# Generating Predicted Class Probabilities from Latent Class Model

Phil Schumm

January 1, 2024

The purpose of this note is to demonstrate how to calculate predicted class probabilities from a fitted latent class model. To start, let $K$ be the number of classes and $\eta_j$ be the prior probability that a randomly chosen individual is in Class $j$ ($\sum_{j=1}^{K} \eta_j = 1$).

## Binary Indicators

Assume that the data consist of $p$ binary indicator (0/1) variables $x_i$ ($i = 1, 2, \ldots, p$), and denote the probability that $x_i = 1$ for an individual in Class $j$ as $\pi_{ij}$. For convenience, we'll refer to the vectors $(x_1, x_2, \ldots, x_p)$ and $(\pi_{1j}, \pi_{2j}, \ldots, \pi_{pj})$ as $\boldsymbol{x}$ and $\boldsymbol{\pi}_j$, respectively. The basic latent class model is

$$
\begin{aligned}
f(\boldsymbol{x}) &= \sum_{j=1}^{K} \eta_j P(\boldsymbol{x}|j, \boldsymbol{\pi}_j) && (1) \\
&= \sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} \pi_{ij}^{x_i} (1 - \pi_{ij})^{1-x_i} && (2)
\end{aligned}
$$

and we can use Bayes theorem to compute the predicted posterior probability of an individual with response vector $\boldsymbol{x}$ belonging to Class $j$ as

$$
\begin{aligned}
P(j|\boldsymbol{x}) &= P(j) P(\boldsymbol{x}|j, \hat{\boldsymbol{\pi}}_j) / \hat{f}(\boldsymbol{x}) && (3) \\
&= \hat{\eta}_j \prod_{i=1}^{p} \hat{\pi}_{ij}^{x_i} (1 - \hat{\pi}_{ij})^{1-x_i} / \hat{f}(\boldsymbol{x}) && (4)
\end{aligned}
$$

where $\hat{f}(\boldsymbol{x})$ is computed by substituting the estimates $\hat{\eta}_j$ and $\hat{\pi}_{ij}$ into equation 1. Note that both $\hat{\eta}_j$ and $\hat{\pi}_{ij}$ are typically generated by software for fitting latent class models, and are what we received from Jennifer Liu.

## Handling Missing Data

While equation 3 can be used to compute the posterior probability of class membership in cases where the values of all of the indicator variables are known, it is likely that in many clinical situations some of these will be unknown. Let $\boldsymbol{x}_{obs}$ be the vector of observed variables, and let $r_i$ be an indicator variable taking the value 1 when $x_i$ is observed and 0 otherwise. Under the assumption of conditional independence of variables given latent class membership (a fundamental part of the latent class model), we can write the distribution of the observed variables as

$$f(\boldsymbol{x}_{obs}) = \sum_{j=1}^{K} \eta_j \prod_{i=1}^{p} [\pi_{ij}^{x_i}(1 - \pi_{ij})^{1-x_i}]^{r_i}. \tag{5}$$

Given this, we may then compute the predicted posterior probability of an individual with observed response vector $\boldsymbol{x}_{obs}$ as

$$P(j|\boldsymbol{x}_{obs}) = \hat{\eta}_j \prod_{i=1}^{p} [\hat{\pi}_{ij}^{x_i}(1 - \hat{\pi}_{ij})^{1-x_i}]^{r_i} / \hat{f}(\boldsymbol{x}_{obs}). \tag{6}$$

As you can see, equations 5 and 6 are effectively the same as the corresponding equations above, simply ignoring the unobserved $x$ variables.

> ⚠ Equation 6 assumes that the missing data are missing at random (MAR); specifically, that the probability of an indicator variable being unobserved is not either a function of class membership or of the true (but unobserved) value of the variable. At least two plausible scenarios would violate this. One would be if a specific class of patients (e.g., Class 3—the cardiovascular class) were monitored more intensively such that their diagnostic profile is more complete (i.e., less likely to have missing data). Another would be if the information about conditions which the patient actually has is more complete than the information about conditions that the patient does not have (i.e., conditions with missing data are more likely to have a true (unobserved) value of 0). Thus, while incorporating missing data into the calculation may in certain cases yield more accurate results, it may not be adequate in these two possibly common scenarios. For this reason, I think that we should add a cautionary note that for best results, users should provide information on as many conditions as possible.

## Inclusion of Covariates

The model in equation 1 treats the population as homogenous in that any randomly selected individual has the same prior probability $\eta_j$ of being in Class $j$. However, our published results contradict this; for example, those aged 65–69 are more likely to be in Class 1 while those 80 or older are more likely to be in Class 3. Thus, if the tool is used in a population whose characteristics differ from those of the sample used to estimate the model (i.e., the Kaiser data), then the predictions will be systematically biased; for example, if it were used exclusively among patients aged 80 or above, it would systematically underestimate the probability of being in Class 3. We could address this by incorporating covariates into the model, which would not only reduce the bias when using the tool in a different population but would also increase the precision of its predictions.

Incorporating covariates would require re-estimating the model, though this could still be done easily in SAS (I'm assuming that the Kaiser folks would need to do it since we don't have access to the data). This would also require adjustments to the equations above, which I can provide if necessary.