**Analysis the Determinants of High School Completion using Bayesian Ordered Probit Approach**

**(Case Study of the High School and Beyond in the US)**

1. **Introduction**

Nowadays, what concern people most is what their future employment would be. People often think what is the return of their education completion on their potential employment. Because most people start to work after or during their high school life, it is interesting to analyse the effect of high school completion on the future employment or unemployment. However, before going far to the mentioned study, it is important to first discover what factors determining the high school completion.

The determinants of high school completion that I want to analyse is high school completion in the US. The high school completion in the US is categorized into four classes. Those level are dropping out of the high school after completing the ninth grade, after completing the tenth grade, after completing the eleventh grade, and completing the high school (Li, 2006). Based on Li (2006), those completion level can be viewed as ordinal data. Hence, as the high school completion outcomes are ordinally ordered, the analysis in this paper will be carried out using statistics methods for ordinal modelling.

Factors that affect high school completion might be an old topic which might have been researched a few times. However, so far, most analysis are based on the frequentist approach. It is commonly known that the frequentist approach treats the unknown parameter as fixed quantity. Hence, as the frequentist estimation method heavily depends on the number of observations, it cause a drawback especially when the data sets are in small size. If the number of the dataset is small, the frequentist methods will probably lead to inaccurate parameter estimate(s).

In order to take account the drawback of the frequentist statistic, the Bayesian approach is proposed to analyse the determinants of the high school completion. The Bayesian approach is preferred as it offers more flexibility compared to the frequentist statistics. Furthermore, The Bayesian statistic allows the unknown parameter to be a random variable in which the estimates produced by the Bayesian has the high probability to be valid. In term of the model fitting, the preliminary information suggest that the dependent variable is naturally an ordered data set. Hence, the relation between variable independent and the variable predictor cannot be modelled as the normal linear model directly (Gelman, 2014). Hence, applying a generalized linear model to analyse the determinants of the high school completion in the US might be useful. Considering the information, the appropriate model to be fitted into the high school completion data set is the Bayesian ordered probit regression.

Given the information given above, the aim of this study is to find the determinants of the of the high school completion in the US with the potential using the Bayesian ordered probit regression. Additionally, as the ordered probit regression is also a classification model, it is of interest to see how precise the fitted model in classifying the high school completion in the US.

1. **Methodology**

**II.(a) Data**

The data set used in this project are extracted from Li (2006) posted in the *Journal of Applied Economics* official site. The author studied the relationship between the high school completion on future youth unemployment using two-stages frequentist regression. At the first stage, the author modelling the relationship between the high school completion with its potential covariates such parental income. The data used in the analysis were obtained from the High School and Beyond (HSB) which is a national survey of US high school students and which was conducted by the National Center for Education Statistics (NCES). In addition, the base data of which the author used come from a survey starting in the spring of 1980. The survey recorded the school completion of each sophomores of 4 categories as those are who dropped out of the high school after completing the ninth grade, after completing the tenth grade, after completing the eleventh grade, and completing the high school. The number of sophomore being surveyed is 5238 (Li, 2016).

In addition, the covariates used in this study are also obtained from the HSB. There are seven predictor variables which potentially affected the high school completion in the US. Those variables are parental income, base year cognitive test, father’s education, mother’s education, number of siblings, female, and minority. Therefore, there are seven predictor variables in this study with the number of observations of 5238. The frequency of each high school completion classes is given below

Table 1. Frequency Table of the US High School Completion

|  |  |  |
| --- | --- | --- |
|  | Frequency | % |
| Dropped out after completing the ninth grade | 76 | 1.45 |
| Dropped out after completing the tenth grade | 219 | 4.18 |
| Dropped out after completing the eleventh grade | 201 | 3.84 |
| Complete the high school | 4742 | 90.53 |

**II.(b)****Identifying the Unknown Parameter**

As previously explained, the interest of this paper is to find the determinants of high school completion in the US using the Bayesian ordered probit regression. Here, as the fitted model is Bayesian ordered probit regression, we can think of ordinal non-numeric variables as arising from some underlying numeric process. In other word, the Bayesian ordered probit regression approach the discrete outcomes with a continuous procedure in term of linear regression using a latent variable.

