

Introduction to Data Analysis Using R

Maria L. Rizzo
Department of Mathematics & Statistics
Bowling Green State University

January 25, 2010
Center for Family and Demographic Research

R Software

- ▶ R is free software - open source
- ▶ Distributed under GNU General Public License Version 2
- ▶ Program and source code available at www.r-project.org [CRAN link]
- ▶ Can be installed to a portable storage device
- ▶ Over 2000 add-on packages can be installed via menu
- ▶ Basic distribution comes with several Recommended packages pre-installed
- ▶ Extensive free documentation installed with program

Outline

Introduction

Getting Started

Data Analysis Example 1: lunatics data

- How to import data from a file
- Exploratory data analysis
- Regression
- Details of Regression Analysis
- Plots

Data analysis Example 2: UCB admissions data

- Objects and Methods in R
- Student Admissions at UC Berkeley

Resources

History

- ▶ The design of R has been heavily influenced by two existing languages: S and Scheme
- ▶ Language is very similar in appearance to S (Becker, Chambers, & Wilks)
- ▶ Implementation and semantics derived from Scheme (Sussman)
- ▶ The core of R is an interpreted computer language
- ▶ Branching and looping, modular programming using functions
- ▶ Additional modules (add-on packages) are available

Why is R named R?

- ▶ Based on the (first) names of the first two R authors (Robert Gentleman and Ross Ihaka)
- ▶ A play on the name of the Bell Labs language 'S'
- ▶ Since mid-1997 there has been a core group (the 'R Core Team') who can modify the R source code archive.

S, R, and Splus

- ▶ R is an open source version of S.
- ▶ S-Plus is a commercial version of S sold by Insightful Corporation, which in 2008 was acquired by TIBCO Software Inc.
- ▶ R offers several graphics features that S-Plus does not.

R GUI Interface

- ▶ Type commands at the command prompt:
 - ▶ Press the up-arrow key to recall commands and edit them.
 - ▶ Use Esc (Escape) key to cancel a command.
- ▶ R scripts: (in plain text .R files)
 1. The script editor in the R GUI.
 2. Paste lines from script to console - can use Ctrl-R.
 3. Source the whole file.

Using the R help system

1. Use the ? operator if you know the keyword: ?seq
2. Use ?? for a wildcard search: ??normal
3. Use help.search() e.g. help.search("binomial")
4. Use html help from the menu
5. data() - a list of data sets available
6. Help files that describe data: ?iris ?sleep ?volcano

Tips for using R Online Help System

- ▶ Classical statistical methods and inference: try e.g.
`help.search("Test", package="stats")`
- ▶ Run the examples in a help file: `example(pairs)`
`example(volcano)`
`example(persp)`
- ▶ Chapter 8 of R-Intro manual

More Probability Functions

1. The `sample` function
2. The `quantile` function

```
> x = sample(1:6, size=10, replace=TRUE)
> x
[1] 5 5 5 5 4 1 4 2 3 3
> quantile(x, .9)
90%
5
```

Probability Functions

Prefix p,q,r,d plus the short name of the distribution.

```
> pnorm(2) # P(Z < 2)
[1] 0.9772499
> qnorm(.95) # 95th percentile
[1] 1.644854
> rnorm(3, 0, 2) # 3 random N(0,2) variates
[1] 2.062226 3.954951 -0.815555
> dnorm(0) # std. normal density at 0
[1] 0.3989423
```

See the R-Intro manual, Chapter 8.

```
> pf(4, 2, 3) #lower tail area of F(2,3) at x=4
[1] 0.8575728
```

DASL *Massachusetts Lunatics* data

<http://lib.stat.cmu.edu/DASL/Datafiles/lunaticsdat.html>

These data are from an 1854 survey conducted by the Massachusetts Commission on Lunacy.

