When Do Latent Class Models Overstate Accuracy for Diagnostic and Other Classifiers in the Absence of a Gold Standard?

Bruce D. Spencer

1Department of Statistics and Institute of Policy Research, Northwestern University, Evanston Illinois 60208, U.S.A.

email: bspencer@northwestern.edu

Summary. Latent class models are increasingly used to assess the accuracy of medical diagnostic tests and other classifications when no gold standard is available and the true state is unknown. When the latent class is treated as the true class, the latent class models provide measures of components of accuracy including specificity and sensitivity and their complements, type I and type II error rates. The error rates according to the latent class model differ from the true error rates, however, and empirical comparisons with a gold standard suggest the true error rates often are larger. We investigate conditions under which the true type I and type II error rates are larger than those provided by the latent class models. Results from Uebersax (1988, Psychological Bulletin 104, 405–416) are extended to accommodate random effects and covariates affecting the responses. The results are important for interpreting the results of latent class analyses. An error decomposition is presented that incorporates an error component from invalidity of the latent class model.

Key words: Classification error; Diagnostic test; Jury; Latent class model; Sensitivity; Specificity; Validity.

1. Introduction

Medical diagnostic tests indicate whether a person has a disease, court trials indicate whether a defendant is guilty, and surveys report whether a respondent favors a candidate. Such procedures classify individuals with regard to a binary state and will be called binary classifiers. A substantial literature is emerging on how to assess the accuracy of the classifications when the truth is unknown and when a gold standard (close approximation treated as truth) is not available. The methods for assessing accuracy without a gold standard rely on latent class models, which posit relationships between the observable or manifest classification $Y$ and the unobservable or latent classification $U$. When the latent class is used as if it were the true class, $V$, estimates of accuracy can be obtained. The error rates defined according to the latent class differ from the true error rates, however. As will be discussed, empirical comparisons with a gold standard suggest the true error rates are larger. This is important, because it is the true rates that matter in practice—the rates from the latent class model are of interest only as approximations to the true rates. We investigate sufficient conditions for the latent class models to overstate accuracy, in the sense that the true type I and type II error rates are as large or larger than those provided by the latent class models.

We denote the classes by 0 or 1, representing, respectively, absence or presence of the condition of interest (e.g., disease, commission of act, attitude), and denote the probabilities of type I and type II error according to the latent class by $\alpha(U) = \operatorname{pr}(Y = 1 | U = 0)$ and $\beta(U) = \operatorname{pr}(Y = 0 | U = 1)$, respectively. The error probabilities according to the true class are $\alpha(V) = \operatorname{pr}(Y = 1 | V = 0)$ and $\beta(V) = \operatorname{pr}(Y = 0 | V = 1)$. As mentioned above, studies comparing latent class model estimates of $\alpha(U)$ and $\beta(U)$ to estimates of $\alpha(V)$ and $\beta(V)$ obtained using a gold standard frequently found that the estimates of $\alpha(U)$ and $\beta(U)$ are smaller than the corresponding estimates of $\alpha(V)$ and $\beta(V)$. This has led to cautions in the medical literature that estimates of accuracy can be distorted when there is not an independent, blind comparison with a gold standard of diagnosis (Sackett and Haynes, 2002), while the fallibility of gold standards is acknowledged as well (Afdhal, 2004; Wells, 2004).

Uebersax (1988, p. 409) presented conditions and a theoretical explanation why $\alpha(U) \leq \alpha(V)$ and $\beta(U) \leq \beta(V)$ for homogeneous latent class models, which are models where, for each $u = 0, 1$,

\[
\operatorname{pr}(Y = 0 | U = u) \text { is constant for all cases.} \tag{1}
\]

We extend those results by finding conditions under which $\alpha(U) \leq \alpha(V)$ and $\beta(U) \leq \beta(V)$ for heterogeneous latent class models, in which $\operatorname{pr}(Y = 1 | U = u)$ varies across cases according to random effects and covariates (Section 3). These inequalities are important for understanding the latent class model estimates, say $\hat{\alpha}$ and $\hat{\beta}$, of $\alpha(V)$ and $\beta(V)$. The total error in the estimate of type I or type II error, $\hat{\alpha} - \alpha(V)$ or $\hat{\beta} - \beta(V)$, decomposes into the sum of statistical error, $\hat{\alpha} - \alpha(U)$ or $\hat{\beta} - \beta(U)$, and invalidity error, $\alpha(U) - \alpha(V)$ or $\beta(U) - \beta(V)$, respectively.

Statistical error is familiar concept. If the estimates $\hat{\alpha}$ and $\hat{\beta}$ have limiting values $\bar{\alpha}$ and $\bar{\beta}$ as sample size increases, the statistical error can be decomposed into the sum of sampling error, $\bar{\alpha} - \alpha(U)$ or $\bar{\beta} - \beta(U)$, and nonsampling error, $\alpha(U) - \alpha(V)$ or $\beta(U) - \beta(V)$. Sampling error arises because only a subset of cases
have been studied, and it may have random and nonrandom components. Nonsampling error arises from misspecification of the estimation model, problems in the sampling frame used for selecting cases, errors in the reporting, recording, and processing of data, etc.

Invalidity error can have an important effect on applications of statistics. Knowing that the invalidity error is less than or equal to zero (as is often the case) provides a useful caution when making inferences and decisions based on latent class model estimates. The measure of invalidity error proposed here differs from the correlation-based measures commonly used in psychometrics, which do not appear well suited for the medical and other kinds of classification problems considered in this article (Section 4).

For recent reviews and contributions to the medical literature on evaluating accuracy of diagnostic tests without a gold standard, see Hui and Zhou (1998), Xu and Craig (2009), Pepe and Janes (2007), Albert and Dodd (2004), and Bronsvoort et al. (2010). Gastwirth and Sinclair (1998) and Spencer (2007) provide estimates of the accuracy of verdicts in criminal cases, where the probability of an erroneous acquittal or erroneous conviction is of interest but the true status of criminals is unknown. Biemer (2011) presents a book length treatment of assessment of the accuracy of responses to survey questions without a gold standard.

Our sufficient conditions for the latent class models to overstate accuracy are of two types. If the latent class model is misspecified, then the estimates based on the model could be so biased that the bias dominates the invalidity error. One set of conditions, then, requires that the model assumptions are correct and, in addition, that certain conditional independence assumptions hold. The conditional independence assumptions are discussed in Section 3.2, and they depend on specific aspects of the study design. The other conditions involve inequalities that can be tested from the data. It is important to know when the latent class models overstate accuracy so that decisions based on the results of the diagnostic tests are not compromised by overconfidence. This is particularly important for medical decisions made about the treatment of patients. It is also important for public health decisions about cost effectiveness of medical screening when screening accuracy is estimated from latent class models without a gold standard.

2. Latent Class Models
2.1 Model Formulation
Let \( J \) denote the number of binary classifiers and let \( Y = (Y_1, \ldots, Y_J)^T \) denote the vector of binary classifications for the case, where \( Y_j \) takes the values 1 or 0 to indicate the presence or absence, respectively, of the condition of interest according to classifier \( j = 1, \ldots, J \). For the most part, it will be convenient and sufficient to focus on a single case at a time. Because error rates on the whole are averages of case-specific error rates, inequalities for the latter thus apply to the former.

