CHAPTER 3

The Impact of Mismeasured Categorical Variables

Of course many explanatory variables encountered in statistical practice are categorical rather than continuous in nature. Mismeasurement arises for such a variable when the actual and recorded categories for subjects can differ. Fundamentally this *misclassification* differs from measurement error as discussed in Chapter 2, as now the surrogate variable cannot be expressed as a sum of the true variable plus a noise variable. Rather, one must characterize the mismeasurement in terms of *classification probabilities*, i.e., given the true classification, how likely is a correct classification. This chapter focusses primarily on the impact of misclassification in *binary* explanatory variables, with some discussion of *polychotomous* explanatory variables at the end of the chapter. Schemes to ameliorate the impact of misclassification are not considered until Chapter 5. Thus, as with Chapter 2, this chapter's role is to instill qualitative understanding of when unchecked mismeasurement can produce very misleading results. While the respective literatures on misclassification of categorical explanatory variables and measurement error of continuous explanatory variables have few points of contact, this chapter demonstrates considerable similarities between the respective impacts of both these kinds of mismeasurement.

3.1 The Linear Model Case

Consider the relationship between a continuous response variable Y and a binary explanatory variable V, with the latter taken to have a zero/one coding. Since V is binary, without loss of generality we can write

$$E(Y|V) = \alpha_0 + \alpha_1 V. \tag{3.1}$$

Now say that for study subjects we observe (Y, V^*) rather than (Y, V), where the zero/one variable V^* is an imperfect surrogate for V. Under the assumption of nondifferential misclassification, whereby V^* and Y are conditionally independent given V, the magnitude of the misclassification can be described by the *sensitivity* and *specificity* of V^* as a surrogate for V. With a view to epidemiological applications, we tend to refer to $V = 0$ as 'negative' or 'unexposed,' and $V = 1$ as 'positive' or 'exposed.' Then the sensitivity $SN = Pr(V^* = 1|V = 1)$ is the probability that a true positive is correctly classified, while the specificity $SP = Pr(V^* = 0|V = 0)$ is the probability that a true negative is correctly classified. Thus the extent to which SN and SP are less than one reflects the severity of the misclassification.

Starting from (3.1) and using the nondifferential property yields

$$\begin{aligned} E(Y|V^*) &= E\{E(Y|V)|V^*\} \\ &= \alpha_0 + \alpha_1 E(V|V^*) \\ &= \alpha_0 + \alpha_1 V^* Pr(V=1|V^*=1) + \\ &\quad \alpha_1(1-V^*)Pr(V=1|V^*=0) \\ &= \alpha_0^* + \alpha_1^* V^*, \end{aligned}$$

where

$$\alpha_0^* = \alpha_0 + \alpha_1 Pr(V=1|V^*=0)$$

and

$$\frac{\alpha_1^*}{\alpha_1} = 1 - Pr(V=0|V^*=1) - Pr(V=1|V^*=0). \qquad (3.2)$$

In terms of estimating α_1 without any correction for misclassification, (3.2) shows that more attenuation results from larger probabilities of misclassification given the apparent classification.

Of course it might seem more natural to express the attenuation the other way around, in terms of sensitivity and specificity which are probabilities of misclassification given the true classification. Towards this, let $r = Pr(V=1)$ and $r^* = Pr(V^*=1)$, which can be thought of as the actual and apparent prevalences of exposure in the population at hand. Clearly r^* is a function of (r, SN, SP) given by

$$\begin{aligned} r^* &= Pr(V=1)Pr(V^*=1|V=1) + Pr(V=0)Pr(V^*=1|V=0) \\ &= rSN + (1-r)(1-SP) \\ &= (1-SP) + (SN+SP-1)r. \end{aligned}$$

Also, from Bayes theorem we have

$$\begin{aligned} Pr(V=0|V^*=1) &= \{Pr(V=0)Pr(V^*=1|V=0)\}/Pr(V^*=1) \\ &= (1-r)(1-SP)/r^*, \end{aligned}$$

and similarly

$$Pr(V=1|V^*=0) = r(1-SN)/(1-r^*).$$

Substituting into (3.2) gives the attenuation factor as

$$\begin{aligned} \frac{\alpha_1^*}{\alpha_1} &= 1 - \frac{(1-r)(1-SP)}{r^*} - \frac{r(1-SN)}{1-r^*} \\ &= (SN+SP-1)\frac{r(1-r)}{r^*(1-r^*)}, \qquad (3.3) \end{aligned}$$

where the second, perhaps more interpretable, equality follows from the first after some algebraic manipulation. In interpreting (3.3) it must be remembered that r^* is a function of (r, SN, SP).

Clearly (3.3) becomes one when $SN = 1$ and $SP = 1$, as must be the case. By taking partial derivatives it is simple to check that (3.3) is increasing in SN for fixed SP and increasing in SP for fixed SN. Thus the effect of misclassification is an attenuating bias, and the attenuation worsens with the severity of the misclassification. To get a sense for the magnitude of the bias, Figure 3.1 gives contour plots of (3.3) as SN and SP vary, for two different values of r. When $r = 0.5$ (left panel), the bias is symmetric in SN and SP, as can be seen by direct inspection of (3.3). The plotted contours reveal that substantial attenuation can occur without the misclassification being very severe. For instance, $SN = SP = 0.9$, which can be interpreted as 10% misclassification, gives $\alpha_1^*/\alpha_1 = 0.8$, interpreted as 20% attenuation. This can be compared to the analogous finding for continuous measurement error in Section 2.1 where 10% measurement error produces only 1% attenuation. Having 10% of all the measurements 'entirely corrupted' in the binary case is clearly much more damaging than having *all* the measurements corrupted by about 10% in the continuous case! We also note from the contour shapes that 'unbalanced' misclassification where SN and SP differ is slightly less damaging than the balanced case of $SN = SP$. For instance, $(SN = 1, SP = 0.8)$ gives $\alpha_1^*/\alpha_1 = 0.833$, corresponding to slightly less attenuation than the $SN = SP = 0.9$ case.