(14 cases, each county is one case)

- ▶ NBR = Number of lunatics, by county
- ▶ DIST = Distance to nearest mental health center
- ▶ POP = County population, 1950 (thousands)
- ▶ PDEN = County population density per square mile
- ▶ PHOME = Percent of lunatics cared for at home

Importing the *lunatics* data into R

- ▶ Data on web is space delimited.
- ▶ Copy-paste into plain text file.
- ▶ Use `read.table` function to read file into data set.
- ▶ Assign the result to a data name:

```
lunatics = read.table('lunatics.txt', header=TRUE)
```

- ▶ `?read.table` for list of options

lunatics data after import

```
> lunatics
      COUNTY NBR DIST    POP PDEN PHOME
1  BERKSHIRE 119  97  26.656  56   77
2  FRANKLIN  84  62  22.260  45   81
3  HAMPSHIRE 94  54  23.312  72   75
4   HAMPDEN 105  52  18.900  94   69
5  WORCESTER 351  20  82.836  98   64
6  MIDDLESEX 357  14  66.759 231   47
7    ESSEX  377  10  95.004 3252  47
8  SUFFOLK  458   4 123.202 3042   6
9  NORFOLK  241  14  62.901  235  49
10 BRISTOL  158  14  29.704  151  60
11 PLYMOUTH 139  16  32.526   91  68
12 BARNSTABLE 78  44  16.692   93  76
13 NANTUCKET  12  77   1.740  179  25
14    DUKES  19  52   7.524   46  79
```

How to import data from comma-delimited files

- ▶ Spreadsheet data can be saved in `.csv` format (comma delimited)
- ▶ Retain only the data and labels.
- ▶ Use `read.table` function with `sep=','`

```
> lunatics = read.table('lunatics.csv', header=TRUE, sep=',')
> head(lunatics)
```

	COUNTY	NBR	DIST	POP	PDEN	PHOME	RDIST
1	BERKSHIRE	119	97	26.656	56	77	0.01030928
2	FRANKLIN	84	62	22.260	45	81	0.01612903
3	HAMPSHIRE	94	54	23.312	72	75	0.01851852
4	HAMPDEN	105	52	18.900	94	69	0.01923077
5	WORCESTER	351	20	82.836	98	64	0.05000000
6	MIDDLESEX	357	14	66.759	231	47	0.07142857

Summary of *lunatics* data after import

```
> attach(lunatics)
> summary(lunatics)
```

COUNTY	NBR	DIST
BARNSTABLE:1	Min. : 12.0	Min. : 4.00
BERKSHIRE :1	1st Qu.: 86.5	1st Qu.:14.00
BRISTOL :1	Median :129.0	Median :32.00
DUKES :1	Mean :185.1	Mean :37.86
ESSEX :1	3rd Qu.:323.5	3rd Qu.:53.50
FRANKLIN :1	Max. :458.0	Max. :97.00
(Other) :8		

POP	PDEN	PHOME
Min. : 1.74	Min. : 45.00	Min. : 6.00
1st Qu.: 19.74	1st Qu.: 76.75	1st Qu.:47.50
Median : 28.18	Median : 96.00	Median :66.00
Mean : 43.57	Mean : 548.93	Mean :58.79
3rd Qu.: 65.79	3rd Qu.: 218.00	3rd Qu.:75.75
Max. :123.20	Max. :3252.00	Max. :81.00

RDIST

How to add a variable to data frame

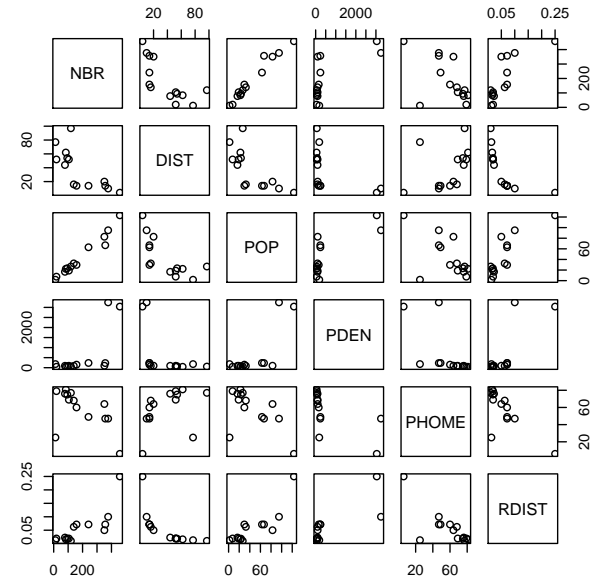
```
> RDIST = 1/DIST
> lunatics = data.frame(lunatics, RDIST)
> head(lunatics) #first few obs.
  COUNTY NBR DIST   POP PDEN PHOME   RDIST
1 BERKSHIRE 119  97 26.656  56   77 0.01030928
2 FRANKLIN  84  62 22.260  45   81 0.01612903
3 HAMPSHIRE 94  54 23.312  72   75 0.01851852
4  HAMPDEN 105  52 18.900  94   69 0.01923077
5 WORCESTER 351  20 82.836  98   64 0.05000000
6 MIDDLESEX 357  14 66.759 231   47 0.07142857
  RDIST.1
1 0.01030928
2 0.01612903
3 0.01851852
4 0.01923077
5 0.05000000
6 0.07142857
```