Let **Y** denote the *n*-vector ***Y*** = ( which is the *n*-ordered categorical outcomes and ***X*** is *n x p* matrix explaining the covariates used in the model. The probit model helps the responses take one of the *k* categories trough a latent variable which is the linear combination of covariates (Sha, 2019). The previous description can be explained more clearly in the term of a regression model below;

where = ( is a *1 x p* vector regression coefficient.

In an ordered probit regression model, the latent variable follow a truncated distribution which need to satisfy that (Hoff, 2009). In this paper, as there are four categories outcomes in the independent variable, the relationship between the observed variable with the latent variable is explained in the following;

= 1 if -

= 2 if (1)

= 3 if

= 4 if

where all the threshold given above are unknown. As the threshold are unknown, those need to be estimated. In addition, the equation given in (1) suggests that if lies at the first interval, then belong to the first category. Additionally, as the value of depend on the coefficient , the quantity of is also unknown means that we also treats as unknown parameter. Therefore, the unknown parameters in the model in this study are **,** }.

**II.(c) Determining Prior Distribution**

To be able to do the analysis using a Bayesian framework, we need to set the prior distribution for each of the unknown parameters previously identified. In this case, in order to form a posterior distribution which can be approximated using the Gibbs sampler, I choose a conjugate prior for each the unknown parameter.

First, I specify a conjugate prior for the ordered probit regression coefficients. As per ordinary regression, I assumed that the coefficient regression has a range from - to +. Hence, setting the normal distribution for the prior distribution for all the coefficients is reasonable. Furthermore, if we set a MVN prior for the coefficients , then the posterior distribution for the is also MVN (Hoff, 2009).

where in this paper, I assumed and At first, it is assumed that all the covariates does have any effect on the high school completion. Hence, the values of is set to be zero. In addition, the variance of the prior distribution is set to be as high as possible to set a weakly prior distribution.

Secondly, I set the prior distribution on the latent variable ***z***. As here we deal with an ordered probit regression model, the density of the latent variable ***z*** are constrained on the interval where *a* equal to - . As the values of are also assumed to be on the interval (- +, I set the prior on ***z*** as truncated normal distribution .

The third step is setting the prior for the which stand as the threshold to classify each corresponding to into a certain class. In this case, the possible values of the threshold could be from - Hence, having the normal distribution as the prior distribution for the is also reasonable. However, when setting up the prior of **,** we need to consider that he threshold given ***Y,X*** are constrained on considering the given constrain, then the appropriate prior distribution for are given below

where *k* is the number of categories in the dependent variable. in this paper, I set the prior for the and .

**II.(d) Identifying Posterior Distributions**

All the prior distributions given in section II.(c) are then updated with the information given by the observed data to produce joint distribution written as follow

where

Based on the joint posterior given above, the distribution is not in the standard form. Hence, to sample the posterior distribution, it might be useful to use the Gibbs sampler algorithm. In order to run the Gibbs sampler, we firstly derive the full conditional distribution of each of the unknown parameter. Here, based on the posterior joint distribution given earlier, the full conditional distribution of each of the unknown parameter is derived as the following;

* Full conditional distribution of

The full conditional distribution of which is expressed in the following only depend on the distribution of ***z*** and on the prior distribution of

* The full conditional distribution of **z**

A the model of **z** is and the , then the full conditional distribution of **z**

where is the normal density function. Hence, the full conditional distribution of **z** is a normal distribution truncated at the interval of (, for a given .