The right panel of Figure 3.1 corresponds to $r = 0.1$, which can be regarded as a 'rare exposure' scenario typifying many epidemiological investigations. Clearly the attenuation worsens much more with declining specificity than with declining sensitivity in this scenario. This makes sense, as there are far more true negatives than true positives in the population. Hence the specificity describing the classification of true negatives has a bigger impact than the sensitivity describing the classification of true positives. When SN and SP are comparable, the attenuation is now stronger than in the $r = 0.5$ case of common exposure. For instance, $SN = SP = 0.9$ now yields a very strong attenuation factor of 0.49, compared to 0.8 in the $r = 0.5$ case. This is of particular concern in light of the general epidemiological interest in rare exposures. Indeed, still considering $SN = SP = 0.9$ we find from (3.3) that when $r = 0.05$ the attenuation factor drops further to 0.32, an exceedingly substantial attenuation. In the context of a rare exposure, even mild nondifferential misclassification can lead to wildly misleading inferences if left unchecked.

3.2 More General Impact

From Chapter 2 we know that the bias induced in the regression coefficient for an explanatory variable subject to continuous measurement error is worsened by the presence of an additional precisely measured explanatory variable, provided the two explanatory variables are correlated. The following result gives a comparable finding for misclassification. In line with Chapter 2 we frame the result in terms of comparing large-sample limiting coefficients for the mismeasured and properly measured cases, without regard to how well

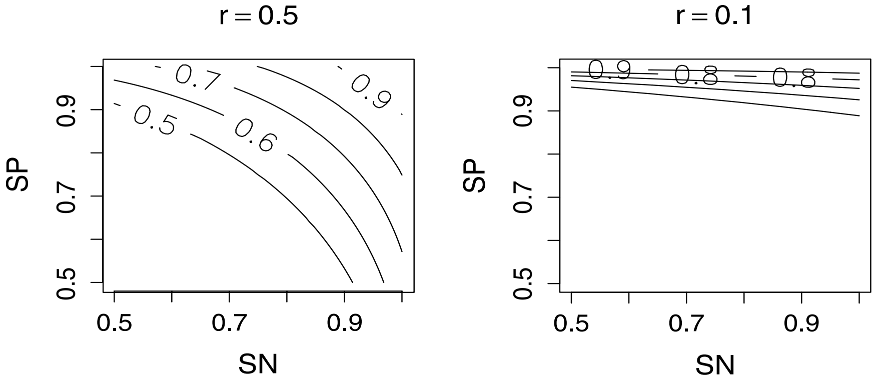

Figure 3.1 *Contours of the attenuation factor as a function of sensitivity and specificity. The left panel corresponds to* $r = Pr(V = 1) = 0.5$ *and the right panel corresponds to* $r = 0.1$*. The contours correspond to an attenuation factor (3.3) of* 0.9*,* 0.8*,* 0.7*,* 0.6*, and* 0.5*.*

the postulated distribution of the response variable given the explanatory variables matches the actual distribution.

Result 3.1 *Suppose that* (Y, V, Z) *jointly have finite second moments, where* V *is a 0-1 random variable. Furthermore, say that* V^* *is another 0-1 random variable defined according to*

$$Pr(V^* = 1|V, Z, Y) \quad = \quad a + bV,$$

where

$$\begin{aligned} a &= 1 - SP, \\ b &= SN + SP - 1. \end{aligned}$$

That is, we have nondifferential misclassification with sensitivity SN *and specificity* SP*. Let* $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ *and* $\alpha^* = (\alpha_0^*, \alpha_1^*, \alpha_2^*)'$ *be the large-sample limiting coefficients from least-squares regression of* Y *on* $(1, V, Z)'$ *and* Y *on* $(1, V^*, Z)'$ *respectively. Then,*

$$\begin{aligned} \frac{\alpha_1^*}{\alpha_1} &= b\left[\frac{r(1-r)(1-\rho^2)}{r^*(1-r^*) - r(1-r)\rho^2 b^2}\right], \qquad (3.4) \\ \alpha_2^* - \alpha_2 &= \alpha_1 \rho \frac{\{r(1-r)\}^{1/2}}{\sigma_z}\left[1 - b\left(\frac{\alpha_1^*}{\alpha_1}\right)\right], \\ \alpha_0^* - \alpha_0 &= r\alpha_1 - r^*\alpha_1^* + \mu_z(\alpha_2^* - \alpha_2), \end{aligned}$$

where $r = Pr(V = 1)$*,* $r^* = Pr(V^* = 1) = a + br$*,* $\mu_z = E(Z)$*,* $\sigma_z = SD(Z)$*, and* $\rho = Cor(V, Z)$*.*

We focus on (3.4), the attenuation factor in estimating the slope α_1 without correction for misclassification. It is straightforward to verify that this factor is increasing in SN for fixed (SP, r, ρ), and increasing in SP for fixed (SN, r, ρ), as is to be expected. In Section 3.8 we also verify that (3.4) is decreasing in $|\rho|$ for fixed SN, SP, and r. That is, the attenuation does worsen with increasing correlation between the two explanatory variables. Note as well that when $\rho = 0$, (3.4) reduces to (3.3). Finally, in Section 3.8 we show that as a function of r for fixed SN, SP, and ρ, the attenuation factor (3.4) is largest when

$$r = \begin{cases} \frac{SP(1-SP)-\{SP(1-SP)-SN(1-SN)\}^{1/2}}{SP(1-SP)-SN(1-SN)} & \text{if } SN \neq SP, \\ 1/2 & \text{if } SN = SP, \end{cases}$$

with (3.4) decreasing to zero as r decreases from this value to zero or increases from this value to one. Particularly, the former case again raises concern about an acute impact of misclassification in rare-exposure scenarios.

