The Final Chapter

Kathleen M. Saul

May 17, 2011

Clarification: Multiple Regression

- R returns both a multiple r^2 and an adjusted r^2 for regression models
  - For simple linear regression (one independent variable), use the multiple r^2 value to determine the percent of variation in y explained by the model.
  - For multiple regression, use the adjusted value of r^2. It takes into account the number of independent variables in your model as well as the number of observations to yield a more accurate value of r^2.

Last Week: Two Way ANOVA

- Original data take the form of a table:

		Factor A		
		A	B	C
	X			
	Y			
Factor B	Z			

- Hypotheses:
  - For the columns:
    - Ho: muA = muB = muC . . .
    - Ha: At least two means differ
  - For the rows (Groups or Blocks):
    - Ho: muX = muY = muZ . .
    - Ha: At least two means differ

- In R: anova(lm(measure~column+row)
- If the p-value for rows or for columns < alpha, perform pairwise comparisons to determine which pairs differ
  - pairwise.t.test(measure,column,
    + p.adj="bonferroni"), Or
  - pairwise.t.test(measure,row,p.adj="bonferroni")

Factorial ANOVA

- Data take the form of a table, with multiple measures for each combination of Hypotheses:
  - For the columns:
    - Ho: muA = muB = muC . . .
    - Ha: At least two means differ
  - For the rows (Groups or Blocks):
    - Ho: muX = muY = muZ . .
    - Ha: At least two means differ

- For the interactions:
  - Ho: The column and row factors do not interact to affect the mean responses
  - Ha: The column and row factors do interact to affect the mean responses
- In R:
  anova(lm(measure~factorA+factorB+factorA*factorB)
- If the p-value for rows or columns< alpha, perform pairwise comparisons to determine which pairs differ

What if my table contains counts and not measurements?

- Example: What if we had the count of the number of trees harvested in given county, in a given year?

	Douglas Fir	Western Hemlock	Cedars	Other Conifer	Red Alder	Other Hardwood
Forest Industry	10109	12254	1998	2866	3352	501
Private Large	10640	20622	1300	2499	2687	740
Private Small	3640	4128	822	10132	1557	2365
State	21984	19235	5495	3295	3845	1100
National Forest	1440	2693	167	737	117	0

Adapted from DNR harvest data, 2000

Use Chi Squared! X^2

- $X^2 = \Sigma \{(\text{observed} - \text{expected})^2 / \text{expected}\}$
  - Expected value = $\frac{(\text{column total} * \text{row total})}{\text{sample size}}$
- df = (rows – 1) * (columns – 1)
- Hypotheses:
  - Ho: The two variables are independent
  - Ha: The two variables are not independent (one affects the other)

- Transform the data by removing the column and row headings (but remember what they are!):

10109	12254	1998	2866	3352	501
10640	20622	1300	2499	2687	740
3640	4128	822	10132	1557	2365
21984	19235	5495	3295	3845	1100
1440	2693	167	737	117	0

- In R: Read in the data and assign the column and row names:

```
 > cut<-read.csv("Timber.csv",sep=",",header=F)
> colnames(cut) <-
c("DougFir","Hemlock","Cedar","OtherFir","Alder","OtherHard")
> rownames(cut)<-
c("Industry","LgPriv","SmPriv","State","NatlFor")
```

```
> cut
          DougFir  Hemlock  Cedar  OtherFir  Alder  OtherHard
Industry    10109    12254   1998      2866   3352        501
LgPriv      10640    20622   1300      2499   2687        740
SmPriv       3640     4128    822     10132   1557       2365
State       21984    19235   5495       329   3845       1100
NatlFor      1440     2693    167       737    117          0
```

- > chisq.test(cut)

Pearson's Chi-squared test

data: cut

X-squared = 37214.60, df = 20, p-value < 2.2e-16

(6 – 1) * (5 – 1)

- Since the p-value < alpha = 0.05, we reject the null that there is no association between the two variables

- Now we look at the pairwise X^2 values
- First, calculate the expected values:

```
> chisq.test(cut)$expected
```