Denote the true status of the case by \( V \), so that \( V = 1 \) or 0 if the condition truly is present or absent.

The true accuracy of classifier \( j \) is defined in terms of \( \text{pr}(Y_j = v|V = v) \), and if this probability varies across cases then its average over the population will be used. In many applications it will be more appropriate to think of the accuracy of a classifier type than of a unique classifier, for example in diagnostic tests we may consider the accuracy of given type of test rather than the lab or people involved in carrying out the test, and in court cases we may consider the accuracy of juries as a group rather than the accuracy of a single jury. The true probability of a type I error is denoted by \( \alpha_j^{(V)} = \text{pr}(Y_j = 1|V = 0) \), and the true specificity is \( 1 - \alpha_j^{(V)} \). The true probability of Type II error is \( \beta_j^{(V)} = \text{pr}(Y_j = 0|V = 1) \), and the true sensitivity is \( 1 - \beta_j^{(V)} \). The marginal distribution of \( V \) is denoted by \( \pi_v = \text{pr}(V = v) \).

When it is not possible to ascertain the true status \( V \), latent class models often are used to assess accuracy. For any random variable \( U \) taking values 0 or 1, the marginal probability distribution of \( Y \) may be expressed as \( \text{pr}(Y = y) = \sum_u \text{pr}(Y = y|U = u) \pi_u \), with \( \pi_u = \text{pr}(U = u) \). In application, the latent variable \( U \) is not directly observed, in contrast to the observed or manifest variables \( Y \). We have a latent class model if two kinds of conditions hold (Clogg, 1995, p. 317). One condition is that the probability distribution of \( Y \), perhaps conditional on observed covariates, is the same for all cases in the same latent class. If there are no covariates, this condition is (1). The second condition is a form of local independence, a simple example being

\[
\text{pr}(Y = y) = \sum_u \text{pr}(U = u) \prod_{j=1}^J \text{pr}(Y_j = y_j|U = u). \tag{2}
\]

Although assumptions such as (2) are important for fitting latent class models, they are not directly needed for the results derived in this article, which address the properties of the latent class being estimated rather than the sampling properties of the estimators. If (2) is believed to be false, then a model allowing for heterogeneity should be used; see Section 3. If (2) is assumed to hold when in fact it does not, or if the model is misspecified in other ways, the estimates will be biased.

In many applications, the estimates of accuracy are misdirected at error rates defined in terms of the latent class \( U \) instead of the true class \( V \), i.e., they are directed at \( \alpha_j^{(U)} = \text{pr}(Y_j = 1|U = 0) \) and \( \beta_j^{(U)} = \text{pr}(Y_j = 0|U = 1) \) instead of \( \alpha_j^{(V)} \) and \( \beta_j^{(V)} \); (Uebersax, 1988, 409ff; Garrett, Eaton, and Zeger, 2002, p. 1291; Bertrand et al., 2005, 697ff). Sections 2.2–2.3 below discuss empirical analyses of the accuracy of medical diagnostic tests based on latent class models with and without a gold standard. Section 2.4 compares estimates of response error for a survey questionnaire based on a latent class model versus a gold standard. In each example the error estimates from the latent class model without a gold standard are as small or smaller than the error estimates based on the gold standard. The latent class model overstates accuracy respect to \( V \) if

\[
\alpha_j^{(V)} \geq \alpha_j^{(U)} , \beta_j^{(V)} \geq \beta_j^{(U)} . \tag{3}
\]

The notation is summarized in Table 1.

**Theorem 1** (Uebersax, 1988): If

\[
\alpha_j^{(V)} + \beta_j^{(V)} \leq 1 , \tag{4}
\]
When Do Latent Class Models Overstate Accuracy?

Table 1

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_j )</td>
<td>Manifest classification of the case by classifier ( j )</td>
</tr>
<tr>
<td>( U )</td>
<td>Latent class for the case</td>
</tr>
<tr>
<td>( V )</td>
<td>True or correct class for the case</td>
</tr>
<tr>
<td>( \alpha_j^{(U)} )</td>
<td>( \text{pr}(Y_j = 1</td>
</tr>
<tr>
<td>( \beta_j^{(V)} )</td>
<td>( \text{pr}(Y_j = 0</td>
</tr>
<tr>
<td>( \pi_U )</td>
<td>( \text{pr}(U = u) )</td>
</tr>
<tr>
<td>( \pi_V )</td>
<td>( \text{pr}(V = v) )</td>
</tr>
<tr>
<td>( \pi_{UV} )</td>
<td>( \text{pr}(U = u, V = v) )</td>
</tr>
<tr>
<td>( \pi_{YjV} )</td>
<td>( \text{pr}(Y = j, U = u, V = v) )</td>
</tr>
</tbody>
</table>

and

\[
\text{pr}(Y | U, V) = \text{pr}(Y | U), \tag{5}
\]

then (3) holds.

Detailed proofs may be found in Web Appendix A.

We can expect (4) to hold if the classifiers are not too inaccurate and \( U \) is not too poor an approximation to the true latent class, \( V \). To see why, note that a classifier based purely on randomization and no information would have \( \alpha_j^{(U)} + \beta_j^{(V)} = \alpha_j^{(U)} + \beta_j^{(V)} = 1 \); for a more informative classifier we will have \( \alpha_j^{(V)} + \beta_j^{(V)} < 1 \) and, if \( U \) is not too different than \( V \), then we will have \( \alpha_j^{(V)} + \beta_j^{(V)} \leq 1 \) as well. Regardless, after the latent class model is fitted and estimates of \( \alpha_j^{(V)} \) and \( \beta_j^{(V)} \) are at hand, one can check whether (4) holds. Condition (5) is implied by (1), which is an assumption underlying the fitting of a homogenous latent class model. Thus, if the model assumptions hold, so does (5). Theorem 1 does not require that the local independence condition (2) holds. Condition (4) is equivalent to the property of latent monotonicity (Holland and Rosenbaum, 1986, p. 1525), specifically that \( \text{pr}(Y_j > y | U = u) \) is nondecreasing in \( u \) for each \( y \), or in the current context of binary variables, \( \text{pr}(Y_j = u | U = u) \geq \text{pr}(Y_j = u | U = \bar{u}) \) for each \( u \). Although the inequalities in (3) do not require that a latent monotonicity property holds between \( Y_j \) and \( V \), the inequalities are tighter (which is desirable) if the property does hold. Given (4) and (5), a sufficient condition for latent monotonicity between \( Y_j \) and \( V \) is latent monotonicity between \( U \) and \( V \).

2.2 Example of Diagnosis of Hearing Impairment

Pepe and Janes (2007, p. 479) present data on three tests for hearing impairment applied to 666 individuals, and calculate maximum likelihood estimates of \( \pi_U^{(1)} \), \( \alpha_j^{(V)} \), and \( \beta_j^{(V)} \), \( j = 1, \ldots, 3 \), with state 1 corresponding to disease and 0 to no disease. In addition, a gold standard was present, yielding estimates of \( \hat{\alpha}_j^{(V)} \) and \( \hat{\beta}_j^{(V)} \). The estimates (denoted with \( \hat{\cdot} \)) of the type I and type II error rates are shown in Table 2. The estimates of type I error and type II based on the latent class model are much lower than the estimates according to the gold standard, in accordance with (3). In addition, \( \hat{\pi}_1^{(U)} = 42\% \) and \( \hat{\pi}_U^{(V)} = 54\% \).

