Lesson 4 Notes: Categorical and Ordinal Regression; Predictor Selection

4.1 Introduction

4.1.1 Many social and health data analysis applications involve ordered or unordered polytomous or categorical variables. Examples of unordered responses include brand, commuting or leisure choices. Ordinal outcomes are illustrated by self-assessed health status based on responses to the survey question "Is your general health poor, reasonable, good, or excellent”.

4.1.2 The multinomial logit model generalises the binary and binomial logit framework to modelling multicategory outcomes which do not involve ordered categories. So just as the logit model for a binary outcome involves the log odds of a positive to a negative response, so a multinomial logit involves stipulating a baseline category (say the first of K outcomes or choices) and comparing the probabilities of outcomes 2,3,..,K against that of the first.

4.1.3 Covariates may be defined for individuals i, for choices k, or particular features of choice k unique to individual i . Thus in a travel mode choice example, the first type of variable might be individual income, the second might be the generic cost of alternative modes, and the third might be individual costs attached to different modes. Observed predictors govern the deterministic component of the utility of choice k for subject i. However, different assumptions about the probability distribution of the unobserved utility give rise to various choice models like multinomial logit, nested logit, multinomial probit, GEV (Generalized Extreme Value), mixed logit etc. Multinomial logit models, in particular, assume that unobserved utility is i.i.d. and follows a Gumbel distribution (e.g. see the online methodology advice from the Transportation Research Board at

http://onlinepubs.trb.org/onlinepubs/nchrp/cd-22/v2chapter5.html)

4.1.4 Consider predictors Xi specific to individuals but not choices, and let rij=1 if option j is chosen, with rik=0 for all k≠j. Then a multiple or multinomial logit model, with category 1 as reference, is specified as





so that

   
Also for j and k both exceeding 1 the odds ratio is



so that choice probabilities are governed by differences in coefficient values between alternatives, independent of all other alternatives in the choice set (the "independence of irrelevant alternatives" assumption).   
  
4.1.5 If instead modelling is based on choice specific attributes, then a conditional logit model is obtained, with impacts of some such attributes (denoted Wij) constant across outcomes, while impacts of others (denoted Zij) vary by outcome. The linear predictor for choice j has the form

ηij=Wijγ+Zijδj

with the probability of the category j then being



4.1.6 So for coefficients constant across alternatives, choice probabilities are determined by differences in the values of characteristics between alternatives. A mixed model, combining features of both multinomial and conditional schemes, includes individual level attributes Xi and alternative specific characteristics (Wij,Zij). Thus the linear predictor is

ηi1=Wi1γ+Zi1δ1 j=1

ηij=Wijγ+Zijδj+αj+Xiβj j≠1



and

log(pij/pik)=(αj-αk)+Xi(βj-βk)+(Wij-Wik)γ+(Zijδj-Zikδk).

4.1.7 The log-likelihood for each observation is

LLi = ri1log(pi1)+….+riKlog(piK)

=

or equivalently LLi= where Lik=riklog(pik). The total log-likelihood LL is



The deviance at case level

di=-2LLi

can be taken as a measure of model discrepancy.

4.1.8 Another possible discrepancy measure involves classification accuracy. This is easier to implement if the model is expressed as a categorical likelihood, equivalent to a multinomial likelihood with samples ni=1. For example, let yi ∈ (1,…,K) be a categorical variable taking value k if rik=1. The classification check involves sampling the replicates yrep,i and comparing them to the actual data yi. The total classification accuracy of the model is then the proportion of cases correctly assigned.

4.1.9 In JAGS/BUGS unordered categorical outcomes can be sampled using a categorical likelihood, as just discussed above. Then with eta[i,j] in the form set out in 4.1.6, and φij=exp(ηij), one could specify

y[i] ~ dcat(p[i,1:K])

for (j in 1:K) {p[i,j] <- phi[i,j]/sum(phi[i,])}

for (j in 2:K) {log(phi[i,j]) <- alpha[j]+beta[j]\*X[i]+delta[j]\*Z[i,j]+gam\*W[i,j]}

log(phi[i,1]) <- delta[1]\*Z[i,1]+gam\*W[i,1]

etc

4.2 Fishing Mode Choice; MNL Model

4.2.1 Consider data on data on choice of recreational fishing mode made by 1182 individuals. There are 4 choices ("beach", "boat", "charter", "pier") available to each individual. Attribute specific predictors are denoted `price' and `catch', respectively the cost of a fishing mode, and (in unspecified units) the expected amount of fish caught. These attributes vary both with individuals and fishing mode. By contrast the third predictor, the income level of an individual, does not vary between choices.