Correlations

```
> corrs = cor(lunatics[,-1]) #omit the county name
> round(corrs, 3)
      NBR   DIST   POP   PDEN   PHOME   RDIST   RDIST.1
NBR    1.000 -0.740  0.980  0.695 -0.577  0.794  0.794
DIST  -0.740  1.000 -0.718 -0.478  0.412 -0.689 -0.689
POP    0.980 -0.718  1.000  0.770 -0.606  0.846  0.846
PDEN   0.695 -0.478  0.770  1.000 -0.638  0.793  0.793
PHOME  -0.577  0.412 -0.606 -0.638  1.000 -0.758 -0.758
RDIST  0.794 -0.689  0.846  0.793 -0.758  1.000  1.000
RDIST.1 0.794 -0.689  0.846  0.793 -0.758  1.000  1.000
```

Scatterplots using pairs

```
> pairs(lunatics[,2:7])
```



Simple Linear Regression

Fit model and display the coefficients of the regression line: estimates are printed.

```
> lm(PHOME ~ RDIST)
```

Call:

```
lm(formula = PHOME ~ RDIST)
```

Coefficients:

```
(Intercept)      RDIST
      73.93      -266.32
```

summary

The summary displays more than the default print method for `lm`:

```
> summary(lm(PHOME ~ RDIST))
Call:
lm(formula = PHOME ~ RDIST)

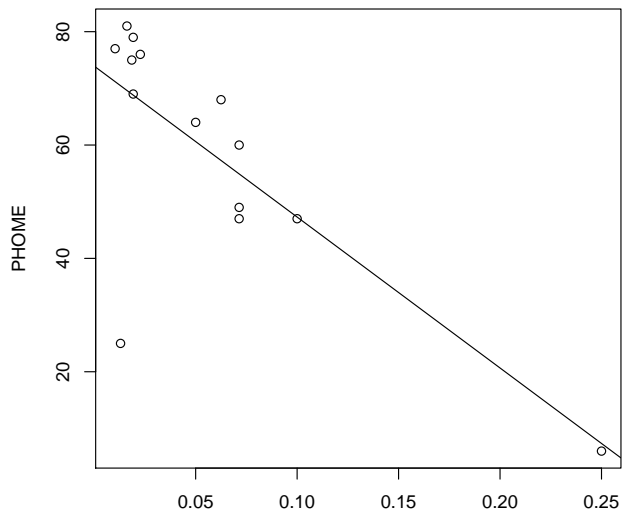
Residuals:
    Min       1Q   Median       3Q      Max
-45.468  -1.083   4.243   7.596  11.369

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  73.927      5.493  13.459 1.33e-08 ***
RDIST       -266.324     66.211  -4.022 0.00169 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.97 on 12 degrees of freedom
Multiple R-squared:  0.5742, Adjusted R-squared:  0.5387
F-statistic: 16.18 on 1 and 12 DF,  p-value: 0.001692
```

Display the fitted regression line

```
> plot(RDIST, PHOME)
> abline(L$coef)
```



The `lm` object

- ▶ Save the fitted model in a `lm` object.
- ▶ Object contains `$fitted.values` `$residuals` `$coefficients` etc.
- ▶ `plot` will display residual plots.

```
> RDIST = 1/DIST
> L = lm(PHOME ~ RDIST)
> summary(L)$coef
            Estimate Std. Error  t value    Pr(>|t|)
(Intercept)  73.92659    5.492683  13.459104 1.333396e-08
RDIST       -266.32414   66.211043  -4.022352 1.692291e-03
> anova(L)
Analysis of Variance Table

Response: PHOME
          Df Sum Sq Mean Sq F value    Pr(>F)
RDIST      1 3624.3   3624.3   16.179 0.001692 **
Residuals 12 2688.1     224.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Identify the outlier

The `identify` function waits for user to identify `n` points on the plot, and optionally labels the points.