2.3 Example of Diagnosis of Chlamydia

Hadgu and Qu (1998) and Qu and Hadgu (1998) compared six diagnostic tests for the sexually transmitted disease chlamydia, including cell culturing and five nonculture tests. They note that cell culturing, the most expensive, time consuming, and technically demanding test, is believed to have a nearly zero type I error rate but a positive type II error rate. They compared the six tests as carried out on specimens collected by six sequential swabs from the endocervix of 4583 women. The cell culture test received the first swab and the assignments of tests to other swabs occurred in random order. Random effects models were fitted when the cell culture was treated as a gold standard (so that \( V \) was assumed known) and when the cell culture test was treated as just another classification subject to both types of error. For the five nonculture tests the type I error estimates satisfied \( \hat{\alpha}_j^{(V)} \leq \hat{\alpha}_j^{(V)} \leq 0.005 \). The preferred estimates of type II error rates for the nonculture tests ranged from about 0.6 to 0.8, and although sometimes \( \hat{\beta}_j^{(V)} > \hat{\beta}_j^{(V)} \), the difference then was small \( \hat{\beta}_j^{(V)} - \hat{\beta}_j^{(V)} < 0.008 \). Both sets of comparisons are subject to sampling error.

2.4 Example of Response Errors to Questionnaires

Kreuter, Yan, and Tourangeau (2008) compared the results of a latent class analysis with a gold standard analysis in order to identify survey questions that would have large response error. Three alternative survey questions were given in 2005 to individuals who, according to official records, had received undergraduate degrees from the University of Maryland between 1989 and 2002: (1) Did you ever receive a grade of for a class? (2) Did you ever receive an unsatisfactory or failing grade? (3) What was the worst grade you ever received in a course as an undergraduate at the University of Maryland? The latent class of interest was whether the student had ever received a grade of D or F as an undergraduate at the University of Maryland, and the true classification was ascertainable from the official records held by the University Registrar. The latent class model was (2) with \( J = 3 \) and with \( Y_j = 1 \) if the individual indicated having received a failing grade in response to question \( j \) and 0 otherwise. Data were obtained from 954 individuals and statistics for the mean of \( Y_j \) and maximum likelihood estimates (denoted with \( \cdot \)) of the type I and type II error rates were calculated under the assumption of conditional independence; see Table 3. Observe that (4) holds, and (3) holds for estimated error probabilities, except that \( \hat{\delta}_3^{(Y)} - \hat{\delta}_3^{(Y)} = 0.03 > 0 \). The difference is so small,
however, that it could just reflect sampling error. The apparent partial failure of (3) could also arise from model misspecification (Biemer, 2011).

2.5 Example of Hui–Walter Model for Accuracy of Trial Verdicts

A criminal trial by jury is conducted to establish the guilt (V = 1) or nonguilt (V = 0) of a defendant. The true guilt or innocence of the defendant typically is not known, and indeed an important issue for understanding the type I and type II error rates concerns the nature of the true state of the defendant. Consider for example a trial of a defendant who committed the crime as charged but for which the evidence is insufficient. Is the correct verdict guilty or not guilty? From an evidentiary or procedural perspective, the correct verdict is not guilty because proof of the crime has not been demonstrated to the standards required by the law. From a material or factual perspective, the correct verdict is guilty because the defendant actually committed the crime (Laudan, 2006; Spencer, 2007).

A study by Kalven and Zeisel (1966) of more than 3500 jury trials in the 1950s across the United States yielded data on 411 trials for burglary and auto theft for which a verdict was reported by both the judge (j = 1) and the jury (j = 2). The data consist of a 2 × 2 table with counts of (Yj1, Yj2), which Gastwirth and Sinclair (1998) used to estimate the parameters in the model (2). The full model has six parameters, and to be able to estimate them the authors adopted a model of Hui–Walter (1980) and used the assumptions that \( \pi_{0j} \alpha_j^{(U)} \) and \( \beta_j^{(U)} \) would be the same for each type of crime. Gastwirth and Sinclair (1998, p. 63) assumed the verdicts were conditionally independent within and across cases given the latent class and they developed the following estimates (with estimated standard error in parentheses): \( \alpha_1^{(U)} = 0.000 (0.399) \), \( \beta_1^{(U)} = 0.012 (0.015) \), \( \alpha_2^{(U)} = 0.009 (0.089) \), \( \beta_2^{(U)} = 0.192 (0.044) \), and hence \( \alpha_1^{(U)} + \beta_1^{(U)} = 0.012 (0.414) \) and \( \alpha_2^{(U)} + \beta_2^{(U)} = 0.201 (0.133) \). These provide a partial confirmation that (4) holds.

One may question whether the model is correctly specified, however, in particular the assumption (1) of constant conditional probabilities of error given U. For example, cases vary in difficulty according to strength of evidence. Failure of (1) causes two problems. First, the latent class model will be misspecified and the estimates will be biased. Second, it is possible that some of the variability in the conditional probabilities is related to V, in which case (5) would fail and Theorem 1 would not imply (3). Section 3 will consider heterogeneity in the error probabilities.

2.6 Extent of Error

The extent of an error refers to the number of such errors as a proportion of all cases. The extent of type I error according to the latent class model equals the average (across cases) of \( \pi_{0j} \alpha_j^{(U)} \), and the true extent equals the average of \( \pi_{0j} \alpha_j^{(U)} \). For type II errors the extent is the average across cases of \( (1 - \pi_{0j}) \beta_j^{(U)} \) or \( (1 - \pi_{0j}) \beta_j^{(U)} \), respectively. If (3) holds then the extent of either Type I or Type II error or both must be understated by the latent class model, because we must have either \( \pi_{0j} \leq \pi_{0j} \) or \( (1 - \pi_{0j}) \leq (1 - \pi_{0j}) \); the former implies that the extent of type I error is understated, and the latter implies that the extent of type II error is understated. Whether the extent of Type I error and the extent of Type II error are both understated will depend on the composition of the population being assessed. If (3) holds then a sufficient condition for the extent of type I and the extent of type II errors to both be understated by the latent class model is that the latent class and the true class have the same marginal distributions, i.e., \( \pi_{0j}^{(U)} = \pi_{0j} \). In both the hearing impairment diagnosis example in Section 2.2 and the response error example of Section 2.4, the estimated extent of each type I error and of each type II error was smaller according to the latent class model than the gold standard, with the sole exception of type I errors for classifier 3 in the response error example, where the estimates were extremely close.

The extents of type I and type II errors are important for public policy purposes. For example, from a public health standpoint the optimal screening rate for a disease depends on the extents and relative costs of type I and type II errors. In criminal justice there has long been concern with the ratio n of the extent of type II error to the extent of type I error, specifically that it is worse to convict one innocent person than to acquit n guilty people. The U.S. Supreme Court has interpreted the U.S. Constitution as stipulating that n be much greater than 1; In re Winship, 397 U.S. 358 (1970). Specific prescriptions for n have ranged from 2 (Voltaire) to 1000 (Maimonides), if not more widely (Volokh, 1997), although the best-known value may be Blackstone’s choice of n = 10: it is better that ten guilty persons escape than that one innocent suffer (Blackstone, 1825, 358). Unfortunately, even if the inequalities in (3) are strict, and even if the extents of error are understated by the latent class model, we do not know the implications for the ratio of the extents.