	DougFir	Hemlock	Cedar	OtherFir	Alder	OtherHard
Industry	9755.961	12024.728	1995.9596	3984.7776	2358.3419	960.2316
LgPriv	12081.321	14890.854	2471.7018	4934.5598	2920.4589	1189.1054
SmPriv	7107.915	8760.873	1454.1991	2903.1951	1718.2205	699.5973
State	17249.971	21261.483	3529.1493	7045.6714	4169.8945	1697.8304
NatlFor	1617.832	1994.062	330.9902	660.7961	391.0841	159.2353

- Next, recall the observed values:

```
> chisq.test(cut)$observed
          DougFir Hemlock Cedar OtherFir Alder OtherHard
Industry    10109   12254  1998     2866  3352       501
LgPriv      10640   20622  1300     2499  2687       740
SmPriv       3640    4128   822    10132  1557      2365
State       21984   19235  5495     3295  3845      1100
NatlFor      1440    2693   167      737   117         0
```

- Calculate the contribution to the overall X^2 statistic for each cell:

```
> E<-chisq.test(cut)$expected
> O<-chisq.test(cut)$observed
> (O-E)*(O-E)/E
```

	DougFir	Hemlock	Cedar	OtherFir	Alder	OtherHard
Industry	12.78	4.37	2.0914e-03	314.11	418.67	219.63
LgPriv	171.95	2205.79	5.55e+02	1202.13	18.66	169.62
SmPriv	1691.98	2449.93	2.75e+02	17999.35	15.13	3964.52
State	1299.19	193.15	1.10e+03	1996.62	25.31	210.50
NatlFor	19.5	244.98	8.13e+01	8.79	192.09	159.24

Cells with low contributions warrant further investigation.

Approach to Data Analyses

- Look at the format of the raw data, think about what tests might be appropriate:
  - Do you have columns? How many?
    - One → z test (given sigma) or t test (given s)
    - Two → t test of independent samples or matched pairs t test
    - Three or more → ANOVA
  - Is it a table? Are the values in the table numeric or categorical?
    - Numeric → ANOVA
    - Categorical → X^2
  - Do you have cause and effect data?
    - → Regression

- Plot the data
- For numeric data:
  - Histogram
    - For a single numeric variable
    - In R: hist(x,freq=F)
      curve(dnorm(x),add=T)
  - Normal Quartile/Quantile plot
    - In R: qqnorm(x)
      qqline(x)
  - Are the data normally distributed? If not, would a transformation result in a normal distribution?
    - Most statistical tests assume an underlying normal population and thus, a normally distributed sample.

For categorical data:

- Box plot
  - For categorical variables
  - In R: boxplot(x)
  - Or, for parallel plots: boxplot(a,b)
- Look for the presence of outliers and for the relative positions of medians and quartiles

- Generate Descriptive Statistics
  - Mean = the arithmetic average
    - For a single variable/vector, in R: mean(x)
    - For one variable of a matrix, in R: mean(name$x)
  - Median = the midpoint of ordered data
    - In R: median(x) or median(name$x)
  - Standard Deviation = the variation around the mean
    - In R: sd(x) or sd(name$x)
  - Summary
    - In R: summary(dataframe)

- Establish the hypotheses to be tested with your data (Ho and Ha)
- Run the tests to generate the statistics and associated p-values
  - If the p-value is less than your alpha, reject the null hypothesis
- R commands:
  - One sample t test:
    - t.test(y,mu=value)
  - T test of two independent samples:
    - t.test(a,b)

- Paired two sample t test:
  - t.test(a,b,paired=T)
- Simple Linear Regression:
  - lm(y~x)
- Multiple Regression:
  - lm(y~a+b+c+ . . . +a*b*c* . . .)
- One Way ANOVA:
  - anova(lm(measure~category)
- Two Way ANOVA:
  - anova(lm(measure~category+category))

- Factorial ANOVA:
  - anova(lm(measure~a+b+c+a*b*c))
- Chi Squared test:
  - chisq.test(table)

- Interpret the results!
  - What does rejecting or not rejecting Ho tell you about your research question?

- **Questions?**