To convey a more concrete sense of the magnitude of the attenuation, Figure 3.2 gives contour plots of the attenuation factor (3.4) as a function of SN and SP, for selected values of r and ρ. The top two panels in which $\rho = 0$ are identical to Figure 3.1. The lower panels then document the further attenuation that results when the second predictor Z is correlated with V. In particular, the $\rho = 0.5$ scenarios yield slightly more bias than the $\rho = 0$ scenarios, while the $\rho = 0.8$ scenarios evidence considerably more bias. Thus the warning in the previous section about mild misclassification having the potential to induce substantial bias is now even more dire: both rare exposure and correlation between the exposure and additional precisely measured covariate are catalysts for this situation.

3.3 Inferences on Odds-Ratios

We now turn attention to situations where the response variable is also binary. In particular, say both the response Y and predictor V are binary, with no other explanatory variables at play. Adopting a common epidemiological scenario, say that $Y = 1$ indicates the presence of a particular disease, while $V = 1$ indicates exposure to a putative risk. Let

$$\begin{aligned} r_0 &= Pr\{V = 1|Y = 0\}, \\ r_1 &= Pr\{V = 1|Y = 1\}, \end{aligned}$$

be the prevalences of exposure amongst disease-free and diseased subjects respectively. As is intuitive and well known, (r_0, r_1) can be estimated from either prospective or retrospective (case-control) studies. Typically, inferential interest would focus on

$$\Psi = \frac{r_1/(1-r_1)}{r_0/(1-r_0)},$$

the odds-ratio describing the association between exposure and disease.

Now say that V is subject to nondifferential misclassification as in the previous section, with the surrogate exposure variable again denoted by V^*.

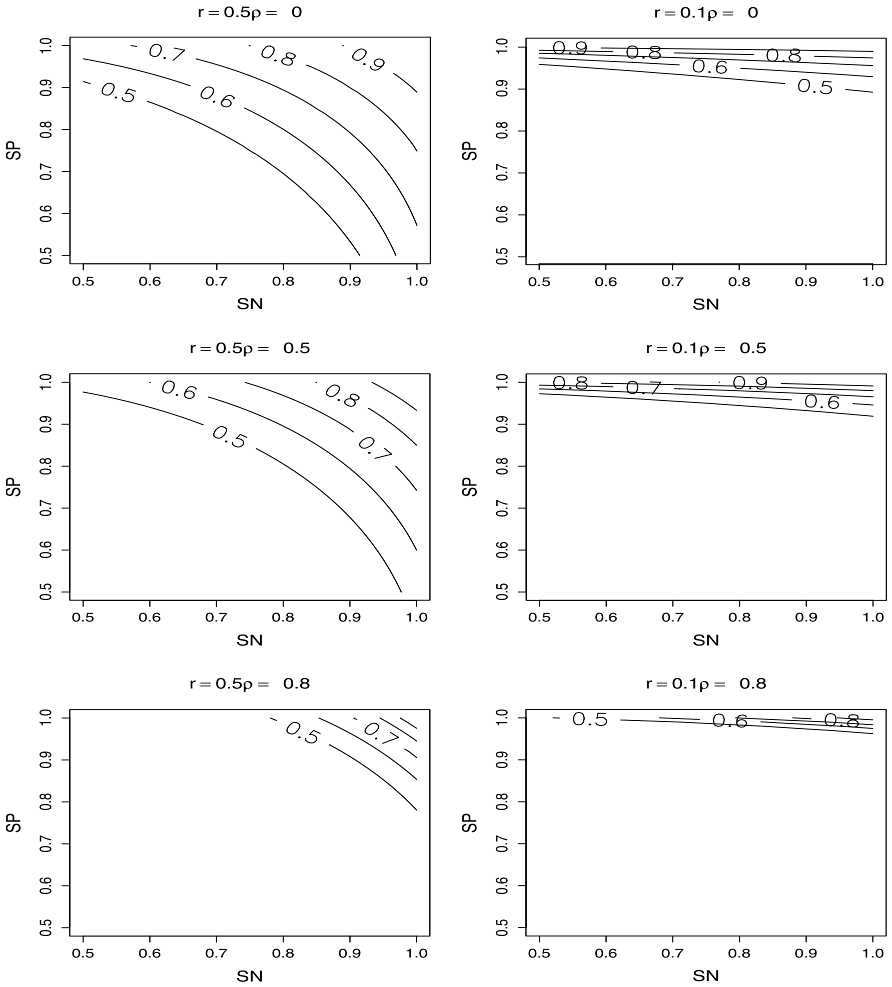

Figure 3.2 *Contours of the attenuation factor as a function of sensitivity and specificity, in the presence of an additional precisely measured explanatory variable. The two columns correspond to* $r = 0.5$ *and* $r = 0.1$*, while the three rows correspond to* $\rho = 0$, $\rho = 0.5$, *and* $\rho = 0.8$ *The contours correspond to an attenuation factor (3.4) of* 0.9, 0.8, 0.7, 0.6, *and* 0.5.