3. Latent Class Models Allowing Heterogeneity and Dependence

3.1 Models without Covariates

More general latent class models have been developed to relax the homogeneity assumption (1) and the local independence condition (2). With regard to the latter, Vacek (1985) and Torrance-Rynard and Walter (1997) directly allowed for pairwise dependence between classifiers, \( \text{cov}(Y_j, Y_{j'})|U = u) \neq 0 \), and Expeland and Handelman (1989) and Yang and Becker (1997) modeled pairwise dependence via log-linear models. Although such dependence complicates the estimation of accuracy, by itself it does not affect the error probabilities \( \alpha_j^{(U)} \), \( \beta_j^{(U)} \), \( \alpha_j^{(U)} \), \( \beta_j^{(U)} \).

The homogeneity assumption (1) is unrealistic when the conditional probabilities of classification given the latent class U vary across cases. In the chlamydia diagnostics in Section 2.3, type II error rates were lower when the test swabs were applied earlier rather than later. For another example, Spencer (2007) found both type I and type II error
When Do Latent Class Models Overstate Accuracy?

probabilities in criminal trials to vary with the strength of evidence presented. To model such heterogeneity, first note that when error probabilities are homogeneous we may, without loss of generality, reexpress the conditional probability of classification for the case as \( \text{pr}(Y_j = 0 | U) = F(-a_j U + b_j) \), with \( F \) taken to be any strictly increasing function taking values on the unit interval. Given a choice of \( F \), the heterogeneity can be modeled parametrically in terms of a case-specific effect \( R \) such that \( \text{pr}(Y_j = 0 | R, U = 0) = F(b_j + c_j R) \) and \( \text{pr}(Y_j = 0 | R, U = 1) = F(-a_j + b_j + d_j R) \), where \( a_j, b_j, c_j, d_j \) are constant across cases. In random effects models (without covariates), the effects \( R \) are random variables distributed independently of the \( U \)’s according to some mixing distribution, the manifest variables \( Y_j \) are conditionally independent given the \( U \)’s and \( R \)’s, and some functional form is specified for \( \text{pr}(Y_j = 0 | U, R) \).

Qu, Tan, and Kutner (1996) analyzed a Gaussian random effects model with \( \text{pr}(Y_j = 0 | U, R) = F(-a_j U + b_j + c_j R + (d_j - c_j) R) U R \), where \( F \) denotes a strictly increasing cumulative distribution function and the \( R \)’s are distributed as Gaussian variates independent of each other and of the \( U \)’s. Albert and Dodd (2004) discuss alternative continuous mixing distributions for the \( R \)’s, as well as discrete mixing distributions as used by Albert, McShane and Shih (2001) and Espeland and Handelman (1989). Hall and Zhou (2003) discuss the theoretical identifiability of such models and Albert and Dodd (2004) discuss practical difficulties of model identification.

Corollary 1, below, implies that (3) still holds if random effects \( R \) are conditionally independent of \( V \) given \( U \), (4) holds, and (5) holds conditionally on the random effect.

3.2 Models with Covariates

In some applications, the probability distribution of the classification of a case depends on a vector of covariates, \( X \). Qu et al. (1996) and Hadgu and Qu (1998) generalized the Gaussian random effects model to allow the probabilities to depend on linear functions of covariates, namely,

\[ \text{pr}(Y_j = 0 | U, R, X) = F(-a_j U + b_j + c_j R + (d_j - c_j) R) U R + e_j^T X + (f_j - e_j^T) U X, \]

where \( X = (x_1, \ldots, x_K)^T \), \( K \) is the number of covariates, and \( e_j \) and \( f_j \) are \( K \times 1 \) vectors that do not vary across cases. Several conditions on independence will be useful. In the construction of latent class models with random effects and covariates, such as (6), it is typically assumed that any random effects \( R \) are independent of the \( U \)’s and the \( X \)’s:

\[ \text{pr}(R | U, V, X) = \text{pr}(R). \]

If the latent class model specifies that the distribution of \( Y \) depends only on \( U \), \( R \), and \( X \), and the specification is correct, then:

\[ \text{pr}(Y | U, V, R, X) = \text{pr}(Y | U, R, X). \]

Thus, conditions (7) and (8) typically are already assumed to hold as part of the latent class analysis. A modification of (5) asserts conditional independence of \( X \) and \( V \) given \( U \):

\[ \text{pr}(X | U, V) = \text{pr}(X | U). \]

If the covariates \( X \) are randomly assigned as part of a designed experiment, as in Section 2.3, then (9) will hold. If the assignment of \( X \) is not controlled, then some consideration will be necessary to consider whether the differences between \( U \) and \( V \) are related to \( X \); if they are, then (9) will not hold. Conditions (7) and (9) jointly imply

\[ \text{pr}(R, X | U, V) = \text{pr}(R, X | U), \]

and (10) implies (9). Of direct importance is that (8) and (10) jointly imply (5), leading to the following corollary to Theorem 1.

Corollary 1: If \( U, V, R, X, Y \) satisfy (4), (8), and (10), then (3) holds.

A variant on property (4) is a conditional version,

\[ \text{pr}(Y_j = 1 | U = 0, R = r, X = x) = \text{pr}(Y_j = 0 | U = 1, R = r, X = x) \leq 1 \text{ for all } r, x. \]

If Simpson’s paradox does not occur, then (11) implies (4), and if (8) and (10) hold then Corollary 1 applies. However, Simpson’s paradox is well known to occur on occasion in the biosciences and demography (e.g., Keyfitz, 1977, pp. 385–391; Julious and Mullee, 1994), and it could occur in a given application of latent class modeling. In any case, (11) can be checked once the latent class model has been fitted. If (11) holds but (4) does not, the following result can be applied.

Theorem 2: If \( U, V, R, X, Y \) satisfy (8), (10), and (11), then

\[ \alpha_j^{(U)} \geq \alpha_j^{(V)} + \min\{0, A_j\}, \quad \beta_j^{(V)} \geq \beta_j^{(U)} + \min\{0, B_j\}. \]

with \( A_j = \sum_r \text{pr}(Y_j = 1 | U = 0, R = r, X = x) g(r, x), B_j = - \sum_r \text{pr}(Y_j = 0 | U = 1, R = r, X = x) g(r, x), \) and \( g(r, x) = \text{pr}(R = r, X = x | U = 1) - \text{pr}(R = r, X = x | U = 0). \)

Note that \( A_j \) and \( B_j \) can be estimated from the data, as illustrated in Section 3.3, below, and the precision of those estimates can also be estimated from the data. Under the conditions of the theorem, if \( A_j \geq 0, B_j \geq 0 \), then (3) holds and the model overstates accuracy with respect to \( V \).

