

2.9.3 A Result on the Impact of Measurement Error in Logistic Regression Estimates

The following result supports the developments in Section 2.6.

Result 2.7 *Let Y be a 0-1 random variable and let T and T^* be random vectors each with d components and finite means. Let β be the large-sample limiting coefficients from logistic regression of Y on T, and let β^* be the large-sample limiting coefficients from logistic regression of Y on T^*. If $E(T^*|T, Y) = MT$ for some fixed $d \times d$ matrix M, then β^* and β satisfy*

$$E\left[T^*\left\{\frac{1}{1+\exp(-\beta' T)} - \frac{1}{1+\exp(-\beta^{*\prime} T^*)}\right\}\right] = 0. \tag{2.24}$$

Proof of Result 2.7. Let $g(z) = \{1 + \exp(-z)\}^{-1}$. Under the stated conditions,

$$\begin{aligned} E\left[T^*\left\{g(\beta' T) - g(\beta^{*\prime} T^*)\right\}\right] &= E\left[T^*\left\{Y - g(\beta' T)\right\}\right] \\ &= E\left[E(T^*|T, Y)\left\{Y - g(\beta' T)\right\}\right] \\ &= M\ E\left[T\left\{Y - g(\beta' T)\right\}\right] \\ &= 0, \end{aligned}$$

where the first equality follows from (2.13) and the last equality follows from (2.12).□

It is straightforward to verify that the conditions in Result 2.7 are met in all three examples discussed in Section 2.6. The form of (2.24) then suggests an easy approach to approximating β^* for a given β. In particular, let $(T_{(1)}, T^*_{(1)}), \ldots, (T_{(m)}, T^*_{(m)})$ denote a Monte Carlo sample of m independent draws from the joint distribution of (T, T^*), with m taken to be very large. For a given β, define $\tilde{Y}_{(i)} = \{1+\exp(-\beta' T_{(i)})\}^{-1}$. Then the value of β^* solving (2.24) can be approximated by the solution to

$$\sum_{i=1}^{m} T^*_{(i)} \left\{ \tilde{Y}_{(i)} - \frac{1}{1+\exp(-\beta^{*\prime} T^*_{(i)})} \right\} \quad = \quad 0. \tag{2.25}$$

This is precisely the form of the logistic regression maximum likelihood equations when $\tilde{Y}_{(i)}$ and $T^*_{(i)}$ are respectively the response variable and the explanatory variables for the i-th case. Of course $\tilde{Y}_{(i)}$ is not binary as per a 'real' logistic regression response variable, but many standard software routines for logistic regression will solve these equations regardless of whether the inputted values of the response variable are binary or not. Thus the value of β^* solving (2.25) is easily obtained from standard software.