CHAPTER 6

# Further Topics

This chapter discusses some more specialized issues in understanding and adjusting for mismeasurement of explanatory variables. Section 6.1 focusses on a link between measurement error for a continuous variable and misclassification for a binary variable, thereby connecting two previously separate threads in this book. Section 6.2 builds on 6.1 in a particular direction, focussing particularly on the interplay between bias due to mismeasurement and bias due to model misspecification. And finally 6.3 takes a closer look at the issue of model identifiability in Bayesian analysis, particularly as it relates to many mismeasured variable situations.

## 6.1 Dichotomization of Mismeasured Continuous Variables

Of course one primary feature of this book is its coverage of both mismeasured continuous variables (Chapters 2 and 4) and mismeasured categorical variables (Chapters 3 and 5). To this point only informal contrasts between the two scenarios have been considered. In Chapter 3, for instance, we compared $\alpha \times 100\%$ binary misclassification, in the sense of sensitivity and specificity equal to $1 - \alpha$, to $\alpha \times 100\%$ continuous measurement error, in the sense of $SD(X^*|X) = \alpha SD(X)$. In particular, it was noted that the former can induce much more bias than the latter. Now we turn attention to more formal connections and comparisons, particularly in scenarios where a mismeasured binary variable is in fact *created* from a mismeasured continuous variable.

The present framework involves a binary variable $V$ which is linked to a continuous variable $X$ via thresholding. That is,

$$V = I\{X > c\}, \tag{6.1}$$

for some threshold $c$. Of course inability to measure $X$ precisely implies inability to measure $V$ precisely. Say $X^*$ is a noisy surrogate for $X$. Then $V^*$, a noisy surrogate for $V$, is defined as

$$V^* = I\{X^* > c\}. \tag{6.2}$$

For instance, say $X$ is serum cholesterol concentration, and physicians regard $c$ as the threshold defining an 'elevated' cholesterol level. Conceptually then it may be of interest to use $V$ rather than $X$ as an explanatory variable in an outcome model. However, if the lab procedure for measuring $X$ actually yields a surrogate $X^*$, then $V^*$ will be recorded in lieu of $V$.

To give some idea of how much misclassification is induced by a given amount of measurement error via (6.2), say that $X$ has a normal distribution,

taken as $X \sim N(0,1)$ without loss of generality. Moreover, say the surrogate $X^*$ arises via $X^*|X \sim N(X, \tau^2)$. Upon expressing $X^* = X - \tau W$ where $(X, W)$ are independent standard normals, the sensitivity of $V^*$ as a surrogate for $V$ can be expressed as

$$\begin{aligned} SN &= Pr(X^* > c | X > c) \\ &= \frac{1}{2} + \{1 - \Phi(c)\}^{-1} \int_0^\infty \{1 - \Phi(c + \tau w)\} \phi(w)\, dw. \end{aligned} \tag{6.3}$$

Here, and throughout this chapter, $\phi()$ and $\Phi()$ denote the standard normal density function and distribution function respectively. The integral in (6.3) must be evaluated numerically. The specificity is determined similarly as

$$\begin{aligned} SP &= Pr(X^* < c | X < c) \\ &= \frac{1}{2} + \{\Phi(c)\}^{-1} \int_0^\infty \{\Phi(c - \tau w)\} \phi(w)\, dw. \end{aligned} \tag{6.4}$$

Figure 6.1 conveys the relationship between the magnitude of measurement error in $X$ and the magnitude of misclassification in $V$. For fixed values of $\tau$, the corresponding values of $(SN, SP)$ arising for various thresholds $c$ are plotted. Clearly $SN = SP$ when $c = 0$, and one can verify that $SN$ decreases with $c$ while $SP$ increases with $c$. Note that in the present sense $\alpha \times 100\%$ measurement error corresponds to much less than $\alpha \times 100\%$ misclassification. When $c = 0$ for instance, $\tau = 0.1$, $\tau = 0.25$, and $\tau = 0.5$ correspond to $SN = SP = 0.968$, $SN = SP = 0.922$, and $SN = SP = 0.853$ respectively.

For a given threshold $c$, (6.3) and (6.4) define a magnitude of misspecification $(SN, SP)$ which matches a given magnitude of measurement error $\tau$. And similarly a matching prevalence for $V$ is $r = Pr(X > c) = 1 - \Phi(c)$ when $X \sim N(0,1)$. An obvious comparison to make is between the bias induced by nondifferential measurement error of a given magnitude and that induced by nondifferential misclassification of the matching magnitude. Some such comparisons are given in Table 6.1, in the case of a linear outcome model and no additional precisely measured covariates. That is, the attenuation factor (2.1) for a given $\tau$ is compared to the attenuation factor (3.3) for the matching $(r, SN, SP)$, for various values of $\tau$ and $c$. In all instances there is more attenuation under binary misclassification than under continuous measurement error, with the difference being extreme when $\tau$ is large and the threshold $c$ is in the tail of the $X$ distribution. Even though 10% measurement error corresponds to much less than 10% misclassification, the latter still leads to more bias than the former.

Table 6.2 considers the scenario where an additional precisely measured covariate is present. Specifically it compares the attenuation factor (2.3) for the $X$ coefficient when $Y$ is regressed on $(X^*, Z)$ instead of $(X, Z)$ to the attenuation factor (3.4) for the $V$ coefficient in the matching misclassification scenario. For illustration $\rho = Cor(X, Z) = 0.8$ is chosen to represent a scenario where the measurement error bias is worsened by substantial correlation between the explanatory variables. To completely determine the matching scenario we take the correlation between the binary predictor $V$ and $Z$ to be

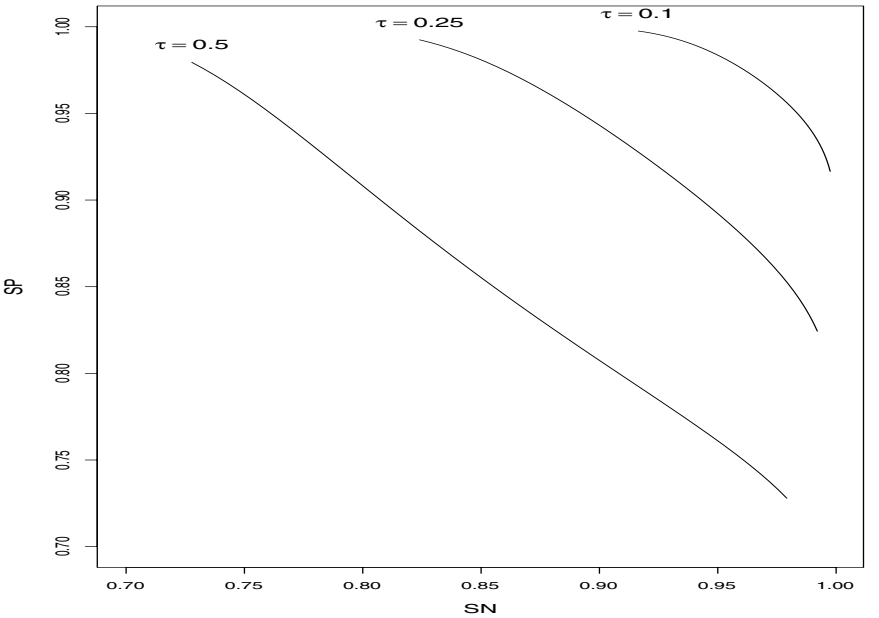


Figure 6.1 *Sensitivity and specificity corresponding to a given magnitude of measurement error under thresholding. For each value of the measurement error standard deviation $\tau$, the plotted values of $(SN, SP)$ correspond to different values of the threshold $c$, ranging from $c = -2$ to $c = 2$.*

| | Meas. Error | Misclassification $c = 0$ | $c = 1$ | $c = 2$ |
|---|---|---|---|---|
| $\tau = 0.1$ | 0.99 | 0.94 | 0.93 | 0.89 |
| $\tau = 0.25$ | 0.94 | 0.84 | 0.81 | 0.71 |
| $\tau = 0.5$ | 0.80 | 0.71 | 0.63 | 0.45 |

Table 6.1 *Attenuation factors under nondifferential measurement error and matching nondifferential misclassification. Different magnitudes of measurement error ($\tau$) and different thresholds ($c$) are considered.*

$\rho^* = Cor(V, Z) = Cor(I\{X > c\}, Z) = \rho\phi(c)$, assuming now that $(X, Z)$ are jointly normal with standardized marginal distributions. Thus the matching correlation, which is given in the table, depends on the threshold $c$. Again more attenuation is manifested under misclassification than under measurement error when $\tau = 0.1$ and $\tau = 0.25$, although the situation is reversed when $\tau = 0.5$ and the threshold is not too extreme.

On the face of it Tables 6.1 and 6.2 might be construed as contrasting the impact of mismeasurement before and after dichotomization of the continu-

| | | Misclassification | | |
|---|---|---|---|---|
| | Meas. Error | $c = 0$ ($\rho^* = 0.32$) | $c = 1$ ($\rho^* = 0.19$) | $c = 2$ ($\rho^* = 0.04$) |
| $\tau = 0.1$ | 0.97 | 0.92 | 0.92 | 0.89 |
| $\tau = 0.25$ | 0.85 | 0.82 | 0.80 | 0.71 |
| $\tau = 0.5$ | 0.59 | 0.67 | 0.61 | 0.44 |

Table 6.2 *Attenuation factors under nondifferential measurement error and 'matching' nondifferential misclassification, with an additional precisely measured explanatory variable.*



ous predictor. This is not the case, however. The tables contrast the impact of nondifferential measurement error and nondifferential misclassification in matched scenarios. However, as first pointed out by Flegal, Keyl and Nieto (1991), if a misclassified dichotomous variable is actually created by thresholding a continuous variable subject to nondifferential measurement error, then in general the misclassification will be differential. See also Irwig, Groeneveld and Simpson (1990) and Delpizzo and Borghes (1995) for discussion of misclassification induced by thresholding a mismeasured continuous variable. Initially it seems counter-intuitive that differential misclassification is manifested. If $X^*$ arises from $X$ in a manner which is blind to the outcome variable $Y$, then $V^* = I\{X^* > c\}$ would also seem to be blind to the value of $Y$. More generally one might suppose that if $X^*$ is a nondifferential surrogate for $X$, then $g(X^*)$ would be a nondifferential surrogate for $g(X)$, for any function $g()$. Mathematically, however, conditional independence of $X^*$ and $Y$ given $X$ does not imply conditional independence of $g(X^*)$ and $Y$ given $g(X)$ for an arbitrary function $g()$.

