Midterm Examination

Biology 607: Introduction to Computational Data Analysis

10/16/2014

Welcome to your mid-term! I hope you enjoy. Note, in all of the questions below, there are easy not so code intensive ways of doing it, and there are longer more involved, yet still workable ways to answer them. I would suggest that before you dive into analyses, you do the following. First, breathe. Second, think about the steps you need to execute to get answer the question. Write them down. Third, for those parts of problems that require code, put those steps, in sequence, in comments in your script file. Use those as signposts to step-by-step walk through the things you need to do. Fourth, go over these steps, and see if there are any that could be easily abstracted into functions, could be vectorized, or otherwise done so that you can expend the minimum amount of effort on the problem to get the correct answer.

You have until 5pm on 10/30 to get it back to me (then enjoy Halloween - maybe you'll go as a p=0.06! Spooky!) And remember, you must work on your own for this one.

Each subquestion is worth 10 points. Partial credit will be given (a mistake early on does not have consequences later when number are wrong). 3 Points of each question are for making your code clean and easy to understand.

1 Sampling Your System

Each of you has a study system your work in and a question of interest. Give an example of one variable that you would sample in order to get a sense of its variation in nature. Describe, in detail, how you would sample for the population of that variable in order to understand its distribution. Questions to consider include, but are not limited to: Just what is your sample versus your population? What would your sampling design be? Why would you design it that particular way? What are potential confounding influences of both sampling technique and sample design that you need to be careful to avoid? What statistical distribution might the variable take, and why?

2 Data Visulization and Sample Properties

Let's get frosty! The data we will use here comes from the Commonwealth Glacier monitored by the McMurdo LTER. We're going to look at snow depths over time. You can grab the data at `http://www.mcmlter.org/queries/glacier/get_glacier_tables.jsp?glacier=Commonwealth#COMAVGLSTK` and see the metadata at `http://metacat.lternet.edu/knb/metacat?action=read&qformat=mcm&docid=knb-lter-mcm.2008&insertTemplate=0&displaymodule=entity&entitytype=dataTable&entityindex=1`. Make sure you get the data for all years from 1993 to 2014. Heck, if you come up with the right query, you can even use RCurl to never have to access a file on your computer - just pull it right from the cloud!

2.1 Visuzlize

Let's first see what's here. Take a gander at average snow depths, splitting them up by month. Note, there are a lot of different ways - from munging text strings to packages that play nicely with date formats - to

separate the date info.

2.2 Thanksiging on the Ice

It looks like November has the most data. Let's focus in on that. Show me the mean and bootstrapped confidence intervals for November in different years.

2.3 Putting it to the Test

2001 looks awfully strange. Let's test the hypothesis that it was an anomolous year relative to the rest of the dataset. But how? One way to ask if two samples differ is to calculate the bootstrapped confidence interval of the difference between take a sample from population a, do the same for population b, take their difference, then rinse and repeat to get an confidence interval on the difference. Write a function to do this, and apply it to the data across years. What is your null hypothesis, and what does the result tell you with regards to your null? Use 10000 simulated draws.

2.4 That was awfully mean of you

So, was the mean really informative? Why or why not? What sample property or properties would you want to look at instead to visually compare years? Oh? You've got an idea? Visualize it for me! With Confidence Intervals (derived via bootstrap, of course - or a formula if you can find one for the property of interest)!

What patterns do you observe now?

2.5 Extra Credit: A 0 bound? That's not Normal! (5-7 points)

So, clearly, this is not a nice normal distributed dataset. Given that we're measuring inches of snow accumulated in each year - and zero is the lower bound - what kind of distribution SHOULD we be using to describe the data in each year? Now that you've figured that out, use likelihood to fit the parameters for that distribution for each year. An extra 2 more points for visualizing the distribution and data for each year.