When Theorem 2 with \( A_j \geq 0, B_j \geq 0 \) or Theorem 1 applies, the latent class model understates the type I and type II error rates, possibly by large amounts, depending on the joint distribution of \( Y, U, V \). Table 4 shows the error rates and joint distribution \( \sigma_{uv}^{(V)} \) of \( U, V, \) for several distributions without random effects \( R \). For these distributions, the underestimation by the latent class model of type I and type II error rates can be quite large. The first row shows that for distribution \( D_1^a \), \( \alpha_j^{(U)} = 0.010 < \alpha_j^{(V)} = 0.941 \) and \( \beta_j^{(V)} = 0.010 < \beta_j^{(U)} = 0.059 \). Switching the labels of 0 and 1 throughout leads to the distribution \( D_2 \), whose type I error rate equals the type II error rate of \( D_1 \), and vice versa, as shown in row 2. In the third row, the error rates with respect to \( U \) and \( V \) are the same. The next three rows have intermediate results. For details about the distributions, note that \( D_1^a \) is given by (6) with \( a_j = 4.1084, b_j = c_j = 2.0542, d_j = 0, \) and \( f_j = -2.0542 \). Additional specifications are that \( Y_j \) is
conditio\ntively independent of $V$ given $U$ and $X$; $V$ is independent of the pair $(U, X)$; $X$ is distributed as Bernoulli (0.5); $U$ and $V$ each are Bernoulli (0.95); $X$ and $U$ are independent. It follows that $D_1$, the joint distribution $\pi_{YUV}$ of $Y, U, V$, is given by $\pi_{000} = 0.002475$, $\pi_{010} = 0.047025$, $\pi_{011} = 0.000475$, $\pi_{100} = 0.009024$, $\pi_{101} = 0.000025$, $\pi_{110} = 0.000475$, $\pi_{111} = 0.893476$. The distribution $D_{12}(\lambda)$ is obtained by picking $D_2$ with probability $\lambda$ and $D_3$ with probability $1 - \lambda$, and $D_{13}(\lambda)$ is constructed analogously. The distribution $D_{123}(\lambda)$ is an equal mixture of $D_{12}$ and $D_{13}$.

The last two rows of Table 4 reflect distributions $D_a$ and $D_b$ such that (9) and (10) fail to hold. Property (3) holds for $D_a$ but fails badly for $D_b$. To construct the distributions, consider an example with $X = Y$ and no random effect. For $D_a$ the joint probability distribution for $Y, U, V$, is given by $\pi_{000} = \pi_{010} = \pi_{011} = 0.25$, $\pi_{100} = \pi_{101} = \pi_{110} = \pi_{111} = 0$. Distribution $D_a$ obtained by switching $U$ and $V$ in $D_b$. The counterexample provided by $D_b$ demonstrates that when covariates and random effects are present, condition (10) must be checked. However, the counterexample is artificial in having the covariate identical with the measurement, and in many practical situations (3) holds even though condition (9) fails.

3.3 A Log-Linear Model for Accuracy of Verdicts
Spencer (2007) analyzed the accuracy of verdicts in a convenience sample of criminal trials in 2000–2001 in four jurisdictions in the United States: Los Angeles, Washington D.C., the Bronx, and Maricopa County, Arizona (which includes the city of Phoenix). Data were available for 271 trials, showing the verdict that the judge reported he or she would have issued had the trial been a bench trial ($Y_1$), the jury’s verdict ($Y_2$), a 3-point rating by the judge of the strength of evidence for conviction ($X_1$), and an analogous 3-point rating by the jury ($X_2$). The observed data come from a table with 36 cells such that each cell includes both values of the unobserved latent variable $U$. The observed 36 cell table was treated as a partially observed 72 cell table derived by splitting each of the 36 observed cells on $U$. The data were treated as arising from a multinomial distribution where the log of the probability of falling in any of the 72 cells was specified by an analysis of variance model involving main effects for $U, X_1, X_2, Y_1, Y_2$, and pairwise interactions between $U$ and $X_1, X_2, Y_1, Y_2$; between $X_1$ and $Y_1$; and between $X_2$ and $Y_2$. The maximum likelihood estimates of the error probabilities were $\hat{\alpha}_1^{(V)} = 0.37$ and $\hat{\beta}_1^{(V)} = 0.02$ for the judge and $\hat{\alpha}_2^{(V)} = 0.25$ and $\hat{\beta}_2^{(V)} = 0.14$ for the jury. The type I error rates are large, and one may wonder how they would change if the latent class represented true guilt or innocence $V$ as discussed in Section 2.5. The maximum likelihood estimates of $A_j$ and $B_j$ are, respectively, $A_1 = -0.18$ and $B_1 = 0.06$ for the judge ($j = 1$) and $A_2 = -0.003$ and $B_2 = 0.12$ for the jury ($j = 2$). Although conditions (8) and (9) are not directly testable, it is plausible that they hold. That is, the correct verdict $V$ would not affect either the conditional distribution of $Y$ given $X$ and $U$, or the conditional distribution of $X$ given $U$, because $X$ is determined by the evaluations by judges and juries of the strength of evidence. $Y$ represents judgments based on evaluations of the evidence, and $U$ is empirically determined by $X$ and $Y$ in the context of the model. Assuming, then, that the conditions of Theorem 2 hold, we infer that the jury’s $\hat{\alpha}_2^{(V)}$ is as large as $\alpha_2^{(V)}$ or not too much smaller (allowing for estimation error in $A_2$). Practically, then, if there is no appreciable bias in $\hat{\alpha}_2^{(V)}$ as an estimate of $\alpha_2^{(V)}$, we may take the jury’s estimated type I error rate from the latent class model, $\alpha_2^{(V)}$, as a downwardly biased estimate of $\alpha_2^{(V)}$, the error rate defined with respect to the true verdict $V$. For the judge, with $A_1 = -0.18$, we see that, because $V$ and $U$ may differ, the large estimated type I error rate ($\hat{\alpha}_1^{(V)} = 0.37$) could be an overstatement of $\alpha_2^{(V)}$ even if $\hat{\alpha}_1^{(V)} \leq \alpha_1^{(V)}$; yet, the possibility that it is an understatement cannot be excluded either. Turning to type II errors, the fact that $B_1 > 0$ suggests that $\beta_2^{(V)} > \beta_1^{(V)}$ for $j = 1, 2$. The reader should not generalize from the estimates, due to the nonrandom nature of the sample, among other considerations (Spencer, 2007, p. 315, 323).

4. Invalidity
The definition of invalidity proposed here is consistent with the perspective adopted in psychometrics, which emphasizes that validity refers to the interpretation of a test or diagnosis for a given use (Linn, 1983, 336ff). For the assessment of the accuracy of error rates in binary classifiers, the invalidity of the latent class constructs, namely, $\alpha^{(V)} - \alpha^{(V)}$ and $\beta^{(V)} - \beta^{(V)}$, depend on the true state $V$, which in turn depends on the use of the classifiers, e.g., how will the diagnosis of the disease be used? Although the perspectives are consistent, the measure of invalidity as formulated here differs from the correlation-based measures commonly encountered in psychometrics. The latter may be inappropriate for medical diagnostics and other binary classifiers discussed in this article.

