STT465\_HW5

Sam Isken

November 8, 2019

Code chunk below reads in nessecary libraries / packages.

An Example: MLR Analysis in Frequentist and Bayesian

We will disscuss most of this today and you should compile the work and submit as Homework #5 (Due Monday 11/11/2019)

1. Load the WAGES dataset into R:

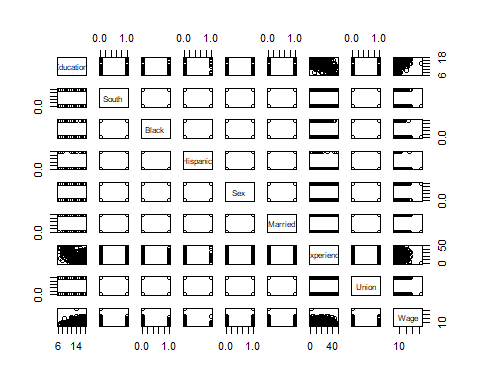
WAGES <- read.table("wages.txt",header = TRUE)  
head(WAGES)

## Education South Black Hispanic Sex Married Experience Union Wage  
## 1 10 0 0 0 0 1 27 0 9.0  
## 2 12 0 0 0 0 1 20 0 5.5  
## 3 12 0 0 0 1 0 4 0 3.8  
## 4 12 0 0 0 1 1 29 0 10.5  
## 5 12 0 0 0 0 1 40 1 15.0  
## 6 16 0 0 0 1 1 27 0 9.0

1. Conduct basic descriptive statistical analysis:

Make comments on descriptive statistics and construct a scatterplot matrix:

pairs(WAGES)



summary(WAGES)

## Education South Black Hispanic   
## Min. : 6.00 Min. :0.0000 Min. :0.0000 Min. :0.00000   
## 1st Qu.:12.00 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.00000   
## Median :12.00 Median :0.0000 Median :0.0000 Median :0.00000   
## Mean :13.09 Mean :0.2917 Mean :0.1269 Mean :0.04735   
## 3rd Qu.:15.00 3rd Qu.:1.0000 3rd Qu.:0.0000 3rd Qu.:0.00000   
## Max. :18.00 Max. :1.0000 Max. :1.0000 Max. :1.00000   
## Sex Married Experience Union   
## Min. :0.0000 Min. :0.0000 Min. : 0.00 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.: 8.00 1st Qu.:0.0000   
## Median :0.0000 Median :1.0000 Median :15.00 Median :0.0000   
## Mean :0.4621 Mean :0.6553 Mean :17.66 Mean :0.1818   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:26.00 3rd Qu.:0.0000   
## Max. :1.0000 Max. :1.0000 Max. :49.00 Max. :1.0000   
## Wage   
## Min. : 1.750   
## 1st Qu.: 5.250   
## Median : 7.790   
## Mean : 9.048   
## 3rd Qu.:11.250   
## Max. :44.500

sapply(WAGES, range)

## Education South Black Hispanic Sex Married Experience Union Wage  
## [1,] 6 0 0 0 0 0 0 0 1.75  
## [2,] 18 1 1 1 1 1 49 1 44.50

Notes on variables: -Education: –Years of education range from 6 to 18 years in this data set –Mean and median are close, roughly normal

-South –Binary variable indicating location

-Black/Hispanic –Binary variables indicating race

-Sex –Binary variable indicating sex

Notes on Correlations: -Education is positively correlated with wages -Theres a slight negative correlation between Education and Experience (logical, opportunity cost of education is time that could be spent working)

These are the main conclusions from this scatter. Due to many of the other variables being indicator variables (Binary 0 or 1) it will be easier to look at those in the context of the regression model.

