Statistical Graphics and Categorical Data Analysis

David B King, Ph.D.

May 12, 2015

Visualization Methods in R

Visualization Methods in R

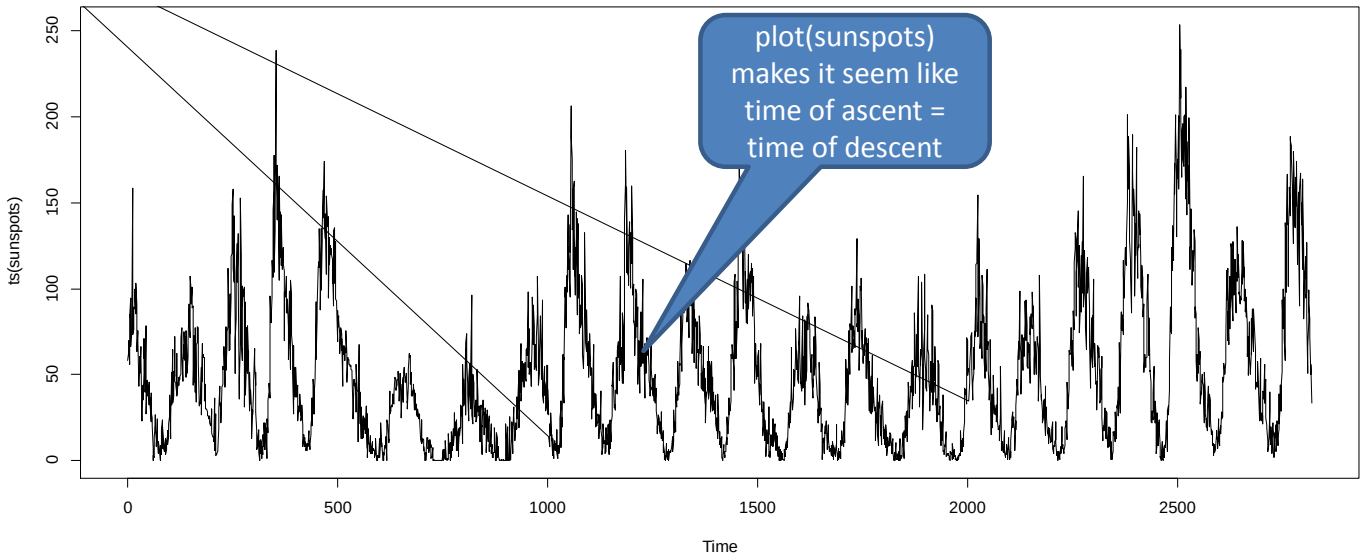
Two great packages in R:

library(lattice)
library(ggplot2)

Visualization Methods in R

Advanced Visualization or Plotting in R

The "sunspots" time series tracks the number of sunspots by month.



Sunspots are caused by intense magnetic fields in the sun.
During solar max period the amount of magnetic turbulence is largest
Time of ascent to solar max is longer than time of descent to solar min.

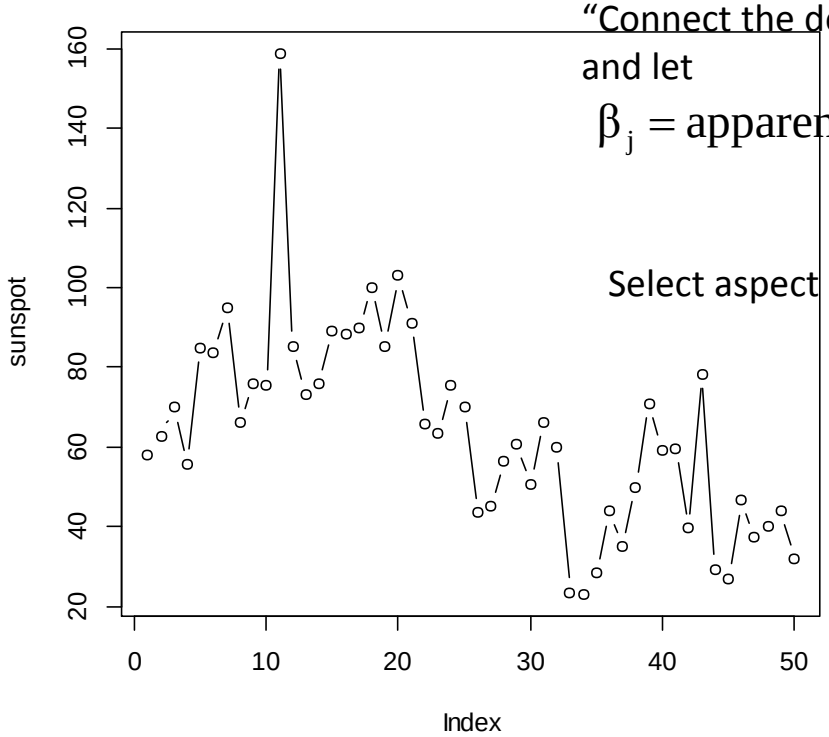
Aspect Ratio is (Height of plot) / (Width of plot)
Problem: Is there a "good" aspect ratio that will allow us to accurately compare time of ascent to time of descent?

Solution: (W. Cleveland) Aspect ratio changed by banking to 45^0

Visualization Methods in R

What is banking a plot?

```
sunspot=sunspots[1:50]  # First 50 data points of sunspots data.
plot(sunspot,type="b")
```