The correlation-based measures are based on a decision-theoretic model for selecting individuals (e.g., for medical

<table>
<thead>
<tr>
<th>Distn.</th>
<th>$\alpha_j^{(V)}$</th>
<th>$\beta_j^{(V)}$</th>
<th>$\pi_{00}^{UV}$</th>
<th>$\pi_{01}^{UV}$</th>
<th>$\pi_{10}^{UV}$</th>
<th>$\pi_{11}^{UV}$</th>
<th>$\hat{\beta}_2^{(V)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_1$</td>
<td>1.0</td>
<td>94.1</td>
<td>1.0</td>
<td>5.9</td>
<td>0.25</td>
<td>4.75</td>
<td>4.75</td>
</tr>
<tr>
<td>$D_2$</td>
<td>1.0</td>
<td>5.9</td>
<td>1.0</td>
<td>94.1</td>
<td>90.25</td>
<td>4.75</td>
<td>4.75</td>
</tr>
<tr>
<td>$D_3$</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>5.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$D_{12}(\lambda)$</td>
<td>1.0</td>
<td>1.0 + 93.1$\lambda$</td>
<td>1.0</td>
<td>1.0 + 4.9$\lambda$</td>
<td>5.00 + 4.75$\lambda$</td>
<td>4.75$\lambda$</td>
<td>4.75$\lambda$</td>
</tr>
<tr>
<td>$D_{13}(\lambda)$</td>
<td>1.0</td>
<td>1.0 + 4.9$\lambda$</td>
<td>1.0</td>
<td>1.0 + 93.1$\lambda$</td>
<td>5.00 + 85.25$\lambda$</td>
<td>4.75$\lambda$</td>
<td>4.75$\lambda$</td>
</tr>
<tr>
<td>$D_{123}(\lambda)$</td>
<td>1.0</td>
<td>1.0 + 49.0$\lambda$</td>
<td>1.0</td>
<td>1.0 + 49.0$\lambda$</td>
<td>5.00 + 40.25$\lambda$</td>
<td>4.75$\lambda$</td>
<td>4.75$\lambda$</td>
</tr>
<tr>
<td>$D_a$</td>
<td>0.0</td>
<td>50.0</td>
<td>0.0</td>
<td>50.0</td>
<td>25.00</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td>$D_b$</td>
<td>50.0</td>
<td>0.0</td>
<td>50.0</td>
<td>0.0</td>
<td>25.00</td>
<td>25.00</td>
<td>25.00</td>
</tr>
</tbody>
</table>
When Do Latent Class Models Overstate Accuracy?

6. Supplementary Materials

Web Appendices referenced in Sections 2 and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org/.

ACKNOWLEDGEMENTS

The author thanks Wei Li, Wesley Johnson, and Shelby Haberman, and Joseph Gastwirth for helpful suggestions and comments.

REFERENCES


treatment) on the basis of a measurement \( Y \). Define the criterion \( Z \) to be the utility derived if the individual is selected, and define \( t(Y) = E(Z|Y) \). In the context of medical diagnoses, \( Z \) can be a measure of the value from accepting for treatment a person diagnosed with disease, and it will tend to be larger if the individual selected for treatment of a disease actually has the disease than if the individual does not have it. An individual is selected if \( t(Y) \) exceeds some threshold, which is determined so that a fixed fraction of individuals is to be selected. If the distribution of \( t(Y) \) belongs to a scale family, then the value of selecting when \( t(Y) \) exceeds the threshold, compared to selecting individuals at random, can be shown to be proportional to the correlation \( \rho_{Zt} \) between \( t(Y) \) and \( Z \) (Cochran, 1951; Cronbach and Gleser, 1957). This correlation is called the criterion validity, and in a sense it measures how closely the utility predicted from the diagnosis corresponds to the actual utility attained. For additional discussion see Web Appendix B.

In the context of latent class models used for estimating accuracy of binary classifiers, the justification for correlation-based measures is problematic. Consider, for example, a diagnostic test used to assign treatment or nontreatment for a disease. First, if the diagnostic test is used in a multistage procedure, where further testing may be done depending on the outcome of a diagnostic test, the value of the test is no longer simply proportional to the correlation \( \rho_{Zt} \) (Cronbach and Gleser, 1957, chapter 6). Second, it typically is unrealistic to suppose that a fixed fraction of individuals are to be selected on the basis of the diagnostic test. Third, the scenario underlying the correlation-based measures stipulates that only the sum of the utilities for the selected individuals matters, and hence the disutility for not treating a diseased person is irrelevant (Berkson, 1947). Fourth, the distribution of \( t(Y) \) might not be in a scale family. For these reasons, the use of correlation-based measures of validity should be approached with caution—although they may be appropriate for a limited set of personnel selection problems, they may be inappropriate for a wide variety of diagnostic and other binary classification problems.

5. Conclusions

When estimates of type I and type II error rates for binary classifiers are based on latent class models, the estimates may have a component of error attributable to the difference between the latent class and the true class of interest. The difference between error rates defined for the true class and the latent class implicit in the model is a type of invalidity and, under certain conditions slightly stronger than those underlying the latent class model, it can be quantified via estimable lower bounds on the difference between the type I or type II error rate for the true class and the corresponding error rate for the latent class as defined by the latent class model. Practical guidance is provided for situations when some of the conditions do not apply (Section 3.2). In many cases, the lower bound is zero, meaning that the invalidity in the latent class tends to reduce the estimates of type I and type II error rates. These results help explain the empirical findings that when latent class model estimates of type I and type II error rates are compared to those based on comparisons to gold standards, the latent class model error rates tend to be too small.


Received July 2010. Revised August 2011. Accepted August 2011.
Proof of Theorem 1

Uebersax (1988) noted that $\alpha_j(V) = \Pr(Y_j = 1|V = 0)$ is a weighted average of $\Pr(Y_j = 1|V = 0, U = 0)$ and $\Pr(Y_j = 1|V = 0, U = 1)$, and by (5) that equals the same weighted average of $\Pr(Y_j = 1|U = 0) = \alpha_j(U)$ and $\Pr(Y_j = 1|U = 1) = 1 - \beta_j(U)$. But $1 - \beta_j(U) \geq \alpha_j(U)$ by (4), and hence $\alpha_j(V) \geq \alpha_j(U)$. Similarly, $\beta_j(V) \geq \beta_j(U)$. Thus, (3) is established.

Proofs of 3 assertions in the last paragraph before Section 2.2

Proof of first assertion. To show that (4) is equivalent to latent monotonicity, suppose first that latent monotonicity holds, and notice that $1 - \beta_j(U) = \Pr(Y_j = 1|U = 1) \geq \Pr(Y_j = 1|U = 0) = \alpha_j(U)$, establishing (4). On the other hand, if (4) holds then

$$0 \leq 1 - \alpha_j(U) - \beta_j(U) = 1 - \Pr(Y_j = 1|U = 0) - \Pr(Y_j = 0|U = 1)$$

which implies $\Pr(Y_j = 0|U = 1) \leq \Pr(Y_j = 0|U = 0)$; similarly,

$$0 \leq 1 - \alpha_j(U) - \beta_j(U) = 1 - \Pr(Y_j = 0|U = 1) - \Pr(Y_j = 1|U = 0)$$

$$= \Pr(Y_j = 1|U = 1) - \Pr(Y_j = 1|U = 0),$$

implying $\Pr(Y_j = 1|U = 0) \leq \Pr(Y_j = 1|U = 1)$. So, $\Pr(Y_j = u|U = u) \geq \Pr(Y_j = u|U = 1 - u)$ for each $u$, and latent monotonicity is established.