Without any adjustment for misclassification, the analyst would infer the exposure prevalences to be $(\tilde{r}_0, \tilde{r}_1)$ in the large-sample limit, where

$$\begin{aligned}
\tilde{r}_i &= Pr(V^* = 1|Y = i) \\
&= E\{Pr(V^* = 1|V, Y = i)|\, Y = i\} \\
&= E\{Pr(V^* = 1|V)|Y = i)\} \\
&= Pr(V^* = 1|V = 1)Pr(V = 1|Y = i) + \\
&\quad\; Pr(V^* = 1|V = 0)Pr(V = 0|Y = i) \\
&= a + br_i,
\end{aligned}$$

where, as before, $a = 1 - SP$ and $b = SN + SP - 1$. Consequently the large-sample limit of the estimated odds ratio would be

$$\tilde{\Psi} = \frac{\tilde{r}_1/(1-\tilde{r}_1)}{\tilde{r}_0/(1-\tilde{r}_0)}.$$

After some algebraic manipulations, the attenuation factor that arises from treating V^* as if it were V is

$$\frac{\tilde{\Psi}}{\Psi} = \frac{\{1 + d/r_1\}/\{1 + c/(1-r_1)\}}{\{1 + d/r_0\}/\{1 + c/(1-r_0)\}}, \qquad (3.5)$$

with

$$c = \frac{1 - SN}{SN + SP - 1}$$

and

$$d = \frac{1 - SP}{SN + SP - 1}.$$

If $SN + SP > 1$, which is a weak condition that might be interpreted as the exposure assessment being more accurate than a coin-flip guess, then it is clear that (3.5) really does correspond to attenuation towards an odds-ratio of unity. That is, either $1 \leq \tilde{\Psi} \leq \Psi$, or $\Psi \leq \tilde{\Psi} \leq 1$.

Though not typically expressed in this fashion, the attenuation factor (3.5), or slight variants thereof, have been derived and discussed by numerous authors; see, for instance, Bross (1954), Goldberg (1975), Barron (1977), and Copeland, Checkoway, McMichael and Holbrook (1977). These articles tend to focus on sample odds-ratios associated with two-by-two exposure-disease tables describing finite samples, rather than large-sample limits. In particular, the odds-ratio for the nominally observed table describing measured exposure and outcome is compared to the nominally unobserved table describing actual exposure and outcome. However, the difference between the finite sample and large-sample views in this context is stylistic at most.

To get a sense for the magnitude of the bias induced by misclassification, Figure 3.3 gives contour plots of the odds-ratio Ψ and the attenuation factor $\tilde{\Psi}/\Psi$, as functions of the prevalences (r_0, r_1). In particular, three such plots for three values of (SN, SP) are given for the attenuation factor. By comparing the plot for the odds-ratio to any of the plots for the attenuation factor, we

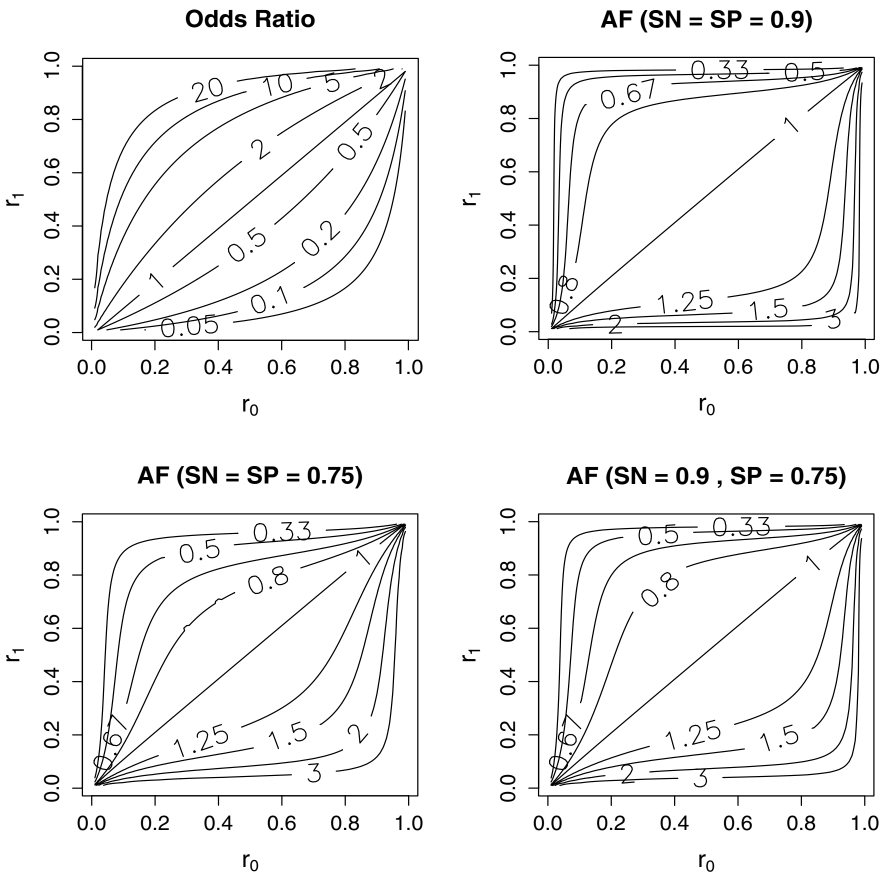

Figure 3.3 *Contours of the odds ratio and attenuation factor as a function of exposure prevalences. The upper-left panel gives contours of the odds-ratio* Ψ *as a function of exposure prevalences* (r_0, r_1)*. The other three panels give contours of the attenuation factor (3.5) for three combinations of sensitivity and specificity. The upper-right panel corresponds to* $(SN, SP) = (0.9, 0.9)$*, the lower-left panel corresponds to* $(SN, SP) = (0.75, 0.75)$*, and the lower right panel corresponds to* $(SN, SP) = (0.9, 0.75)$*.*

see the attenuation is much more pronounced when the true odds-ratio is far from one. In fact, if we consider a scenario where Ψ tends to infinity (because either r_0 tends to zero for fixed r_1 or r_1 tends to one for fixed r_0), we see that the attenuation factor tends to zero, with $\tilde{\Psi}$ tending to a constant.