3.Regress wage on education, race, experience, region, sex and marital status via OLS using lm():

lin\_model\_1 <- lm(Wage~Education+Black+Hispanic+Experience+South+Sex+Married,data=WAGES)

4.Using the output of summary(lm(y~….)) answer the following questions:

summary(lin\_model\_1)

##   
## Call:  
## lm(formula = Wage ~ Education + Black + Hispanic + Experience +   
## South + Sex + Married, data = WAGES)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.897 -2.609 -0.594 1.905 37.580   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.94354 1.28195 -3.076 0.00221 \*\*   
## Education 0.93549 0.08319 11.245 < 2e-16 \*\*\*  
## Black -0.78938 0.58513 -1.349 0.17790   
## Hispanic -0.41694 0.91673 -0.455 0.64943   
## Experience 0.10405 0.01751 5.944 5.11e-09 \*\*\*  
## South -0.89492 0.43096 -2.077 0.03833 \*   
## Sex -2.33784 0.38810 -6.024 3.23e-09 \*\*\*  
## Married 0.56781 0.42077 1.349 0.17777   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.42 on 520 degrees of freedom  
## Multiple R-squared: 0.2714, Adjusted R-squared: 0.2616   
## F-statistic: 27.67 on 7 and 520 DF, p-value: < 2.2e-16

1. What of the factors / variables have a significant effect on wages?

Education, Experience, Region (located in South or not, binary variable 1 or 0) and Sex are the 4 variables (besides the intercept) that are statistically significant in this model.

1. How much do you expect wage to increase per year of education?

Given a 1 year increase in education we estimate hourly wages to increase by $ .93 (93 cents).

1. What is the average wage difference between male and female in the sample?

$2.33784, so roughly $2.33. Women earn, on average, $2.33 less than men in this data and model.

1. Is the difference in c) statistically different from 0?

Yes, it is very statistically significant since the P-Value is far below .05.

1. Load the Gibbs Sampler Function in R:

### Gibbs Sampler #######  
  
gibbsMLR=function(y,X,nIter=10000,df0=4,S0=var(y)\*0.8\*(df0-2),b0=0,varB=1e12,verbose=500){  
   
 ## Objects to store samples  
 p=ncol(X); n=nrow(X)  
 B=matrix(nrow=nIter,ncol=p,0) # create a matrix to store the gibbs sample for beta  
 varE=rep(NA,nIter) # .. for error variance  
   
 ## Initialize  
 B[1,]=0 # initial values for slopes  
 B[1,1]=mean(y) # initial value for y-intercept  
 b=B[1,]  
 varE[1]=var(y) # initial error variance  
 resid=y-B[1,1] # centered y (orthogonal)  
   
 ## Computing sum x'x for each column  
 SSx=colSums(X^2)  
   
 for(i in 2:nIter){  
 # Sampling regression coefficients  
 for(j in 1:p){  
 A=SSx[j]/varE[i-1]+1/varB  
 Xy= sum(X[,j]\*resid)+SSx[j]\*b[j] # equivalent to X[,j]'(y-X[,-j]%\*%b[-j])  
 rhs=Xy/varE[i-1] + b0/varB # Numerator of beta^tilda\_k  
 condMean=rhs/A  
 condVar=1/A  
 b\_old=b[j]  
 b[j]=rnorm(n=1,mean=condMean,sd=sqrt(condVar))  
 B[i,j]=b[j]   
 resid=resid-X[,j]\*(b[j]-b\_old) # updating residuals  
 }  
 # Sampling the error variance   
 RSS=sum(resid^2)  
 DF=n+df0  
 S=RSS+S0  
 varE[i]=S/rchisq(df=DF,n=1)  
   
 ## if(i%%verbose==0){ cat(' Iter ', i, '\n') }  
 }  
   
 out=list(effects=B,varE=varE)  
 return(out)  
}

1. Collect, for the same model specified above 15,000 samples.

Hint: gibbsMLR(y,X,..) takes as inputs a numeric vector with the response (WAGES) and an incidence matrix for effects (sex,race,education,experience,…). lm() creates the incidence matrix internally. For this you can use X=model.matrix(~a+b+…,date=WAGES) where a,b are the predictors that are included in data. This will create your incidence matrix and then you will use it in gibbsMLR(y,X).