Proof of second assertion. To see that the inequalities in (3) are tighter if the latent monotonicity property holds between $Y_j$ and $V$, suppose that (4), (5), and hence (3) hold but latent monotonicity

does not hold between \( Y_j \) and \( V \). Then latent monotonicity holds between \( Y_j \) and \( \tilde{V} = 1 - V \), and we have \( \beta_j^{(V)} = 1 - \Pr\{Y_j = 1|\tilde{V} = 0\} \geq 1 - \Pr\{Y_j = 1|\tilde{V} = 1\} = \beta_j^{(\tilde{V})} \geq \beta_j^{(U)} \); similarly \( \alpha_j^{(V)} \geq \alpha_j^{(\tilde{V})} \geq \alpha_j^{(U)} \).

**Proof of third assertion.** Finally, we show if (4) and (5) hold, latent monotonicity between \( U \) and \( V \) implies latent monotonicity between \( Y_j \) and \( V \). First observe that for the current context with binary variables there is latent monotonicity between \( Y_j \) and \( V \) if and only if \( \Pr\{Y_j = u|V = u\} - \Pr\{Y_j = u\} \geq 0 \). Notice that

\[
\Pr\{Y_j = u|V = u\} = \Pr\{Y_j = u|U = u, V = u\}\Pr\{U = u|V = u\} + \Pr\{Y_j = u|U = 1 - u, V = u\}\Pr\{U = 1 - u|V = u\} = \Pr\{Y_j = u|U = u\}\Pr\{U = u|V = u\} + \Pr\{Y_j = u|U = 1 - u\}[1 - \Pr\{U = u|V = u\}]
\]

and

\[
\Pr\{Y_j = u\} = \Pr\{Y_j = u|U = u\}\Pr\{U = u\} + \Pr\{Y_j = u|U = 1 - u\}\Pr\{U = 1 - u\} = \Pr\{Y_j = u|U = u\}\Pr\{U = u\} + \Pr\{Y_j = u|U = 1 - u\}[1 - \Pr\{U = u\}].
\]

Subtracting, we have

\[
\Pr\{Y_j = u|V = u\} - \Pr\{Y_j = u\} = \Pr\{Y_j = u|U = u\}\Pr\{U = u|V = u\} \geq \Pr\{Y_j = u|U = u\}\Pr\{U = u\} + \Pr\{Y_j = u|U = 1 - u\}[1 - \Pr\{U = u|V = u\}] = \Pr\{Y_j = u|U = u\}[\Pr\{U = u|V = u\} - \Pr\{U = u\}] + \Pr\{Y_j = u|U = 1 - u\}[\Pr\{U = u\} - \Pr\{U = u|V = u\}]
\]
\[
\begin{align*}
= & \left[ \Pr \{ Y_j = u | U = u \} - \Pr \{ Y_j = u | U = 1 - u \} \right] \\
& \times \left[ \Pr \{ U = u | V = u \} - \Pr \{ U = u \} \right] \\
\geq & \ 0.
\end{align*}
\]

Thus, there is latent monotonicity between \( Y_j \) and \( V \).

**Proof that (7) and (9) jointly imply (10)**

\[
\begin{align*}
\Pr \{ R, X | U, V \} &= \Pr \{ R | U, V, X \} \Pr \{ X | U, V \} \\
&= \Pr \{ R \} \Pr \{ X | U \} \\
&= \Pr \{ R | X, U \} \Pr \{ X | U \} \\
&= \Pr \{ R, X | U \}.
\end{align*}
\]

**Proof that (8) and (10) jointly imply (5)**

Let \( E_{R,X|U,V} \) denote expectation with respect to the conditional distribution of \( R, X \) given \( U, V \), and let \( E_{R,X|U} \) denote expectation with respect to the conditional distribution of \( R, X \) given \( U \).

\[
\begin{align*}
\Pr \{ Y | U, V \} &= E_{R,X|U,V} \Pr \{ Y | R, U, V, X \} \\
&= E_{R,X|U,V} \Pr \{ Y | R, U, X \} \\
&= E_{R,X|U} \Pr \{ Y | R, U, X \} \\
&= \Pr \{ Y | U \}.
\end{align*}
\]

**Proof of Theorem 2**

We use a simple conditioning argument, as in Uebersax (1988) and Bertrand et al. (2005). We will use the notation \( E_{R,X|V=0} \) to denote expectation with respect to the conditional joint distribution of \( R \) and \( X \) given \( V = 0 \). Observe first that

\[
\begin{align*}
\alpha_j^{(V)} &= \Pr \{ Y_j = 1 | V = 0 \} \\
&= E_{R,X|V=0} \left[ \Pr \{ Y_j = 1 | V = 0, R, X \} \right]
\end{align*}
\]
\[ \begin{align*}
&= E_{R,X|V=0} \left[ \sum_{u=0}^{1} \Pr\{Y_j = 1|U = u, V = 0, R, X\} \Pr\{U = u|V = 0, R, X\} \right] \\
&= E_{R,X|V=0} \left[ \sum_{u=0}^{1} \Pr\{Y_j = 1|U = u, R, X\} \Pr\{U = 0|V = 0, R, X\} \\
&\quad + \Pr\{Y_j = 1|U = 1, R, X\} \Pr\{U = 1|V = 0, R, X\} \right] \\
&\geq E_{R,X|V=0} \left[ \Pr\{Y_j = 1|U = 0, R, X\} \Pr\{U = 0|V = 0, R, X\} \\
&\quad + \Pr\{Y_j = 1|U = 0, R, X\} \Pr\{U = 1|V = 0, R, X\} \right] \\
&= E_{R,X|V=0} \left[ \Pr\{Y_j = 1|U = 0, R, X\} \right].
\end{align*} \]

In contrast, \( \alpha_j^{(U)} = E_{R,X|U=0} \left[ \Pr\{Y_j = 1|U = 0, R, X\} \right] \), and so we have

\[ \begin{align*}
\alpha_j^{(V)} - \alpha_j^{(U)} &\geq E_{R,X|V=0} \left[ \Pr\{Y_j = 1|U = 0, R, X\} \right] - E_{R,X|U=0} \left[ \Pr\{Y_j = 1|U = 0, R, X\} \right] \\
&= \sum_{r,x} \Pr\{Y_j = 1|U = 0, R = r, X = x\} \\
&\quad \times \left[ \Pr\{R = r, X = x|V = 0\} - \Pr\{R = r, X = x|U = 0\} \right]
\end{align*} \]

where for simplicity of exposition we treat the covariates as discrete. Notice that

\[ \begin{align*}
\Pr\{R = r, X = x|V = 0\} &= \Pr\{R = r, X = x|U = 0, V = 0\} \Pr\{U = 0|V = 0\} \\
&\quad + \Pr\{R = r, X = x|U = 1, V = 0\} \Pr\{P(U = 1|V = 0\} \\
&\overset{(10)}{=} \Pr\{R = r, X = x|U = 0\} \Pr\{U = 0|V = 0\} \\
&\quad + \Pr\{R = r, X = x|U = 1\} \Pr\{U = 1|V = 0}\)
\end{align*} \]

