Chapter 13.7 Latent Class Modeling

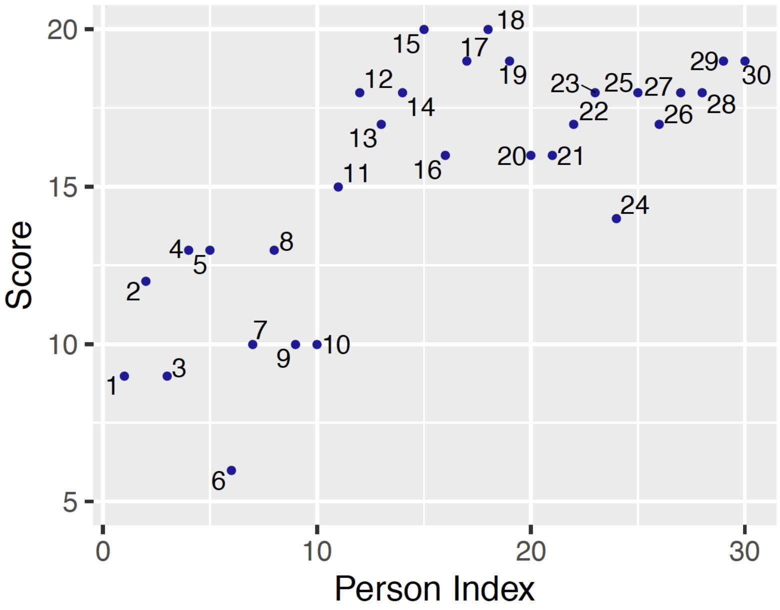
Jim Albert and Monika Hu

Chapter 13 Case Studies

Two classes of test takers

- Suppose thirty people are given a 20-question true/false exam (see Figure on next slide)
- Note that test takers 1 through 10 appear to have a low level of knowledge about the subject matter.
- The remaining test takers 11 through 30 seem to have a higher level of knowledge.

Two classes of test takers (figure)



Two Groups?

- Are there two groups of test takers, a random-guessing group and a knowledgeable group?
- If so, how can one separate the people in the two ability groups?
- How can one make inferences about the correct rate for each group?
- Is it possible to have more than two groups of people by ability level?

A Classification Problem

- In this testing example one believes the people fall in two ability groups.
- However one does not observe the actual classification of the people into groups.
- Itt is assumed that there exists **latent** or unobserved classification of observations.
- The class assignments of the individuals are unknown and can be treated as random parameters in our Bayesian approach.

Group Assignment Parameters

- If there exists two classes, the class assignment parameter for the i-th observation z_i is unknown and assumed to follow a Bernoulli distribution.
- Assume π is the probability of belonging to the first class, i.e. $z_i = 1$.
- With probability $1 - \pi$ the i-th observation belongs to the second class, i.e. $z_i = 0$.

Response Variable

- Once the class assignment z_i is known, the response variable Y_i, the number of correct answers, follows a binomial distribution with a group-specific parameter.
- The response variable Y_i conditional on the class assignment variable z_i is assigned a Binomial distribution with probability of success p_{z_i}.

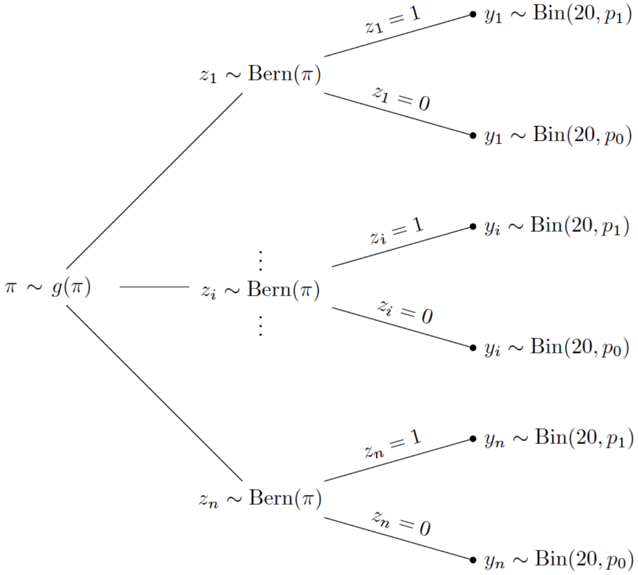
$$Y_i = y_i \mid z_i, p_{z_i} \sim \text{Binomial}(20, p_{z_i}). \tag{1}$$

- For the guessing group, the number of correct answers is Binomial with parameter p_1, and for the knowledgeable group the number of correct answers is Binomial with parameter p_0.

Latent Class Modeling

- The fundamental assumption is that there exists unobserved two latent classes of observations, and each latent class has its own sampling model with class-specific parameters.
- All observations belong to one of the two latent classes and each observations is assigned to the latent classes one and two with respective probabilities π and $(1 - \pi)$.
- Once the latent class assignment is determined, the outcome variable y_i follows a class-specific data model.