* the full conditional distribution of **g**

In section II.(c), it is mentioned that the prior distribution is constrained at . Then, the full conditional density of is written as follow;

constrained on the interval ( where

As said in the earlier part, it is known that all the prior distribution here is a conjugate prior. Hence, the Gibbs sampling algorithm can be applied to draw the univariate ordered probit to obtain the posterior distribution of each of the unknown parameter. The MCMC simulation using the Gibbs sampler approach are explained in the following steps;

1. Setting the initial value of as zero
2. Setting up the initial value of and to update the threshold **.** For this paper, I set the 1 x *k-1* vector with zero as its entries and the matrix *k-1 x k-1* matrix variance covariance with 100 as its diagonal entries.
3. Inside the MCMC simulation using the Gibbs sampler algorithm, the latent variable, the boundary parameter **,** the latent variable **z,** and the regression coefficient
4. Update the threshold parameter from its full conditional density given in section II.(d). The threshold is updated from the truncated normal density at
5. Update the regression coefficients from its full conditional distribution given in section II.(d).
6. Update the latent variable from its full conditional density given in section II.(d). In this case, to sample a value z from a normal truncated on the interval (a, , the following steps are being performed
7. Sample u ~ uniform()
8. Set z =

where and are the cdf and the inverse cdf of the standard normal distribution respectively.

For a more complete processes, see the algorithm in term of the R-codes attached in the Appendix.

In order to obtain the posterior estimates for the unknown parameters in the Bayesian ordered probit model, I generated 20000 draws for the MCMC simulation with the Gibbs sampler. The draws from the MCMC simulations will be further processed to obtain the posterior estimates. Beforehand, it is important to see the behaviour of the first few draws of each of the unknown parameter. If some of the first few draws deviate from the general trend, it is better to discard them and set the burn-in time. After setting up the burn-in time, the diagnostic checks is applied to see whether the draws already achieved stationarity and do not have any autocorrelation. The diagnostics check will be done by constructing the trace plot, the boxplot and the ACF of each set produced by the MCMC simulations. Furthermore, the effective sample sizes are also to be computed using the package “coda”.

After constructing the diagnostic check and confirming that the samples generated by the MCMC simulations using the Gibbs sampler already produced reliable posterior estimates, several analysis to answer the aims for the paper can be done. Here, to answer the research questions, some further analyses such as building the histogram of the posterior estimates of , constructing the 95% CI for all the coefficient , computing the posterior means, plotting the prior and posterior distribution, and computing the misclassification measurement performed by the fitted model.

In this paper, analysis of the Bayesian ordered probit regression is run using the reference R-codes written in Hoff (2009), some R-package such as “coda” and “MCMCpack”. The “coda” package was used to find the effective size of the of the MCMC samples generated using the Gibbs Sampler. The MCMCpack package is used to fit the model.

1. **Results**

**III (a). Diagnostic Checks**

As explained in the second part of the paper, prior to using the samples drawn from the MCMC simulation for further analysis, it is needed to construct diagnostic checks on the generated samples. If the draws from the MCMC simulation are convergence and independent, then the draws produce reliable posterior estimates for the unknown parameters in the model. First, to see the stationary of the MCMC samples, the trace plot and the boxplots of the samples for each of the seven coefficient are constructed displayed at Figure 1(a) and Figure 1(b).

|  |  |
| --- | --- |
| Figure 1(a). Trace plot for the posterior coefficients | Figure 1(b). Boxplot for the posterior coefficients |

Based on Figure 1(a) and 1(b), it can be said that the samples generated by the MCMC simulation using Gibbs sampler with 20000 draws are already convergence. At this state, we can use the draws to obtain the posterior estimate for each of the parameter . However, the boxplot given in Figure 1(b) suggests that most draws for the posterior estimate still does not converge perfectly. Some of those plots still shows some fairly unstable deviation. The only plot that shows almost perfectly convergence is the draws for the and . Hence, the outputs suggest that the samples for the MCMC simulation should be increased to actually achieved stationarity for all the .

