ES 207 Spring 2016

Homework 4: Linear Models Continued (ANOVAs & ANCOVAs)

Updated February 2016

Table of Contents

Homework Exercise (3 Points total)

These exercises rely on the lab data found in the resource section of CatCourses, `riparian_HW4.csv` (which was derived from the last 2 labs).

Your document should have the following sections, and provide written explanations formatted in RMarkdown that explains your code, output and graphics in the following format:

NAME
CLASS
DATE

Homework Assignment 4

Objective Statement: [What are you trying to accomplish?]

Methods: [In general terms, what analyses are you doing?]
- **Data**: [What are the data and where did they come from?]
- **Code**: [In specific terms, what is the code that was used to conduct the analysis?]

Results: [What do the results show? Numerical evidence and graphic evidence are required.]

Discussion: [What do the results mean?]

Limitations: [What are the limitations, caveats, and assumptions of the analysis?]

OBJECTIVE

You have been asked to further your analysis of riparian tree data to determine if a model can adequately characterize the amount of biomass found in these habitats. You have explored a dataset of field measurements and observations taken at various project sites throughout northern California that are intended to estimate aboveground carbon stocks in riparian areas. Some of these project sites vary in terms of the condition of the forest, and now you are being asked to determine which sites and genera vary in tree heights compared to others

To further this investigation, you extend your analysis from last week (Homework 4) that developed a subset of promising data into a linear regression, but created separate models for each genus. In order to make reasonable estimates of standing carbon, you decide to develop a full linear model that relates tree height as a function of its diameter (at breast height) or dbh and either project site or genus. In this exercise, you build a new report showing model results for the enhanced model of `ht ~ dbh * (factor)` where factor is either project site and/or genus.

Step 1 - Using ANOVA for future model selection

Load your latest riparian data frame from your last homework. Create a summary subset that uses plot (see Plot.Names) level means and standard deviations to determine if height variability differs among sites or among genera. If there is no systematic bias in height variation by site (where all sites are different from each other), use genus as your desired factor.

```
#riparian data loaded as rip
rip$projplot <- as.factor(paste(rip$ProjCode,rip$Plot.Name))

#use tapply() to cycle through each project plot and generate stats
#where 'htcm' is height in cm
ripsum <-
data.frame(cbind(tapply(rip$htcm,rip$projplot,mean),tapply(rip$htcm,r
ip$projplot,sd),tapply(rip$htcm,rip$projplot,length)))

#add column names
#(height mean, height standard deviation, number of plots)
colnames(ripsum) <- c("htcmmn","htcmsd","plot.n")

#add a projplot column (from row names) to ripsum
ripsum$projplot <- as.factor(rownames(ripsum))
```

We only want to concern ourselves with observations that have a number of plots greater than one. Additionally, let's create a new column for proj for labeling purposes:

```
#subset for plots with more than one measurement
ripsum <- ripsum[ripsum$plot.n > 1,]

#create proj column by stripping out the first 5 characters
ripsum$proj <- as.factor(substr(ripsum$projplot,1,5))
```

For loops are useful for conducting a repetitive task, however list apply (`lapply`) is much more powerful for vectorization (and also more efficient as compared to a "`for … loop`" in R). Compare using a "`for … loop`" to the utilization of the `lapply` function:

```
#create list of project sites
projlevels <- levels(ripsum$proj)

#compare a 'for' loop of summary
for (p in 1:length(projlevels)) print(summary(ripsum[ripsum$proj ==
projlevels[p],]))

#with a summary using lapply() (known as list apply)
lapply(projlevels, function(x) summary(ripsum[ripsum$proj == x,]))
```

Now use `lapply` to randomly select 6 samples plot summaries from each project site:

```
#note that there are several ways of doing this task, but
#lapply() uses a list and a function to execute list items
#>> lapply(list, function(x) myfunction(x))
#here we subset using which() where we select for x in list
#and then subset again randomly using sample()
#because these row subsets are from the vectorized data.frame
#it looks like this: data[which(),][sample(),] where [rows,cols]

nsamples <- 6
ripres <- lapply(projlevels, function(x) ripsum[which(ripsum$proj ==
x),][sample(nrow(ripsum[which(ripsum$proj == x),]),nsamples),])

# combine samples by row using rbind()
# and by calling ripres lapply function from do.call()

ripsample <- do.call(rbind,ripres)
summary(ripsample$proj)
```

Add coefficient of variation (CV) to the summary table:

```
#calculate CV using with(data,calc)
ripsample$cv <- with(ripsample, htcmsd / htcmmn)
```

Run a one-way ANOVA with CV and Project Codes (proj):

```
rip.proj.cv.aov = aov(cv~proj,data=ripsample)
summary(rip.proj.cv.aov)
#compare it against
summary.lm(rip.proj.cv.aov)
#what are the differences?
```

And check for significant differences between sites using the Tukey test:

```
rip.aov.hsd <- TukeyHSD(rip.aov)
rip.aov.hsd
```

Is there a significant difference between sites? In other words, does height variability -- as measured by the coefficient of variation -- differ between project? What metrics do you use to decide?

Advanced users: Repeat the random sampling and anova process using genus instead of project site (ie. create `rip.genus.cv.aov`). Compare the results.

Step 2 - ANCOVA

Run analysis of covariance using height as a function of dbh and genus as a factor (run twice: once with single terms, and once with terms plus interaction). Technically ANCOVA is a linear model, thus use the `lm()` function as opposed to `aov()`. Run summaries using both `summary()` and `summary.lm()`. How well did this model do compared to the "minimal adequate" (global) model of ht~dbh? Which of the two ANCOVAs performed better?

Install the "HH" package and library.

```
install.packages("HH")
library(HH)
```

Run `ancovaplot()` for the two model formulations (with and without interaction). Present each with your diagnostic reports for the ANCOVA runs. Decide which model is best and report your findings.

Resources

Data

- `riparian_HW4.csv` on CatCourses (Files→Homework→HW4)

References

See Crawley for sections on ANOVA and ANCOVA.

Keywords

`lm(), aov(), lappply(), do.call(), with()`