3 Power and α

In their 2012 paper, *S*etting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests (`http://dx.doi.org/10.1371/journal.pone.0032734`), Mudge et al outline a procedure where one uses both the type I and type II error rate to calculate a third quantity, ω. For any data set, we can calculate β given α, a sample size, a measure of effect size for an estimated parameter that we deem critical, and variation as measured in our data. Once we have obtained α and β, we can calculate ω as

$$\omega = \frac{\alpha + \beta}{2}$$

and then plot a curve of the relationship between α and ω. The value of α at the minimum value of ω is the 'optimal α' that balances type I and type II error against one another. For example, here's a plot of α versus omega for one particular test with a dashed line at the minimum value of ω to highlight the optimal value of alpha.

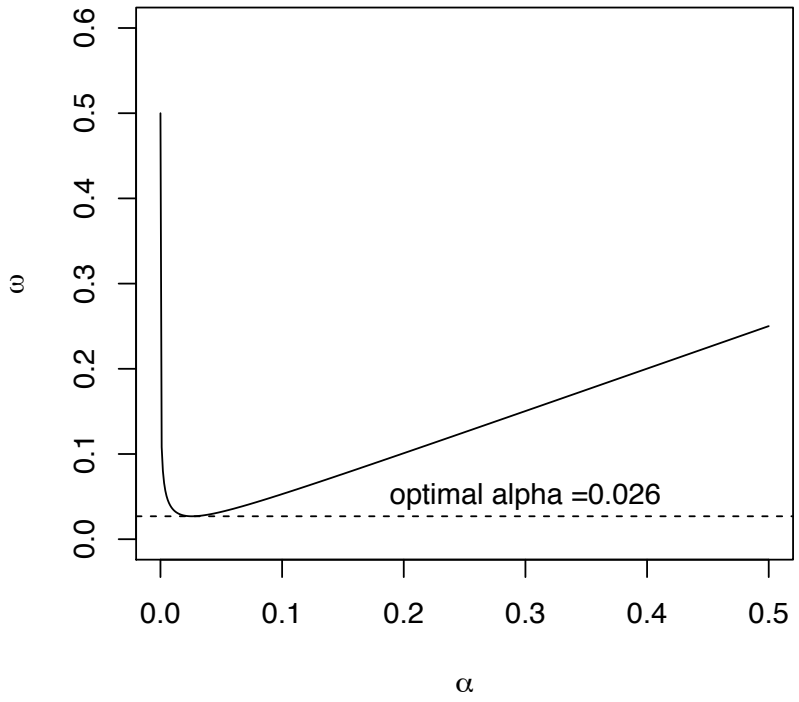
```
n<-100
set.seed(698)
z<-rnorm(n, 2, 5)

alpha <- seq(0,0.5, 0.001)
```

```r
beta <- sapply(alpha, function(x) 1-power.t.test(n, 2, sd(z),
                                                 sig.level=x,
                                                 alternative="two.sided",
                                                 type="one.sample")$power)
omega = (alpha+beta)/2

plot(omega ~ alpha, type="l",
     xlab=expression(alpha), ylab=expression(omega),
     ylim=c(0,0.6))

#the baseline
abline(h=min(omega), lty=2)
text(0.3, 0.05, paste("optimal alpha =",
                      alpha[which(omega==min(omega))], sep=""))
```



The other great property of this is that we can calculate this optimal alpha after sampling our data. We can use the variation observed in our data in the calculation of power. Only the effect size, sample size, and α levels need to be specified *a* priori.

3.1 Find your α

Let's assume you're interested in testing whether the observed temperature anomoly (the difference from the long-term average) around the globe is different from 0. To appease critics, you're assessment of a critical effect size (difference between a new temperature and the baseline) is 1.5 degrees C. You know from looking at all of your observed temperatures that the standard deviation from temperatures across the globe is 5 degrees C. Using simulation to calculate β, what is your optimal alpha for 100 sample temperature readings? Note, using functions to help you avoid heavy lifting are going to be pretty key from here on out.

3.2 Optimal α in a variable world

How does this relationship between alpha and beta change if the standard deviation across all of the temperature sensors was 10 degrees C?

3.3 Redesigning your climate sampling

Back to SD=5, how does your optimal alpha change with sample sizes from 10 to 1000?