For the second simulation, I increased the draws by generating 40000 draws. Similar to the first simulation, the diagnostic check are also applied in the second simulation to see whether the increasing simulation bringing some improvement in term of the samples’ stationarity. The second diagnostic checks are shown in the following figure.

|  |  |
| --- | --- |
| Figure 2(a). Trace plot for the posterior coefficients | Figure 2(b). Boxplot for the posterior coefficients |
| Figure 2(c). The AC plot for the posterior coefficients | |

Table 2. Effective Sample Sizes for each Parameter

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 251.9143 | 4213.1534 | 432.7904 | 354.2840 | 233.5368 | 226.5879 | 851.5473 |

After increasing the number of samples, the convergence of the draws for each of the parameter seems experience a fairly improvement. As for the independency between each draw, we can see that the ACF plots move quickly to zero. The result suggest that the relationship between each draw are independent. Therefore, the assumption of independent between each draw is satisfied. However, although the convergence shown in Figure 2 seems reasonable, the latest MCMC simulations still show some fairly fluctuation in the mean. As the fluctuation is not that obvious, we just accept the latest MCMC draws as reliable sample to answer the aim of this paper as the latest simulation still produces similar result to that of the first MCMC simulation. In addition, I obtained the sample sizes of each parameter given in Table 2. The results suggest that the draws for and are more reliable compared to the other draws as the effective sample sizes of those two draws are greater than 1000. The number displayed by the effective sample sizes consistent with the trace plot given in Figure 2. The more reliable the draws, the darker the trace plot as it indicates a true convergence.

**III(b). Posterior Estimates for the High School Completion in the US.**

After finishing the diagnostic checks for the MCMC simulations using the Gibbs sampler with 40000 iterations, now it is of interest to summarise the simulations to obtain the posterior estimate of each coefficient given below;

Table 3. Summary of Bayesian Ordered Probit on the High School Completion in the US

|  |  |  |  |
| --- | --- | --- | --- |
| Parameter | Mean | Variance | 95% CI |
|  | 0.09209 | 0.00390 | -0.01921, 0.22504 |
|  | 0.54936 | 0.01387 | 0.35015, 0.81241 |
|  | 0.01479 | 0.00012 | -0.00676, 0.0364 |
|  | 0.02661 | 0.00078 | -0.02865, 0.08065 |
|  | -0.07189 | 0.00117 | -0.13996, -0.00508 |
|  | 0.01430 | 0.00794 | -0.16242, 0.18663 |
|  | 0.37419 | 0.01320 | 0.16906, 0.61461 |

Based on Table 3, we can see which variable significantly affect the high school completion in the US based on the 95% confidence interval. Since only three out of the seven possible covariates has no zero in their confidence interval, then only the three of them significantly affect the high school completion in the US based on the Bayesian ordered probit regression. Those variables are cognitive test, the number of siblings, and being minority. The posterior coefficient estimates of the cognitive variable has positive sign. The sign indicate that the time when the students took their cognitive test positively correlated to the probability of them not finishing the study. As in the number of siblings in the family, this covariate has negative sign. The negative sign suggests that those who have more siblings has larger probability of not completing the high school study. This is reasonable as the more people in the family member, the more financial necessity being needed which sometimes force the member of the family to drop out from the school to find work. Furthermore, being minority high schooler does not make those students dropping out from the school earlier compared to the non-minority students as the sign in the coefficient regression corresponding to the variable has positive value.