```
identify(RDIST, PHOME, n=1, labels=COUNTY)
[1] 13
```

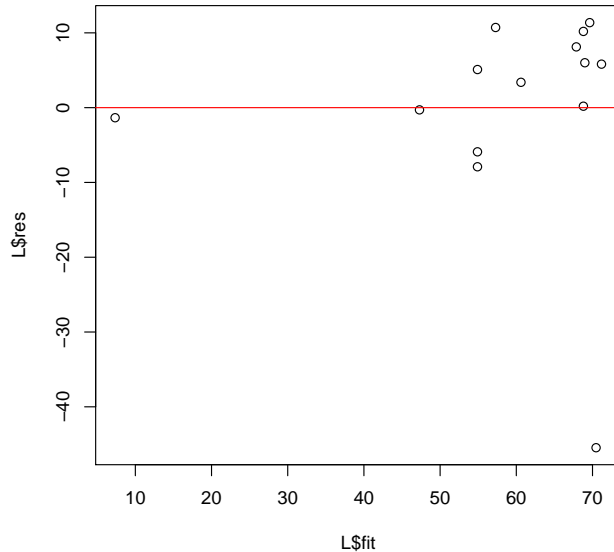
This corresponds to NANTUCKET county:

```
> lunatics[13,]
      COUNTY NBR DIST  POP PDEN PHOME      RDIST  RDIST.1
13 NANTUCKET  12   77  1.74  179   25 0.01298701 0.01298701
```

Note: the `plot` method will display residual plots with outliers labeled by number.

Display a residuals vs fits

```
> plot(L$fit, L$res)
> abline(h=0, col=2)
```



objects and methods

- ▶ R has different types of objects
- ▶ R objects may have methods - such as plot, summary, etc.
- ▶ Methods are specific to and appropriate for the object
- ▶ e.g. plot a linear model fit object? Residual plots
- ▶ e.g. plot a table? Mosaic plot

Background for this example

The data are from a 1854 study involving the percentage of lunatics cared for at home and a number of associated factors for 14 counties in Massachusetts. A study of the relationship between the percentage of lunatics cared for at home and distance to the nearest health center is not linear but the relation to the reciprocal of distance is essentially linear. In viewing this data the observer should keep in mind that Suffolk County includes Boston, the largest population center in the state, that Berkshire is the westernmost county in the state, and that Duke County (Martha's Vineyard) and Nantucket are islands offshore. Nantucket is a clear outlier to the pchome vs distance relationship as well as the relation to the reciprocal of distance. It is best indexed with an indicator (or dummy) variable. The student might also look at the pchome vs population density relation. Note that population density and distance are highly correlated.

Hunter comments that 'Jarvis' discovery and exposition of distance decay in hospital services remains important today. The commission recommended that numerous small mental hospitals be erected at scattered locations rather than huge centralized facilities." (See Datafile reference and description.) Remember that regression methods were not formalized until 30-40 years after 1854.

Student Admissions at UC Berkeley

- ▶ data set UCBA admissions is data available in R
- ▶ Usage: UCBA admissions
- ▶ Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.
- ▶ At issue is whether the data show evidence of sex bias in admission practices.
- ▶ Format: A 3-dimensional array resulting from cross-tabulating 4526 observations on 3 variables.

No	Name	Levels
1	Admit	Admitted, Rejected
2	Gender	Male, Female
3	Dept	A, B, C, D, E, F

UCBAdmissions

```
> UCBAdmissions
, , Dept = A

      Gender
Admit  Male Female
Admitted  512    89
Rejected  313    19
```

```
, , Dept = B

      Gender
Admit  Male Female
Admitted  353    17
Rejected  207     8
```

```
, , Dept = C

      Gender
Admit  Male Female
Admitted  120   202
```

UCBAdmissions - summary method

summary is a generic method, it produces an appropriate type of summary for the type of object. Here the data is a three-way table.

```
> summary(UCBAdmissions)
Number of cases in table: 4526
Number of factors: 3
Test for independence of all factors:
Chisq = 2000.3, df = 16, p-value = 0
```

Different objects have different summary methods.
e.g. summary of numeric variable is the 5-number summary and mean.

UCBAdmissions - extracting sub-tables

UCBAdmissions[i,j,k] contains the count for decision i, gender j, and department k.