As a more focussed demonstration of the issue, let $V$ and $V^*$ be obtained by thresholding $X$ and $X^*$ at $c$, as per (6.1) and (6.2), while $X^*$ and $Y$ are conditionally independent given $X$. We demonstrate that the conditional distribution of $V^*$ given $(V, Y)$ does in fact vary with $Y$. In particular consider the conditional sensitivity of $V^*$ as a surrogate for $V$, given $Y = y$. Clearly

$$\begin{aligned} Pr(V^* = 1|V = 1, Y = y) &= Pr(X^* > c|X > c, Y = y) \\ &= E\{Pr(X^* > c|X, Y)|X > c, Y = y\} \\ &= E\{Pr(X^* > c|X)|X > c, Y = y\} \\ &= \int Pr(X^* > c|X = x) f_{X|X>c,Y}(x|y)dx. \end{aligned} \tag{6.5}$$

Of course the probability inside the integral in (6.5) will increase with $x$, provided $X^*$ is a reasonable surrogate for $X$. Now say that $X$ and $Y$ are positively associated, i.e., there is some positive relationship between the exposure and the response. Consequently the distribution of $(X|X > c, Y = y)$ will put

more mass on larger values of $X$ as $y$ increases, so that (6.5) increases with $y$. That is, the misclassification is differential.

To give a concrete example say

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\begin{pmatrix} 1 & \omega \\ \omega & 1 \end{pmatrix},$$

while $X^*|X, Y \sim N(X, \tau^2)$. Then the probability inside (6.5) is simply $1 - \Phi\{(c-x)/\tau\}$, while the integration over $x$ is with respect to a $N(\omega y, 1-\omega^2)$ distribution truncated to values larger than $c$. As a numerical example say $c = 0$ and $\tau = 0.5$, giving an unconditional sensitivity (6.3) of 0.843. Now say that there is substantial dependence between $X$ and $Y$, with $\omega = 0.5$. Then numerical integration of (6.5) gives the conditional sensitivity as 0.75, 0.79, 0.83, 0.88, and 0.93, for $y = -2, \ldots, y = 2$ respectively. Clearly the misclassification is differential to a substantial extent. Moreover, this will be further magnified when the association between $X$ and $Y$ is stronger. For instance when $\omega = 0.9$ the conditional sensitivities range from 0.57 when $y = -2$ to 0.96 when $y = 2$. While the misclassification does arise in a manner which is qualitatively blind to the outcome variable, the end result is a substantial departure from the nondifferential scenario. Following Gustafson and Le (2002) we describe this kind of misclassification as *differential due to dichotomization* (DDD), to distinguish it from the nondifferential situation.

Having established that a mismeasured predictor arising from dichotomization involves differential misclassification, it is clear that the impact of this misclassification may differ from that of nondifferential misclassification as given in Tables 6.1 and 6.2. To investigate this in a simple setting, say that ideally one wishes to fit a linear regression of $Y$ to $V = I\{X > c\}$, but in actuality $Y$ is regressed on $V^* = I\{X^* > c\}$, where $X^*$ is a nondifferential surrogate for $X$. Clearly

$$E(Y|V^*) = E(Y|V^* = 0) + V^*\{E(Y|V^* = 1) - E(Y|V^* = 0)\},$$

and similarly

$$E(Y|V) = E(Y|V = 0) + V\{E(Y|V = 1) - E(Y|V = 0)\},$$

so the attenuation factor for estimating the $V$ coefficient is simply

$$AF = \frac{E(Y|V^* = 1) - E(Y|V^* = 0)}{E(Y|V = 1) - E(Y|V = 0)}.$$

Upon writing the conditional expectations in iterated fashion, and noting that $E(Y|X, V^*)$ does not depend on $V^*$ since $X^*$ is nondifferential, we have

$$AF = \frac{E\{m(X)|X^* > c\} - E\{m(X)|X^* < c\}}{E\{m(X)|X > c\} - E\{m(X)|X < c\}}, \tag{6.6}$$

where

$$m(x) = E(Y|X = x)$$

is the actual regression function. A first comment about (6.6) is simply that it

does depend on the actual distribution of $Y$ given $X$, via $m(X)$. In contrast, it was stressed in Chapters 2 and 3 that the bias due to nondifferential mismeasurement does not depend on the actual response distribution. Moreover, we know that the bias due to nondifferential measurement error depends on the distribution of $X$ only through the variance of $X$. In contrast, it is evident that (6.6) depends on the distribution of $X$ in a more detailed manner. Thus claims about the impact of DDD misclassification are likely to be less general than claims about nondifferential mismeasurement.

To consider a special case of (6.6), say that $X$ is normally distributed, assuming without loss of generality that $X \sim N(0,1)$. Also assume that $X^*|X,Y \sim N(X,\tau^2)$, and that $m(X) = \beta_0 + \beta_1 X$, i.e., a linear response model holds. It is clear from the form of (6.6) that $\beta_0 = 0$ and $\beta_1 = 1$ can be assumed without loss of generality. Then the assumed joint normality of $(X, X^*)$ leads easily to

$$AF = \frac{\lambda R(c)}{R(\lambda c)}, \tag{6.7}$$

where $\lambda = (1+\tau^2)^{-1/2}$ and

$$R(z) = \frac{\Phi(z)\{1-\Phi(z)\}}{\phi(z)},$$

where again $\phi()$ and $\Phi()$ denote the standard normal density function and distribution function respectively. Since $R(z)$ is symmetric and decreasing in $|z|$, it follows that the multiplicative bias (6.7) is a decreasing function of $\tau$, as expected. Also, (6.7) is a decreasing function of $|c|$, so that the bias worsens as the threshold used for dichotomization moves away from the centre of the $X$ distribution.

An obvious question is whether dichotomization makes the impact of mismeasurement better or worse. That is, how does the attenuation factor (6.7) compare to the attenuation factor when $Y$ is regressed on $X^*$ in lieu of $X$. A particularly clear comparison results when $c = 0$, i.e., the threshold is at the centre of the $X$ distribution. Then the DDD attenuation factor (6.7) is simply $(1+\tau^2)^{-1/2}$, as opposed to $(1+\tau^2)^{-1}$ for nondifferential measurement error. That is, changing the explanatory variable from $X$ to $V = I\{X > 0\}$ involves a *reduction* in bias due to mismeasurement. This is surprising given that generally nondifferential misclassification of a binary variable tends to yield *more* bias than nondifferential measurement error in a continuous variable, as witnessed by Tables 6.1 and 6.2. As a specific example, say $\tau = 0.5$ and $c = 0$. Then the DDD attenuation factor is 0.894 compared to the nondifferential measurement error attenuation factor of 0.8. On the other hand, Table 6.2 gives the nondifferential misclassification attenuation factor as 0.63.

Of course the comparisons above pertain to a threshold $c = 0$ in the centre of the $X$ distribution. As already noted, the DDD attenuation is worse when the threshold is further from the centre of the distribution. It is interesting to note, however, that $R(z) \sim 1/|z|$ as $|z| \to \infty$, so that the attenuation factor (6.7) tends to $(1+\tau^2)^{-1}$ as $|c|$ tends to infinity. Even in the worst-case of a

threshold in the extreme tail of the $X$ distribution, the attenuation due to DDD misclassification is only as bad as the attenuation due to continuous measurement error.

The determination of the DDD attenuation factor is readily extended to situations involving an additional precisely measured explanatory variable. Gustafson and Le (2002) give an expression for the attenuation factor when a linear regression of $Y$ on $(1, V^*, Z)$ is used in lieu of $(1, V, Z)$. In the special case that the actual response relationship is linear, i.e., $E(Y|X, Z)$ is linear in $(X, Z)$, and $(X^*, X, Z)$ have a multivariate normal distribution, this expression specializes to

$$AF = \lambda \left\{ \frac{R(c) - \rho^2 \phi(c)}{R(\lambda c) - \rho^2 \lambda \phi(\lambda c)} \right\}, \tag{6.8}$$

which clearly reduces to (6.7) when $\rho = Cor(X, Z) = 0$. Gustafson and Le demonstrate that the previously elucidated relationship still holds in this situation, namely (6.8) corresponds to less bias than nondifferential measurement error, which in turn corresponds to less bias than nondifferential misclassification. In addition, they demonstrate similar behaviour in logistic regression scenarios, using numerical techniques akin to those in Chapters 2 and 3.

Gustafson and Le also point out several curious features about the behaviour of (6.8). In particular, for some fixed values of $c$ (6.8) is not monotonically decreasing in $|\rho|$, and for some fixed values of $\rho$ (6.8) is not monotonically decreasing in $|c|$. Thus the impact of DDD misclassification is less generally predictable than that of nondifferential mismeasurement. Some insight into the behaviour is gleaned by noting that for any fixed $\rho$ (6.7) tends to $\lambda^2$ as $|c|$ tends to infinity, matching the attenuation due to nondifferential measurement error in the special case that $\rho = 0$. In some sense then an extreme threshold diminishes the impact of correlation between predictors on the mismeasurement bias.