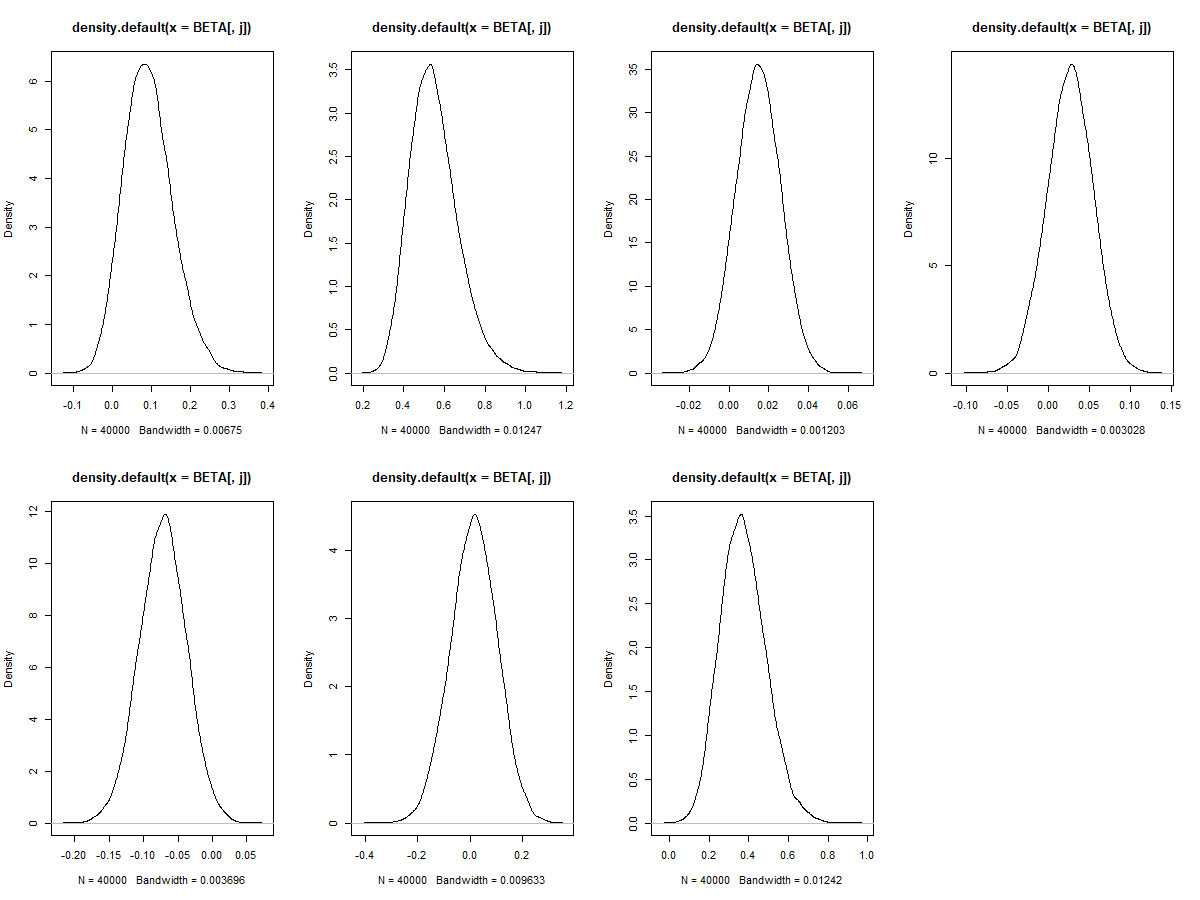


Figure 3. Density plot of the posterior estimate

However, the significance of each posterior estimate is a bit weird here. Although, almost all the posterior variance are relatively small, number of them are not important variable toward the high school completion in the US. Afterwards, it might be useful to visualize the distribution of each posterior . Based on the density plot of the posterior , most of them centred around zero. Therefore, although their variance are relatively small which indicate that the Bayesian approach produces precise estimation, it is reasonable for some of them not being a significant variable for the US high school completion. The condition is obviously different compared to the significant variables. All the significant variable show that their distribution are not centred around zero which make them being the important variable in this study.

|  |  |
| --- | --- |
| Figure 4(a). Histogram and prior and posterior density of | Figure 4(b). Histogram and prior and posterior density of |

|  |
| --- |
| Figure 4(b). Histogram and prior and posterior density of |

Afterward, it is also important to see how the Bayesian framework improve the prior distribution specifically for the significant covariates. Based on the plots given above, it is obvious that all the posterior distribution for the significant parameters form a very narrow range. For all the posterior estimates parameter , we have set a very high variances which equal to . Here, the Bayesian framework shift almost all the variance to a narrow range. The results is consistent with the 95% confidence intervals for the posterior parameter regression which also shows narrow interval. All the outputs both form the 95% confidence interval suggest that the Bayesian method produce precise results.