4 Quailing at the Prospect of Linear Models

I'd like us to walk through the three different 'engines' that we have learned about to fit linear models. To motivate this, we'll look at Burness et al.'s 2012 study "Post-hatch heat warms adult beaks: irreversible physiological plasticity in Japanese quail" `http://rspb.royalsocietypublishing.org/content/280/1767/20131436.short` the data for which they have made available at Data Dryad at `http://datadryad.org/resource/doi:10.5061/dryad.gs661`. We'll be looking at the morphology data.

4.1 Fitting, Three Ways

To begin with, I'd like you to fit the relationship that describes how Tarsus (leg) length predicts upper beak (Culmen) length. Fit this relationship using least squares, likelihood, and Bayesian techniques. For each fit, demonstrate that the necessary assumptions have been met. Note, functions used to fit with likelihood and Bayes may not behave well when fed NAs.

4.2 Interpreting, Three Ways

OK, now that we have fits, take a look! Do the coefficients and their associated measures of spread match? How would we interpret the results from these different analyses differently? Or would we? Note, confint works on `lm` objects as well.

4.3 Every Day I'm Profilin'

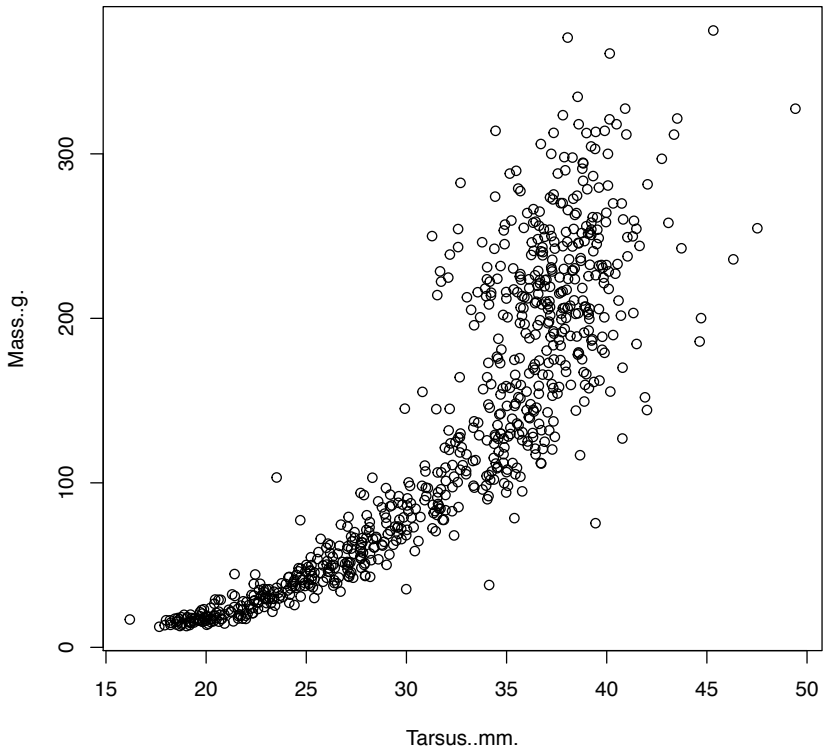
Generate the profile 95% and 75% confidence intervals by Brute Force for the slope and intercept from the Likelihood model. Check yourself against the results from the mle2 fit. The `logLik` function may help you out.

4.4 Extra Credit on Visualization (5 points)

Visually show the likelihood surface from the fit of the linear model. I honestly don't have a good answer for this one (4-D - fun!), and am curious to see what you cook up.

4.5 The world is not a straight line (no, not extra credit - but a challenge!)

One hypothesis one might have is that birds with longer legs can support more mass. Why not! But when we plot this relationship, it is distinctly nonlinear!



You actually have all of the tools you need to fit a nonlinear model to this relationship using Likelihood without any data transformation. Let's assume for the moment that the error distribution is normal and additive. Fit a model you think should work here based on the shape of the data with a normal error distribution. N.B. You might want to have the residual SD fall out from observed - fitted rather than fit an additional parameter. Not necessary, but it does simplify things a bit.