Of course (6.7) and (6.8) describe the attenuation arising in quite narrow settings, with a linear regression function being assumed in particular. In fact the dependence of the attenuation factor on the actual regression function relates to reasons why an analyst might choose to dichotomize a predictor variable in the first place. One may dichotomize to avoid postulating a particular form for the regression relationship. Whereas $E(Y|X)$ may or may not be linear in $X$, $E(Y|V)$ is necessarily linear in $V$, since $V$ is binary. On the other hand, of course, information is lost by analyzing $(Y, V)$ rather than $(Y, X)$. Also, dichotomization does not necessarily lead to a correct model in the presence of additional covariates. Nonetheless, dichotomization of predictors is relatively common in statistical practice, even when one can reasonably assume that $Y$ varies with $X$ in a 'smooth' manner. On the other hand, sometimes predictors are dichotomized because it is plausible to posit a threshold regression effect rather than a smooth effect. That is, one may think that $E(Y|X)$ being linear in $V = I\{X > c\}$ is more plausible *a priori* than $E(Y|X)$ being linear in $X$.

To investigate how the DDD attenuation depends on the actual regression

function more thoroughly, say the real regression relationship is

$$m(X) \quad = \quad (1-\omega)X + \omega I\{X > c\},$$

involving a combination of a linear effect and a threshold effect. This is a special case of a scenario investigated by Gustafson and Le (2002) involving an additional precisely measured covariate. The DDD attenuation factor (6.6) is then readily computed as

$$AF \quad = \quad \frac{(1-\omega)a_1 + \omega a_3}{(1-\omega)a_2 + \omega}, \tag{6.9}$$

where

$$a_1 \quad = \quad E(X|X^* > c) - E(X|X^* < c),$$

$$a_2 \quad = \quad E(X|X > c) - E(X|X < c),$$

and

$$a_3 \quad = \quad 1 - Pr(V = 0|V^* = 1) - Pr(V = 1|V^* = 0).$$

Of course (6.9) reduces to (6.7) when $\omega = 0$, as must be the case. As $\omega$ increases to one, however, the attenuation factor increases monotonically to $a_3$, which is identically the attenuation factor (3.3) in the matching nondifferential misclassification scenario. That is, if $E(Y|X)$ is actually linear in $V = I\{X > c\}$, then the bias matches that incurred by nondifferential misclassification. In fact, under the slightly stronger assumption that the distribution of $Y|X$ depends on $X$ only through $V = I\{X > c\}$, one can verify that $V^* = I\{X^* > c\}$ is actually a nondifferential surrogate for $V$. Put another way, the extent to which thresholding yields differential misclassification depends strongly on how the conditional distribution of $Y|X$ depends on $X$. The simple scenario above alludes to a tradeoff. In situations where fitting a threshold variable is more appropriate ($\omega$ larger), the mismeasurement bias is closer to that incurred under nondifferential misclassification, i.e., larger. The interplay between the appropriateness of the model and the mismeasurement bias is explored more generally in Section 6.2.

In summary, dichotomization of predictors via thresholding is a scenario where misclassification might appear to be nondifferential, but can in fact be substantially differential. This certainly raises flags about 'off-the-cuff' assumptions of nondifferential mismeasurement. Since the DDD misclassification bias is typically quite modest relative to the nondifferential misclassification bias, there is a risk of overcorrecting for misclassification by incorrectly making the nondifferential assumption. The situation is further muddied given our demonstration that the bias due to DDD misclassification can vary considerably with the actual regression relationship between the outcome variable and the underlying continuous explanatory variable.

We close this section by noting that DDD misclassification can arise without explicit construction of a binary variable from a continuous variable. In many questionnaire-based studies, subjects are asked to dichotomize variables

implicitly. For instance, a subject might be asked to place his cumulative exposure to an agent into one of several ordered categories. In a rough sense the mental processes leading to the 'noisy' response for such a question must involve thresholding of an 'internal' $X^*$. Consequently DDD misclassification is not limited to scenarios whereby an investigator explicitly constructs the dichotomous predictor by thresholding the continuous predictor.

### 6.2 Mismeasurement Bias and Model Misspecification Bias

In the previous section, the bias induced by DDD misclassification was seen to vary with the actual relationship between the outcome variable and the underlying continuous explanatory variable. This touches on the much broader point that most statistical analyses are susceptible to multiple sources of bias. Of course the emphasis in this book is on understanding and reducing the bias resulting from mismeasured variables. It may not always be reasonable, however, to view this in isolation from other potential biases. The previous findings for DDD misclassification motivate more formal consideration of how biases due to mismeasurement and model misspecification might interact with one another in practice.

There seems to be very little consideration of mismeasurement and model misspecification simultaneously, with the notable exceptions of Lagakos (1988) and Begg and Lagakos (1992, 1993) in a hypothesis-testing context. In a linear model setting, Gustafson (2002a) gives a simple framework for considering bias due to mismeasurement and bias due to model misspecification simultaneously. Consider an outcome variable $Y$ and a pool of $k$ possible predictor variables $X = (X_1, \ldots, X_k)'$, with independent $(Y, X)$ realizations for $n$ subjects. Now say that in fact the 'correct' predictor variables are $S = \{S_1(X), \ldots, S_q(X)\}'$, in the sense that

$$E(Y|X) \quad = \quad \nu_1 S_1(X) + \ldots \nu_q S_q(X), \tag{6.10}$$

for some coefficients $\nu = (\nu_1, \ldots, \nu_q)'$. Thus an ideal analysis involves linear regression of $Y$ on $S$. Note that this formulation is very general, as $q$, the number of correct predictors, need not be the same as $m$, the number of available predictors. Also, a given $S_j$ can depend on the available predictors in an arbitrary manner, and in particular the actual regression function (6.10) might involve interaction terms or nonlinear terms.

To entertain the possibility of fitting a misspecified model, say that $T = \{T_1(X), \ldots, T_p(X)\}'$ are the regressors chosen for analysis. Moreover, let $\beta = (\beta_1, \ldots, \beta_p)'$ be the large-$n$ limiting coefficients from linear regression of $Y$ on $T$. Thus with enough data the analysis incorrectly indicates the regression function to be

$$\tilde{E}(Y|X) \quad = \quad \beta_1 T_1(X) + \ldots \beta_p T_p(X). \tag{6.11}$$

In trying to estimate the regression function then, the difference between (6.11) and (6.10) is the large-sample bias incurred because of model misspecification. We choose to summarize the magnitude of this bias by the average

squared-error (ASE), defined as

$$ASE_{SPC} \quad = \quad E\left[\{\beta' T(X) - \nu' S(X)\}^2\right]. \tag{6.12}$$

In particular, the average is with respect to the distribution of the available predictors $X$, with the view that errors in estimating the regression function at common values of $X$ are more damaging than errors at rare values.

Now say that in fact the chosen predictors $T$ cannot all be measured precisely. Rather $T^* = (T_1^*, \ldots, T_p^*)'$ is the observed surrogate for $T$. Also, let $\beta^* = (\beta_1^*, \ldots, \beta_p^*)'$ be the large-sample limiting coefficients from linear regression of $Y$ on $T^*$. If the mismeasurement is ignored or undetected, then with enough data the analysis indicates

$$\tilde{E}^*(Y|X) \quad = \quad \beta_1^* T_1(X) + \ldots \beta_p^* T_p(X) \tag{6.13}$$

to be the true regression function, i.e., the analysis treats the coefficients from regression of $Y$ on $T^*$ as if they are the coefficients from regression of $Y$ on $T$. The total bias in estimating the regression function is then the difference between (6.13) and (6.10). In keeping with (6.12) we summarize this via

$$ASE_{TOT} \quad = \quad E\left[\{\beta^{*\prime} T(X) - \nu' S(X)\}^2\right]. \tag{6.14}$$

Similarly, the difference between (6.13) and (6.11) is naturally regarded as the bias attributed to mismeasurement alone. Thus we also define

$$ASE_{MSR} \quad = \quad E\left[\{\beta^{*\prime} T(X) - \beta' T(X)\}^2\right]. \tag{6.15}$$

To this point, (6.14) summarizes the joint impact of model misspecification and mismeasurement, while (6.12) and (6.15) describe the respective individual impacts. Perhaps somewhat predictably, it is shown in Gustafson (2002a) that an additive relationship exists, i.e.

$$ASE_{TOT} \quad = \quad ASE_{SPC} + ASE_{MSR}. \tag{6.16}$$

Also, computable expressions for $ASE_{SPC}$ and $ASE_{MSR}$ are given. These follow easily upon noting that both $\beta$ and $\beta^*$ are functions of $\nu$, since the limiting coefficients under a misspecified model necessarily depend on the true relationship between $Y$ and $X$. In particular, standard large-sample theory for misspecified models (e.g., White 1982) gives

$$\beta \quad = \quad \{E(TT')\}^{-1} E(TS')\nu,$$

and similarly

$$\beta^* \quad = \quad \{E(T^*T^{*\prime})\}^{-1} E(T^*S')\nu.$$

This leads easily to

$$ASE_{SPC} = \nu' A \nu,$$

where

$$A \quad = \quad E(SS') - E(ST')E(TT')^{-1}E(TS').$$

Similarly,

$$ASE_{MSR} = \nu' B \nu,$$

where

$$B = C' E(TT') C,$$

with

$$C = E(TT')^{-1} E(TS') - E(T^*T^{*\prime})^{-1} E(T^*S').$$



### *6.2.1 Further Consideration of Dichotomization*

We consider an example where the decomposition (6.16) sheds light on the interplay of model misspecification and mismeasurement. In particular, we build upon the discussion of dichotomization in Section 6.1. Assume that the two available predictors $(X, Z)$ follow a bivariate normal distribution, assuming without loss of generality that both $X$ and $Z$ are standardized to have mean zero and variance one. Also, let $\rho = \mathrm{Cor}(X, Z)$. Now say that the analyst is contemplating whether or not to dichotomize the continuous predictor $X$. That is, he has two choices of regressors in mind, namely $T^{(C)} = (1, X, Z)'$ and $T^{(D)} = (1, I\{X > 0\}, Z)'$. Note that in the latter case the threshold is at the centre of the $X$ distribution. Also, say that $X$ is subject to nondifferential measurement error which is both unbiased and normally distributed, so that the measured regressors are either $T^{*(C)} = (1, X^*, Z)$ or $T^{*(D)} = (1, I\{X^* > 0\}, Z)$, where $X^*|X, Z, Y \sim N(X, \tau^2)$.