Considering all the posterior outputs, the following is the Bayesian ordered probit regression model;

with the threshold given below

= 1 if -

= 2 if

= 3 if

= 4 if

To see how precise the Bayesian ordered probit regression at classifying the US high school completion, the posterior frequency table is given below;

Table 4. Frequency Table of the US High School Completion based on the posterior estimate for

|  |  |  |
| --- | --- | --- |
|  | Frequency | % |
| Dropped out after completing the ninth grade | 17 | 0.324 |
| Dropped out after completing the tenth grade | 59 | 1.126 |
| Dropped out after completing the eleventh grade | 219 | 4.180 |
| Complete the high school | 4742 | 94.368 |

Based on Table 4, the overall misclassification is 9.145% which can be considered low. However, when we examine precise classification at the individual class, the misclassification in the first and the second category are quite high. In this case, the Bayesian ordered probit poorly performed when classifying the lower categories. The poor performance by the Bayesian ordered probit in classifying the lower categories might be because the data set is predominant by the fourth category which is around 90%. Hence, the fitted model perform significantly good at predict the class of the dominant group.

1. **Conclusion**

In this study, I proposed a Bayesian framework to predict the ordinal outcomes of the high school completion in the US. The main finding is the explanatory variables that significantly affecting the high school completion in the US are when the cognitive test of each sophomore, the number of siblings in the family and dummy variable of being minority. Furthermore, the performance of Bayesian ordered probit in classifying the US high school completion generally is good. Although, it does not perform quite well in predicting the minority class contain in the independent variable.

1. **Self-Critism**

* Range of the data

As we can see in the summary table, the fourth category significantly predominate the data set. Around 90% of observations are belong to the fourth class. This might be the reason why the fitted Bayesian ordered probit regression poorly predict the first two categories as they just consist 1.5% and 4% respectively out of the whole independent variable. Considering the condition, it might better to fit a model which can take account a data set which consist of significantly predominant by a certain category.

* Model Specification

Based on the information given in the data set, the data used in this study is a result from a survey on 1980 born students conducted in the spring 1982, 1984 and in 1986. In my previous analysis, I assumed that the data can be treat as a cross-section data by assuming that the data is simply a survey data and the author of the paper where I borrow the data assumed so. However, as in this study I kind of ignoring the time effects, this might be a reason why the posterior draws of each coefficients parameter does not converge perfectly even after the draws being increased significantly. Hence. For the future research, it might be better to separate the observations between different year and fitting an ordinal panel model.

**References**

Gelman, Andrew., Carlin, John B., Stern, Hal S., Dunson, David B. 2014. *Bayesian Data Analysis: Third Edition*. Taylor and Francis Group.

Hoff, Peter D. 2009. *A first course in Bayesian statistical methods*. London; New York: Springer.

Li, Mingliang. 2006. High School Completion and future youth unemployment: new evidence from High School and Beyond. *Journal of Applied Econometrics*, 21: 23-53/

Sha, Naijun. 2019. A Bayes Inference for Ordinal Response with Latent Variable Approach. *MDPI. Stats*, 2(2): 321-331