4.2.2 The package mnlogit is used to estimate a classical MNL model for these data (you will need to install that package). For JAGS the data are arranged by individual with the ‘price’ and ‘catch’ variables over the four choices named as price1,...,price4 and catch1,…,catch4 (see “fish data.xlsx” which are part of Supplied Materials, and also fish.txt). The JAGS data takes y in category form with values ∈ 1,..,4.

4.2.3 The sequence of commands is then

require(mnlogit)

data(Fish, package = 'mnlogit')

fm <- formula(mode ~ price | income | catch)

CM <- mnlogit(fm, Fish, "alt")

summary(CM)

library(rjags); library(jagstools)

setwd("C://R files")

# columns y, income, price1, price2, price3, price4, catch1, catch2, catch3, catch4

# modes 1,2,3,4: beach, boat, charter, pier

D <- read.table("fish.txt",header=T)

cat("data {K <- 4}

model { for (i in 1:1182) {y[i] ~ dcat(p[i,1:4])

for (k in 1:K) {p[i,k] <- phi[i,k]/sum(phi[i,])}

log(phi[i,1]) <- gam\*prices[1,i] +delta[1]\*catchs[1,i]

for (k in 2:K) { log(phi[i,k]) <- alph[k]+ gam\*prices[k,i]+beta[k]\*incomes[i] +delta[k]\*catchs[k,i]}

prices[1,i] <- price1[i]/100; prices[2,i] <- price2[i]/100;prices[3,i] <- price3[i]/100;

prices[4,i] <- price4[i]/100;

catchs[1,i] <- catch1[i]; catchs[2,i] <- catch2[i];catchs[3,i] <- catch3[i]; catchs[4,i] <- catch4[i]

incomes[i] <- income[i]/1000}

# priors

alph[1] <- 0; for (j in 2:K) {alph[j] ~ dnorm(0,0.001)}

beta[1] <- 0; for (j in 2:K) {beta[j] ~ dnorm(0,0.001)}

for (j in 1:K) {delta[j] ~ dnorm(0,0.001)}

gam ~ dnorm(0,0.001)}

", file="MNL.jag")

INI <- list(list(alph=c(NA,0,0,0), beta=c(NA,0,0,0), delta=c(0,0,0,0),gam=0),

list(alph=c(NA,1,1,1), beta=c(NA,0,0,0), delta=c(1,1,1,1),gam=-2))

M <- jags.model(data=D,inits=INI,n.chains=2,n.adapt=500, file="MNL.jag")

R <- coda.samples(M,c("alph[2:4]","beta[2:4]","delta","gam"),n.iter=5000)

summary(R); gelman.diag(R)

Questions

Q1: Include a log-likelihood and deviance calculation in the JAGS code at case level (LLik and di in 4.1.7). The indicators rik can be obtained using the equals command

r[i,k] <- equals(y[i],k).

Run the model for 2500 iterations. Which two cases have the highest deviance?

Q2: Calculate the interaction between price and catch variables and fit the classical MNL model (using mnlogit) but now including the price-catch interaction, and assuming choice varying coefficients for that interaction predictor. How far does the log-likelihood increase: by under 15 or over 15.

Q3 Extend the JAGS model to have a second attribute specific predictor with choice varying effects (at the moment only catchs[j,i] has an impact varying by choice). Specifically include code for an extra predictor formed by the interaction of ‘price’ and ‘catch’, and assume this new predictor also has varying effects between choices. Run the model for 2500 iterations. Does the posterior mean deviance fall (a) by more than 40, or (b) by less than 40.

4.3 Ordered Category Responses

4.3.1 The multinomial logit model makes no assumptions about the ordering of a categorical outcome. However, ordinal response or choice data are frequently encountered in the social and health sciences. The regression model for an observed response



where categories 1 to K are ordered, may be viewed as reflecting an underlying continuous random variable W such that

   
where εi has cumulative distribution function F, such as (normal cdf) or  (logistic cdf), and corresponding density function f.

4.3.2 The observed yi is defined according to the location of Wi. Specifically yi=j when



where the θj are cut-points on the latent scale. So defining rij=1 when yi=j and rik=0 for k≠j, and setting ri=(ri1,…riK), one has











   
where

   
are cumulative probabilities over the ranked categories,

.

Conversely the probabilities of an observation being in the jth ranked category are given by differencing the cumulative probabilities.