Following the example in Section 6.1, it proves fruitful to consider what happens when the true relationship between the response and predictors falls in between the possibilities considered by the analyst. Specifically, say that

$$\begin{aligned} E(Y|X, Z) &= \nu^T S \\ &= \nu_1 + \nu_2 \left\{ (1-\omega)X + \omega\sqrt{2\pi} \left( I\{X > 0\} - \frac{1}{2} \right) \right\} + \\ &\quad \nu_3 Z, \end{aligned} \tag{6.17}$$

for some $\omega \in [0, 1]$. As $\omega$ increases then, $T^{(D)}$ becomes a more appropriate choice of regressors while $T^{(C)}$ becomes a less appropriate choice. As an aside, the centering and scaling of the indicator function in (6.17) appears for a technical reason. In particular, for any value of $\omega$ it leads to $\beta = \nu$ when considering $T^{(C)}$ as the predictors. Thus $\omega$ is more readily interpreted as the weight given to the dichotomous component in (6.17).

With moderate numerical effort the moments needed to determine $ASE_{SPC}$ and $ASE_{MSR}$ can be computed in this scenario. General considerations discussed in Gustafson (2002a) imply that both terms will depend on $\nu$ only through $\nu_2$. Roughly speaking, this follows since only the second of the three regressors is misspecified and subject to mismeasurement. The decomposition

(6.16) thus simplifies to $ASE_{SPC} = A_{22}\nu_2^2$ and $ASE_{MSR} = B_{22}\nu_2^2$. In particular, the ratio of either $ASE_{SPC}$ or $ASE_{MSR}$ to $ASE_{TOT}$ does not depend on $\nu$.

Figure 6.2 plots the terms in the decomposition as functions of $\omega$. Both choices of predictors $T$ are considered, as are different values of $\rho$ and $\tau$. The observed behaviour of the model misspecification term $ASE_{SPC}$ is entirely predictable. If the continuous predictor is used then this term increases with $\omega$, as the true regression function moves away from the postulated form. Similarly, $ASE_{SPC}$ decreases with $\omega$ when the dichotomous predictor is used. The mismeasurement term behaves differently for the two choices of predictors. With the continuous predictor, $ASE_{MSR}$ does not depend on $\omega$. This extends the notion from Chapter 2 that the bias due to continuous nondifferential measurement error does not depend on the actual distribution of the outcome variable given the response variable. With the dichotomized predictor, however, $ASE_{MSR}$ increases with $\omega$. This quantifies the tradeoff noted in Section 6.1. With DDD misclassification, the mismeasurement bias increases as the model misspecification bias decreases. This phenomenon is particularly acute in the $\tau = 0.75$ scenarios shown in Figure 6.2. Here $ASE_{TOT}$ actually increases with $\omega$ for larger values of $\omega$, so that the overall error *increases* as the true relationship moves closer to the postulated model. That is, the gain in improved model fit is more than offset by the increased bias due to mismeasurement. Generally speaking this tradeoff indicates why it is valuable to consider mismeasurement bias in tandem with other biases rather than in isolation.

### *6.2.2 Other Examples*

Two different examples of using the decomposition (6.16) are given in Gustafson (2002a). The first involves a situation where the analysis is missing a regressor. Specifically the analysis uses $T = (1, X, Z)$ as regressors when in fact either $S = (1, X, Z, X^2)$ or $S = (1, X, Z, XZ)$ are the true regressors. Thus the analysis is ignoring either curvature or interaction which is actually present. Moreover, a nondifferential surrogate $X^*$ is measured in lieu of $X$. Predictably it turns out that $ASE_{SPC}$ increases with the coefficient on the term in $S$ which is omitted from $T$. It turns out, however, that $ASE_{MSR}$ decreases with the magnitude of this coefficient. That is the damaging impact of more missed curvature or interaction is partly offset by a reduction in bias due to mismeasurement. This constitutes another situation where $ASE_{SPC}$ and $ASE_{MSR}$ are inversely related as the underlying regression relationship changes.

The other example in Gustafson (2002a) involves forming $m$ binary variables from a continuous variable $X$. That is, the analysis proceeds with

$$T = (I_{A_1}(X), \ldots, I_{A_m}(X), Z)'$$

taken to be the predictors, where the disjoint intervals $(A_1, \ldots A_m)$ partition the real line. Because of measurement error, however, the recorded predictors

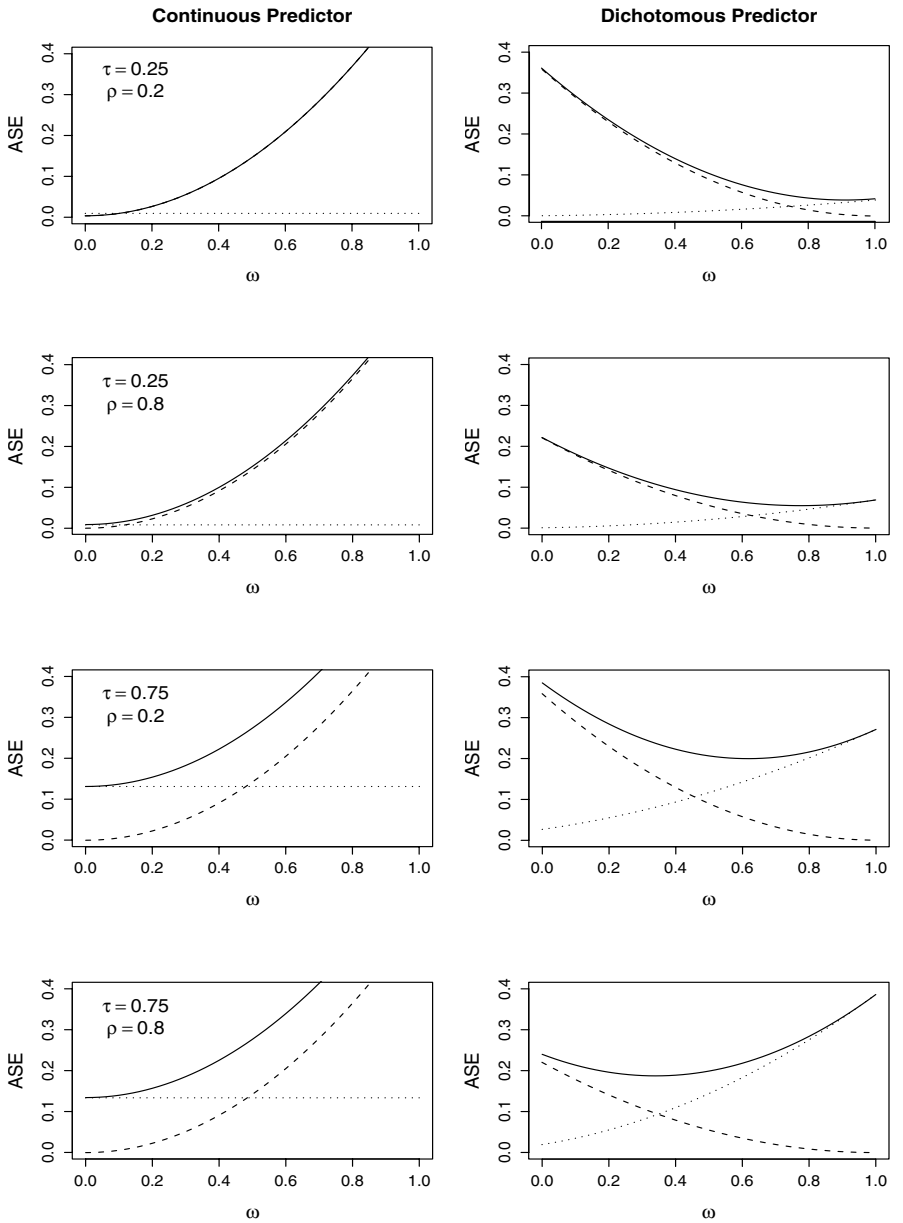


Figure 6.2 *$ASE_{TOT}$ (solid curve), $ASE_{SPC}$ (dashed curve), and $ASE_{MSR}$ (dotted curve) as a function of $\omega$ which indexes the true relationship (6.17). The four pairs of plots correspond to combinations of $\tau = 0.25$ or $\tau = 0.75$ and $\rho = 0.2$ or $\rho = 0.8$. For each pair, the left plot corresponds to $T^{(C)} = (1, X, Z)'$ as the postulated predictors, while the right plot corresponds to $T^{(D)} = (1, I\{X > 0\}, Z)'$ as the postulated predictors. The vertical scaling of the plots is based on $\nu_2 = 1$.*

are

$$T^* \quad = \quad (I_{A_1}(X^*), \ldots, I_{A_m}(X^*), Z)',$$

where $X^*$ is an unbiased and nondifferential surrogate for $X$. The true regression function is taken to be simply linear, i.e., $S = (1, X, Z)$. With lots of data one can argue for selecting a large number $m$ of indicator functions, to better approximate the unknown shape of the actual regression function. Indeed, $ASE_{SPC}$ decreases with $m$. This is partially offset, however, by $ASE_{MSR}$ which increases with $m$. Once again the two biases are inversely related. It almost seems as if a 'no-free-lunch' principle is at play, as steps taken to reduce the impact of model misspecification may well lead to increased error due to mismeasurement, and vice versa.