"Connect the dots" with line segments…

and let

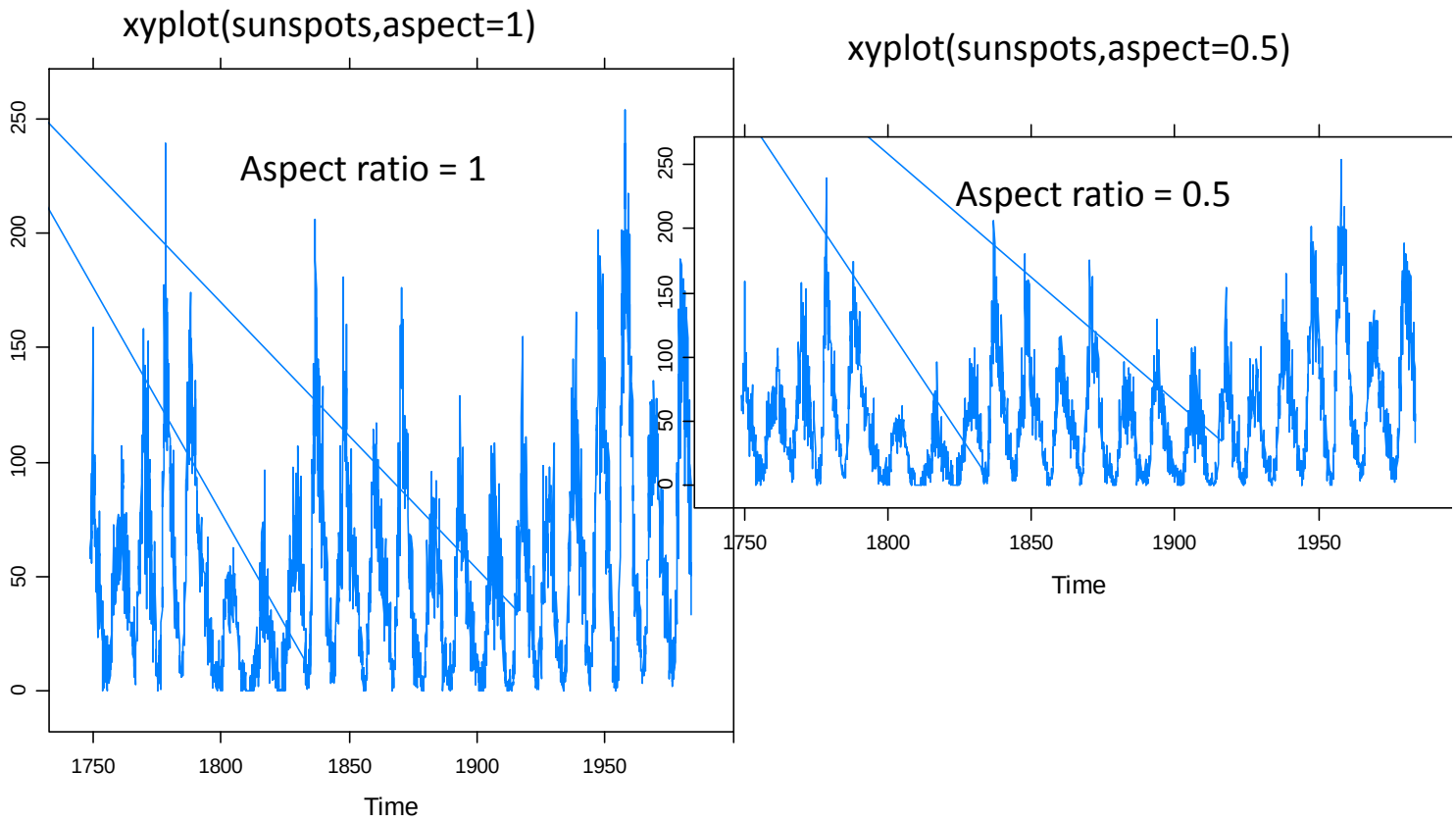
$$\beta_j = \text{apparent slope of line segment } j$$

Select aspect ratio in graph such that

$$\frac{1}{K}\sum_{i=1}^{K}\left|\beta_j\right| = 1$$

Visually want average |slope| = 45^0

Visualization Methods in R



This command adjusts the plot aspect ratio to give one
A correct visual interpretation by banking plots to 45^0.

xyplot(sunspots,aspect="xy")

Visualization Methods in R

Lattice Package

Object=graph_type(*Y ~ X | A * B*, data=)

Basic command

Y = dependent variable
X = independent (abscissa) variable
A, B = conditioning variables (usually categorical)

- Above gives you a separate plot for each level of A*B
- Called a Trellis plot because the structure of panels resembles a garden trellis work.
- Each plot area within is called a panel.

The "workhorse" basic function for Trellis plots is Obj = xyplot(y ~ x | grp , data)

The panel function controls what the plot does inside each panel.

Visualization Methods in R

Plotting Functions within the Lattice Library

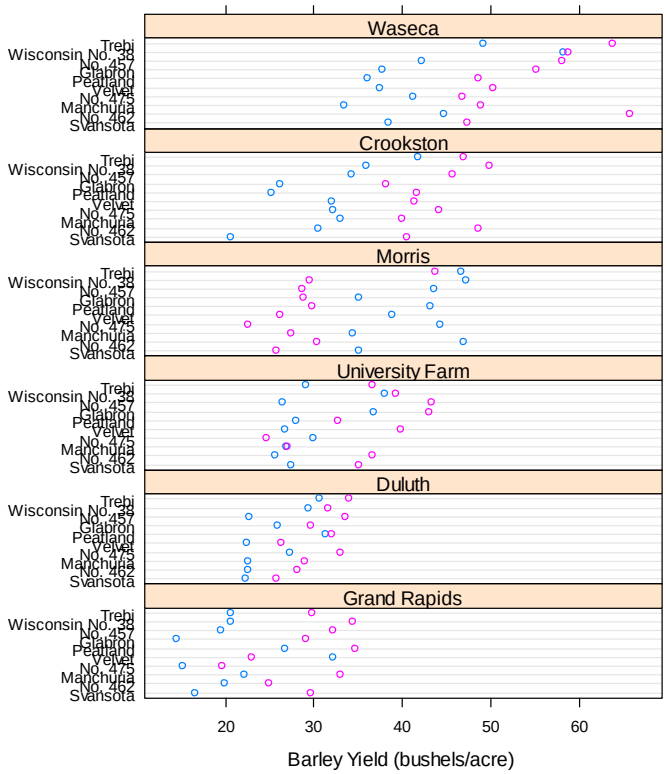
graph_type	description	formula examples
barchart	bar chart	x~A or A~x
bwplot	boxplot	x~A or A~x
cloud	3D scatterplot	z~x*y|A
contourplot	3D contour plot	z~x*y
densityplot	kernal density plot	~x|A*B
dotplot	dotplot	~x|A
histogram	histogram	~x
levelplot	3D level plot	z~y*x
parallel	parallel coordinates plot	data frame
splom	scatterplot matrix	data frame
stripplot	strip plots	A~x or x~A
xyplot	scatterplot	y~x|A
wireframe	3D wireframe graph	z~y*x

In the above
y,x,z = continuous variables
A,B = categorical variables

Visualization Methods in R

The Dot Plot

```
dotplot(variety ~ yield|
site,data=barley,groups=year,auto.key=list(space="right"), xlab =
"Barley Yield (bushels/acre)",aspect="xy", layout = c(1,6))
```



Basic format:

Category ~ Numeric | categorical condition

Separate panels for each level of condition var

1932
1931

Group variable controls color of dots within each panel

Layout = c(m,n) means m columns of panels and n rows.

This plot reveals the famous Morris mistake, the yields in 1931 were higher everywhere except university farm, in which the reverse was true.

Visualization Methods in R

Plotting Examples with mtcars

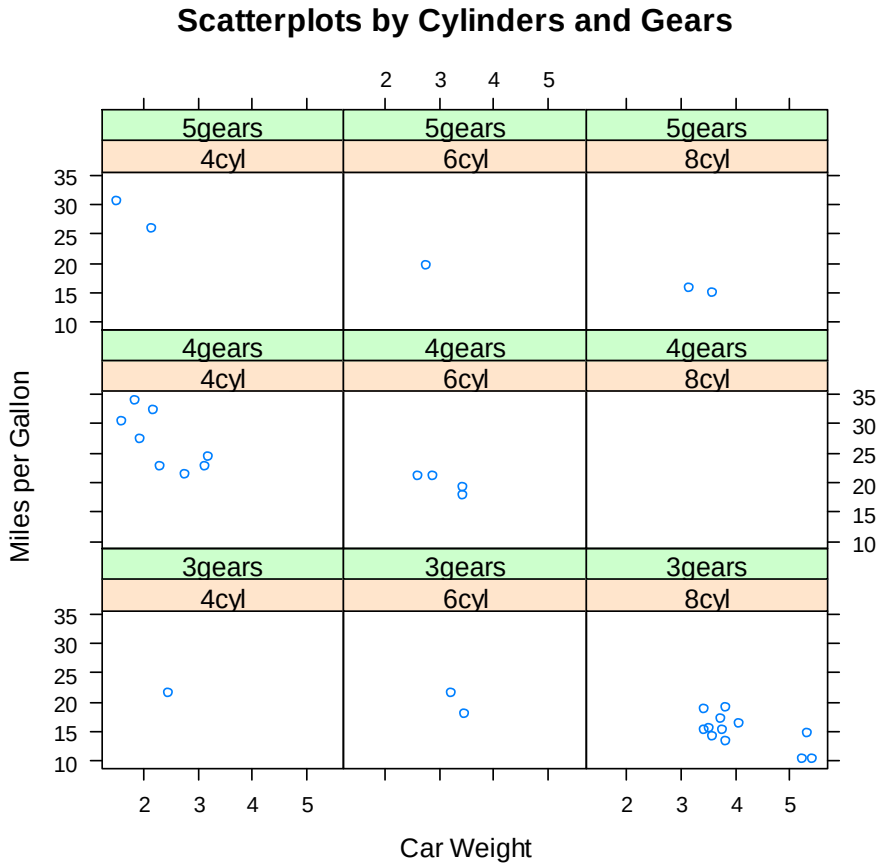
This dataset gives us a lot of both categorical and continuous variables...

```
> attach(mtcars)
> mtcars
                  mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
.....
```

```
gear.f<-factor(gear,levels=c(3,4,5),labels=c("3gears","4gears","5gears"))
cyl.f <-factor(cyl,levels=c(4,6,8),  labels=c("4cyl","6cyl","8cyl"))
```