and so

\[ \begin{align*}
\Pr\{R = r, X = x|V = 0\} - \Pr\{R = r, X = x|U = 0\} \\
&= \Pr\{R = r, X = x|U = 0\} \left[ \Pr\{U = 0|V = 0\} - 1 \right] \\
&\quad + \Pr\{R = r, X = x|U = 1\} \Pr\{U = 1|V = 0\} \\
&= \Pr\{R = r, X = x|U = 0\} \left[ -\Pr\{U = 1|V = 0\} \right] \\
&\quad + \Pr\{R = r, X = x|U = 1\} \Pr\{U = 1|V = 0\}
\end{align*} \]
\[ \text{with } g(r, x) = \text{pr}\{R = r, X = x|U = 1\} - \text{pr}\{R = r, X = x|U = 0\}. \] It follows that
\[ \alpha_j^{(V)} - \alpha_j^{(U)} \geq \text{pr}\{U = 1|V = 0\}\]
\[ \times \sum_{r, x} \text{pr}\{Y_j = 1|U = 0, R = r, X = x\}g(r, x) \]
\[ = \text{pr}\{U = 1|V = 0\}A_j \]
and hence \( \alpha_j^{(V)} - \alpha_j^{(U)} \geq 0 \) if \( A_j \geq 0 \). If \( A_j < 0 \) then \( \alpha_j^{(V)} - \alpha_j^{(U)} \geq A_j \). Thus, \( \alpha_j^{(V)} \geq \alpha_j^{(U)} + \min\{0, A_j\} \). The condition for \( \beta_j^{(V)} \geq \beta_j^{(U)} \) follows similarly.

**Counterexample to Theorem 2: Failure of Property (3) When Condition (9) Fails**

Consider \( X = Y \). Suppose there is no random effect, so that (7) holds vacuously. Let the joint probability distribution for \( Y, U, V \), say \( D_0 \), be given by \( p_{000} = p_{010} = p_{101} = p_{111} = 0.25, p_{001} = p_{011} = p_{100} = p_{110} = 0 \), where \( p(y, u, v) \) denotes \( \text{pr}\{Y_j = y, U = u, V = v\} \). Condition (10) fails because \( \text{pr}\{X = 0|U = 1, V = 0\} = 1 \) but \( \text{pr}\{X = 0|U = 1, V = 1\} = 0 \). Condition (8) holds because \( Y \) is always equal to \( X \), and (4) holds as well. Observe that \( \beta_j^{(U)} = \alpha_j^{(U)} = 0.5 > 0 = \alpha_j^{(V)} = \beta_j^{(V)} \), so that the latent class model grossly overstates both the type I and the type II error rate and (3) fails to hold.

**Web Appendix B: Additional Discussion of Correlation-Based Validity Measures**

There are other correlation-based measures of validity in addition to criterion validity. Let \( \theta \) denote a (possibly unobserved) variable for the individual such that \( Z \) and \( E(Z|Y) \) are conditionally uncorrelated given \( \theta \). Define the construct \( \tau \) as \( \tau(\theta) = E\{t(Y)|\theta\} \). For example, it may be helpful to imagine \( Y \) and \( Z \) each equal to some function of the construct plus independent noise. The construct validity of the measure \( t \) is defined as the correlation between \( \tau \) and \( t(Y) \), say \( \rho_{\tau t} \). The criterion validity of the construct is defined as the correlation between \( Z \) and \( \tau \), say \( \rho_{Z\tau} \). The
criterion validity factors into the product of the criterion validity of the construct and the construct validity, $\rho_{Zt} = \rho_{Z} \rho_{t}$. The proof follows.

**Proof of Factorization of Correlation**

There is some joint probability distribution for $(Z, Y, \theta)$. We have by definition $t(Y) = E_{Z|Y}(Z|Y)$ and

$$
\tau(\theta) = E_{Y|\theta}(t(Y)|\theta)
= E_{Y|\theta}E_{Z|Y,\theta}(Z|Y, \theta)
= E_{Z|\theta}(Z|\theta).
$$

We also have

$$
E_{Z,Y|\theta}(t(Y)|\theta) = E_{Z,Y|\theta} \left[ E_{Z|Y,\theta}(Z|Y, \theta) \right]
= E_{Z|\theta}(Z|\theta)
= \tau(\theta).
$$

We have by assumption that

$Z$ and $t$ are conditionally uncorrelated given $\theta$.

It follows that

$$
E_{Z,Y,\theta}(Zt(Y)) = E_{\theta}E_{Z,Y|\theta}(Zt(Y)|\theta)
= E_{\theta} \left[ E_{Z,Y|\theta}(Z|\theta)E_{Z,Y|\theta}(t(Y)|\theta) \right]
= E_{\theta} \left[ E_{Z|\theta}(Z|\theta)E_{Z,Y|\theta}(t(Y)|\theta) \right]
= E_{\theta} \left[ \tau(\theta)^2 \right].
$$

Also,

$$
E_{Z,Y,\theta}(Z\tau(\theta)) = E_{Z,\theta}(Z\tau(\theta))
= E_{Z,\theta} \left[ Z E_{Z|\theta}(Z|\theta) \right].
$$
\[
\begin{align*}
&= E_\theta E_{Z|\theta} \left[ Z E_{Z|\theta}(Z|\theta) \right] \\
&= E_\theta \left[ E_{Z|\theta}(Z|\theta)^2 \right] \\
&= E_\theta \left[ \tau(\theta)^2 \right]. \\
\end{align*}
\]

Finally,
\[
E_{Z,Y,0}(\tau(\theta) t(Y)) = E_{Y,Z,0} \left[ E_{Y|\theta}(t(Y)|\theta) t(Y) \right] \\
= E_{Y,Z,0} \left[ E_{Z|\theta}(Z|\theta) t(Y) \right] \\
= E_\theta E_{Y,Z|\theta} \left[ E_{Z|\theta}(Z|\theta) t(Y) \right] \\
= E_\theta \left[ E_{YZ|\theta} E_{Z|\theta}(Z|\theta) E_{Y,Z|\theta} t(Y) \right] \\
= E_\theta \left[ \tau(\theta)^2 \right].
\]

Next, observe that \(EZ = Et(Y) = E\tau(\theta)\), which together with the previous results implies
\[
\begin{align*}
\rho_{Zt} &= \frac{E\tau^2 - EZ Et}{\sqrt{V(Z)V(t)}} = \frac{V(\tau)}{\sqrt{V(Z)V(t)}}, \\
\rho_{Z\tau} &= \frac{E\tau^2 - EZ E\tau}{\sqrt{V(Z)V(\tau)}} = \frac{\sqrt{V(\tau)}}{\sqrt{V(Z)}}, \\
\rho_{\tau t} &= \frac{E\tau^2 - E\tau Et}{\sqrt{V(\tau)V(t)}} = \frac{\sqrt{V(\tau)}}{\sqrt{V(t)}}.
\end{align*}
\]

It follows immediately that \(\rho_{Zt} = \rho_{Z\tau}\rho_{\tau t}\).