Another aspect of the interplay between model misspecification and mismeasurement emphasized in Gustafson (2002a) is that mismeasurement impairs the ability of standard diagnostic procedures to detect model misspecification. In situations where clear model violations would be evident from residual plots based on $(Y, X)$, the analogous plots based on $(Y, X^*)$ may appear to be in order. This can arise simply because the noise associated by observing $X^*$ rather than $X$ blurs the pattern in the plot. Hence it seems reasonable to develop flexible Bayesian response models for mismeasurement adjustment. Indeed, flexible response models may prove to have greater utility than flexible exposure models. This topic has not received much attention in the literature, and is ripe for further research. Berry, Carroll and Ruppert (2002) and Mallick, Hoffman and Carroll (2002) are recent papers with interesting work in this direction.

### 6.3 Identifiability in Mismeasurement Models

Consider a statistical model for observable data $y$ given a parameter vector $\theta$, expressed via a density function $f(y|\theta)$. The model is said to be *identifiable* if $f(y|\theta_1) = f(y|\theta_2)$ for all $y$ implies that $\theta_1 = \theta_2$. Intuitively, a *nonidentifiable* model lacking this property cannot yield consistent estimators of $\theta$ as a whole. In particular say $\theta_1$ and $\theta_2$ are distinct values of $\theta$ violating the property above, and say that in fact $\theta_1$ is the true parameter vector. Clearly data alone cannot shed light on whether $\theta_1$ is a more plausible parameter value than $\theta_2$, even with a very large sample size.

Often issues of parameter identifiability are quite central to models which account for mismeasurement, since a great deal must be known about the mismeasurement process in order to obtain a fully identified model. This is particularly true in scenarios where neither gold-standard measurements or pure replicate measurements can be obtained. We have already alluded to problems of nonidentifiability in Chapter 5. For instance, the model in Section 5.2 involving partial knowledge of misclassification probabilities is formally nonidentifiable, though the findings there suggest that reasonably good partial knowledge can still lead to useful inferences. In this section we take a closer

look at the utility of mismeasurement models that lack formal parameter identifiability.

### *6.3.1 Estimator Performance in Nonidentified Models*

We start with a general discussion of how much can be learned about a parameter of interest from a Bayes analysis with a nonidentifiable model. The general treatment follows Gustafson (2002b), which in turn builds on earlier investigations in specific nonidentifiable models (Neath and Samaniego 1997, Gustafson, Le and Saskin 2001). Say the model in question is initially parameterized by a vector $\theta$ with $p$ components. This can be thought of as the 'obvious' parameterization for scientific purposes, and the parameterization under which the prior distribution is specified in particular. Thus a model density $f(\text{data}|\theta)$ and a prior density $f(\theta)$ are at hand.

To gain insight regarding identifiability, we seek to reparameterize from the original parameter vector $\theta$ to $\phi = (\phi_I, \phi_N)$, in such a way that $f(\text{data}|\phi) = f(\text{data}|\phi_I)$. That is, the distribution of the data depends only on $\phi_I$, which we call the identifiable component of $\phi$, and not on $\phi_N$, the nonidentifiable component. We call such a parameterization *transparent*, as it is intended to make apparent the impact of nonidentifiability. Of course the prior distribution $f(\theta)$ as specified in the original parameterization can be transformed to $f(\phi)$ in the transparent parameterization. Indeed, along the lines of Dawid (1979), it is useful to think of the prior for $\phi$ in terms of the marginal density $f(\phi_I)$ and the conditional density $f(\phi_N|\phi_I)$. Then immediately Bayes theorem gives

$$f(\phi_I|\text{data}) \quad \propto \quad f(\text{data}|\phi_I)f(\phi_I), \tag{6.18}$$

while

$$f(\phi_N|\phi_I, \text{data}) \quad = \quad f(\phi_N|\phi_I). \tag{6.19}$$

Thus (6.18), the posterior marginal distribution for $\phi_I$, is typically governed by the regular asymptotic theory that applies to identifiable models. In particular, the posterior marginal distribution of $\phi_I$ converges to a point mass at the true value of $\phi_I$ as the sample size increases. Alternately, (6.19), the posterior conditional distribution for $\phi_N|\phi_I$, is identical to the prior conditional distribution. There is no Bayesian learning whatsoever about this conditional distribution.

Often a simple and reasonable choice of prior for $\theta$ will involve independent components, i.e., $f(\theta) = f(\theta_1) \ldots f(\theta_p)$. However, the induced prior $f(\phi)$ may involve substantial dependence between components, and dependence between $\phi_I$ and $\phi_N$ specifically. This has implications, as the posterior marginal density for $\phi_N$ is

$$f(\phi_N|\text{data}) \quad = \quad \int f(\phi_N|\phi_I)f(\phi_I|\text{data})d\phi_I,$$

which does not in general reduce to the prior marginal $f(\phi_N)$, unless $\phi_I$ and $\phi_N$ are *a priori* independent. Whereas conditionally there is no learning or

updating about $\phi_N$ given $\phi_I$, marginally there can be some learning about $\phi_N$ from the data. We refer to this as *indirect learning*, as it is learning about $\phi_N$ that results only because of, or via, learning about $\phi_I$.

Now consider a scalar parameter of interest $\psi$, which, while initially defined in terms of the $\theta$ parameterization, can be expressed as a function of $\phi$ equally well. Thus we write $\psi = g(\phi)$. We can estimate $\psi$ by its posterior mean, given as

$$\begin{aligned} E(\psi|\text{data}) &= E\{g(\phi_I, \phi_N)|\text{data}\} \\ &= \int\int g(\phi_I, \phi_N) f(\phi_I, \phi_N|\text{data})\, d\phi_N\, d\phi_I \\ &= \int\int g(\phi_I, \phi_N) f(\phi_N|\phi_I)\, d\phi_N\, f(\phi_I|\text{data})\, d\phi_I \\ &= E\{\tilde{g}(\phi_I)|\text{data}\}, \end{aligned}$$

where

$$\tilde{g}(\phi_I) = \int g(\phi_I, \phi_N) f(\phi_N|\phi_I) d\phi_N .$$

Thus the posterior mean of $\psi = g(\phi_I, \phi_N)$ is identically the posterior mean of $\tilde{g}(\phi_I)$, which of course is a function of the identifiable parameter component alone.

Given the re-expression of the posterior mean in terms of $\phi_I$ alone, under weak regularity conditions the large-sample behaviour of the estimator $E(\psi|\text{data})$ is described by the usual asymptotic theory applied to the identifiable model $f(\text{data}|\phi_I)$. That is, if the model is correct and a sample of size $n$ yields $\hat{\psi}^{(n)} = E(\psi|\text{data})$, then

$$n^{1/2}\left\{\hat{\psi}^{(n)} - \tilde{g}(\phi_I)\right\} \quad \Rightarrow \quad N\left[0, \{\tilde{g}'(\phi_I)\}' I(\phi_I)^{-1}\{\tilde{g}'(\phi_I)\}\right]$$

in distribution, as $n \to \infty$. Here $I(\phi_I)$ is the Fisher information matrix based on $f(\text{data}|\phi_I)$ for a single datapoint. Of course the estimator is biased, as it converges to $\tilde{g}(\phi_I)$ rather than $g(\phi_I, \phi_N)$ as desired. Indeed, considering both bias and variance, the mean-squared-error ($MSE$) incurred when estimating $\psi$ by $\hat{\psi}^{(n)}$ can be approximated as

$$\begin{aligned} MSE \approx\ & \{\tilde{g}(\phi_I) - g(\phi_I, \phi_N)\}^2 + \\ & n^{-1}\{\tilde{g}'(\phi_I)\}' I(\phi_I)^{-1}\{\tilde{g}'(\phi_I)\}, \end{aligned} \tag{6.20}$$

where the first term is the squared asymptotic bias and the second term is large-sample approximate variance. Despite being very easy to establish, this approach to quantifying the frequentist performance of a posterior mean in a nonidentified model does not seem to have been pursued prior to Gustafson (2002b).

Given a prior and true value for $\theta$, (6.20) indicates how well a parameter of interest can be estimated. Thus we can investigate how much prior information is needed to overcome the lack of identifiability and obtain useful inferences. In many mismeasurement models $\phi_N$ will be of dimension one or two, making

it feasible to compute $\tilde{g}()$ and its derivatives numerically if no closed-form exists. While indirect learning is discussed less formally in the next example, the subsequent example of Section 6.3.3 uses (6.20) explicitly.

### *6.3.2 Example: Partial Knowledge of Misclassification Probabilities Revisited*

We return to the scenario of Section 5.2, to examine the impact of nonidentifiability more closely when only partial information about misclassification probabilities is available. Recall that the goal is to make inferences about the odds-ratio in a case-control scenario with nondifferential misclassification of the binary exposure. The initial parameterization is $\theta = (r_0, r_1, p, q)$, where $r_i$ is the exposure prevalence in the $i$-th population ($i = 0$ corresponding to controls, $i = 1$ corresponding to cases), while $(p, q)$ are the sensitivity and specificity of exposure misclassification. As alluded to in Section 5.2, a transparent parameterization involves $\phi_I = (\theta_0, \theta_1)$ and $\phi_N = (p, q)$, where $\theta_i$ is the prevalence of apparent exposure in the $i$-th population.