Visualization Methods in R

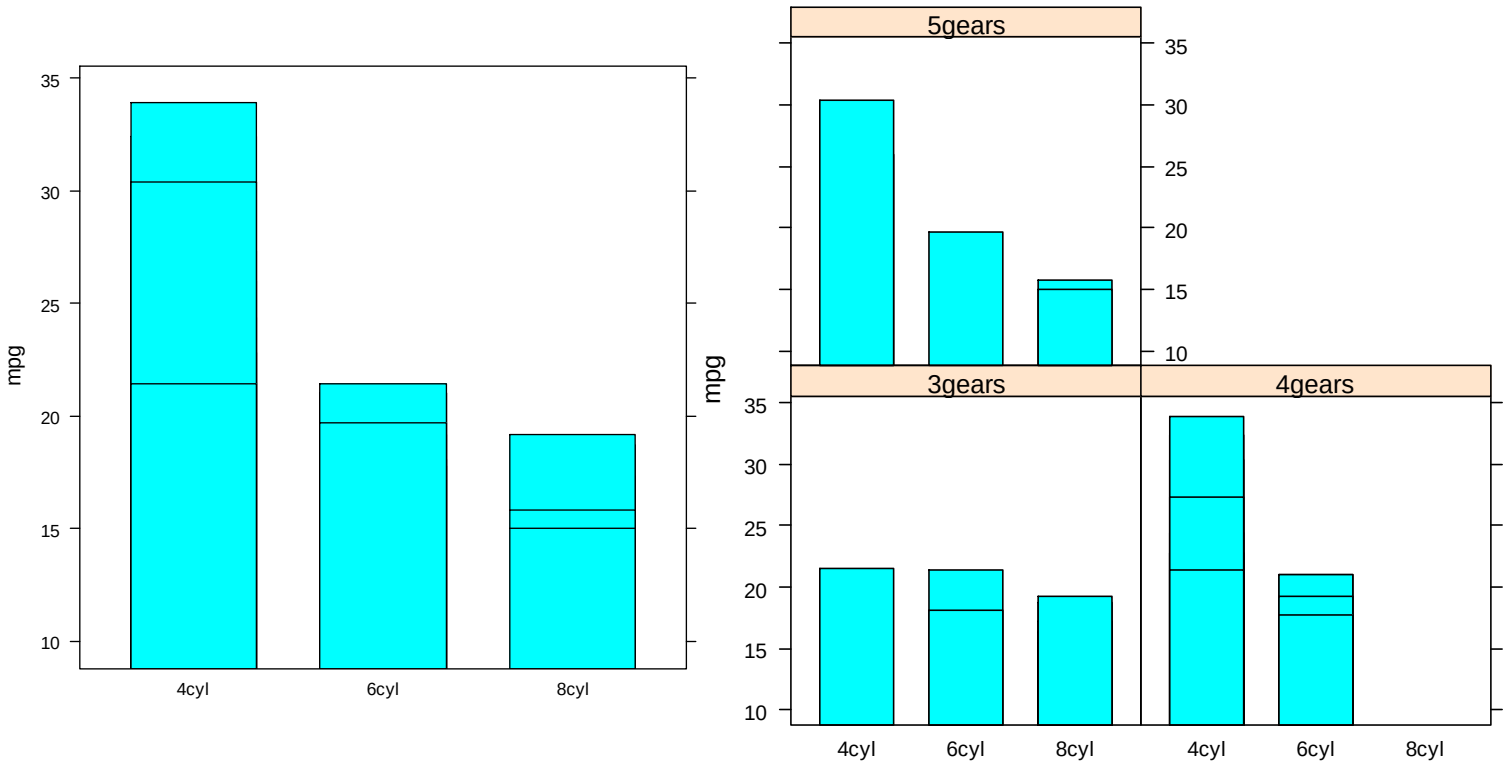
xyplot(mpg~wt|cyl.f*gear.f, main="Scatterplots by Cylinders and Gears", ylab="Miles per Gallon", xlab="Car Weight")



Visualization Methods in R

Bar Charts

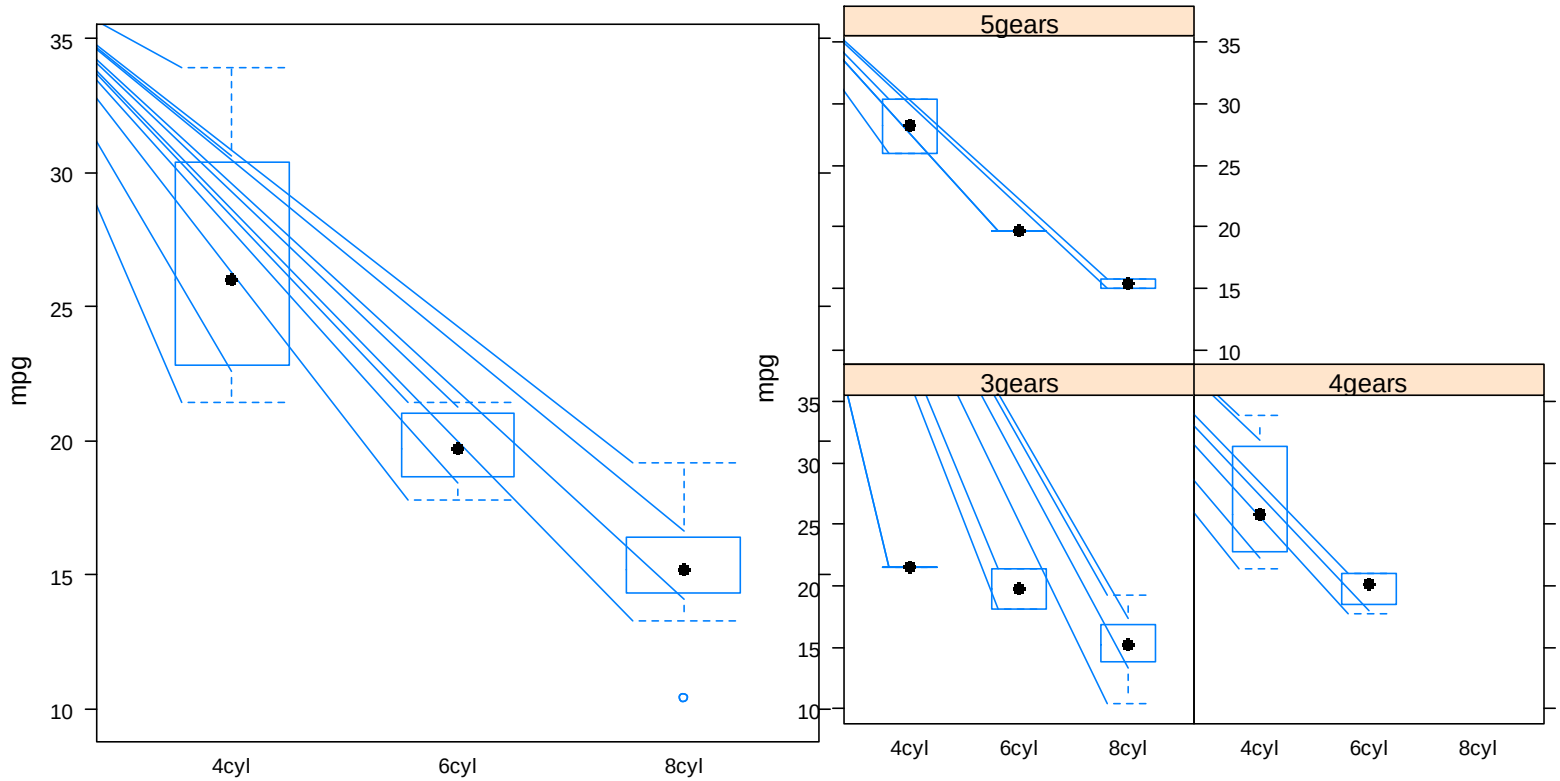
```
> barchart(mpg~cyl.f)
> barchart(mpg~cyl.f|gear.f)
```