   
  
4.3.3 If F is a logistic cdf, with

 ,  
and β is invariant across response categories j, then the θj are the logits of belonging to categories (1,..,j) up to and including the jth (as against categories j+1,..,K) for subjects with X=0. The difference in cumulative logits for different X values, say X1 and X2, is independent of j. This is known as the `proportional odds' property

   
  
4.3.4 Assuming Xi excludes an intercept, the K-1 thresholds  are unknowns subject to the order constraint

 .

One possible Bayesian prior involves re-parameterisation



   
with the Δ parameters assumed normal. A constrained prior is avoided in the procedure used in the program JAGS whereby a set of K-1 unranked parameters, say θ0=(θ01,..,θ0K), are sampled (e.g. using a diffuse normal prior), and then the required ranked θj are obtained by sorting the θ0j.

4.3.5 The log-likelihood for each observation is the same as for the MNL model, namely

LLi = ri1log(pi1)+….+riKlog(piK)



The deviance at case level, di=-2LLi, can be taken as a measure of model discrepancy. As for MNL regression, another discrepancy measure involves classification accuracy. The classification check involves sampling the replicates yrep,i and comparing them to the actual data yi. The total classification accuracy of the model is then the proportion of cases correctly assigned.

4.4 Wine Data; Ordinal Logit Regression

4.4.1 Consider data on wine bitterness available in the object wine in package ordinal. The data represent a factorial experiment on factors determining the bitterness of wine with 1 =“least bitter” and 5 =“most bitter”. Two treatment factors (temperature and contact) each have two levels. Temperature and contact between juice and skins can be controlled when crushing grapes during wine production. Nine judges each assessed wine from two bottles from each of the four treatment conditions, hence there are 72 observations. The objective is to examine the effect of contact and temperature on the perceived bitterness of wine.

4.4.2 We use the sort option in JAGS to set the prior for the unknown cut-points θ1,..,θ4. As noted above, Xi does not include an intercept. The sequence of R commands (some to be completed) is

require(ordinal)

CM <- clm(rating ~ contact + temp, data = wine)

library(rjags)

library(jagstools)

setwd("C://R files")

# columns y,temp, contact,judge

D <- read.table("wine.txt",header=T)

# JAGS

cat(" data { K <- 5; KM <- 4}

model{ for(i in 1:72){ eta[i] <- beta[1]\*contact[i] + beta[2]\*temp[i]

logit(Q[i,1]) <- theta[1]-eta[i]

p[i,1] <- Q[i,1]

for(j in 2:KM) {logit(Q[i,j]) <- theta[j]-eta[i]

p[i,j] <- Q[i,j] - Q[i,j-1]}

p[i,K] <- 1 - Q[i,KM]

y[i] ~ dcat(p[i,1:K])

yrep[i] ~

match[i] <- equals(y[i],yrep[i])}

Classif.acc <-

# prior for cut-points

for(r in 1:4){ theta0[r] ~dnorm(0,1.0E-3)}

theta <- sort(theta0)

for (j in 1:2){beta[j] ~ dnorm(0,1.0E-3)}}

", file="ologit.jag")

INI <- list(list(theta0=c(-0.6,0,0.6,1.2),beta=c(0,0)),

list(theta0=c(-0.5,0,0.5,1),beta=c(0.5,0.5)))

M <- jags.model(data=D,inits=INI,n.chains=2,n.adapt=500, file="ologit.jag")

R <- coda.samples(M,c("beta","theta"),n.iter=5000)

summary(R)

gelman.diag(R)

Both classical and Bayesian fits show that both contact and high temperature lead to higher probabilities of observations in the high categories.

Questions

Q4. Complete the above code to (a) sample replicate data yrep,i (b) check whether replicate values equal actual values yi and (c) calculate the total classification accuracy. The latter is the number of cases correctly classified divided by the sample size. Is the classification accuracy rate (a) above 25%, or (b) below 25%.

Q5. There may be rater (judge) effects on the ordinal responses. We can expand the model to include a rater effect. This will involve 9 extra parameters, say omega[1:9], of which one, say omega[1], is fixed at zero, a corner constraint obtained via the code omega[1] <- 0. The remaining unknowns, omega[2],…omega[9], can be assigned normal priors with mean 0 and variance 1000. The linear regressor would now be

eta[i] <- beta[1]\*contact[i] + beta[2]\*temp[i]+omega[judge[i]]

Is the classification accuracy now (a) above 30% or (b) below 30%.