Say the investigator starts with an independence prior in the original parameterization, assigning uniform priors to the prevalences $r_0$ and $r_1$, and priors with densities $f(p)$ and $f(q)$ to $p$ and $q$ respectively. Then a slight extension of the developments in Section 5.6.2 gives the conditional prior density of $\phi_N$ given $\phi_I$ as

$$f(p, q|\theta_0, \theta_1) \propto \frac{1}{(p+q-1)^2} f(p) f(q) I_{A(\theta_0,\theta_1)}(p, q), \tag{6.21}$$

where $A(\theta_0, \theta_1)$ is the union of two rectangles in the unit square, as defined in (5.25). In particular, $A$ consists of $(p, q)$ values which are componentwise greater than $(p, q) = (\max\{\theta_0, \theta_1\}, 1 - \min\{\theta_0, \theta_1\})$, and values which are componentwise less than $(p, q) = (\min\{\theta_0, \theta_1\}, 1 - \max\{\theta_0, \theta_1\})$. Note also that the $(p + q - 1)^{-2}$ term in (6.21) arises from the Jacobian associated with reparameterization.

Bearing in mind that as the sample size grows the posterior distribution on $(p, q)$ will converge to (6.21) evaluated at the true values of $(\theta_0, \theta_1)$, we see that indirect learning is manifested in two ways. First, the shape of the density on $(p, q)$ is modified by the multiplicative term $(p + q - 1)^{-2}$. On first glance one might think this to have great impact, as this term tends to infinity along the $q = 1 - p$ line in the unit square. However, this line necessarily lies entirely outside the support set $A(\theta_0, \theta_1)$, thereby limiting the influence of this term. Second, the support set may exclude regions with substantial probability under the marginal prior $f(p)f(q)$, thereby making the conditional prior (6.21) more concentrated than the marginal prior.

As a specific example, say the investigator selects identical Beta priors for $p$ and $q$, both with mean 0.8 and standard deviation 0.075. In particular this yields the $Beta(21.96, 5.49)$ distribution. Again this sort of prior might be viewed as a quantification of quite crude knowledge that the exposure assessments are good but not perfect. A $(p, q)$ scatterplot of a simulated sample from this prior is given in the upper-left panel of Figure 6.3. Now say that in

| scenario | $r_0$ | $r_1$ | $\theta_0$ | $\theta_1$ |
|---|---|---|---|---|
| (i) | 0.0625 | 0.9375 | 0.15 | 0.85 |
| (ii) | 0.375 | 0.625 | 0.4 | 0.6 |
| (iii) | 0.0625 | 0.1250 | 0.15 | 0.2 |

Table 6.3 *Three scenarios for prevalence of actual and apparent exposure in a case-control study. The prevalences of apparent exposure* $(\theta_0, \theta_1)$ *are obtained from the prevalences of actual exposure* $(r_0, r_1)$ *via nondifferential misclassification with sensitivity* $p = 0.9$ *and specificity* $q = 0.9$.



fact the true values of both $p$ and $q$ are 0.9, i.e., each is slightly more than one prior SD away from the prior mean. This typifies a situation where the prior distribution is neither spectacularly good or bad as a representation of the actual truth. Finally, consider three possible values of the exposure prevalences $(r_0, r_1)$. These are given along with the implied apparent exposure prevalences $(\theta_0, \theta_1)$ in Table 6.3. Note in particular that scenario (iii) might be quite realistic for case-control studies.

For prevalences (i) the upper-right panel of Figure 6.3 displays a sample from the limiting posterior marginal distribution of $(p, q)$, i.e., the prior conditional distribution (6.21) evaluated at the true values of $(\theta_0, \theta_1)$. Note that truncation has a very considerable and desirable impact, forcing the limiting posterior distribution to be very concentrated near the true values of $(p, q)$ relative to the prior distribution. Thus considerable learning about $(p, q)$ can occur. For prevalences (ii), however, the impact of truncation is almost negligible (lower-left panel of Figure 6.3). In this instance the multiplicative effect of the $(p+q-1)^{-2}$ term is apparent from the plot, though the discrepancy between the prior and limiting posterior distributions is quite mild. Thus little can be learned about $(p, q)$ from the data in this scenario. Finally, prevalences (iii) lead to considerable learning about the specificity $q$ but virtually no learning about the sensitivity $p$ (lower-right panel of Figure 6.3).

The overall message from Figure 6.3 is that the the efficacy of indirect learning about $\phi_N = (p, q)$ is quite unpredictable. For a given prior distribution and true value for $\phi_N$, the amount of indirect learning varies dramatically with the true value of $\phi_I$, from very considerable with prevalences (i), to very slight with prevalences (ii). We find this sort of unpredictability to be the norm with nonidentified mismeasurement models. In essence it seems desirable to formulate the best prior possible to permit the *possibility* of benefiting from indirect learning, but in advance one cannot count on such learning taking place to a substantial extent.

In the present inferential scenario, Gustafson, Le and Saskin (2001) compare using a prior on $(p, q)$ to treating $(p, q)$ as known. More precisely they consider 'best guess' values for $(p, q)$, and compare the posterior distribution arising from a prior distribution centred at the guessed values to the posterior distribution arising from pretending that the guessed values are known true values. The latter approach gains identifiability at the cost of model misspecification.

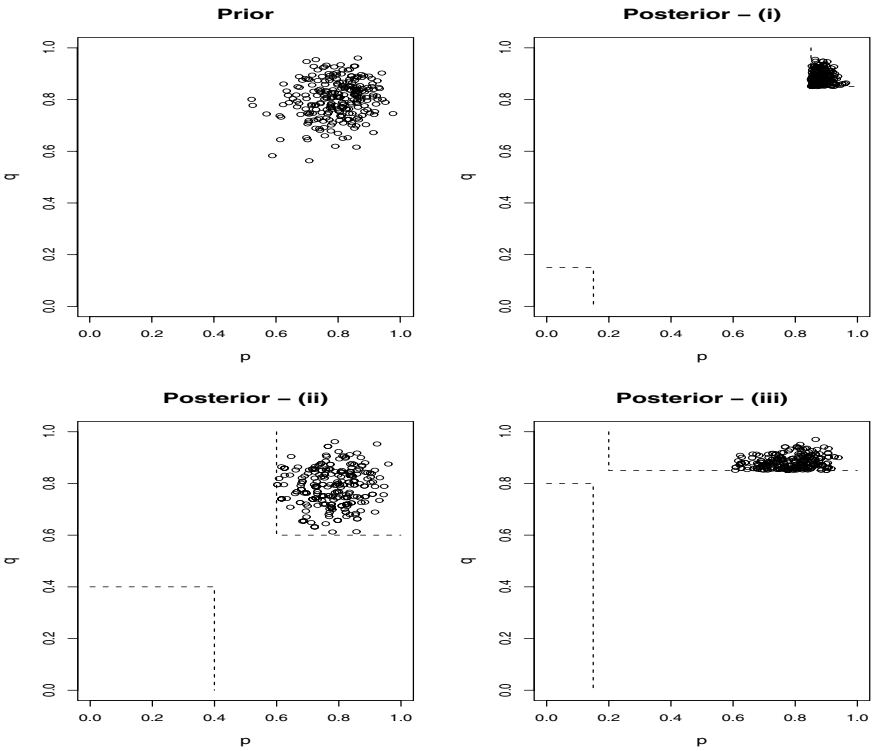


Figure 6.3 *Prior and limiting posterior distributions for* $(p, q)$. *Each panel gives* $(p, q)$ *scatterplots of a sample from the distribution in question. The upper-left panel gives the prior distribution, while the remaining panels give the large-sample limiting posterior distributions under prevalences (i), (ii), and (iii) as described in the text. The true values are* $(p, q) = (0.9, 0.9)$ *throughout. The dashed lines delineate the rectangles* $A(\theta_0, \theta_1)$ *comprising the support of the limiting posterior distribution.*

Of course there is a generic argument in favour of assigning prior distributions to quantities which are genuinely unknown, to more fully account for all the uncertainties at play. In the present setting then, one might expect that admitting uncertainty about $(p, q)$ would result in wider posterior distributions for parameters of interest. Indeed, for some underlying parameter values Gustafson, Le and Saskin (2001) find this to be the case. That is, credible intervals arising from a prior on $(p, q)$ are wider on average, but have higher empirical coverage probabilities than those arising from fixed $(p, q)$. For other underlying parameter values, however, using a prior gives credible intervals that are *both* shorter on average and have higher empirical coverage than their fixed $(p, q)$ counterparts. In particular, this occurs when substantial indirect learning occurs. Surprisingly, even though the data do not inform us directly about $\phi_N$, there are circumstances where admitting uncertainty about $\phi_N$ *a*

*priori* can actually lead to decreased uncertainty about parameters of interest *a posteriori*!

### *6.3.3 Linear and Normal Models Revisited*

Here we revisit the combination of linear and normal measurement, outcome, and exposure models, as discussed in Chapter 4. In particular, say the regression of $Y$ on $X$ is of interest, but only $(X^*, Y)$ can be observed, where $X^*$ is a surrogate for $X$. The measurement, outcome, and exposure models are taken to be

$$\begin{aligned} X^*|X,Y &\sim N(X, \gamma\lambda^2), \\ Y|X &\sim N(\beta_0 + \beta_1 X, \sigma^2), \\ X &\sim N(\alpha_0, \lambda^2). \end{aligned}$$

Note that the parameter $\gamma$ is used to describe the ratio of $Var(X^*|X)$ to $Var(X)$, i.e., $\gamma$ is interpretable as the measurement error magnitude expressed in relative terms. Note also that this parameterization is particularly relevant in the common scenario that $X$ and $X^*$ are logarithmic transformations of positive variables thought to be related via a multiplicative measurement error relationship.