Visualization Methods in R

Box and Whisker Plots

```
> bwplot(mpg~cyl.f)
> bwplot(mpg~cyl.f|gear.f)
```



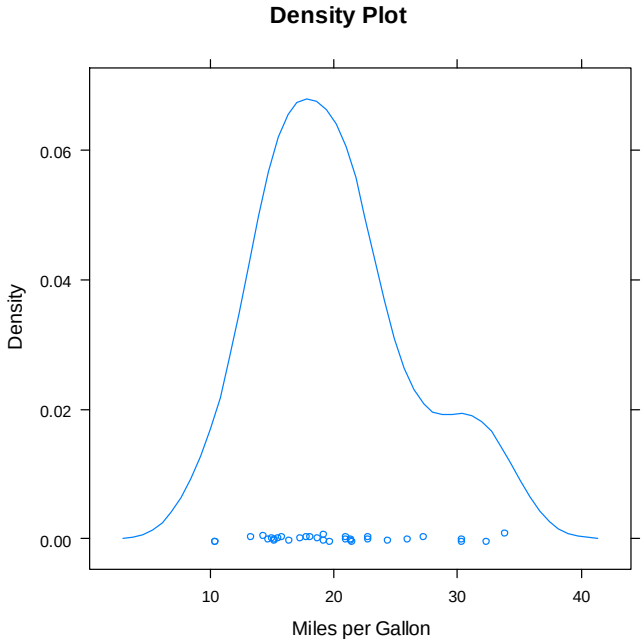
Visualization Methods in R

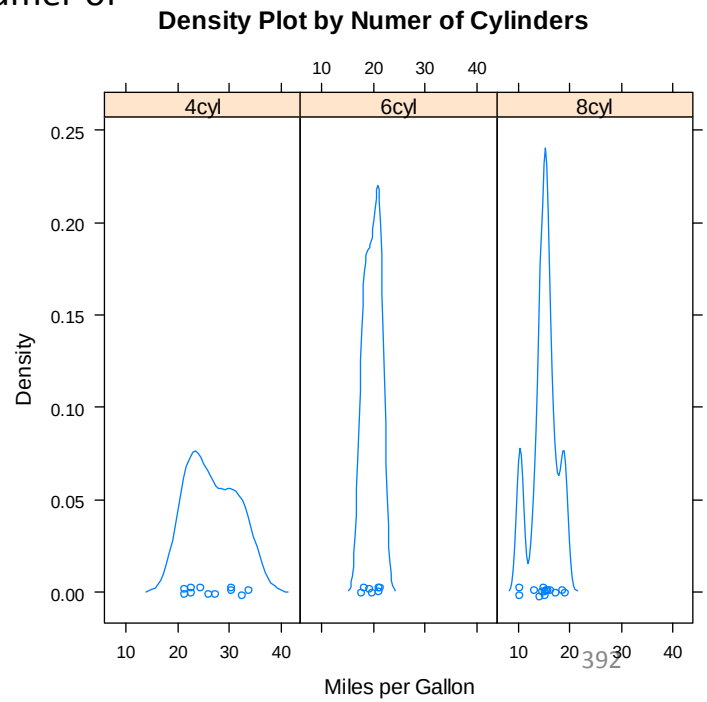
Kernel Density Plots

densityplot(~mpg,main="Density Plot",xlab="Miles per Gallon")

>\# kernel density plots by factor level
densityplot(~mpg|cyl.f,main="Density Plot by Number of Cylinders",xlab="Miles per Gallon")

> # kernel density plots by factor level (alternate layout)
densityplot(~mpg|cyl.f,main="Density Plot by Numer of Cylinders",xlab="Miles per Gallon",layout=c(1,3))