Illustration of the Latent Class Model



Why is Latent Class Modeling Useful?

- Latent class models provide the flexibility of allowing unknown class assignments of observations and the ability to cluster observations with similar characteristics.
- In this exam example, the fitted latent class model will pool one class of observations with a lower success rate and pool other class with a higher success rate.
- The fitted model also estimates model parameters for each class, providing insight of features of each latent class.

Details of the Model

- Suppose the true/false exam has m questions and y_i denotes the score of observation i.
- Assume there are two latent classes and each observation belongs to one of the two latent classes.
- Let z_i be the class assignment for observation i and π be the probability of being assigned to class 1.
- Given the latent class z_i for observation i, the score Y_i follows a Binomial distribution with m trials and a class-specific success probability.

Choice of Priors

- One does not know the class assignment probability π, the class assignments $z_1, ..., z_n$, and the probabilities p_1 and p_0 for the two Binomial distributions.
- A natural choice for the probability of class membership π is a Beta prior with shape parameters a and b.
- The parameters p_1 and p_0 are the success rates in the Binomial model in the two classes. If one believes the test takers in class 1 are simply random guessers, then one can fix p_1 to the value of 0.5.
- In general, if one is uncertain about the values of p_1 and p_0, one assumes the success rates are random and assign prior distributions.

Scenario 1: known parameter values

- We begin with a simplified version of this latent class model.
- Consider use of the fixed values $\pi = 1/3$ and $p_1 = 0.5$, and a random p_0 from a Uniform distribution between 0.5 and 1.
- This indicates that one believes strongly that one third of the test takers belong to the random-guessing class, while the remaining two thirds of the test takers belong to the knowledgeable class.
- One is certain about the success rate of the guessing class, but the location of the correct rate of the knowledgeable class is unknown in the interval (0.5, 1).

JAGS Model Script

- One introduces a new variable `theta[i]` that indicates the correct rate value for observation `i`.
- In the sampling section, the first block is a loop over all observations, where one first determines the rate `theta[i]` based on the classification value `z[i]`.
- As π is considered fixed and set to $1/3$, the variable `z[i]` is assigned a Bernoulli distribution with probability $1/3$.
- In the prior section the guessing rate parameter `p1` is assigned the value 0.5 and `p0` is assigned a Beta(1, 1) distribution truncated to the interval (0.5, 1).

JAGS Script

```
modelString<-"
model {
## sampling
for (i in 1:N){
   theta[i] <- equals(z[i], 1) * p1 + equals(z[i], 0) *
y[i] ~ dbin(theta[i], m)
}
for (i in 1:N){
   z[i] ~ dbern(1/3)
}
## priors
p1 <- 0.5
p0 ~ dbeta(1,1) T(0.5, 1)
}"
```

Inference

- One performs inference for `theta` and `p0` by looking at their posterior summaries.
- There are $n = 30$ test takers, each with an associated `theta` indicating the correct success rate of test taker `i`.
- The variable `p0` is the estimate of the correct rate of the knowledgeable class.

Interpretation

- Let's revisit the earlier scatterplot
- Among the test takers with lower scores, it is obvious that test taker # 6 with a score of 6 is likely to be assigned to the random-guessing class, whereas test takers # 4 and # 5 with a score of 13 are probably assigned to the knowledgeable class.
- Among test takers with higher scores, test takers # 15 and # 17 with respective scores of 20 and 19 are most likely to be assigned to the knowledgeable class, and test taker # 24 with a score of 14 is also likely assigned to the knowledgeable class.

Posterior summaries of the correct rates θ_i of six selected test takers

Test Taker	Score	Mean	Median	90% Credible Interval
# 4	13	0.553	0.500	(0.500, 0.876)
# 5	13	0.555	0.500	(0.500, 0.875)
# 6	6	0.500	0.500	(0.500, 0.500)
# 15	20	0.879	0.879	(0.841, 0.917)
# 17	19	0.878	0.879	(0.841, 0.917)
# 24	14	0.690	0.831	(0.500, 0.897)