It is easy to verify that if all six parameters $\theta = (\beta_0, \beta_1, \alpha_0, \sigma^2, \lambda^2, \gamma)$ are unknown then the model is nonidentifiable. Essentially the model assumes that $(X^*, Y)$ have a bivariate normal distribution, so that at most five parameters can be estimated consistently. Toward applying the ideas of Section 6.3.1, note that a transparent parameterization obtains by taking $\phi_I = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\alpha}_0, \tilde{\lambda}^2, \tilde{\sigma}^2)$ and $\phi_N = \gamma$, where

$$\begin{aligned} \tilde{\beta}_0 &= \beta_0 + \alpha_0\beta_1/(1+\gamma), \\ \tilde{\beta}_1 &= \beta_1/(1+\gamma), \\ \tilde{\alpha}_0 &= \alpha_0, \\ \tilde{\sigma}^2 &= \sigma^2 + \beta_1^2\lambda^2\gamma/(1+\gamma), \\ \tilde{\lambda}^2 &= \lambda^2(1+\gamma), \end{aligned}$$

so that

$$\begin{aligned} Y|X^* &\sim N(\tilde{\beta}_0 + \tilde{\beta}_1 X^*, \tilde{\sigma}^2), \\ X^* &\sim N(\tilde{\alpha}_0, \tilde{\lambda}^2). \end{aligned}$$

Now say a prior of the form

$$f(\beta_0, \beta_1, \alpha_0, \sigma^2, \lambda^2, \gamma) = f_1(\beta_0) f_2(\beta_1) f_3(\alpha_0) f_4(\sigma^2) f_5(\lambda^2) f_6(\gamma)$$

is assigned in the original parameterization. One can verify that the $\theta \to \phi$ reparameterization has unit Jacobian, and consequently the conditional prior for $\phi_N|\phi_I$ takes the form

$$f(\gamma|\beta_0^*, \beta_1^*, \mu_*, \sigma_*^2, \lambda_*^2) \propto f_2\left(\tilde{\beta}_1(1+\gamma)\right) \times$$

$$f_4\left(\tilde{\sigma}^2 - \gamma\tilde{\beta}_1^2\tilde{\lambda}^2\right) \times$$
$$f_5\left(\tilde{\lambda}^2/(1+\gamma)\right) \times$$
$$f_6(\gamma) I_{\left(0, m(\tilde{\beta}_1, \tilde{\sigma}^2, \tilde{\lambda}^2)\right)}(\gamma), \qquad (6.22)$$

where

$$m\left(\tilde{\beta}_1, \tilde{\sigma}^2, \tilde{\lambda}^2\right) = \min\left\{\frac{\tilde{\sigma}^2}{\tilde{\beta}_1^2\tilde{\lambda}^2}, 1\right\}.$$

Thus there is potential for indirect learning here, through both the multiplicative terms which act on $f_6(\gamma)$ in (6.22), and the truncation that arises if $m()$ is less than one.

The nature of indirect learning in this context is investigated by elaborating on an example from Gustafson (2002b). Consider the true parameter values $\gamma = 0.3$, $\beta_0 = 0$, $\alpha_0 = 0$, $\sigma^2 = 0.25$, and $\lambda^2 = 1$. Three different true values for $\beta_1$ are considered, namely $\beta_1 = 0.5$, $\beta_1 = 1$, and $\beta_1 = 1.5$. Prior independence in the initial parameterization is assumed, with the priors for $\beta_0$, $\beta_1$, and $\alpha_0$ taken to be $N(0,1)$ and the priors for $\sigma^2$ and $\lambda^2$ taken to be $IG(0.5, 0.5)$. Such priors might be appropriate if the data are standardized before analysis. Finally a prior of the form $\gamma \sim Beta(a,b)$ is specified, with four possible values for the hyperparameters considered. In case (i), $(a,b) = (1,1)$, i.e., $\gamma$ has a uniform prior distribution. In the remaining instances $(a,b)$ are selected to attain a desired prior mean and standard deviation. Specifically prior (ii) is taken to yield $E(\gamma) = 0.4$, $SD(\gamma) = 0.1$, while prior (iii) gives $E(\gamma) = 0.2$, $SD(\gamma) = 0.1$. In both cases the true value of $\gamma = 0.3$ is one prior standard deviation away from the prior mean, again representing priors of 'typical' quality in representing the truth. Prior (iv) is also of typical quality in this sense, with $E(\gamma) = 0.35$, $SD(\gamma) = 0.05$. This smaller prior standard deviation corresponds to sharper prior information than in (ii) or (iii).

For each choice of prior and each true value of $\beta_1$, Figure 6.4 plots the conditional density (6.22) evaluated at the true value of $\phi_I$. Recall that this is the large-sample limiting posterior marginal distribution for $\gamma$, and by comparing this density to the prior density we see how much can be learned about $\gamma$ from the data. The figure shows that the amount of learning varies greatly across the different priors and across the underlying value of $\beta_1$. Generally, when the updating is appreciable it is also desirable. That is, when the data do influence the prior they do 'push' toward the true value of $\gamma$. Also, there tends to be more updating when the underlying value of $\beta_1$ is larger. In comparing priors (ii) and (iii) we see more ability to push prior (ii) to the left than to push prior (iii) to the right. This is due to the truncation in (6.22), which can eliminate larger values of $\gamma$. In comparing priors (ii) and (iv) more updating is seen with the former, presumably because it is a less concentrated prior distribution initially. Overall the findings are in keeping with our general impression in the previous subsection — the extent to which indirect learning occurs is quite unpredictable.

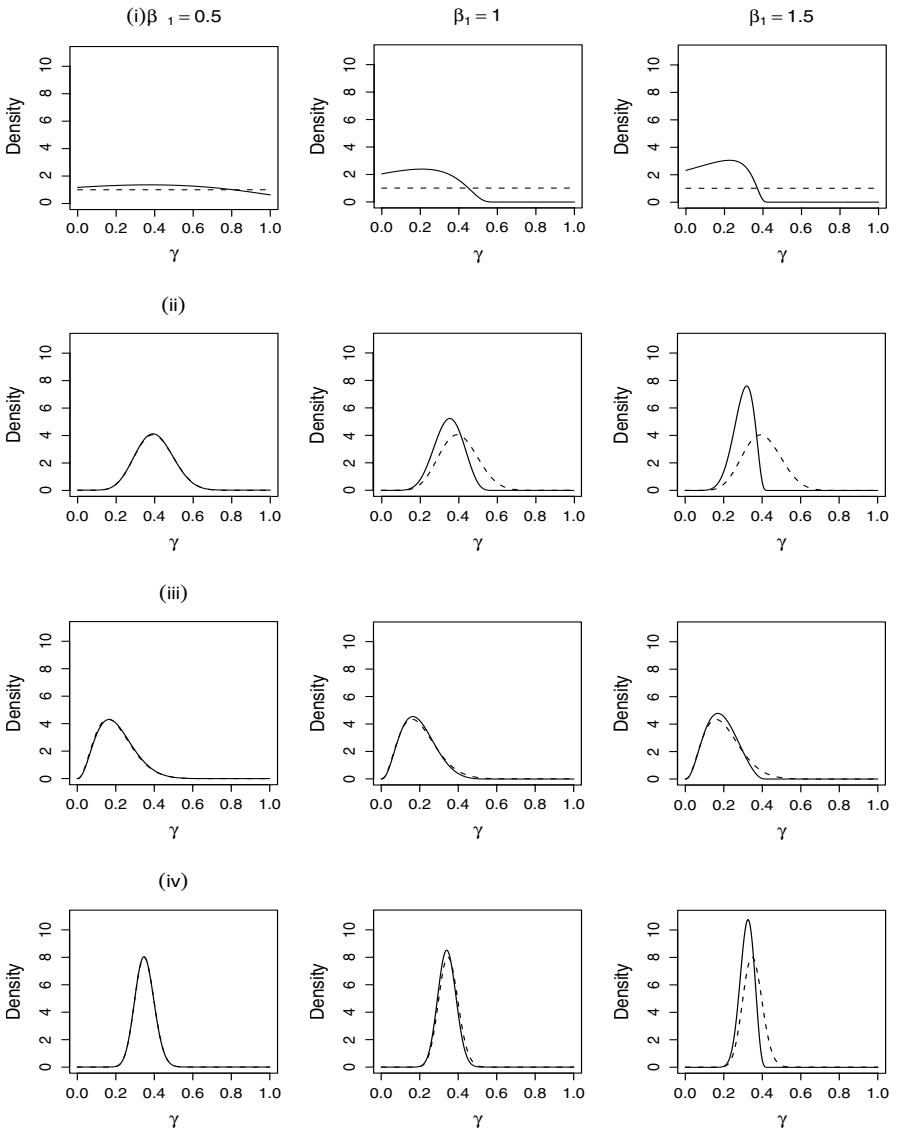


Figure 6.4 *Prior and limiting posterior densities for $\gamma$. Each panel gives the prior marginal density as a dotted curve and the limiting posterior density (equivalently the prior conditional density) as a solid curve. The rows correspond to Beta priors (i) through (iv) described in the text, while the columns correspond to different underlying values of $\beta_1$.*

| prior | $\beta_1 = 0.5$ | $\beta_1 = 1$ | $\beta_1 = 1.5$ |
|---|---|---|---|
| (i) | 0.78 | 0.39 | 0.58 |
| (ii) | 0.99 | 0.50 | 0.01 |
| (iii) | 0.97 | 1.06 | 1.13 |
| (iv) | 1.00 | 0.83 | 0.38 |

Table 6.4 *Ratio of absolute asymptotic bias in using a prior on $\gamma$ versus pretending $\gamma$ is known, when estimating $\beta_1$. The columns correspond to priors (i) through (iv) as described in the text. The rows correspond to different underlying values of $\beta_1$. Note that the layout matches that of Figure 6.4*



With some knowledge about $\gamma$, one alternative to using a prior distribution is simply to fix $\gamma$ at a 'best guess' value. This is the notion of 'adjusting with false certainty' first discussed in Section 5.2. While this strategy leads to identifiability, the model will be somewhat misspecified unless the guess is correct. To explore this tradeoff we consider the same scenarios as portrayed in Figure 6.4, with the guessed value of $\gamma$ matching the prior mean for $\gamma$ in all instances. For estimating $\beta_1$, Table 6.4 gives the ratio of absolute asymptotic bias when using the prior on $\gamma$ to the absolute asymptotic bias when $\gamma$ is fixed. Recall from Section 6.3.1 that the bias in the former case is given as $\tilde{g}(\phi_I) - g(\phi_I, \phi_N)$, and note that $\tilde{g}(\phi_I)$ is readily computed via one-dimensional numerical integration with respect to (6.22).