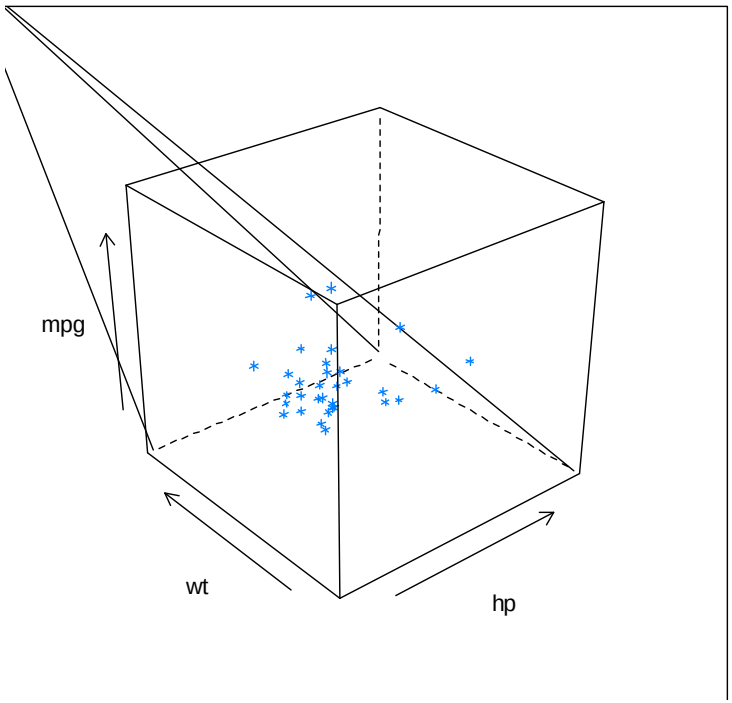


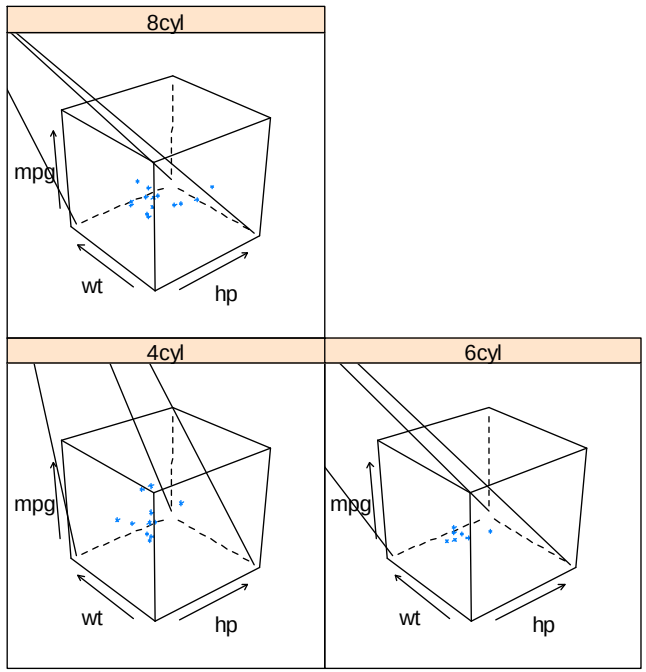


Visualization Methods in R

Cloud Plots

```
> cloud(mpg~hp*wt)
> cloud(mpg~hp*wt|cyl.f)
```

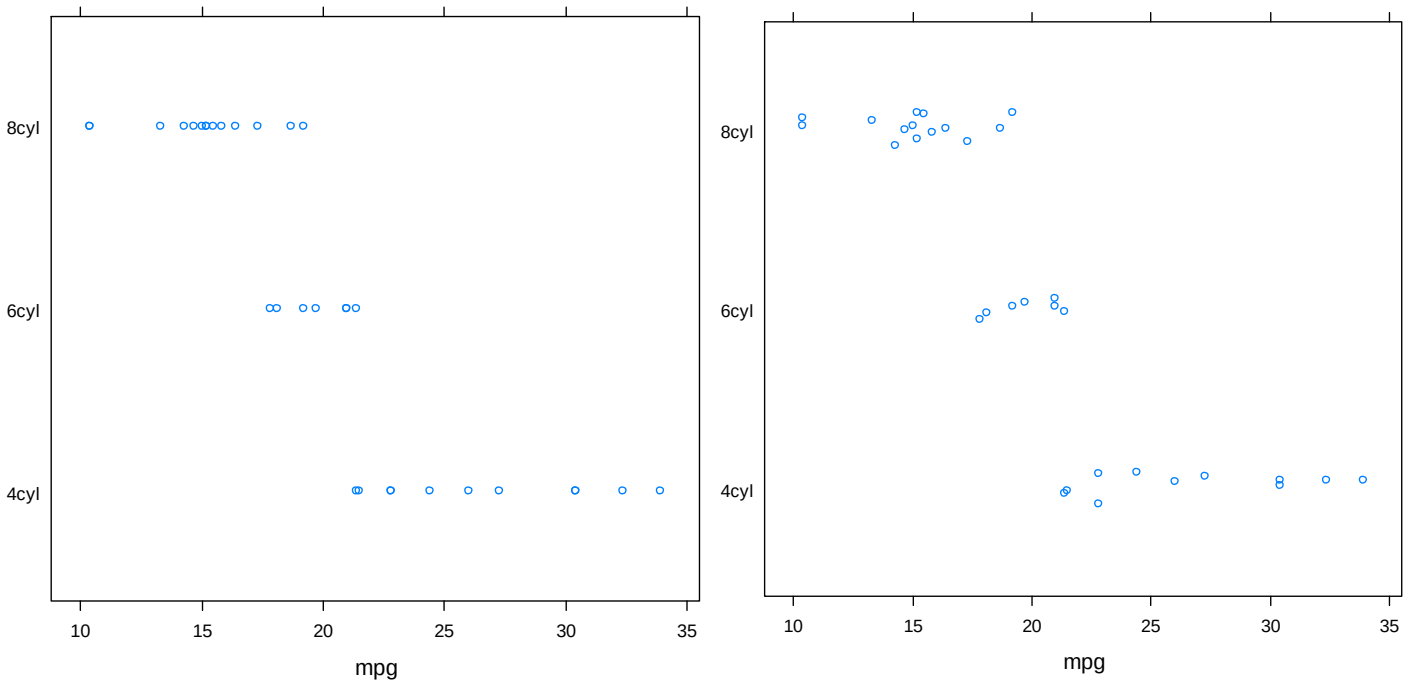




Visualization Methods in R

Strip Plots

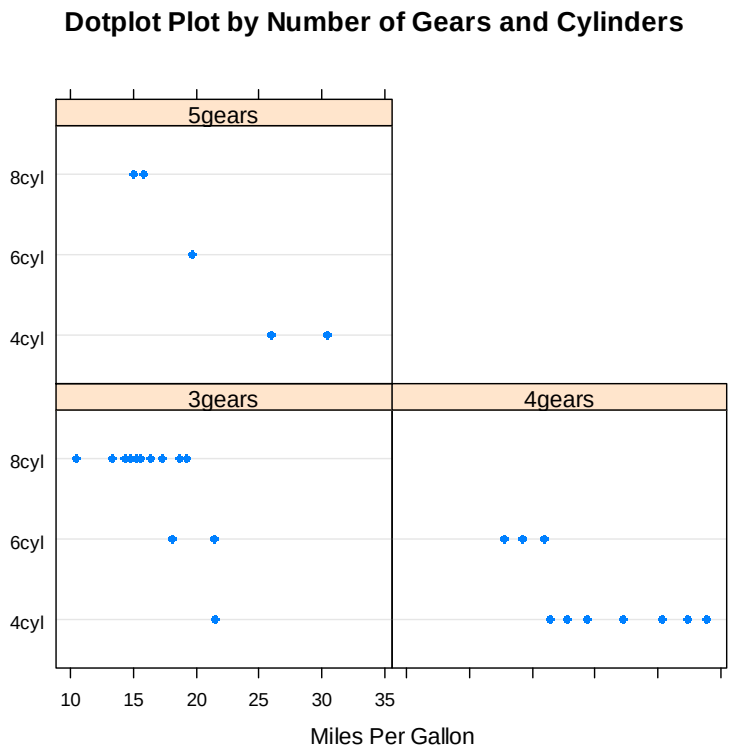
```
> stripplot(cyl.f~mpg)
> stripplot(cyl.f~mpg,jitter=TRUE)
```



