OUTLINE
MARCH 2, 2015

multiple regression

-using Galapagos and Wage Gap walkthroughs

permutation test

-using Titanic, revisited walkthrough

key concepts reviewed:

* test statistic* null hypothesis* p-value

new key concepts:

* relative risk* odds ratios* permutation tests in 2x2 contingency tables.

FEEDBACK

1. Scripts from HW Exercises now on website::
   -his take on the problem.
   -not necessarily the only way to go about the problem
   -compare what you have done with what he did
   -do not share with classmates taking this course in the future

2. Work Load
   -should be three:one ratio of "work done outside:hours in the classroom"
   -some of the feedback wasn't necessarily time spent, but about the level
   -his aim is a design of frustration, but not in the way that you need to memorize all the commands for R. The course is about analyzing data, interpreting results, etc. (these sorts of things are good to be frustrated on)

3. Connecting the Reading/ what is coming out of lectures to the R script
   -it is really important that you find this challenging
   -ask questions if one of these connections is not being made

CLASS LECTURE

Multiple Regression

Galapagos Island Walkthrough:
Three pieces of information (Size of island, how many species on island, and elevation)

Questions we ask:

How would double the size of the tallest mountain change the number of species on the island?
-we must hold everything else (size/area on island) constant to answer this question correctly
-we cannot get these effects just by running a simple linear regression model

Lets prove that a simple linear regression model will not work::
-power law:
-area-only model:
species = e^29 (Area)^.39
slope = .388

-elevation-only model:
species= e^-23 + (Elevtion)^1.1

-these are correct descriptions of how one variable changes as a function of another
BUT it does not answer our question because we are not holding the other variables constant (it is a failure to adjust for confounders)

-so, to fix this, lets run a Multiple Regression Model:

log(species) = Bo + [B1 * log(area)] * [B2 * log(elevation)]

both B1 and B2 terms are
-we are fitting a model for the vertical dimension using two different 2D variables (makes a point cloud in 3D)(finds us the best fitting plane using least squares to relate log species to the two variables on the right).
log(species) = Bo + [B1 * log(area)] * [B2 * log(elevation)]
B1= .48
B2= -.31

*note the elevation coordinate flipped from a positive to a negative. Shows that elevation and area were confounding for each other

Wage Gap Walkthrough:
-assess whether there is a wage gap between men and women once you have taken out other variables (have to make sure all else is equal)

-we must adjust for all other qualifications (educational experience)

Variables in the data set are :

Salary (annual salary in dollars)
Education (years of grad school)
Experience
Months (of work experience)
Sex

Without taking into account confounders (such as experience)

-we see a boxplot that looks like females earn a higher salary than males

But we must take into account the experience variable

-once we adjust for experience, it looks like the men are earning a higher salary
-this model (as well) suggests that there may be a wage gap (but it is actually proving it the other way)

Need a multiple regression model to handle this:

-we need to adjust for experience and for education
-look at residuals for education, sex, and experience all in one model
-here is the equation

Salary = Bo + Bi * Experience + B2 * Education + B3 * 1{Male}
43,000 + 439 * Experience + 1533 * Education + 3544 *1{Male}

-how do we interpret this?
-the coefficient of the dummie variable for Sex tells us that males would be on average paid $3,544 more than a female, if we were to hold experience and education constant

- note confidence intervals and bootstrapping all carry over to multiple regression models

Permutation Test

-Titanic, Revisited Walkthrough

-answers question of "how do we know whether a given effect is statistically significant?"

-contingency table:
-lets do a permutation test to make it very obvious there is an effect
-the Null Hypothesis is that association due to survival based on sex is that it was completely random.

-relative risk:

-ratio between Probability of dying if male: Probability of dying if female
-if you are three times as likely to die as a male, the relative risk = 3
-if there is no association between sex and dying, the relative risk should be 1
-so our null hypothesis should be a ratio of one.

-Step 1:

-compute relative risk (you can do this by hand)

-divide probability of dying as male (.8) by risk as a female (.27)
-relative_risk = 2.96

Permutation Test comes in "Shuffle Test":

EX) sesame street and army = perfect association between the two sets. But, if you don't want it perfect, you can just shuffle one of the sets of cards → you have just broken any association. Could have been the case that your shuffled cards could come out as " ABCDE" and it would look like you have a perfect association.

-so what does this mean in the titanic walkthrough,
-if you were to shuffle the "sexes," a relative risk of 2.96 easily could have arisen
-if there were an association originally between the two variables, I would be breaking it here (by shuffling the cards)
-with a permutation test =
-the association between sex and survival status is broken
-we now see an association much closer to 1

```
permtest1 = do(1000)*{
  t1_shuffle = xtabs(~shuffle(sex) + survived,
data=TitanicSurvival)
  relrisk(t1_shuffle)
}
head(permtest1)
```

-we find that it is centered around 1
-this is the range of plausible set of numbers around the Null Hypothesis
- we didn't get a relative risk of 1 every time (bc there is chance involved)
-when you shuffle the cards, we see that the null hypothesis is probably wrong (unless we are seeing an absurd miracle here)