Q6. Extend the model in Q5 to include an interaction between temperature and contact. Is the classification accuracy now (a) above 40% or (b) below 40%.

4.5 **Predictor Choice**

4.5.1 Regression model uncertainty may involve different aspects of model specification, such as the error structure (e.g. Normal vs Student t errors in linear regression), whether transformations should be applied to predictors and/or response, and whether there are non-linear regression effects. However, an important and interrelated consideration in regression concerns choice of the best predictor subset. While approaches based on marginal likelihood comparison have been successfully applied to regression model selection, distinctive selection indicator methods have been proposed for Bayesian predictor choice.

4.5.2 A major motivation for predictor selection occurs if there is collinearity between predictors, (x1,…,xP). If y is regressed on a single predictor xj, the coefficient βj may have a clearly posterior in line with subject matter knowledge (e.g. a 95% credible interval confined to positive values, assuming xj was expected to have a positive effect on y). However, with several predictors operating together, coefficients on particular xj may be reduced to `insignificance' (95% credible intervals neither clearly positive nor negative), or even take signs opposite to expectation.

4.5.3 Choice among P predictors to reduce such parameter instability leads to selection schemes for including/excluding each predictor, involving prior probabilities



for binary inclusion or retention indicators γj. An alternative are priors designed to shrink unnecessary regression effects towards zero. Under binary inclusion, the most common choice (but not the only possible one) is

   
implying 2P equally probable models, and that about half the predictors are to be retained a priori. The intercept is usually retained by default. Posterior marginal retention probabilities



are estimated by the proportion of MCMC iterations when γj=1.

4.5.4 Posterior model probabilities (e.g. of the models 1+x1+x4, or 1+x2+x5, where 1 is the intercept) are based on the sampled frequency of different combinations of retained predictors. A Bayes factor for marginal retention can be obtained by comparing the posterior retention odds



with the prior odds



Interest generally lies with predictors having posterior retention probabilities exceeding their prior probability. For example, assuming



Barbieri and Berger (2004) define the median probability model as that defined by predictors with posterior inclusion probabilities exceeding 0.5 (equivalent to the posterior median for γj being 1).

4.5.5 Regarding the Bayes factor, this is a Bayesian measure of the plausibility of a hypothesis or model (denote as H), comparing the prior odds on H (before seeing the data) with the odds on H after observing (or taking account) of the actual data. In the case of predictor selection in regression models, one particular model/hypothesis under examination is that the predictor Xj is relevant to the regression (equivalently that γj=1). The prior

γj ~ Bern(0.5)

above is equivalent a prior probability of 0.5 that the predictor be included. So the prior odds on the hypothesis that γj=1 is 1.If the posterior inclusion probability Pr(γj=1|y) is (say) 0.98, then the posterior odds is 0.98/0.02 is 49. Then the Bayes factor on the particular hypothesis that γj=1 is also 49.

We could set prior odds differently, for example γj ~ Bern(0.25) is more sceptical regarding the relevance of the predictor Xj with prior odds now 0.25/0.75=1/3. If we again found the posterior inclusion probability Pr(γj=1|y) was 0.98, then the Bayes factor would be 217. According to the rules mentioned at https://en.wikipedia.org/wiki/Bayes\_factor, these Bayes factor values (i.e. 49 and 217) are strong/very strong.

4.5.6 George and McCulloch (1997) propose a stochastic search variable selection scheme (SVSS), whereby βj has a vague prior centred at zero (or some other value) when γj=1, but when γj=0 is selected, the prior is centred at zero with high precision (i.e. βj is effectively zero). So for γj=1 (retention option), one might set a relatively large prior variance τ for βj, but multiply this variance by a small constant 0<cj<1 (so the implied impact of xj on y is limited) when γj=0. Possible values are cj=0.001, or cj=0.0005, while George and McCulloch (1997, p. 344) recommend cj > 0.0001.

4.5.7 This leads to a conditional prior with



where



  
The Bernoulli parameters rj can be preset, for example, to a relatively low value such as 0.25 (favouring parsimonious models), or to 0.5, the 'indifference' value. They can also be taken as unknowns, for example

rj ~ Beta(1,1).