Visualization Methods in R

Dot Plots

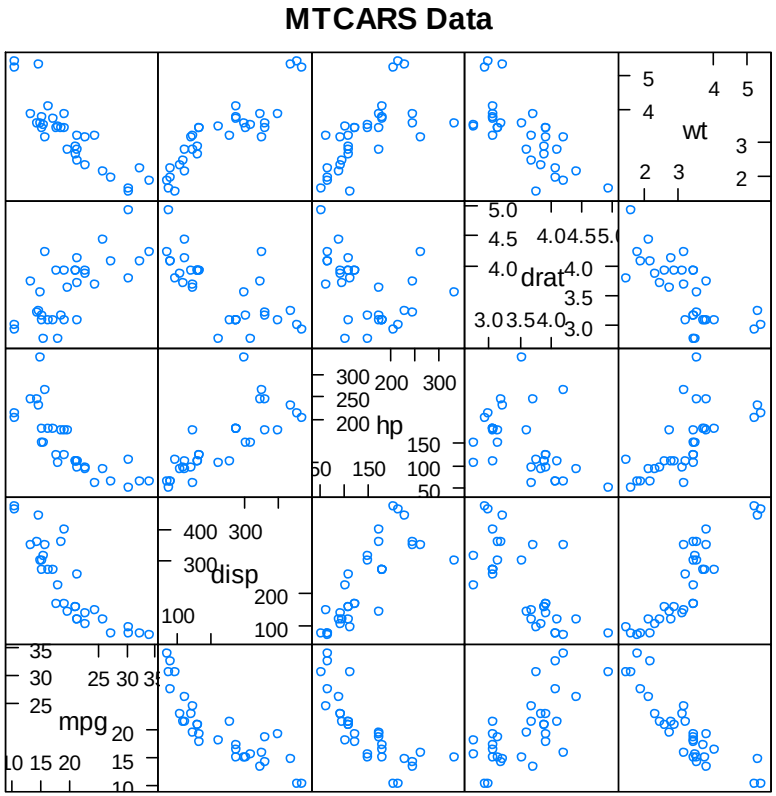
dotplot(cyl.f~mpg|gear.f, main="Dotplot Plot by Number Gears and Cylinders",xlab="Miles Per Gallon")



Visualization Methods in R

Scatter Plot Matrix (splom)

splom(mtcars[c(1,3,4,5,6)],main="MTCARS Data")

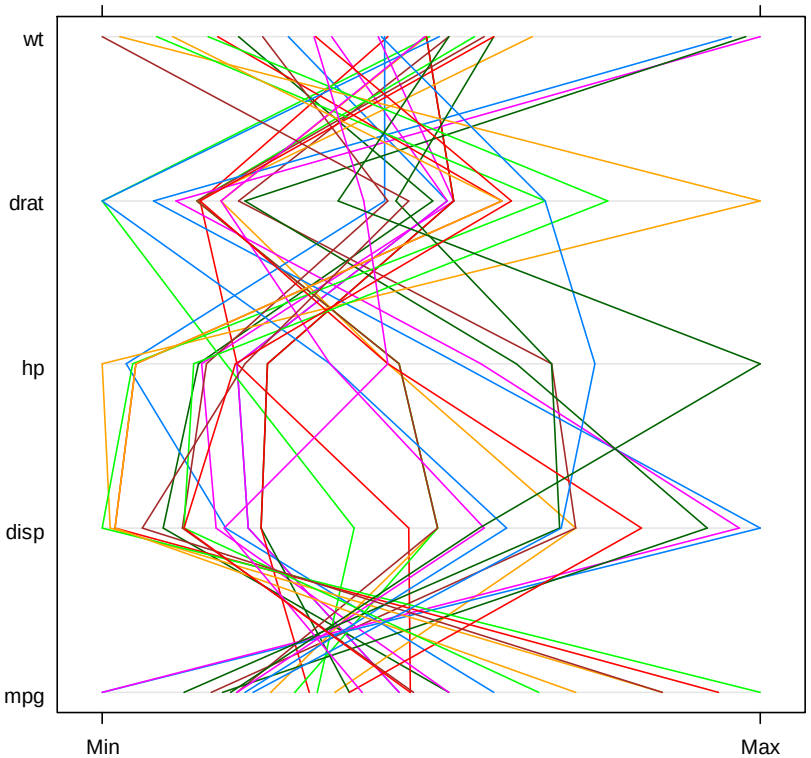


Scatter Plot Matrix

Visualization Methods in R

Parallel Plots

parallelplot(mtcars[c(1,3,4,5,6)])



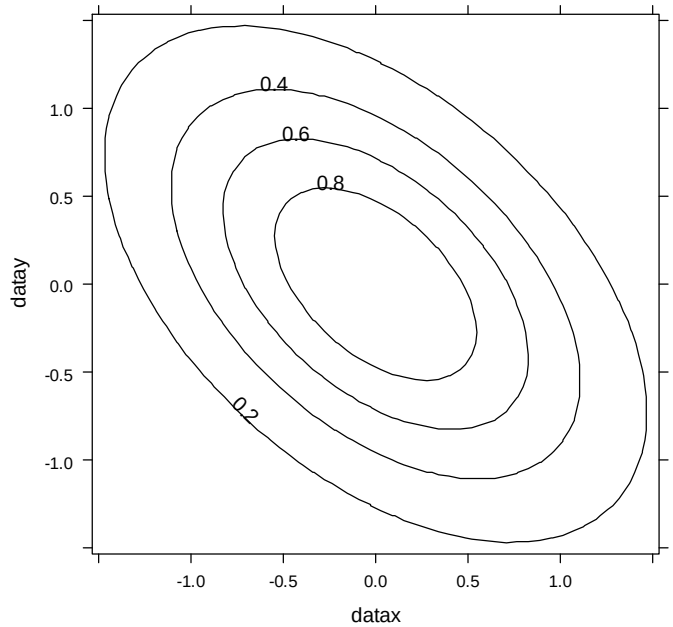
Visualization Methods in R

Example of 3D plots in lattice package

The Contour Plot

```
datax <- rep(seq(-1.5, 1.5, length=50), 50)
datay <- rep(seq(-1.5, 1.5, length=50), rep(50, 50))
dataz <- exp(-(datax^2 + datay^2 + datax*datay))
gauss <- data.frame(datax, datay, dataz)
```

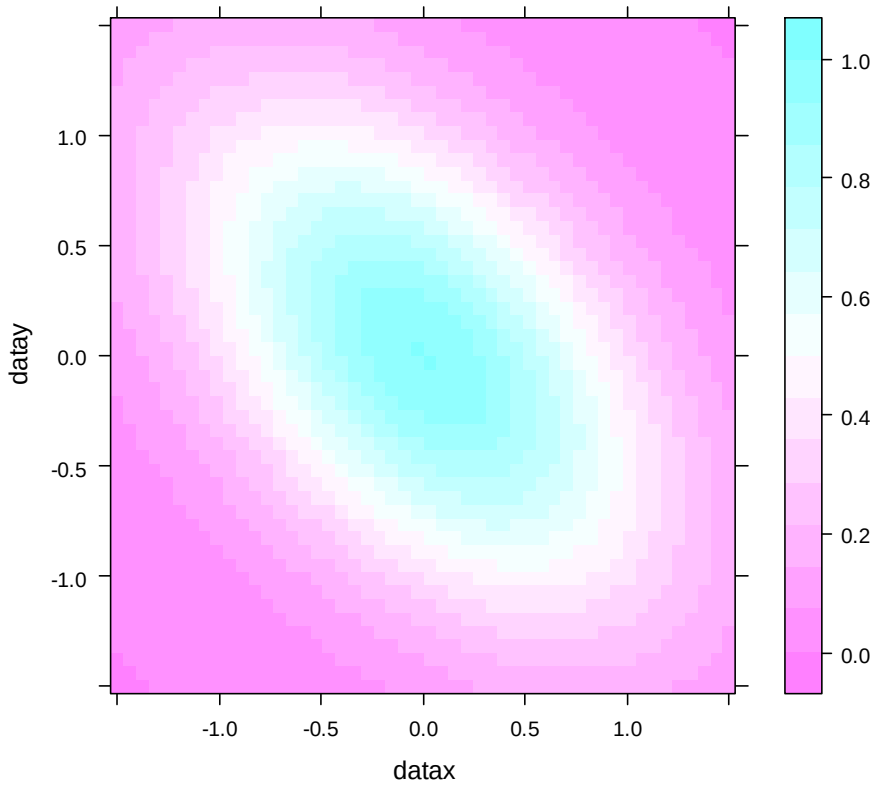
```
> contourplot(dataz~datax * datay, data = gauss)
```