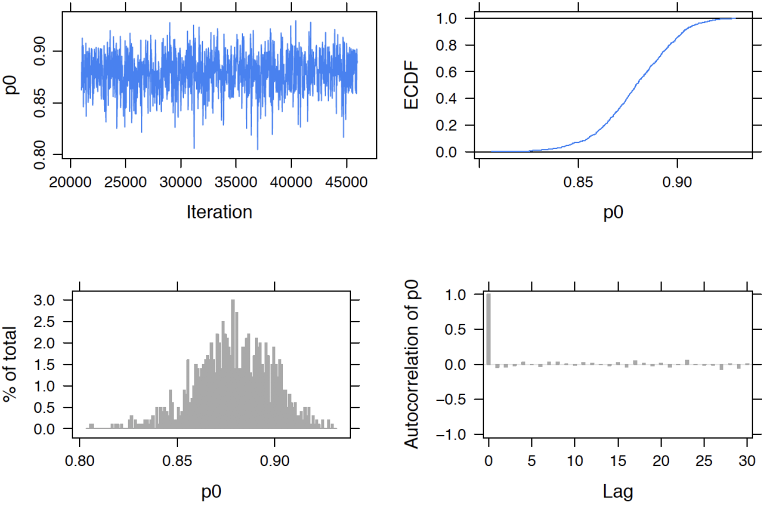
Discussion

- Posterior summaries of the correct rate of test taker # 6 indicate that the model assigns this test taker to the random-guessing group and the posterior mean of the correct rate is at 0.5.
- Test taker # 24 has a higher posterior mean than the test takers # 4 and # 5. But with a posterior mean 0.69, the posterior probability is split between random guessing and knowledgeable states.
- Test takers # 15 and # 17 are always classified as knowledgeable with posterior mean and median of correct rate around 0.88.

Posterior of Success Rates

- Focus on the posterior draws of p_0 corresponding to the success rate for the knowledgeable students.
- Figure on the next slide provides MCMC diagnostics for p_0. Its posterior mean and 90% credible interval are 0.879, and (0.841, 0.917). These estimates are very close to the correct rate of test takers # 15 and # 17.
- These test takers are always classified in the knowledgeable class and their correct rate estimates are the same as p_0.

MCMC Diagnostic Plots for Correct Rate of Knowledgeable Class



Scenario 2: all parameters unknown

- It is more realistic to assume that the probability of assigning an individual into the first class π is unknown.
- Assume little is known about this classification parameter and so π is assigned a $\mathrm{Beta}(1, 1)$
- Assume both p_0 and p_1 are unknown.
- Assign the success rate p_1 a Uniform prior on the interval (0.4, 0.6). ALso assume p_0 is Uniform in the interval $(p_1, 1)$.

JAGS Script

- Introduce the class assignment parameter q as π and assign it a Beta distribution with parameters 1 and 1.
- The prior distributions for p1 and p0 are modified to reflect the new assumptions.

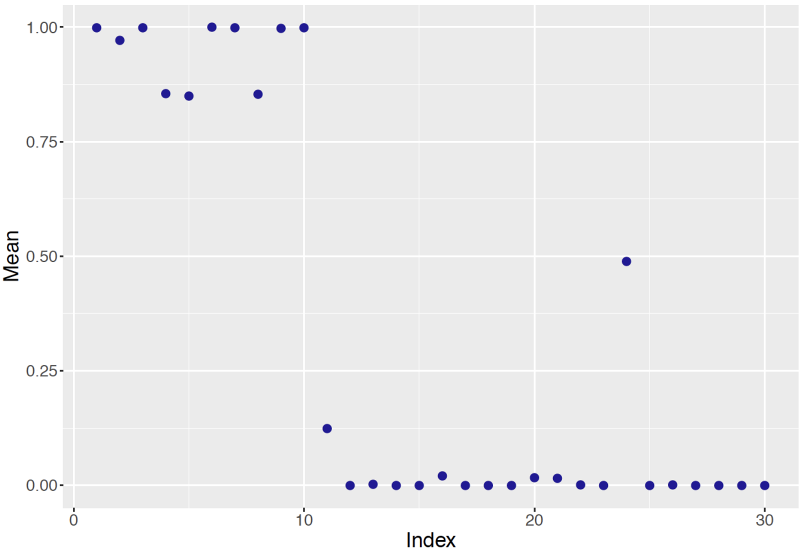
JAGS Script

```
modelString<-"
model {
## sampling
for (i in 1:N){
   theta[i] <- equals(z[i], 1) * p1 + equals(z[i], 0) *
   y[i] ~ dbin(theta[i], m)
}
for (i in 1:N){
   z[i] ~ dbern(q)
}
## priors
p1 ~ dbeta(1, 1) T(0.4, 0.6)
p0 ~ dbeta(1,1) T(p1, 1)
q ~ dbeta(1, 1)
}
"
```

Posterior Analysis

- Focus on the posterior distributions of the classification parameters `z[i]` where `z[i] = 1` indicates a person classified into the random-guessing group.
- Figure on next slide displays the posterior means of the z_i for all individuals.
- As expected, individuals #1 through # 10 are classified as guessers and individuals with labels 12 and higher are classified as knowledgeable.
- Individuals # 11 and # 24 have posterior classification means between 0.25 and 0.75 indicating some uncertainty about the correct classification .

Posterior Means of Classification Parameters



Posteriors of Class Assignment and Rate Parameters

- Figure on next slide displays density estimates of the simulated draws from the posterior distributions of the class assignment parameter π and the rate parameters p_1 and p_0.
- The posterior distributions of p_1 and p_0 are centered about values of 0.54 and 0.89.
- There is some uncertainty about the class assignment parameter as reflected in a wide density estimate for π (q in the figure).

Posteriors of Class Assignment and Rate Parameters