Mary Claire Adams
2 March 2015

OVERVIEW FROM WEBSITE

Today we finish up our discussion of multiple regression. Then we will introduce the idea of a permutation test in the context of the Titanic, revisited walkthrough. This will review some key concepts from your first statistics class, including:

- Test statistic
- Null hypothesis
- P-value

**Find this information in an old statistics book or in chapter 4.3 of OpenIntro: Statistics found under the Resource tab

This will also involve some new concepts, including:

- Relative risk
- Odds ratios
- Permutation tests in 2x2 contingency tables

ORGANIZATION OF NEXT TWO CLASSES

1. Multiple regression (return to the 3-D plot from gala and salary gap data)
2. New material on hypothesis testing (in the context of the simplest test = 2 by 2)
3. Permutation
4. You will be expected to be familiar with hypothesis testing (look over 4.3 – understand to this level in the context of regression models)
5. Next Monday – review (structured for the first half and questions for the second half)

TEST – You a pen a blue book that's all you need

FEEDBACK REVIEW

- Common themes
    1. Request to put R-script up on the web (posted on the website now)
        a. Caveat – there are many ways to approach these problems. What you see on these scripts is "my" approach to the product.
        b. Do not look at what I have done and take it for gospel
        c. Treat this as the beginning of a conversation
        d. Please to do not share these with people who may be taking this course in the future

2. Volume of Work
    a. General rule of thumb (reading, walkthroughs…) should be a 3 to 1 ratio. For every hour we are in class you should be spending 3 hours outside of class working with the material
    b. It is good to be pushed to the edge of your intelligence
    c. Please send me emails about R commands (it will take me 90 seconds) – if you come to a situation where you already know what you want to do but you cannot figure out how to do it in R please ask
3. Concern between connecting the readings and R
    a. This is a challenge I want you to face
    b. Ask – other people will have the same question
4. Perceived the workings of the course to bifurcate
5. Please come ask questions – there is nothing more important to me than that you learn

MULTIPLE REGRESSION

Gala data – How would change affect the area holding everything else constant?

- We can't do this with a one variable regression model

  ```
  lm2 = lm(log(Species) ~ log(Area), data = gala)
  abline(lm2)
  summary(lm2)
  ```

- Now species versus elevation

  ```
  lm3 = lm(log(Species) ~ log(Elevation), data = gala)
  ```

  Species = e^ 2.9 (Area^.39)
  Species = e^-2.3 (Elevation^1.1)

  This is wrong in regard to holding everything else constant. This is a failure to adjust for confounders (log elevation and log area are high correlated with each other).

- Now try with multiple regression

  ```
  Lm10 = lm(log(Species) ~ log(Area) + log(Elevation), data = gala)
  ```

- Log species = Bo + B1(log(area)) + B2(log(Elevation))

  B1 = .48
  B2 = -.31

- 3-D graph shows the relationship between the multiple regression model
    - We are taking this information to create a plane
    - A least squares plane
    - Also called a hyper plane if you have more than two variables on the right hand side
    - If we want to adjust for the systematic effect of area we must rotate the plane to look at it where Area has no variation (negative slope)
    - If you want to look at the regression model adjusted for elevation then you will notice that the slope is positive

WAGE GAP WALKTHROUGH

The variables in the data set are:
* Salary: annual salary in dollars
* Education: years of post-second education
* Experience: months of experience at the particular company
* Months: total months of work experience, including all previous jobs
* Sex: whether the employee is male or female

```
#distibution of salary by sex
mean(Salary~Sex,data=salary)
##       0       1
## 62610.45 59381.90
```

0 = female
1 = male

If there were more dummy variables they would have to be coded differently

```
boxplot(Salary~Sex,data=salary, names=c("Female", "Male"))
```

```
#does the story change if we adjust for work experience
plot(Salary~Experience, data=salary)
```

```
lm1 = lm(Salary~Experience, data=salary)
summary(lm1)
boxplot(resid(lm1)~salary$Sex)
```

this is

Remember what the residuals represent = the y adjusted for the x variable
Now we want to see effect of adjusting for Education
This is evidence for a wage gap.

```
#Multiple-regression model that accounts for education
lm2 = lm(Salary~Experience+Education, data=salary)
summary(lm2)
boxplot(resid(lm2)~salary$Sex)
```

Lets try to adjust for multiple predictors.
The more plausible thing to do would be to add sex to the model

```
#Build a model that accounts for both these factors and includes a dummy variable for
the sex of the employee
lm3= lm(Salary~Experience+Education+Sex, data=salary)
summary(lm3)

Call:
lm(formula = Salary ~ Experience + Education + Sex, data = salary)

Residuals:
    Min      1Q   Median      3Q     Max
-18002.9  -5330.1   -293.9   7276.1  20560.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42922.4     6878.0   6.241  2.4e-07 ***
Experience     439.4      135.0   3.255  0.00235 **
Education     1533.8     1435.7   1.068  0.29196
Sex           3544.6     3525.3   1.005  0.32087
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9093 on 39 degrees of freedom
Multiple R-squared:  0.2641,  Adjusted R-squared:  0.2075
F-statistic: 4.665 on 3 and 39 DF,  p-value: 0.007029
```