Does your resulting fit look good? Assess the qq plot and the fitted v. residuals.

What problems might still be going on here with this new fit? How might you fix them?

4.6 Extra Credit: It's not a normal world either (5 points)

OK, fit a nonlinear model properly using likelihood here such that you meet the assumptions of good model fit.

4.7 The Power of the Prior

This data set is pretty big. After excluding NAs, it's over 766 lines of data! Now, a lot of data can overhwelm a strong prior. But only to a point. Show first that there is enough data here that a prior for the slope with an estimate of 0.4 and a variance of 0.0001 is overwhelmed by the data, and produces similar results to our already fit flat prior. Second, see if a very small sample size would at least include 0.4 in it's 95% Confidence Limit. Last, demonstrate at what sample size that 95% CL first begins to include 0.4 when we have a strong prior. How much data do we really need to overcome our prior belief?

5 Extra Credit (variable)

It's election season! Hooray! Create a model that uses available public data to predict the outcome of one election. Include full documentation of how and why you created this model - what inputs it takes, and what kinds of predictions it makes in the end. Using this model, make a prediction for a race of interest. 10 points for a good model. An extra 10 if it proves correct. Here's a totally lame example using an R package that works with the Huffington Post Elections API (i.e., no, you can't do this for your answer, it's already taken!). There are others, and there are weird and wild sources of data out there. Have at it using whatever framework you want! (Seriously, there are a TON of data sources out there - have fun hunting around)

```
# For my predictions, I'll use the Huffington Post
# model with average poll data. You can dig deeper, but I won't.
library(pollstR)
```

I want to know the name of the Coakley v. Baker dataset. pollstr_polls tells me all of the polls available by different names - so let's find the indices of the one with coakley in it - I looked at `pollstr_polls(state="MA")` first to figure out where I needed to look.

```
chartIDX <- grep("coakley", pollstr_polls(state="MA")$questions$chart)[1]

chartNeeded <- pollstr_polls(state="MA")$questions$chart[chartIDX[1]]

chartNeeded

## [1] "2014-massachusetts-governor-baker-vs-coakley"

#get the chart with average polling values
#and fits from the HuffPo Model
election <- pollstr_chart(chartNeeded)
```

Now I'm going to gratuitously eyeball the data and fit some naive linear trends

```
#Let's Eyeball the data
library(ggplot2)
electionPlot <- qplot(date, value, color=choice, data=election$estimates_by_date)

#well, there are some interesting trends...
electionPlot + stat_smooth(method="lm")
```

Allright. First off, what's up with the old data? Oh, it's projected. OK, so, I'm going to filter out just the data for Martha Coakley. Then I'm going to filter it to the average model only after August, when the data becomes meaningful. Then I'm going to fit a simple linear model to it. I'll use that simple linear model to issue my prediction.

```
#ok, based on an LM, who will win on Nov 4
avg_poll <- subset(election$estimates_by_date, election$estimates_by_date$choice != "Undecided")
avg_poll <- subset(avg_poll, avg_poll$choice != "Baker")

#convert date to a continuous variable
avg_poll$day <- as.numeric(gsub("-", "", avg_poll$date))

#filter out data older than August 2014
avg_poll <- subset(avg_poll, avg_poll$day>20140801)
fitTrendCoakley <- lm(value ~ day, data=avg_poll)

#OK, where will she be on Nov 4th, 2014?
predict(fitTrendCoakley, data.frame(day=20141014), interval = c("prediction"))
```

```
##     fit   lwr  upr
## 1 45.24 42.88 47.6
```

```
predict(fitTrendCoakley, data.frame(day=20141014), interval = c("confidence"))
```

```
##     fit   lwr   upr
## 1 45.24 44.75 45.73
```

OH - it does not look good. Both the fit and prediction intervals show her losing if life continues to obey a linear trend.

But when is the world ever linear? When are all polls equally good? When is poll data the only source of truth? Have a crack at it!