# Data Pre-Processing and Sampling  
# Set wage to be just the Wage column of data set WAGES   
wage <- WAGES$Wage  
# Create incidence matrix to use in sampling function  
X <- model.matrix(~Education+Black+Hispanic+Experience+South+Sex+Married,data=WAGES)  
# Pull sample using gibbsMLR(y,X)  
SAMPLES <- gibbsMLR(y=wage,X=X,nIter = 15000)

dim(SAMPLES$effects) #Columns = effects, rows = Samples

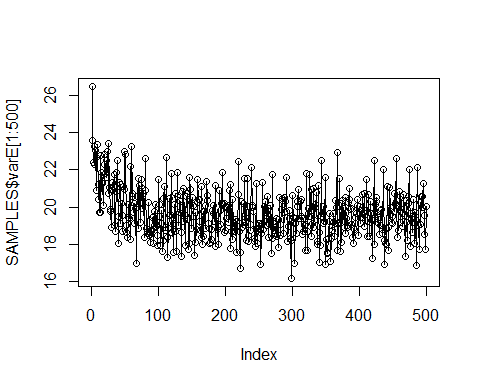
## [1] 15000 8

head(SAMPLES$varE)

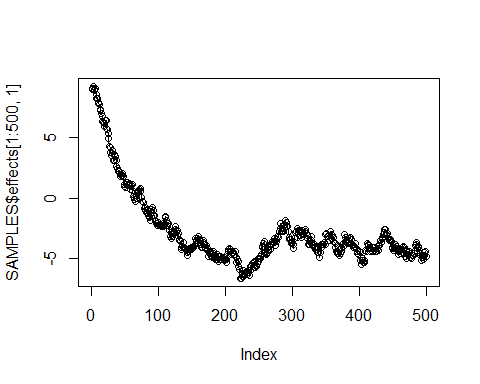
## [1] 26.46158 23.57790 22.40492 23.22906 22.29866 23.05627

1. Conduct post-gibbs analysis (trace plot, auto-correlation, decide on burin-in and thinning, provide posterior means, posterior SDs and posterior credibility regions, estimate and report MC errror).

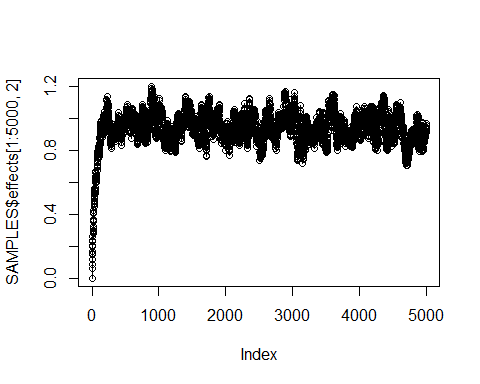
# Trace plot of all variables   
plot(SAMPLES$varE[1:500],type='o')



# Intercept trace plot  
plot(SAMPLES$effects[1:500,1],type='o')



# The plot below allows us to decide on burn-in quantity, since the data centers around index = 1000 I will burn-n (remove) the first 1000 samples  
plot(SAMPLES$effects[1:5000,2],type='o')



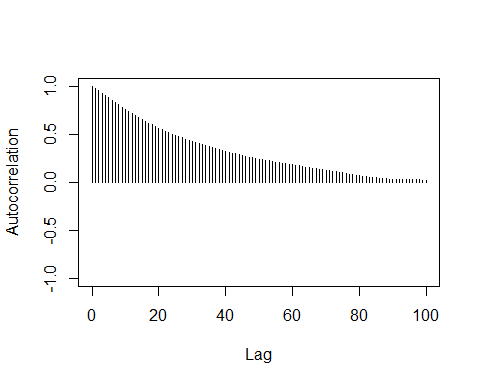
# Discarding Burn - In  
B <- SAMPLES$effects[-(1:1000),];colnames(B)=colnames(X)  
varE <- SAMPLES$varE[-(1:1000)]  
  
summary(B)

