PCB 6466 - Methods in Experimental Ecology Fall 2013 Semester
Dr. Pedro F. Quintana-Ascencio, James Angelo and Matt Tye 10/15/2013

R Demonstration – Multiple Regression

Objective: The purpose of this week's session is to demonstrate how to perform multiple linear regressions (i.e. linear regression models with two or more predictor variables) in R and OpenBUGS. In the first part, we will examine the issue of collinearity in multiple regression models. In the second part, we will investigate various issues related to multiple regression and model selection.

Part I. Multiple Regression and the Issue of Collinearity

NOTE: This part of the exercise assumes that you have downloaded the dataset from Paruelo & Lauenroth (1996) and saved it in your `PCB6466` folder as a tab-delimited text file named *`paruelo.txt`*. You also need to download the *`multiRegression.R`* script and save it in your `PCB6466` folder.

After starting R, change the directory to your `PCB6466` folder and open the *`multiRegression.R`* script. The first two lines of the script read and attach the Paruelo & Lauenroth (1996) dataset:

```
## load the Paruelo & Lauenroth (1996) dataset from file
paruelo_data <- read.table("paruelo.txt", header=T)
attach(paruelo_data)
```

Next, we use the *`par`* function with the `mfrow=` argument to create a plotting area comprised of two rows and two columns. After doing the appropriate transformations, we can plot the histograms of the `C3, log (C3), log10(C3)` and `sqrt(C3)` response variables, and then use the *`par`* function to restore the plotting area to a single frame:

```
## plot histograms of the C3, log(C3) LOG10(C3) and sqrt(C3) variables
par(mfrow=c(2,2))
hist(C3,10)
log_10_C3 <- log(C3,10)
hist(log_10_C3,10)
log_C3 <- log(C3)
hist(log_C3,10)
SQRT_C3 <- sqrt(C3)
hist(SQRT_C3,10)
par(mfrow=c(1,1))
```

From the histograms, it appears that `sqrt(C3)`somewhat improves the spread of the `C3` variable.

Now we will check for possible collinearity issues by creating a new data frame object that contains only our potential predictor variables:

```
## to assess the potential for collinearity, we create a subset of
## the data containing only the potential predictor variables
subset_data <- data.frame(LAT, LONG, MAP, MAT, JJAMAP, DJFMAP)
```

We can now check for collinearity both graphically and numerically. First, we will use the `pairs` function to create a matrix of scatterplots and then we will use the `cor` function to display the correlation matrix for all of the predictor variables:

```
## use the pairs() function to plot all the variables against each other
pairs(subset_data,panel=panel.smooth)

## generate a correlation matrix using the cor() function
cor(subset_data)
```

From the scatterplots and the correlation matrix, we see that LAT and MAT have a strong negative correlation (-0.839), and so do LONG and MAP (-0.734). As discussed in the lecture, based on our scientific knowledge we expect the temperature to decrease as latitude increases and we also expect precipitation to decrease as you move from east to west across the United States. Thus, these variables will exhibit collinearity and we will want to choose the geographic pair (LAT and LONG) or the climatic pair (MAP and MAT), but not both, in our regression model.

To see how collinearity inflates the variance of a multiple regression, we first create a "full model" with the `lm` function based on all of our predictor variables:

```
## create a full regression model and summarize the results
model <- lm(SQRT_C3 ~ MAP+MAT+LONG+LAT+JJAMAP+DJFMAP)
summary(model)
```

Notice that, while the overall model is highly significant (F = 12.78, r^2=0.54, p < 0.001), only the `LAT` and the `JJAMAP` coefficients are significant at the 0.05 level. Furthermore, we can use the `vif` function to compute the variance inflation factor (VIF) values for all of the coefficients in our estimated model.

Because the `vif` function is not built into the base version of R (it is included in an add-on library known as a *car*, which stands for "Companion to Applied Regression") you will previously need to download, install, and then load the package into your R session using the `library` function:

```
## use the vif() function to calculate the variance inflation factors
library(car)
vif(model)
```

Notice that the VIF values for all of the predictor variables are well over one (some are over 5), indicating that collinearity has caused the model variance to be inflated.

<u>Part II. Selecting the "Best" Multiple Regression Model</u>

Based on the analysis in the previous section, assume that we have decided to avoid potential problems with collinearity by using only `LAT` and `LONG` as our predictor variables. Now we will create a series of linear models based on these predictors and determine which of the models has the best fit:

```
model1 <- lm(SQRT_C3~1)
model2 <- lm(SQRT_C3~LONG)
model3 <- lm(SQRT_C3~LAT)
model4 <- lm(SQRT_C3~LONG+LAT)
model5 <- lm(SQRT_C3~LONG+LAT+LONG:LAT)
model6 <- lm(SQRT_C3~LONG*LAT)
## NOTE: the LONG*LAT term is equivalent to LONG+LAT+LONG:LAT
```

The first model has no predictor variable, `model2` and `model3` are simple linear regressions based on a single predictor (`LONG` and `LAT`, respectively), and `model4` contains both of the predictor variables. The next model, `model5`, adds an interaction term between the `LONG` and `LAT` predictors, which is indicated in R by the `LONG:LAT` notation. You can also use the `LONG*LAT` notation to indicate both predictor variables and their interaction term (i.e., `LONG+LAT+LONG:LAT`), as was done for `model6` above. Thus, `model5` and `model6` are identical models.