Focussing on the mild misclassification scenario involving $SN = SP = 0.9$ in Figure 3.3, we see an attenuation factor of 0.8 when $\Psi > 1$ (or equivalently 1.25 when $\Psi < 1$) can arise without either prevalence or the odds ratio being

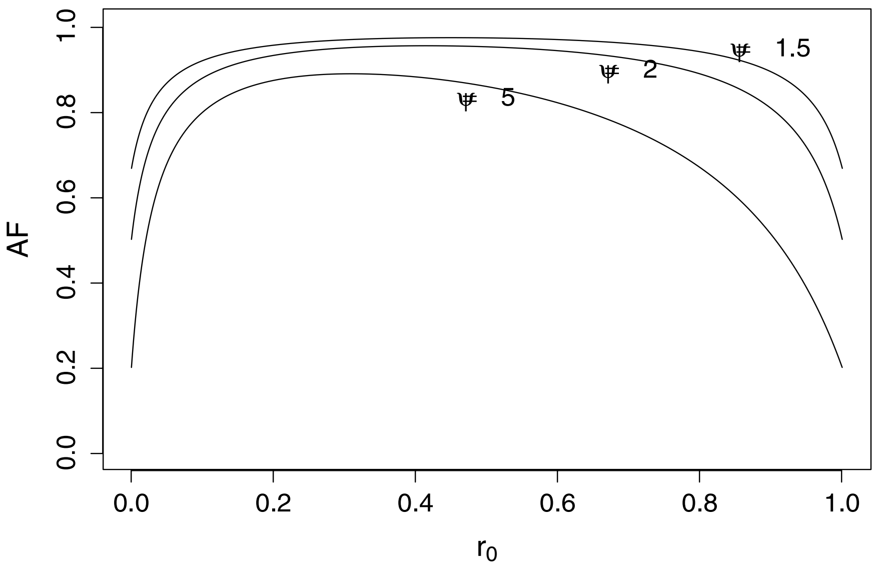

Figure 3.4 *Attenuation factor (3.5) as a function of the prevalence r_0, for some fixed values of the true odds-ratio Ψ. The misclassification is based on $SN = SP = 0.9$. Values $\Psi = 1.5$, $\Psi = 2$, and $\Psi = 5$ are considered.*

particularly extreme. Thus fairly mild misclassification can yield a sizeable inferential bias under plausible conditions. And of course even more substantial bias can arise in the other two scenarios involving larger misclassification probabilities.

To elaborate further, in the $SN = SP = 0.9$ scenario Figure 3.4 displays the attenuation factor (3.5) as a function of r_0, for some fixed values of the true underlying odds ratio Ψ. It is easy to verify from (3.5) that for fixed values of Ψ, SN, and SP, the attenuation factor tends to $1/\Psi$ as r_0 tends to zero or one. That is, perfect attenuation occurs in this limit, with $\tilde{\Psi}$ tending to 1 regardless of the value of Ψ or the extent of the misclassification as reflected by (SN, SP). Moreover, the figure suggests this limit is approached quite rapidly as r_0 goes to zero. Hence inference about the odds ratio can be exquisitely sensitive to even relatively modest misclassification in a rare exposure scenario. There is less attenuation for mid-range prevalences, although the effect is still sizeable. For instance, when $\Psi = 5$ the *largest* value of the attenuation factor is about 0.85 as r_0 varies.

3.4 Logistic Regression

Of course the relatively simple form of (3.5) describes attenuation when the analysis incorporates only the binary outcome variable and the misclassified binary exposure variable. In practice most investigations involve other explanatory variables as well. Thus we turn our attention to scenarios where an additional precisely-measured explanatory variable Z is recorded. For simplicity we consider the scenario where Z is also binary, though the treatment when Z is continuous would be virtually the same.

One approach to investigating the impact of misclassification in this scenario is to consider separate two-by-two (Y, V) tables for the two levels of Z, and then determine the resulting 'typical' (Y, V^*) tables stratified by Z. In particular, odds-ratios in the latter tables can be compared to those in the former. This approach is pursued by Greenland (1980), under various assumptions about the misclassification mechanism. Our related approach is to compare logistic regression of Y on (V^*, Z) to logistic regression of Y on (V, Z), in line with the investigation in Section 2.6 for a continuous exposure variable.

Again we consider nondifferential misclassification, expressed as $Pr(V^* = 1|V, Z, Y) = a + bV$, where $a = 1 - SP$ and $b = SN + SP - 1$. Also, say that the logistic regression model

$$\text{logit}Pr(Y = 1|V, Z) \quad = \quad \alpha_0 + \alpha_1 V + \alpha_2 Z \qquad (3.6)$$

is postulated for $Y|V, Z$. In particular, this involves the nontrivial assumption that an interaction effect is not manifested.