Visualization Methods in R

The Level Plot

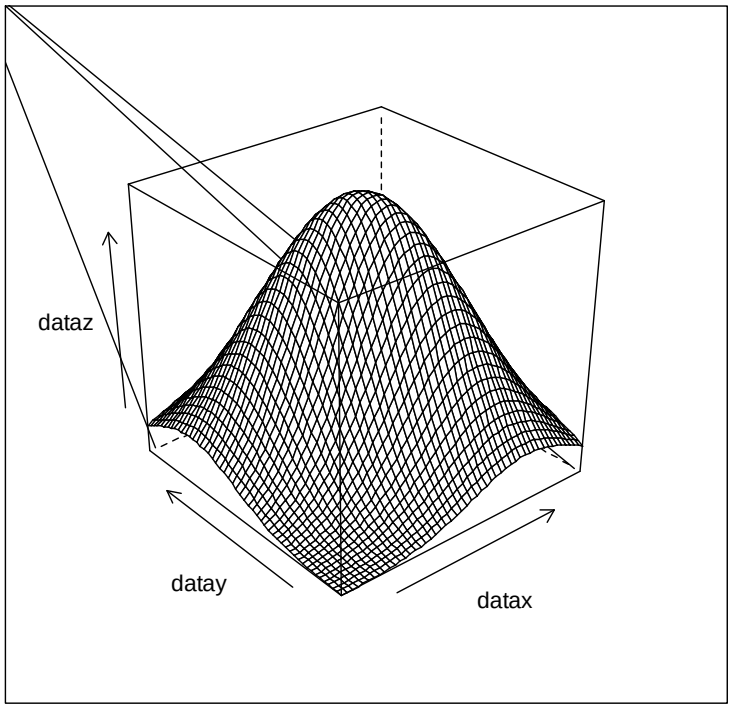
levelplot(dataz~datax * datay, data = gauss)

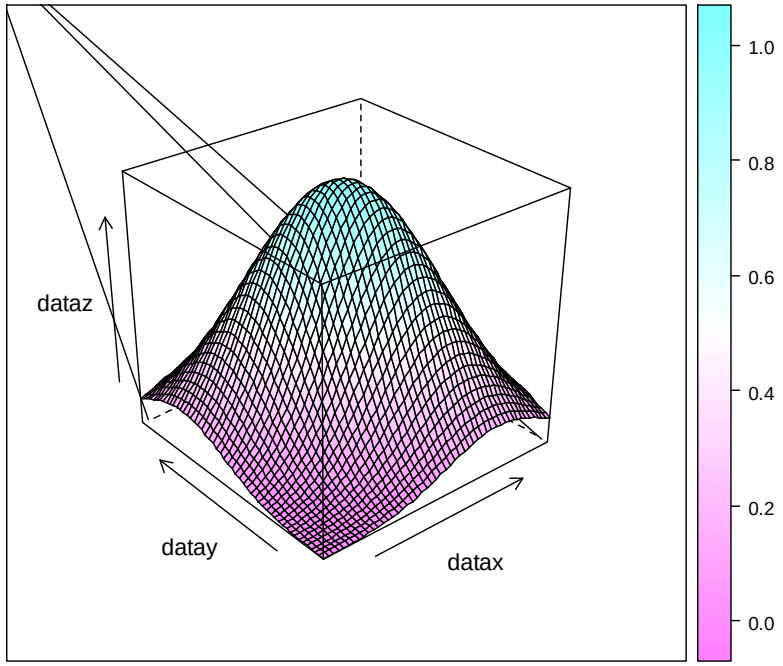


Visualization Methods in R

The Wire Frame Plot

> wireframe(dataz~datax*datay,data=gauss,drape=TRUE)
> wireframe(dataz~datax*datay,data=gauss,drape=FALSE)

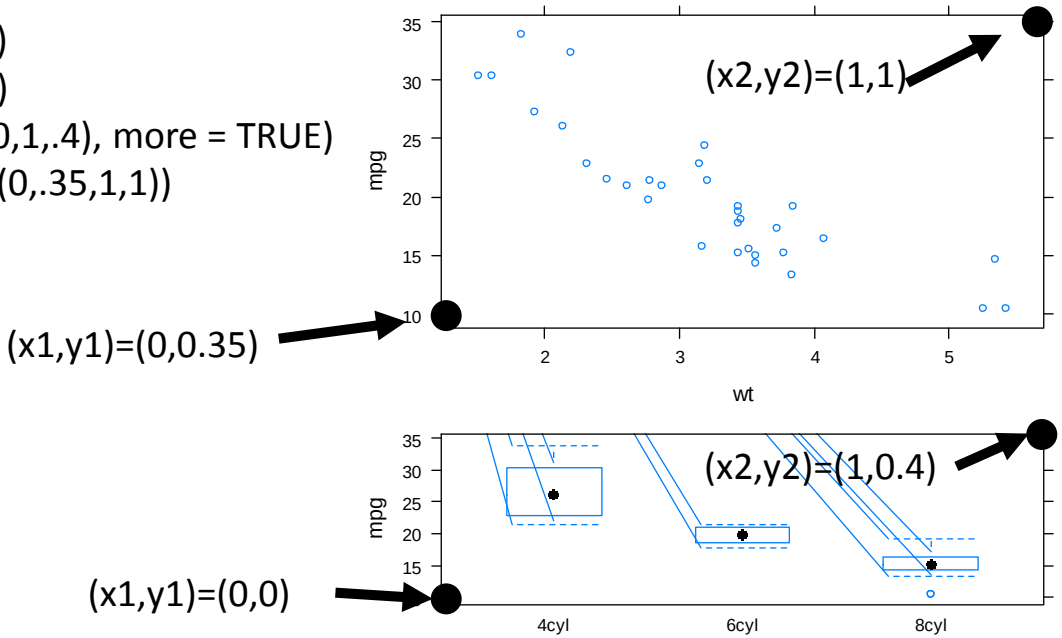




Visualization Methods in R

Printing plots and controlling their position on the page

```
box.plot <-  bwplot(mpg~cyl.f)
scatter.plot <- xyplot(mpg~wt)
print(box.plot, position = c(0,0,1,.4), more = TRUE)
print(scatter.plot, position = c(0,.35,1,1))
```