Thus:

Salary = 43,000 + 439(Experience) + 1533(Education) + 3544(1{Male})

439 and 1533 are partial slopes
3544 is a coefficient on a dummy variable

```
#based on this model - men make $3544 more per year than similarly qualified women
#however, standard error is $3525 which makes it hard to rule out 0
```

```
confint(lm3)
#this confidence interval includes 0

#Check with a bootstrap sample
myboot = do(1000)*{
  lm_boot = lm(Salary~Experience+Education+Sex, data=resample(salary))
  coef(lm_boot)
}

hist(myboot$Sex)
confint(myboot)
confint(lm3)
                 2.5 %     97.5 %
(Intercept) 29010.3689 56834.3760
Experience    166.3287   712.5244
Education   -1370.2657  4437.8634
Sex         -3586.0132 10675.2388
```

This is our best guess 3500; however, our confidence interval is large that it indicates that it is not a very effective model. This value could vary from -3500 to 10,000.

Is there a point where you look at the two variable plot and notice that there is no correlation? This is a good visual tool; however, using just a linear plot is crude and should not be the only measure of testing correlation

TITANIC TEST - Permutation tests in 2x2 contingency tables

```
head(TitanicSurvival)
##                                  X survived    sex     age passengerClass
## 1   Allen, Miss. Elisabeth Walton      yes female 29.0000            1st
## 2  Allison, Master. Hudson Trevor      yes   male  0.9167            1st
## 3    Allison, Miss. Helen Loraine       no female  2.0000            1st
## 4 Allison, Mr. Hudson Joshua Crei       no   male 30.0000            1st
## 5 Allison, Mrs. Hudson J C (Bessi       no female 25.0000            1st
## 6             Anderson, Mr. Harry      yes   male 48.0000            1st
```

Null would be that there is no association what so ever. What do we mean by association? Can we reduce association to a single number?

Relative risk = Probability of dying if male/Probability of dying if female

Thus, the null hypothesis for relative risk should be 1

```
t1 = xtabs(~sex + survived, data=TitanicSurvival)
```

prop.table(t1, margin=1)

```
##        survived
## sex            no       yes
##   female 0.2725322 0.7274678
##   male   0.8090154 0.1909846
```

0.8090154/0.2725322 = 2.968513

or use the relrisk function

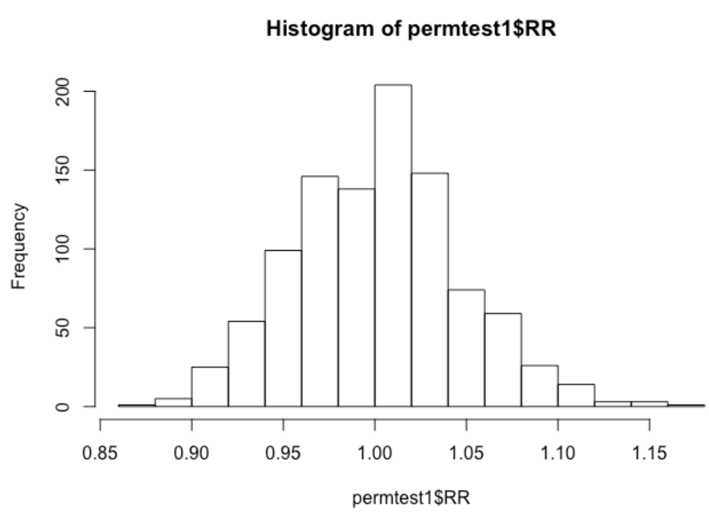
How do we operationalize the operation of random chance?

Analogy of shuffling the cards
There are two alphabet systems (Sesame street vs. military). If I put both of these up there would be perfect association between the two sets. If I wanted to break the association just shuffle the Sesame street alphabet. It is possible that you could have shuffled the cards to get perfect association.

t1_shuffle = xtabs(~shuffle(sex) + survived, data=TitanicSurvival) relrisk(t1_shuffle) ## [1] 1.021879

The relative risk is much closer to 1

permtest1 = do(1000)*{ t1_shuffle = xtabs(~shuffle(sex) + survived, data=TitanicSurvival) relrisk(t1_shuffle) } head(permtest1) ## RR ## 1 0.9529538 ## 2 1.0108670 ## 3 1.0558642 ## 4 1.0443738 ## 5 0.9892993 ## 6 0.9529538 hist(permtest1$RR)



This is the range of plausible values we got under the null hypothesis. Therefor the null hypothesis is probably wrong considering we got ~2.6

What is the range for your test statistic then you check to see if it is consistent with the range of null hypothesis

PARTY AFILIATOIN SCRIPT

```
t1 = xtabs(~Sex + Party, data=partyaffil)
prop.table(t1, margin=1)
relrisk(t1)

Party_shuffle = data.frame(shuffle(partyaffil$Sex), partyaffil$Party)
head(Party_shuffle)

permtest1 = do(1000)*{
  t1_shuffle = xtabs(~shuffle(Sex) + Party, data=partyaffil)
  relrisk(t1_shuffle)
}
head(permtest1)

hist(permtest1$RR)
```

The histogram shows us the association of relative risk