```
> UCBAdmissions[, , 2] #department B

      Gender
Admit  Male Female
Admitted  353    17
Rejected  207     8
```

```
> UCBAdmissions[2, , ] #decision reject

      Dept
Gender  A  B  C  D  E  F
Male   313 207 205 279 138 351
Female  19  8 391 244 299 317
```

```
> UCBAdmissions[2, , 2] #decision reject in Dept. B

      Male Female
      207      8
```

UCBAdmissions - chisquare tests

To test if gender and acceptance are independent for Dept. D:

```
> deptD = UCBAdmissions[, , 4]
> chisq.test(deptD)
```

Pearson's Chi-squared test with Yates' continuity correction

```
data: deptD
X-squared = 0.2216, df = 1, p-value = 0.6378
```

For output similar to 'crosstabs' in other software, use the CrossTable function in gmodels package.

UCBAdmissions - chisquare tests, cont.

```
> library(gmodels) #install it first
> CrossTable(UCBAdmissions[,4], chisq=TRUE, format="SPSS")
```

Cell Contents	
Count	
Chi-square contribution	
Row Percent	
Column Percent	
Total Percent	

Total Observations in Table: 792

Continued on next two slides ...

UCBAdmissions - chisquare tests, cont.

Admit	Gender		Row Total
	Male	Female	
Admitted	138	131	269
	0.093	0.104	
	51.301%	48.699%	33.965%
	33.094%	34.933%	
	17.424%	16.540%	
Rejected	279	244	523
	0.048	0.053	
	53.346%	46.654%	66.035%
	66.906%	65.067%	
	35.227%	30.808%	
Column Total	417	375	792
	52.652%	47.348%	

UCBAdmissions - chisquare tests, cont.

Statistics for All Table Factors

Pearson's Chi-squared test

Chi² = 0.2979776 d.f. = 1 p = 0.5851531

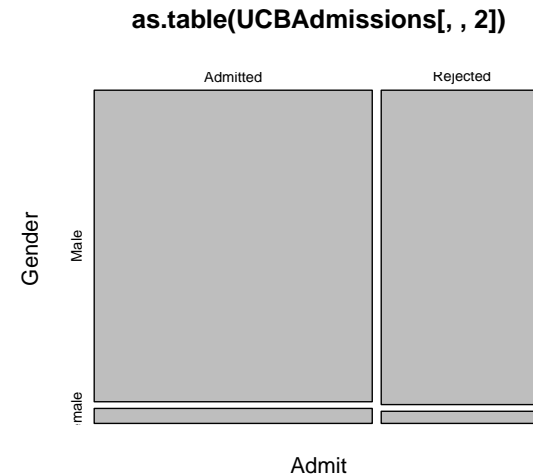
Pearson's Chi-squared test with Yates' continuity correction

Chi² = 0.2215937 d.f. = 1 p = 0.6378283

Minimum expected frequency: 127.3674

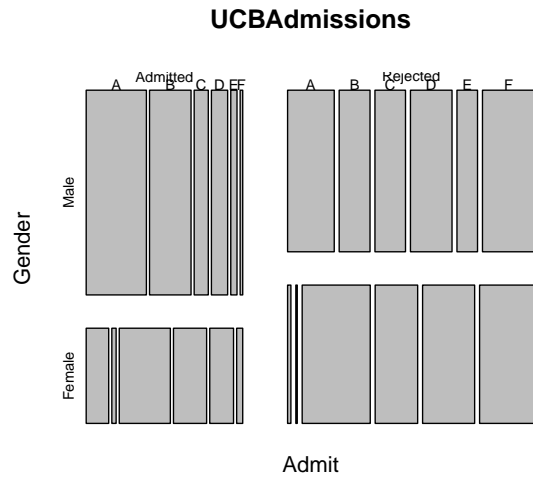
UCBAdmissions - plot Dept. B

```
> plot(as.table(UCBAdmissions[,2])) #compare with mosaicpl
```



UCBAdmissions - plot all

```
> plot(UCBAdmissions) #compare with mosaicplot
```



Useful resources for R

- ▶ 'Quick R for SAS/SPSS/Stata Users' - excellent resource - <http://www.statmethods.net/>
- ▶ www.r-project.org - see Manuals, contributed documentation.
- ▶ "R Reference Card" by Tom Short
- ▶ By Jim Albert at <http://bayes.bgsu.edu/eda/>
 - ▶ Creating and reading a dataset into R
 - ▶ Creating and reading a dataset into R using OpenOffice
 - ▶ A simple data analysis on R
- ▶ For linear models see the online PDF book "Practical Regression and Anova using R" by Julian Faraway