## (Intercept) Education Black Hispanic   
## Min. :-7.985 Min. :0.5952 Min. :-3.4180 Min. :-4.5258   
## 1st Qu.:-4.735 1st Qu.:0.8772 1st Qu.:-1.1945 1st Qu.:-1.0433   
## Median :-3.861 Median :0.9318 Median :-0.7996 Median :-0.4192   
## Mean :-3.875 Mean :0.9313 Mean :-0.8000 Mean :-0.4246   
## 3rd Qu.:-3.030 3rd Qu.:0.9863 3rd Qu.:-0.4036 3rd Qu.: 0.1921   
## Max. : 1.583 Max. :1.2098 Max. : 1.6892 Max. : 2.8042   
## Experience South Sex Married   
## Min. :0.03201 Min. :-2.6574 Min. :-3.9296 Min. :-1.1379   
## 1st Qu.:0.09189 1st Qu.:-1.2027 1st Qu.:-2.6004 1st Qu.: 0.2911   
## Median :0.10329 Median :-0.9079 Median :-2.3373 Median : 0.5756   
## Mean :0.10344 Mean :-0.9058 Mean :-2.3386 Mean : 0.5723   
## 3rd Qu.:0.11500 3rd Qu.:-0.6101 3rd Qu.:-2.0806 3rd Qu.: 0.8503   
## Max. :0.17191 Max. : 0.7758 Max. :-0.9045 Max. : 2.0742

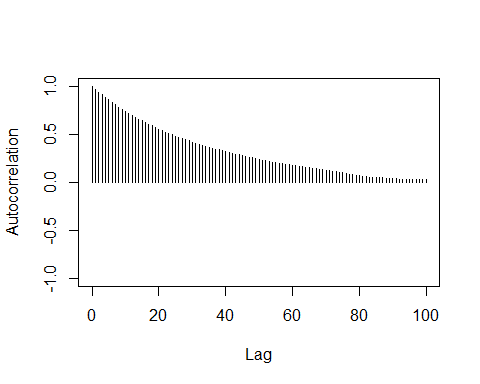
Let’s now convert B and varE to mcmc objects in order to examine autocorrelation using the coda libray.

# Convert to a MCMC object   
B <- as.mcmc(B)  
varE <- as.mcmc(varE)

autocorr.plot(B[,1],lag.max=100)



autocorr.plot(B[,2],lag.max=100)



In both instances the autocorrelation approaches 0 as Lag increases to 100. Thus we should sample 1 value every 100 indexes to control for autocorrelation. Thus, we should resample with value 100, 200, …. to n hundred (end of our data set). This process is called thinning our data.

Let us now create a HPD interval for our B mcmc object. Since this is an MCMC object I read in an additional formula from the cran GitHub for coda:

HPDinterval <- function(obj, prob = 0.95, ...) UseMethod("HPDinterval")  
  
HPDinterval.mcmc <- function(obj, prob = 0.95, ...)  
{  
 obj <- as.matrix(obj)  
 vals <- apply(obj, 2, sort)  
 if (!is.matrix(vals)) stop("obj must have nsamp > 1")  
 nsamp <- nrow(vals)  
 npar <- ncol(vals)  
 gap <- max(1, min(nsamp - 1, round(nsamp \* prob)))  
 init <- 1:(nsamp - gap)  
 inds <- apply(vals[init + gap, ,drop=FALSE] - vals[init, ,drop=FALSE],  
 2, which.min)  
 ans <- cbind(vals[cbind(inds, 1:npar)],  
 vals[cbind(inds + gap, 1:npar)])  
 dimnames(ans) <- list(colnames(obj), c("lower", "upper"))  
 attr(ans, "Probability") <- gap/nsamp  
 ans  
}  
  
HPDinterval.mcmc.list <- function(obj, prob = 0.95, ...)  
 lapply(obj, HPDinterval, prob)  
# source: https://github.com/cran/coda/blob/master/R/HPDinterval.R

print("The below shows the posterior mean and standard deviation")

## [1] "The below shows the posterior mean and standard deviation"

summary(B)

