Bayesian Models for Categorical Variables

Bayes Theorem and Bayesian Statistics

Recall that Bayes theorem is the basis for what is referred to as Bayesian statistics and is at the conceptual heart of many different applications. It is a way of conceptualizing a conditional probability for an event, such as the probability that A occurs given B has occurred, $P(A|B)$, as a function of the marginal probabilities of two events, $P(A)$ and $P(B)$, as well as the reverse conditional probability, $P(B|A)$.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

The example I used in the contingency chi-square handout as the probability that an individual would have a drug relapse (A) given that the person had completed a treatment center program (B). In more general Bayesian terms, we can call the conditional probability on the left side, $P(A|B)$, a *posterior probability* that is estimated based on a known, assumed, or stated prior probability (or often just "prior"), $P(A)$, multiplied by a likelihood, which is $P(B|A)$, in this simple case. This serves as a rudimentary example of more complex sorts of analyses in which we estimate the posterior probability of a parameter, given the data. By contrast, the traditional hypothesis testing approach is to estimate the likelihood of the data given some parameter value (e.g., what is the likelihood of the sample estimate given a null hypothesis value).

If we give this a little bit more general, formal statistical notation, we usually refer generically to an estimate of a parameter, such as logistic regression coefficient, using the Greek theta θ. The general Bayesian equation is then a conditional probability of θ, given some observed data. Stated in terms of distribution functions (e.g., Lynch, 2007), using the symbol $f(\ )$ for a distribution function, the Bayesian estimation problem is expressed as:

$$f(\theta \mid data) = \frac{f(data \mid \theta \mid) f(\theta)}{f(data)}$$

We can call $f(data|\theta)$ a likelihood, which could be a maximum likelihood estimate, and then $f(\theta)$ is known as a *distributional prior*. A useful translation is to think about the whole equation as stating that the posterior probability is proportional to the likelihood, weighted (multiplied) by the prior. If the prior adds nothing, then the posterior distribution is simply equal to the likelihood estimate. In such a case, the Bayesian estimation adds nothing to the maximum likelihood estimate. If the prior provides some useful information, then we could improve upon the maximum likelihood estimate.

Bayesian statistical analysis is often thought of as a philosophical alternative to traditional hypothesis testing. Traditional hypothesis testing assumes that a parameter is a kind of fixed definite estimate, whereas, from the Bayesian perspective, there is a probability distribution for the parameter and thus an uncertainty that is built into the posterior estimate—called a *credible interval*. If we use a 95% credible interval, the Bayesian result is to say that the true parameter has a 95% probability of being between the lower and upper credible intervals. This is in contrast to how we interpret confidence intervals with classical (or "frequentist") statistical testing, in which the confidence interval represents the idea that 95% of the intervals from all of the samples from a sampling distribution are contained within the lower and upper confidence intervals. In the classical statistical approach, we should not interpret the confidence interval as representing the probability that the true population parameter is contained within the confidence limits. The Bayesian

interpretation of the credible interval is more akin to saying that we estimate that the true value is within the credible intervals.

Implementation of the Bayesian approach depends on the type of analysis and the software. All analyses however, involve specifying distributional priors. The choice of priors is up to the researcher and may be dictated in part by the type of analysis (e.g., normal distribution priors or a continuous outcome model or a Poisson distribution for analysis of count variables)[1] and can vary within a specific analysis depending on whether the researcher desires stronger (*informative*) or weaker (*uninformative*) prior information. An example of an informative prior might the normal distribution with a certain specified mean and variance and an example of a noninformative prior might be a uniform distribution from $-\infty$ to $+\infty$. One statistical application might be to estimate the posterior probability for the binomial parameter for Bernoulli independent trials using a prior distribution (often a beta probability distribution is used for this) and the binomial function for the likelihood, $\pi^k (1-\pi)^{n-k}$, based on the number of success (k) and failure trials as exponents.

In practice, applying Bayesian statistical analyses nearly always involves a Markov chain Monte Carlo (MCMC) process. This is an iterative, random draw process that uses values drawn from a distribution rather than an empirical resampling like bootstrapping. There are different algorithms of the MCMC process, such as the Gibbs sampler and the Metropolis-Hastings algorithms. A broad description of the MCMC process is that the iterations begin with initial start values for the parameters, then selects at random values from the given prior distribution, and then computes the posterior estimates. This general process is repeated and continues moving to new estimates as long as the new estimates of the proposed distribution of the coefficient continues to improve. There are several approaches to help decide when to end the process or whether it has sufficiently "converged" (e.g., potential scale reduction factor), a decision that is less simple than convergence with maximum likelihood which uses more universal standards of concluding the process.

Bayesian Methods for Logistic Regression

Bayesian logistic regression is possible,[2] although may not be implemented frequently since researchers are most familiar with the classical approach and it performs reasonably well under many conditions.

The Bayesian logistic regression model could be written as (Hosmer et al., 2013):

$$f(\beta_0, \beta_1 \mid \mathrm{y}) = \frac{f(\beta_0, \beta_1) f(\mathrm{y} \mid \beta_0, \beta_1)}{f(\mathrm{y})}$$

On the left side is the posterior distribution for the intercept and slope logistic regression parameters and, on the right side, is the Bayes prior distribution and the likelihood in the numerator and the distribution for y in the denominator. The distribution for y must be estimated using an integral, a normal approximation, or a sampling method, such as the MCMC.

Other Bayesian Categorical Analyses

There are a variety of possible categorical analyses that might use a Bayesian approach. Sometimes, researchers might turn to a Bayesian estimation approach because the computations for the classical approach is burdensome (e.g., marginal maximum likelihood for a large measure in IRT analysis) or because the performance of a certain classical approach does not perform optimally

[1] These might both be examples of so-called *conjugate* priors, in which the posterior and prior distributions are in the same family.
[2] For example, logistic regression could be conducted with `bayesreg` package in R or PROC MCMC in SAS.

in some circumstance (e.g., multilevel models with small sample sizes). In regression analysis, an increasingly popular exploratory model building approach is Bayesian additive regression trees (BART; see Tan & Roy, 2019). This is an exploratory model testing strategy that chooses the most important variables and can simultaneously consider interactions and nonlinear effects. Bayesian estimation can be used for item response theory (IRT) psychometric analyses (e.g., Levy & MisLevy, 2016). For structural equation modeling with categorical variables, the most common estimation approach is diagonal weighted least squares (the WLSMV estimator in Mplus and the lavaan R package), but a Bayesian estimation is one possible option (Muthén & Asparouhov, 2012). Multiple imputation for missing data commonly employs a Bayesian estimation approach (Enders, 2022; see the "Missing Data with Categorical Analysis").

References

Enders, C.K. (2022). *Applied missing data analysis, second edition.* Guilford.

Levy, R., & Mislevy, R. J. (2017). Bayesian psychometric modeling. Chapman and Hall/CRC.

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists* (Vol. 1). New York: Springer.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods, 17*(3), 313.

Tan, Y. V., & Roy, J. (2019). Bayesian additive regression trees and the General BART model. Statistics in medicine, 38(25), 5048-5069.