**Appendix. Main Code**

|  |
| --- |
| #-------------Ordered Probit Codes from the Lecture Notes ----------------#  rm(list=ls())  setwd("H:/STAT7016/Final Project/Final Project")  #setwd("C:/SEMESTER 4/STAT7016/Final Project")  #jae.4486.5.data.2 <- read.table("C:/SEMESTER 4/STAT7016/Final Project/jae-4486-5-data-2.txt", quote="\"", comment.char="")  jae.4486.5.data.2 <- read.table("H:/STAT7016/Final Project/Final Project/jae-4486-5-data-2.txt", quote="\"", comment.char="")  data.probit <- jae.4486.5.data.2  names(data.probit) <- c("school\_ID", "parental\_income", "cog\_test","fathers\_edu", "unknown\_fathedu", "mothers\_edu",  "unknown\_motheredu", "num\_siblings", "unknown\_num\_siblings", "female", "minority", "employment\_rate",  "age\_on\_january","elig\_dropout", "postsecond\_edu", "grade\_completion", "unemployed\_stat")  names(data.probit)  #y <- data.probit[,17]  #y.new <- ifelse(y > 0.5,1,0)  completion <- rep(1,length(data.probit[,16]))  completion <- ifelse(data.probit[,16]==0,4,completion)  completion <- ifelse(data.probit[,16]==1,1,completion)  completion <- ifelse(data.probit[,16]==2,2,completion)  completion <- ifelse(data.probit[,16]==3,3,completion)  completion  y <- completion  X <- data.probit[, c(2,3,4,6,8,10,11)] # 7 covariates  X <- as.matrix(X)  X <- X  y <- y  keep<- (1:length(y))[ !is.na( apply( cbind(X,y),1,mean) ) ]  X<-X[keep,] ; y<-y[keep]  #ranks<-match(y,sort(unique(y))) ;  ranks<-y  uranks<-sort(unique(ranks))  K<-length(uranks)  n<-dim(X)[1] ; p<-dim(X)[2]  iXX<-solve(t(X)%\*%X) ; V<-iXX\*(n/(n+1)) ; cholV<-chol(V)  #assumed prior on g's  mu<-rep(0,K-1) ; sigma<-rep(100,K-1)  ###starting values  set.seed(1)  beta<-rep(0,p)  z<-qnorm(rank(y,ties.method="random")/(n+1)) # divide by n+1 so that ranks are between 0 and 1.  g<-rep(NA,length(uranks)-1) #K-1 thresholds if there are K unique ranks  K<-length(uranks)  #objects to store results  BETA<-matrix(NA,40000,p) ; Z<-matrix(NA,40000,n) ; ac<-0  S<-80000  for(s in 1:S)  {    #update g  for(k in 1:(K-1))  {  a<-max(z[y==k]) #constraints on g\_k  b<-min(z[y==k+1])  u<-runif(1, pnorm( (a-mu[k])/sigma[k] ), #inverse CDF methods to draw a random g\_k  pnorm( (b-mu[k])/sigma[k] ) )  g[k]<- mu[k] + sigma[k]\*qnorm(u)  }    #update beta  E<- V%\*%( t(X)%\*%z )  beta<- cholV%\*%rnorm(p) + E    #update z  ez<-X%\*%beta  a<-c(-Inf,g)[ match( y-1, 0:K) ] #match returns a vector of the positions of (first) matches of its first argument in its second.  #1 is matched with zero; 2 is matched with 1.....(defines lower limit)  b<-c(g,Inf)[y] #assign upper limit to each z\_i (i=1,..n)  u<-runif(n, pnorm(a-ez),pnorm(b-ez) )  z<- ez + qnorm(u)      #help mixing  c<-rnorm(1,0,n^(-1/3))  zp<-z+c ; gp<-g+c #add extra random component to zp and gp  lhr<- sum(dnorm(zp,ez,1,log=T) - dnorm(z,ez,1,log=T) ) +  sum(dnorm(gp,mu,sigma,log=T) - dnorm(g,mu,sigma,log=T) )  if(log(runif(1))<lhr) { z<-zp ; g<-gp ; ac<-ac+1 }    if(s%%(S/40000)==0)  {  cat(s/S,ac/s,"\n")  BETA[s/(S/40000),]<- beta  Z[s/(S/40000),]<- z  }    }  par(mfrow=c(2,4))  blabs<-c(expression(beta[1]),expression(beta[2]),expression(beta[3]),expression(beta[4]), expression(beta[5]),  expression(beta[6]), expression(beta[7]))  #thin<-c(1,(1:1000)\*(S/1000))  thin<-c(1,(1:500)\*(40000/500))  #Diagnostic check for posterior beta\_j  for (j in 1 : p ){  plot(thin,BETA[thin,j],type="l",xlab="iteration",ylab=blabs[j])  abline(h=mean(BETA[,j]) )  }  # Boxplot for beta-j  par(mfrow=c(2,4))  for (j in 1 : p ){  boxplot(BETA[1:2500,j],BETA[2501:5000,j],BETA[5001:7500,j],BETA[7501:10000,j],  BETA[10001:12500,j],BETA[12501:15000,j],BETA[15001:17500,j],BETA[17501:20000,j],  BETA[20001:22500,j],BETA[22501:25000,j],BETA[25001:27500,j],BETA[27501:30000,j],  BETA[30001:32500,j],BETA[32501:35000,j],BETA[35001:37500,j],BETA[37501:40000,j],  xlab='iteration', ylab=blabs[j])  }  # Boxplot for beta-j  par(mfrow=c(2,4))  for (j in 1 : p ){  boxplot(BETA[1:5000,j],BETA[5001:10000,j],BETA[10001:15000,j],BETA[15000:15001,j],  BETA[10001:12500,j],BETA[12501:15000,j],BETA[15001:17500,j],BETA[17501:20000,j],  BETA[20001:22500,j],BETA[22501:25000,j],BETA[25001:27500,j],BETA[27501:30000,j],  BETA[30001:32500,j],BETA[32501:35000,j],BETA[35001:37500,j],BETA[37501:40000,j],  xlab='iteration', ylab=blabs[j])  }  par(mfrow=c(2,4))  for (i in 1:p){  acf(BETA[thin,i],xlab="iteration",ylab=blabs[j] )  }  #Diagnostic check for posterior beta\_j  for (j in 1 : p ){  plot(thin,BETA[thin,j],type="l",xlab="iteration",ylab=blabs[j])  abline(h=mean(BETA[,j]) )  }  #Density Plot  for (j in 1 : p ){  plot(density(BETA[,j]))  }  #Efectove sample size  library(coda)  effectiveSize(BETA)  beta.post <- apply(BETA, 2, mean)  round(beta.post,5)  var.beta.post <- apply(BETA,2,var)  round(var.beta.post,5)  # Confidence Interval  apply(BETA,2,function(x) quantile(x,prob=c(.025,.5,.975)))  ZPM<-apply(Z,2,mean)  ZPM  # Posterio, prior beta  par(mfrow=c(1,2))  hist(BETA[,2], xlab=expression(beta[2]), main="")  plot(density(BETA[,2],adj=2),lwd=2,xlim=c(-1,2),main="",  xlab=expression(beta[2]),ylab="density")  sd<-sqrt( solve(t(X)%\*%X/n)[2,2] )  x<-seq(-1,6,length=100)  lines(x,dnorm(x,0,sd=sqrt(100)),lwd=2,col="gray")  legend(0.8,3,legend=c("prior","post"),lwd=c(2,2),col=c("gray","black"),bty="n")  hist(BETA[,5], xlab=expression(beta[5]), main="")  plot(density(BETA[,5],adj=2),lwd=2,xlim=c(-1,2),main="",  xlab=expression(beta[5]),ylab="density")  sd<-sqrt( solve(t(X)%\*%X/n)[2,2] )  x<-seq(-1,6,length=100)  lines(x,dnorm(x,0,sd=sqrt(100)),lwd=2,col="gray")  legend(0.8,3,legend=c("prior","post"),lwd=c(2,2),col=c("gray","black"),bty="n")  hist(BETA[,7], xlab=expression(beta[7]), main="")  plot(density(BETA[,7],adj=2),lwd=2,xlim=c(-1,2),main="",  xlab=expression(beta[7]),ylab="density")  sd<-sqrt( solve(t(X)%\*%X/n)[2,2] )  x<-seq(-1,6,length=100)  lines(x,dnorm(x,0,sd=sqrt(100)),lwd=2,col="gray")  legend(0.8,3,legend=c("prior","post"),lwd=c(2,2),col=c("gray","black"),bty="n")  # Posterior predicitive check  y.post <- rep(1, length(y))  y.post <- ifelse(ZPM < g[1] & ZPM > Inf, 1, y.post)  y.post <- ifelse(ZPM < g[2] & ZPM > g[1], 2, y.post)  y.post <- ifelse(ZPM < g[3] & ZPM > g[2], 3, y.post)  y.post <- ifelse(ZPM < Inf & ZPM > g[3], 4, y.post)  (sum(y.post==1)/length(y))\*100  (sum(y.post==2)/length(y))\*100  (sum(y.post==3)/length(y))\*100  (sum(y.post==4)/length(y))\*100  (sum(y==1)/length(y))\*100  (sum(y==2)/length(y))\*100  (sum(y==3)/length(y))\*100  (sum(y==4)/length(y))\*100  ##Missclasification check  1-(sum(y.post==y)/length(y)) |