##   
## Iterations = 1:14000  
## Thinning interval = 1   
## Number of chains = 1   
## Sample size per chain = 14000   
##   
## 1. Empirical mean and standard deviation for each variable,  
## plus standard error of the mean:  
##   
## Mean SD Naive SE Time-series SE  
## (Intercept) -3.8746 1.25236 0.0105844 0.0895466  
## Education 0.9313 0.08149 0.0006887 0.0056867  
## Black -0.8000 0.58494 0.0049436 0.0057860  
## Hispanic -0.4246 0.92523 0.0078196 0.0091973  
## Experience 0.1034 0.01737 0.0001468 0.0004765  
## South -0.9058 0.43893 0.0037096 0.0067376  
## Sex -2.3386 0.38921 0.0032894 0.0055121  
## Married 0.5723 0.41730 0.0035268 0.0072559  
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%  
## (Intercept) -6.25527 -4.73502 -3.8612 -3.0300 -1.41540  
## Education 0.77064 0.87718 0.9318 0.9863 1.08774  
## Black -1.93607 -1.19450 -0.7996 -0.4036 0.34673  
## Hispanic -2.21541 -1.04328 -0.4192 0.1921 1.39419  
## Experience 0.06929 0.09189 0.1033 0.1150 0.13719  
## South -1.76573 -1.20266 -0.9079 -0.6101 -0.04728  
## Sex -3.10035 -2.60044 -2.3373 -2.0806 -1.57204  
## Married -0.24709 0.29114 0.5756 0.8503 1.38809

summary(varE)

##   
## Iterations = 1:14000  
## Thinning interval = 1   
## Number of chains = 1   
## Sample size per chain = 14000   
##   
## 1. Empirical mean and standard deviation for each variable,  
## plus standard error of the mean:  
##   
## Mean SD Naive SE Time-series SE   
## 19.54717 1.21561 0.01027 0.01042   
##   
## 2. Quantiles for each variable:  
##   
## 2.5% 25% 50% 75% 97.5%   
## 17.29 18.70 19.49 20.33 22.08

print("The below shows the poterior credibility interval")

## [1] "The below shows the poterior credibility interval"

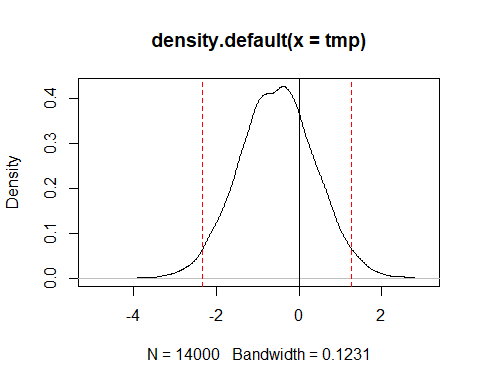
HPDinterval.mcmc(B, prob=.95)

## lower upper  
## (Intercept) -6.31940512 -1.49462502  
## Education 0.77706608 1.09271076  
## Black -1.95126706 0.32678367  
## Hispanic -2.23698181 1.36538206  
## Experience 0.06926171 0.13717029  
## South -1.77919916 -0.06296253  
## Sex -3.11340290 -1.58950460  
## Married -0.25433399 1.37839277  
## attr(,"Probability")  
## [1] 0.95

1. Test whether there is an ethnic disparity between black and Hispanic workers and report the posterior probability.

#### Bayesian Linear Hypotheseis

contrast=c(0,0,0,1,-1,0,0,0)  
tmp=B%\*%contrast  
plot(density(tmp));abline(v=0)  
abline(v=HPDinterval.mcmc(as.mcmc(tmp),p=.95),col=2,lty=2)



print(paste("The ethnic disparity between black and hispanic is > 0", mean(tmp>0)\*100))

## [1] "The ethnic disparity between black and hispanic is > 0 27.9357142857143"

print(paste("The ethnic disparity between black and hispanic is< 0", mean(tmp<0)\*100))

## [1] "The ethnic disparity between black and hispanic is< 0 72.0642857142857"

print("Both in percent of the time")

## [1] "Both in percent of the time"