Not surprisingly, Table 6.4 shows a ratio of absolute biases close to one in scenarios where little indirect learning is manifested in Figure 6.4. If the data contain virtually no information about $\gamma$, then similar point estimates arise from either using a prior on $\gamma$ or fixing $\gamma$ at a commensurate value. On the other hand, scenarios where appreciable indirect learning occurs lead to much less bias when a prior is assigned to $\gamma$. That is, the use of a prior allows the data to push toward the true value of $\gamma$, which of course cannot happen if $\gamma$ is fixed in advance. In general then, the use of a reasonable prior may or may not lead to a substantial reduction in bias, but it is very unlikely to produce a substantial increase in bias.

To consider both bias and variance, Table 6.5 gives the approximate $RMSE$ when estimating $\beta_1$ with a sample of size $n = 500$ and a prior distribution on $\gamma$. This is computed as the square-root of (6.20). Again the different priors on $\gamma$ and different underlying values of $\beta_1$ match those in Figure 6.4. The overall impression is that the $RMSE$ tends to be quite small in absolute terms, to the extent that inferences of useful precision can be drawn even with a uniform prior distribution on $\gamma$, or in scenarios where virtually no indirect learning occurs. As a point of comparison, the last row of Table 6.5 gives the $RMSE$ for the naive estimator which doesn't account for mismeasurement, i.e., the slope coefficient from regression of $Y$ on $X^*$. In all cases these $RMSE$ values are considerably larger than those arising from the mismeasurement model, indicating that the attempted adjustment is well worthwhile even without good prior information on $\gamma$. For instance, over the four priors and three

| | $\beta_1 = 0.5$ | | $\beta_1 = 1$ | | $\beta_1 = 1.5$ | |
|---|---|---|---|---|---|---|
| prior | RMSE | (%DTB) | RMSE | (%DTB) | RMSE | (%DTB) |
| (i) | 0.065 | (86%) | 0.065 | (82%) | 0.138 | (94%) |
| (ii) | 0.048 | (62%) | 0.049 | (61%) | 0.039 | (0%) |
| (iii) | 0.045 | (68%) | 0.087 | (88%) | 0.136 | (93%) |
| (iv) | 0.035 | (30%) | 0.047 | (47%) | 0.045 | (25%) |
| NAIVE | 0.117 | (97%) | 0.232 | (99%) | 0.348 | (99%) |

Table 6.5 *Approximate Root-Mean-Squared-Error (RMSE) in estimating $\beta_1$ with a sample size of $n = 500$. The fraction of the mean-squared-error due to bias, i.e., the ratio of squared bias to MSE, is given as the percentage due to bias (%DTB). As previously, the rows and columns correspond to different priors and different underlying values of $\beta_1$. The last row corresponds to naive estimation which ignores the measurement error.*

| prior | $\beta_1 = 0.5$ | $\beta_1 = 1$ | $\beta_1 = 1.5$ |
|---|---|---|---|
| (i) | 0.78 | 0.41 | 0.58 |
| (ii) | 0.98 | 0.57 | 0.31 |
| (iii) | 0.97 | 1.04 | 1.11 |
| (iv) | 1.00 | 0.88 | 0.61 |

Table 6.6 *Ratio of approximate RMSE using a prior on $\gamma$ to pretending $\gamma$ is known, when estimating $\beta_1$ with a sample size of $n = 500$. The columns correspond to priors (i) through (iv) as described in the text. The rows correspond to different underlying values of $\beta_1$. Note that the layout matches that of Figure 6.4.*

values of $\beta_1$, the largest ratio of $RMSE$ to $\beta_1$ occurs with the uniform prior (i) and $\beta_1 = 0.5$. Even in this worst-case, however, the $RMSE$ is an order of magnitude smaller than that achieved by the naive estimator.

Table 6.5 also reports the percentage of the $MSE$ that is due to squared-bias rather than variance. It is interesting to note that in most cases the bias contribution does not dominate the variance contribution, despite the relatively large sample size. Given this, we revisit the comparisons of Table 6.4 between using a prior on $\gamma$ and fixing $\gamma$ at the corresponding best guess. In particular, Table 6.6 reports the ratio of $RMSE$ using a prior to $RMSE$ with $\gamma$ fixed. Again the table indicates ratios much less than one in instances where substantial indirect learning occurs, and ratios near one otherwise.

### *6.3.4 Dual Exposure Assessment in One Population*

Section 5.3 considered dual assessment of a binary exposure in a case-control setting, and the DUAL-IND model described there is applicable more generally to problems with two misclassified surrogates measured in two populations. In particular, recall that an identifiable six-parameter model results under the strong assumptions that the exposure assessments are nondifferential and conditionally independent of one another given the true exposure.

On the other hand, it is easy to verify that identifiability is lost if the assessments are made in only one population. Notwithstanding this, Joseph, Gyorkos and Coupal (1995) consider dual exposure assesesment in a single population, focussing on Bayesian inference about the exposure prevalence and the four misclassification probabilities. They use a uniform prior on the exposure prevalence, but relatively informative priors on the misclassification probabilities.

Johnson, Gastwirth and Pearson (2001) criticize the approach of Joseph *et al.*, particularly because it lacks identifiability. They argue that one should always split the population under study into two distinct sub-populations or strata, as a route to the identifiable DUAL-IND model. While they are not very prescriptive about how this splitting might be done in practice, demographic variables that are routinely collected in epidemiological investigations could be used for this purpose. For instance, one can imagine using gender or age to define two strata.

On the face of it the argument of Johnson *et al.* seems compelling. Why work with a nonidentifiable model if the data can be augmented easily to yield an identifiable model? As noted in Section 5.3, however, the identifiability breaks down in the special case that both strata have identical exposure prevalences, i.e., $r_0 = r_1$ in the notation used previously. Intuitively, this confirms that identifiability cannot be obtained by flipping coins to split the sampled units into two strata. Further in this vein, it may also be difficult to obtain precise inferences if the prevalences differ only modestly from one another. For instance, unless there is reason to believe that the exposure prevalence varies markedly with gender, stratifying according to gender may be a poor strategy. Perhaps then there are circumstances where it is reasonable to proceed as in Joseph *et al.* with a single population.

To explore this issue, Gustafson (2002b) uses the ideas of 6.3.1 to assess how well the exposure prevalence can be estimated in a single population, using subjective, though not terribly precise, prior distributions on the misclassification probabilities. In this setting the five-dimensional initial parameterization $\theta = (r, p_1, q_1, p_2, q_2)$ involves the exposure prevalence and the four misclassification probabilities for the two assessments. The reparameterization to $\phi = (\phi_I, \phi_N)$ involves taking $\phi_N$ to be three cell probabilities which determine the joint distribution of the two assessments, while $\phi_N$ is conveniently taken to be the exposure prevalence $r$ plus any one of the four misclassification probabilities. Two-dimensional integration can then be used to compute $\tilde{g}(\phi_I)$ and its derivatives, as needed to compute the approximate MSE using (6.20).

At least for the particular choices of priors and underlying parameter values investigated in Gustafson (2002b), reasonably precise inferences can result from relatively crude priors. In particular, scenarios where the same Beta prior is assigned to each of $p_1$, $q_1$, $p_2$, $q_2$ are considered, as representations of a rough belief that the two assessments are 'good but not perfect.' Note in particular that such a prior does not attempt to distinguish between the two assessments, or between the sensitivity and specificity for a given assessment.

The striking finding is that the resulting estimator performance can be better than that achieved by dichotomizing the population and using the DUAL-IND model, even if the sub-population prevalences differ to some extent. In short, consideration of the nonidentifiable model performance leads to a tempering of the recommendation in Johnson *et al.* Splitting the population to achieve identifiability will not always lead to improved estimator performance.

### 6.4 Further Remarks

The topics discussed in this chapter appear outwardly to be somewhat idiosyncractic. In all cases, though, they relate to broad issues in understanding and correcting for mismeasurement of explanatory variables. And generally speaking, all three topics are ripe for further research. The consideration of DDD misclassification in Section 6.1 points to how differential mismeasurement can arise easily and subtly, and reinforces the notion that differential mismeasurement can have quite a different impact than nondifferential mismeasurement. In Section 6.2 the decomposition of error in estimating a regression function reminds us that statistical analysis is usually subject to more than one potential bias. Moreover, the demonstrated interplay between the terms in the decomposition underscores the need to view the biases together rather than in isolation from one another. Finally, the focus on identifiability in Section 6.3 is very relevant, as sometimes not enough is known about the mismeasurement process to yield an identifiable model. The common solution to this problem is to pretend to know more than is justified, in terms of distributional assumptions and parameter values describing the mismeasurement. The result is a model which is identifiable but probably somewhat misspecified. The discussion in 6.3, however, elucidates the value of admitting what is not known via appropriate prior distributions, and proceeding with a nonidentifiable model.