Whether or not (3.6) is correct, let $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$ be the large-sample limiting coefficients resulting from logistic regression of Y on $T = (1, V, Z)'$. Correspondingly, let α^* be the large-sample limiting coefficients resulting from logistic regression of Y on $T^* = (1, V^*, Z)'$. Adopting the approach detailed in Chapter 2, it is clear that the relationship between α^* and α is governed by the system of three equations

$$g(\alpha) \quad = \quad g^*(\alpha^*), \qquad (3.7)$$

where

$$g(\alpha) \quad = \quad E\left\{ \begin{pmatrix} 1 \\ a + bV \\ Z \end{pmatrix} \frac{1}{1 + \exp(-\alpha' T)} \right\},$$

and

$$g^*(\alpha^*) \quad = \quad E\left\{ \begin{pmatrix} 1 \\ V^* \\ Z \end{pmatrix} \frac{1}{1 + \exp(-\alpha^{*\prime} T^*)} \right\}.$$

Note, in particular, that the relationship between α and α^* is the same regardless of the actual response distribution of $Y|V, Z$. Of course similar observations were made in Chapter 2 concerning continuous measurement error.

Given (SN, SP) and the joint distribution of (V, Z), it is straightforward to

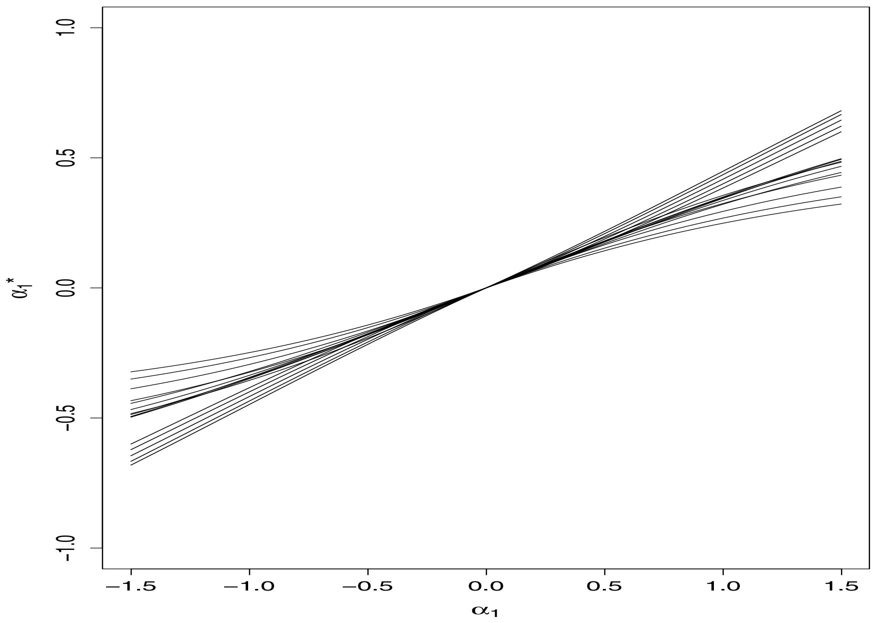

Figure 3.5 *The large-sample coefficient* α_1^* *for* V^* *as a function of the large-sample coefficient* α_1 *for* V*. The scenario involves* $Pr(V = 1) = Pr(Z = 1) = 0.25$*,* $Cor(V, Z) = 0.5$*,* $SN = 0.9$*,* $SP = 0.7$*. Each curve is* α_1^* *plotted against* α_1*, for fixed values of* α_0 *and* α_2*. These values are combinations of* $\alpha_0 = -2, 0, 2$ *and* $\alpha_1 = -1, -0.5, 0, 0.5, 1$.

evaluate $g(\alpha)$. Moreover, the joint distribution of (V^*, Z) is easily determined, so that g^* and its first derivative can be computed. Consequently, for a given α one can find the value of α^* solving (3.7) via the Newton-Raphson method.

Of course for linear models with either continuous nondifferential measurement error or binary nondifferential misclassification we have seen the effect on the coefficient of interest to be the same multiplicative attenuation regardless of the underlying values of the coefficients. For instance, the attenuation factor (α_1^*/α_1) given in (3.4) does not depend on α. Based on numerical investigation we find this to be *approximately* true in the present situation. To illustrate, consider a scenario under which $Pr(V = 1) = Pr(Z = 1) = 0.25$, $Cor(V, Z) = 0.5$, $SN = 0.9$, $SP = 0.7$. For a variety of values of α_0 and α_2 we plot α_1^* as a function of α_1 in Figure 3.5. We see that the relationship passes through the origin and is quite close to linear, with some modest curvature for large values of α_1. We also see that the slope of the near-linear relationship varies only slightly with α_0 and α_2. Thus the attenuation factor only varies slightly with $\alpha = (\alpha_0, \alpha_1, \alpha_2)'$.