Now that we have defined all of our regression models, we will use the *`AIC`* function to determine which model fits the data best:

```
AIC(model1,model2,model3,model4,model5,model6)
```

As discussed on page 285 of the Gotelli & Ellison text, AIC (Akaike information criterion) takes into account the number of predictor variables (i.e., parameters) when calculating the model fit. Also, somewhat counter-intuitively, a *lower* AIC value indicates a better-fitting model. Thus, the results for our models indicate that `model5`, which has the lowest AIC (-22.95), has the most information. Note that `model6`, which is identical to `model5`, has the exact same AIC value.

Now that we have identified `model5` as our best-fitting model, we can use the *`vif`* function to again test for variance inflation in the predictor variables:

```
## since model5 appears "best" so far, check its VIF values
vif(model5)
```

Yikes! Now the VIF values for all of our predictors are much greater than one, with the smallest VIF being close to 67. As mentioned in the lecture, we can "center" the predictor variables by subtracting their means in order to try to reduce the variance inflation:

```
## to reduce the VIFs, center the predictor variables
cLAT <- LAT-mean(LAT)
cLONG <- LONG-mean(LONG)
```

Next, we will create 4 new linear models based on the centered predictor values and then compute the AIC values for all of these models:

```
## run the models again, this time using the centered values
modelc1 <- lm(SQRT_C3~cLONG)
modelc2 <- lm(SQRT_C3~cLAT)
modelc3 <- lm(SQRT_C3~cLONG+cLAT)
modelc4 <- lm(SQRT_C3~cLONG+cLAT+cLONG:cLAT)

## re-compute the AIC values for the centered models
AIC(modelc1,modelc2, modelc3,modelc4)
```

As before, our model with both predictor variables and the interaction term (`modelc4` here) has the lowest AIC. Now when we compute the VIFs for this model, however, we get values that are all close to one:

```
## since the full model still appears best, recalculate the VIFs
vif(modelc4)
```

Thus, we conclude that this model (`SQRT_C3~cLONG+cLAT+cLONG:cLAT`) has the best fit for our data. We can use the *`summary`* function on this model:

```
summary(modelc4)
```

Notice that the overall regression is highly significant (F = 27.03, Adjusted r^2=0.52, $p <$ 0.001) and that all of the model parameters except `cLONG` are significant at the 0.05 level. Finally, with its adjusted r^2 value of 0.52, we can conclude that this model explains almost 50% of the variation in the response variable. Finally, we use the last lines of code to plot the best model in several 3D scatterplots.

Part III. Bayesian Multiple Regression

We begin by defining sample size, as well as assigning variables for our two predictor and our response variable (note the square root transformation on the response).

```
library(R2OpenBUGS)
n <- 73
x <- paruelo_data$LONG
w <- paruelo_data$LAT
y <-sqrt(paruelo_data$C3)
```

We then write the overall model for the multiple regression. Note that we define outputs both for the transformed data as well as on back-transformed variables in which we have removed the transformations.

```
# Write model
linreg<-function()
{
  mx <- mean(x[])   # calculate mean of the two explanatory variables
  mw <- mean(w[])
  for (i in 1:73)   # for each of the 73 sites
  {
  Y[i] <-y[i] # the dependent variable
  Y[i] ~ dnorm(mean[i], prec) # assume normal distribution
  mean[i] <- a + b[1]*(x[i]-mx) + b[2]*(w[i]-mw) + b[3]*(x[i]-mx)*(w[i]-mw)
  }
  # uninformative priors
  a ~ dnorm(0, 1.0E-6)
  for (i in 1:3)
  {
    b[i] ~ dnorm(0, 1.0E-6)
  }
  prec ~ dgamma(0.001, 0.001)
  for (i in 93:120)
  {
    predlat35[i] <- a + b[1]*(i-mx) + b[2]*(35-mw) + b[3]*(i-mx)*(35-mw)
    predlat45[i] <- a + b[1]*(i-mx) + b[2]*(45-mw) + b[3]*(i-mx)*(45-mw)
    # back-transformed prediction at latitude 35 and 45
    predrichlat35[i] <- (predlat35[i]*predlat35[i])
    predrichlat45[i] <- (predlat45[i]*predlat45[i])
  }
}
write.model(linreg, "linreg.txt")

# Bundle data
win.data <- list("x", "w","y")
# Inits function
inits <- function(){ list(a=runif(1), b=c(runif(1),runif(1),runif(1)),prec =
100)}
# Parameters to estimate
params <-
c("a","b","prec","predlat35","predlat45","predrichlat35","predrichlat45")
# MCMC settings
nc = 1
ni=1000
nb=100
nt=100
# Start Gibbs sampler
out <- bugs(data = win.data, inits = inits, parameters = params, model =
"linreg.txt",
n.thin = nt, n.chains = nc, n.burnin = nb, n.iter = ni, codaPkg=T)

library(coda)
reg.coda<-read.bugs(out)
results<-summary(reg.coda)
restat<-results$statistics
resquant<-results$quantiles
p <- 93:120
```

This final section of the script plots lines and confidence intervals of the relationship between longitude and C3 relative abundance at two different latitudes (**35N** in blue, **45N** in red) using both Bayesian and frequentist methods.