4.5.8 We can set up a model index based on which of the γj are 0 or 1. If there are P predictors then there are M=2P possible models. A model index can be obtained using the expression

m <- 1+γ12P-1+γ22P-2+…+γP-12+γP

The following shows how this works out for the case P=4 (see the spreadsheets “Template for Indexing Models” in supplied materials for the cases P=4 and P=6).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Retain X1 | Retain X2 | Retain X3 | Retain X4 | Model index |
| g1 | g2 | g3 | g4 |
| 0 | 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 | 2 |
| 0 | 0 | 1 | 0 | 3 |
| 0 | 0 | 1 | 1 | 4 |
| 0 | 1 | 0 | 0 | 5 |
| 0 | 1 | 0 | 1 | 6 |
| 0 | 1 | 1 | 0 | 7 |
| 0 | 1 | 1 | 1 | 8 |
| 1 | 0 | 0 | 0 | 9 |
| 1 | 0 | 0 | 1 | 10 |
| 1 | 0 | 1 | 0 | 11 |
| 1 | 0 | 1 | 1 | 12 |
| 1 | 1 | 0 | 0 | 13 |
| 1 | 1 | 0 | 1 | 14 |
| 1 | 1 | 1 | 0 | 15 |
| 1 | 1 | 1 | 1 | 16 |

4.6 Employee Satisfaction

4.6.1 From a survey of clerical employees in a large financial organization, the data are aggregated from the questionnaires of the approximately 35 employees for each of 30 (randomly selected) departments. The numbers give percent proportions of favourable responses to seven questions in each department. The data for n=30 departments are

Y rating Overall rating

X1 complaints (Handling of employee complaints)

X2 learning (Opportunity to learn)

X3 advance (Advancement opportunities)

X4 privileges (Does not allow special privileges)

X5 raises (Raises based on performance)

X6 critical (Management too critical)

4.6.2 The following code shows the SSVS approach applied with unstandardized predictors, and with rj=r preset at 0.5. The settings on cj and τ (see item 4.5.5) are τ=10, and cj=1000.

library(rjags)

library(jagstools)

setwd("C://R files")

# columns rating, complaints,learning,advance,privileges,raises,critical

D <- read.table("attitude.txt",header=T)

# conventional least squares

LM <- lm(rating ~complaints+learning+advance+privileges+raises+critical, data=D)

summary(LM)

**#**

**# SSVS fixed r[j]**

**#**

cat("data {r <- 0.5; tau2[1] <- 0.01; tau2[2] <- 10; p <- 6}

model { for (i in 1:30) {rating[i] ~ dnorm(mu[i],tau)

e[i] <- rating[i] - mu[i]

mu[i] <- beta0+beta[1]\*complaints[i]+ beta[2]\*learning[i]+ beta[3]\*advance[i]+ beta[4]\*privileges[i]+ beta[5]\*raises[i]+ beta[6]\*critical[i]}

for (j in 1:6) {beta[j] ~ dnorm(0, 1/tau2[G[j]])

G[j] <- gam[j]+1

gam[j] ~ dbern(r)}

tau ~ dgamma(1,0.001); beta0 ~ dnorm(0,0.001)

M <- 1+gam[1]\*pow(2,p-1) +gam[2]\*pow(2,p-2) +gam[3]\*pow(2,p-3)

+gam[4]\*pow(2,p-4) +gam[5]\*pow(2,p-5) +gam[6]

for (m in 1:64) {mod[m] <- equals(m,M)}}

", file="linreg.jag")

INI <- list(list(beta=c(0,0,0,0,0,0),tau=1, beta0=0),

list(beta=c(0,0,0,0,0,0),tau=0.1, beta0=10))

M <- jags.model(data=D,inits=INI,n.chains=2,n.adapt=500, file="linreg.jag")

R <- coda.samples(M,c("beta","gam"),n.iter=25000)

Q7 Which predictors have posterior probability of inclusion exceeding 0.95?

Q8 Modify the code so that all predictors are standardised before being include in the regression mean mu[i]. For example

z1[i] <- (complaints[i]-mean(complaints[]))/sd(complaints[])

etc.

Which predictors now have posterior probabilities of inclusion exceeding 0.75.

Q9 Monitor the model selection indicators using the code developed in Q8, and extract them using jagresults, as in

R1 <- coda.samples(M,c("mod"),n.iter=2500)

mod.R1 <- jagsresults(R1, c("mod"))

Express models as 1+X2+X4 (e.g. model retaining X2 and X4 only). Which model has the highest posterior probability?

Q10 Modify the code developed in Q8 to allow unknown rj as in



rj ~ Beta(1,1).

Which posterior mean r[j] exceed 0.5?

**References (optional background reading, not required)**

Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. Annals of Statistics, 32, 870-897.

George, E, McCulloch, R (1997) Approaches for Bayesian variable selection. Statistica Sinica, 7(2), 339-373.