In light of the insensitivity of the attenuation factor to the underlying co-

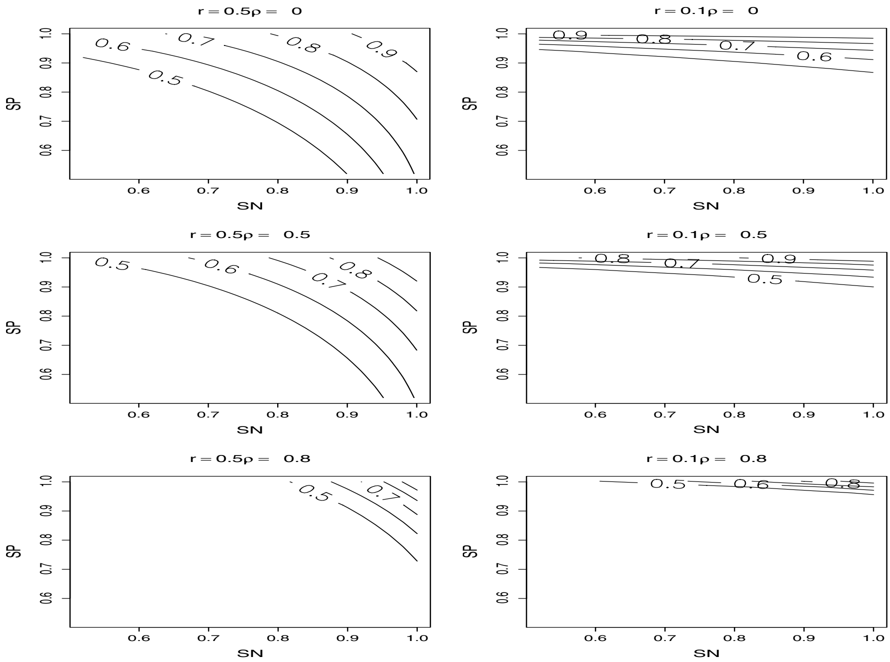

Figure 3.6 *Contours of the attenuation factor* (α_1^*/α_1) *as a function of sensitivity and specificity under logistic regression. The columns correspond to values of* 0.5 *and* 0.1 *for* $q = Pr(V = 1) = Pr(Z = 1)$. *The rows correspond to values of* 0, 0.5, *and* 0.8 *for* $\rho = Cor(V, Z)$. *Within each panel the contours correspond to attenuation factors of* 0.9, 0.8, 0.7, 0.6, *and* 0.5. *The underlying large-sample limiting coefficients are* $\alpha = (-1, 0.5, 0.25)'$.

efficients, we arbitrarily fix $\alpha = (-1, 0.5, 0.25)$ and give contour plots of the attenuation factor (α_1^*/α_1) as a function of SN and SP. This is done for combinations of $Pr(V = 1) = Pr(Z = 1) = 0.5$ or $Pr(V = 1) = Pr(Z = 1) = 0.1$, and $Cor(V, Z) = 0$, $Cor(V, Z) = 0.5$, or $Cor(V, Z) = 0.8$. These plots, which appear in 3.6, bear a striking similarity to those in Figure 3.2. That is, the bias induced by nondifferential misclassification in logistic regression is very similar to that induced in linear regression. A similar finding for nondifferential continuous measurement error was commented upon in Chapter 2.

3.5 Differential Misclassification

Of course the assumption that misclassification is nondifferential is not benign. In case-control analysis particularly there is concern about recall bias inducing misclassification which is actually differential. Say, for instance, that the binary exposure measurement is obtained from a questionnaire administered to subjects, perhaps with the potential exposure occurring years prior to the administration of the questionnaire. Under such circumstances a subject's disease status might have a subtle influence on his recall about the possible past exposure. While it is almost impossible to say anything general about the bias arising in such differential settings, one can give some example scenarios. We do so in the simple setting of having no additional precisely measured explanatory variables. In particular, the apparent exposure prevalences $\tilde{r}_0$ and $\tilde{r}_1$ given by (3.8) are modified to

$$\tilde{r}_i \quad = \quad r_i SN_i + (1 - r_i)(1 - SP_i), \qquad (3.8)$$

so that both the sensitivity and specificity of the exposure assessment may differ for controls ($i = 0$) and cases ($i = 1$). As before, of course, an analysis which ignores misclassification will be estimating the odds-ratio based on $(\tilde{r}_0, \tilde{r}_1)$ rather than the desired odds-ratio based on (r_0, r_1).

To investigate further we consider two scenarios. The first involves true exposure prevalences $(r_0, r_1) = (0.15, 0.22)$, while the second involves $(r_0, r_1) = (0.025, 0.0395)$. In both cases the true odds ratio is $\Psi = 1.6$, but clearly the latter scenario involves a much rarer exposure than the former. Each prevalence scenario in considered in tandem with three misclassification scenarios. The first is simply mild nondifferential misclassification with $SN_0 = SN_1 = 0.9$ and $SP_0 = SP_1 = 0.9$. The second, labelled as the 'denial' scenario, involves cases being slightly more reticent than controls to admit being exposed. Clearly this might arise if the exposure is of a self-inflicted nature, as cases may be wary of interpretations whereby their actions have contributed to their disease. Particularly, we take $SP_0 = SP_1 = 0.9$, but $SN_0 = 0.95$ and $SN_1 = 0.85$ for this scenario. On the other hand, scenarios where cases might be differentially tempted to falsely claim exposure, so as to explain or attribute blame for their plight, are also plausible. In this 'blame' scenario we take $SN_0 = SN_1 = 0.9$, but $SP_0 = 0.95$ and $SP_1 = 0.85$.

Limiting odds-ratios under these scenarios appear in Table 3.1. Clearly differential misclassification can either attenuate or accentuate the actual odds-ratio, with the denial misclassification leading to stronger attenuation than the nondifferential misclassification, but the blame misclassification leading to accentuation of the effect. It is interesting to note that for all three misclassification scenarios there is more bias when exposure is rarer. This is very much in keeping with the other scenarios considered in this chapter.

3.6 Polychotomous Variables

Another situation where the impact of misclassification is complex and hard to intuit is when a polychotomous (i.e., categorical with more than two levels) exposure variable is subject to misclassification. As pointed out by Dosemeci, Wacholder and Lubin (1990), even the impact of nondifferential misclassification can be quite unpredictable in this scenario. Birkett (1992) and Weinberg, Umbach and Greenland (1994) shed further light on this issue. Here we provide a very simple example to illustrate that nondifferential misclassification of a

prevalences	misclassification	apparent OR
$(0.15, 0.22)$	none	1.60
	nondifferential	1.35
	differential - “denial”	1.23
	differential - “blame”	2.13
$(0.025, 0.0395)$	none	1.60
	nondifferential	1.11
	differential - “denial”	1.08
	differential - “blame”	2.85

Table 3.1 *Bias in odds-ratio for case-control study with various misclassification scenarios. The nondifferential scenario involves* $SN = SP = 0.9$*. The “denial” scenario involves* $SP = 0.9$*, but* $SN_0 = 0.95$ *for controls and* $SN_1 = 0.85$ *for cases. The “blame” scenario involves* $SN = 0.9$*, but* $SP_0 = 0.95$ *for controls and* $SP_1 = 0.85$ *for cases.*

polychotomous explanatory variable does not necessarily lead to attenuation in the exposure-disease relationship.