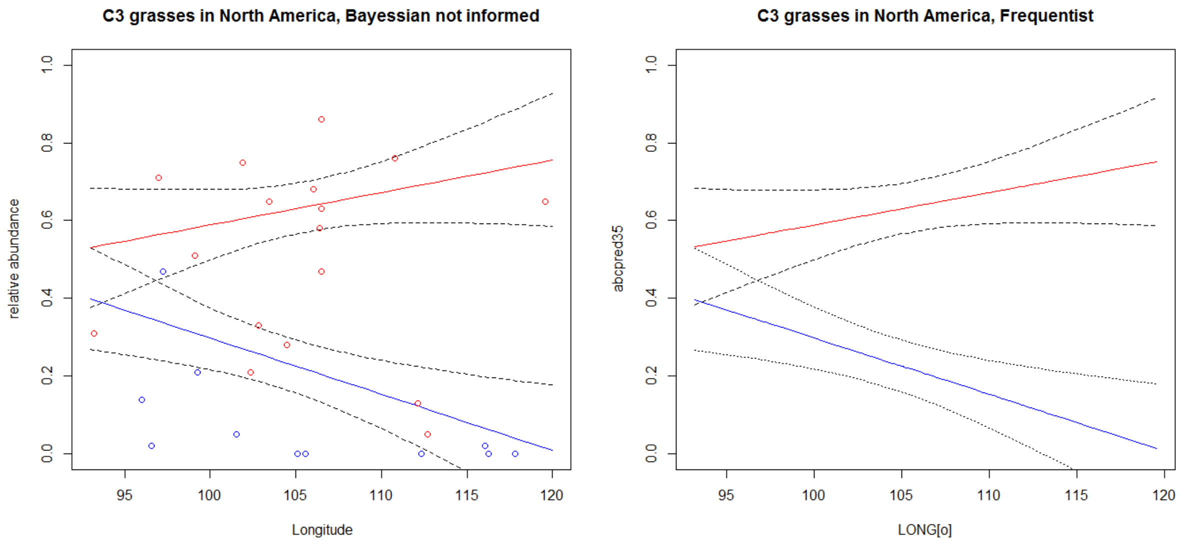
```
par(mfrow=c(1,2))
###Plotting predicted values of C3 grasses by LONG at different LATs in
Bayesian
p <- 93:120
```

```
plot(p,restat[7:34,1], type="l",ylim=c(0,1.0),main= "C3 grasses in North
America, Bayessian not informed",cex = 1.2, ylab= "relative abundance", xlab ="
Longitude", col="blue")
lines(p,resquant[7:34,1],lty=2)
lines(p,resquant[7:34,5],lty=2)
lines(p,restat[35:62,1],type = "l", col = "red", lwd = 1)
lines(p,resquant[35:62,1],lty=2)
lines(p,resquant[35:62,5],lty=2)
points(paruelo_data$LONG[paruelo_data$LAT>33 &
paruelo_data$LAT<37],paruelo_data$C3[paruelo_data$LAT>33 &
paruelo_data$LAT<37], col="blue")
points(paruelo_data$LONG[paruelo_data$LAT>43 &
paruelo_data$LAT<47],paruelo_data$C3[paruelo_data$LAT>43 &
paruelo_data$LAT<47], col="red")


###Plotting predicted values of C3 grasses by LONG at different LATs in
Frequentist
o <- order(LONG)
a<-sqrt(C3[o])
abc<-lm(a~LAT[o]*LONG[o])
abcpred35<-predict(abc, list(LAT=rep(35, length(LAT)), LONG=LONG),
type="response")
abcpred35CI<-predict(abc, list(LAT=rep(35, length(LAT)), LONG=LONG),
interval="confidence")
plot(LONG[o],abcpred35, ylim=c(0,1), type="l", col="blue",main= "C3 grasses in
North America, Frequentist")
lines(LONG[o],abcpred35CI[,2], lty=3)
lines(LONG[o],abcpred35CI[,3], lty=3)
abcpred45<-predict(abc, list(LAT=rep(45, length(LAT)), LONG=LONG),
type="response")
abcpred45CI<-predict(abc, list(LAT=rep(45, length(LAT)), LONG=LONG),
interval="confidence")
lines(LONG[o],abcpred45, col="red")
lines(LONG[o],abcpred45CI[,2],lty=2)
lines(LONG[o],abcpred45CI[,3],lty=2)

par(mfrow=c(1,1))

## detach the data
detach(paruelo_data)
```


C3 grasses in North America, Bayessian not informed
relative abundance
Longitude
C3 grasses in North America, Frequentist
abcpred35
LONG[o]