Consider a continuous response variable Y and a polychotomous exposure variable V which takes on one of k values. For simplicity we use $1, \ldots, k$ to label these levels. Now say that nondifferential misclassification gives rise to V^* as a surrogate for V. Let $\gamma_j = E(Y|V = j)$ and $\gamma_j^* = E(Y|V^* = j)$. A straightforward probability calculation then yields

$$\gamma_j^* = \frac{\sum_{i=1}^k \gamma_i Pr(V^* = j|V = i)Pr(V = i)}{\sum_{i=1}^k Pr(V^* = j|V = i)Pr(V = i)}. \quad (3.9)$$

Numerical experimentation with (3.9) quickly indicates that the impact of misclassification is no longer very predictable or intuitively clear. In particular, one can construct plausible scenarios where the impact of nondifferential misclassification is not necessarily attenuation. For instance, say $k = 3$, with the ordered levels $V = 1$, $V = 2$, and $V = 3$ corresponding to no exposure, mild exposure, and substantial exposure respectively. Say that V is uniformly distributed across the three categories, while $V^*|V$ is distributed as

$$Pr(V^* = j|V = i) = \begin{pmatrix} 0.7 & 0.2 & 0.1 \\ 0.15 & 0.7 & 0.15 \\ 0.1 & 0.2 & 0.7 \end{pmatrix},$$

where i indexes rows and j indexes columns of the matrix. Clearly this distribution of $V^*|V$ has plausible characteristics. In particular, when $V = 1$ or $V = 3$ the chance of the classification being off by two categories is less than the chance of being off by just one category. Also, for each value of V the probability of correct classification is relatively large.

Now say that $E(Y|V)$ is given by $\gamma = (0, 0.25, 3)'$. Then calculation via (3.9) gives $\gamma^* = (0.355, 0.705, 2.25)'$. Note, in particular, that $\gamma_2^* - \gamma_1^* = 0.349$

is larger than $\gamma_2 - \gamma_1 = 0.25$. That is, the *apparent* effect of mild exposure relative to no exposure is larger than the actual effect. Clearly attenuation does not always result from nondifferential misclassification of polychotomous explanatory variables. As noted by Carroll (1997), situations where plausible measurement error leads to accentuation of effects are quite unusual and curious!

3.7 Summary

In some broad sense this chapter has conveyed a similar story to that of Chapter 2. Nondifferential misclassification of a binary explanatory variable yields attenuated estimates of associated effects, as does nondifferential measurement error in a continuous explanatory variable. While binary misclassification seems to be more damaging than continuous measurement error in general, they share some key features. In both cases a primary determinant of how bad the bias will be is the strength of correlation between the mismeasured explanatory variable and other precisely measured explanatory variables. Also, the notion of the mismeasurement bias not depending on the actual distribution of the response variable carries over from the previous chapter.

One feature not in play in Chapter 2 is the prevalence of exposure. Generally speaking the bias due to mismeasurement worsens as the proportion of subjects exposed gets close to zero (or close to one). In epidemiological contexts it is often quite natural to conduct studies with low exposure prevalences, so there is a clear need for methods which can adjust inferences to account for misclassification. Such methods are considered in Chapter 5, after methods which adjust for continuous measurement error are discussed in the next chapter.

3.8 Mathematical Details

To show that the attenuation factor (3.4) decreases with $|\rho|$, note that direct differentiation of (3.4) with respect to ρ^2 shows this to be the case provided that

$$b^2 r(1-r) \quad < \quad r^*(1-r^*).$$

Using $r^* = (1-SP) + (SN+SP-1)r$, the inequality can be rewritten as

$$(SN+SP-1)(SN-SP)r \quad < \quad SP(1-SP). \qquad (3.10)$$

We focus on the nonpathological case of better-than-chance assessment, i.e., $SN + SP > 1$, while also assuming that both SN and SP are strictly less than one. Then (3.10) holds trivially if $SN \leq SP$. If $SN > SP$ then the left-hand side of (3.10) as a function of SN and r is largest when $r = 1$ and SN tends to one, with equality obtained in the limit. Thus we have determined that the attenuation factor is decreasing in $|\rho|$, i.e., the attenuation worsens as correlation between the two explanatory variables increases.

Also we wish to establish the behaviour of the attenuation factor (3.4) as a function of prevalence r for fixed SN, SP, and ρ. Upon noting that (3.4) is a ratio of quadratic functions of r, direct differentiation reveals the derivative of (3.4) with respect to r has the same sign as

$$g(r) \quad = \quad (SN + SP - 1)(SN - SP)r^2 - 2SP(1 - SP)r + SP(1 - SP),$$

regardless of the value of ρ. It is straightforward to check that $g()$ is decreasing and must cross zero on $(0, 1)$. If $SN = SP$ then clearly $r = 1/2$ is the solution to $g(r) = 0$. If $SN \neq SP$ then the quadratic formula gives the solution as

$$r \quad = \quad \frac{SP(1 - SP) - \{SP(1 - SP)SN(1 - SN)\}^{1/2}}{SP(1 - SP) - SN(1 - SN)},$$

as desired, where we have used the algebraic fact that $(SN + SP - 1)(SN - SP) = SP(1 - SP) - SN(1 - SN)$ in obtaining this expression.