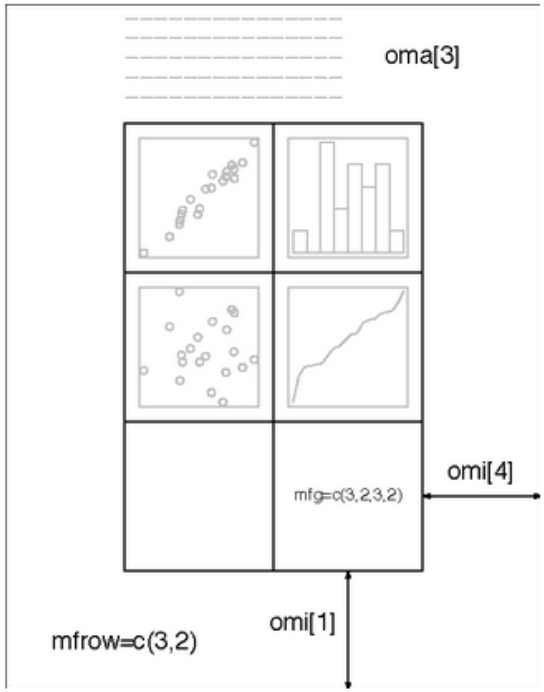
The argument `position` specifies the position of each graph on the page using a page coordinate system in which the lower left corner of the page is (0, 0) and the upper right corner is (1, 1). The *graph rectangle* is the portion of the page allocated to a graph. `position` takes a vector of four numbers; the first two numbers are the coordinates of the lower left corner of the graph rectangle, and the second two numbers are the coordinates of the upper right corner. The argument `more=` has been give a value of `T`, which says that more drawing is coming.

Visualization Methods in R

Printing plots and controlling their position on the page

par(oma = c(x,y,z,w)) controls the size of the outer margins in lines

A vector of the form `c(bottom, left, top, right)` giving the size of the outer margins in lines of text.



par.settings=list(layout.heights=list(top .padding=-2)) in command controls the Inner margins within each panel

Visualization Methods in R

Splitting the Screen and controlling position on the page

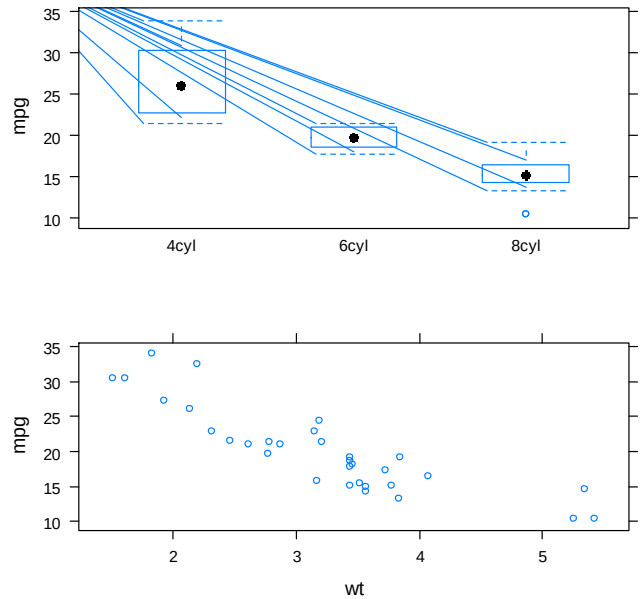
```
print(box.plot, split = c(1,1,1,2), more = TRUE)
print(scatter.plot, split = c(1,2,1,2))
```

Want a c(1,2) array of plots

split = c(1,1,1,2),

(1,1)

(1,2)



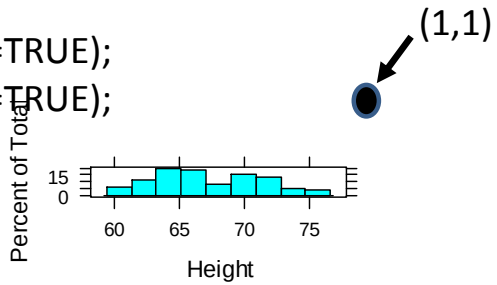
`split=` takes a vector of four values. The last two define an array of subregions in the graphics region. In our example, the array has one column and two rows for both plots. The first two values of `split=` prescribe the subregion in which the current plot is to be drawn.

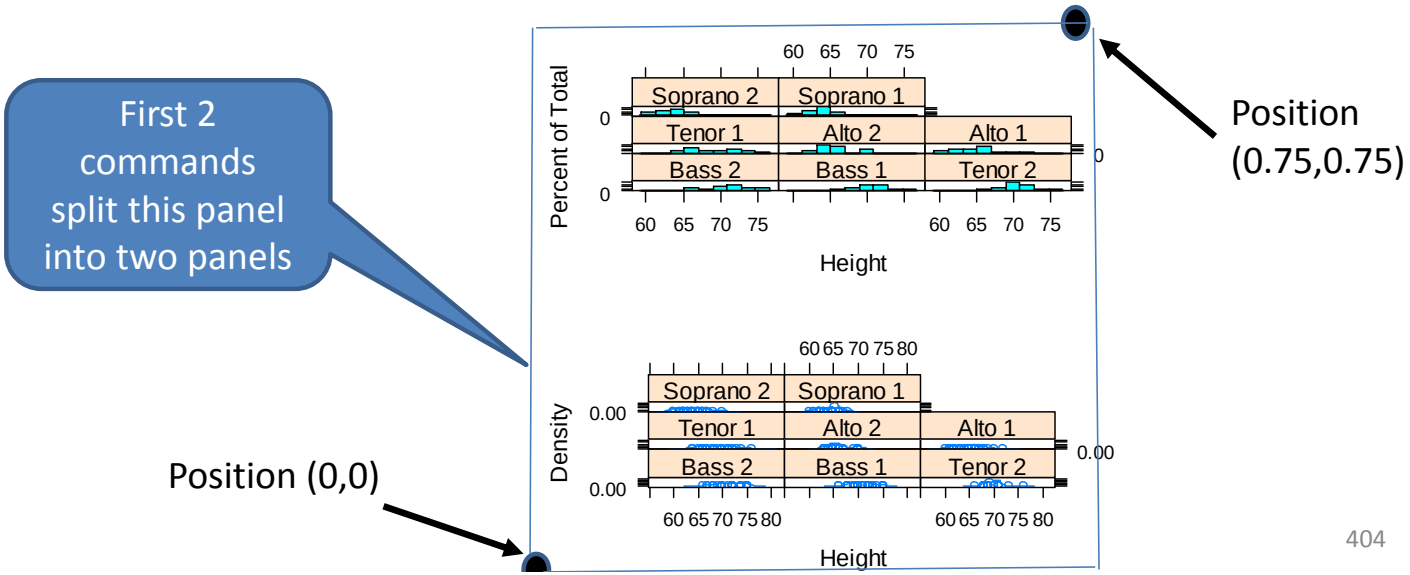
Visualization Methods in R

Example of Using Both Position and Split

```
p11 <- histogram( ~ height | voice.part, data = singer, xlab="Height");
p12 <- densityplot( ~ height | voice.part, data = singer, xlab = "Height");
 p2 <- histogram( ~ height, data = singer, xlab = "Height")

print(p11, position = c(0,0,.75,.75), split=c(1,1,1,2), more=TRUE);
print(p12, position = c(0,0,.75,.75), split=c(1,2,1,2), more=TRUE);
print(p2, position = c(.5,.75,1,1), more=FALSE)
```





Visualization Methods in R

Conditioning on a Continuous Variable: The idea of the shingle

The other major contribution to Visualization by W. Cleveland was the concept of the shingle which enables us to condition on continuous variables

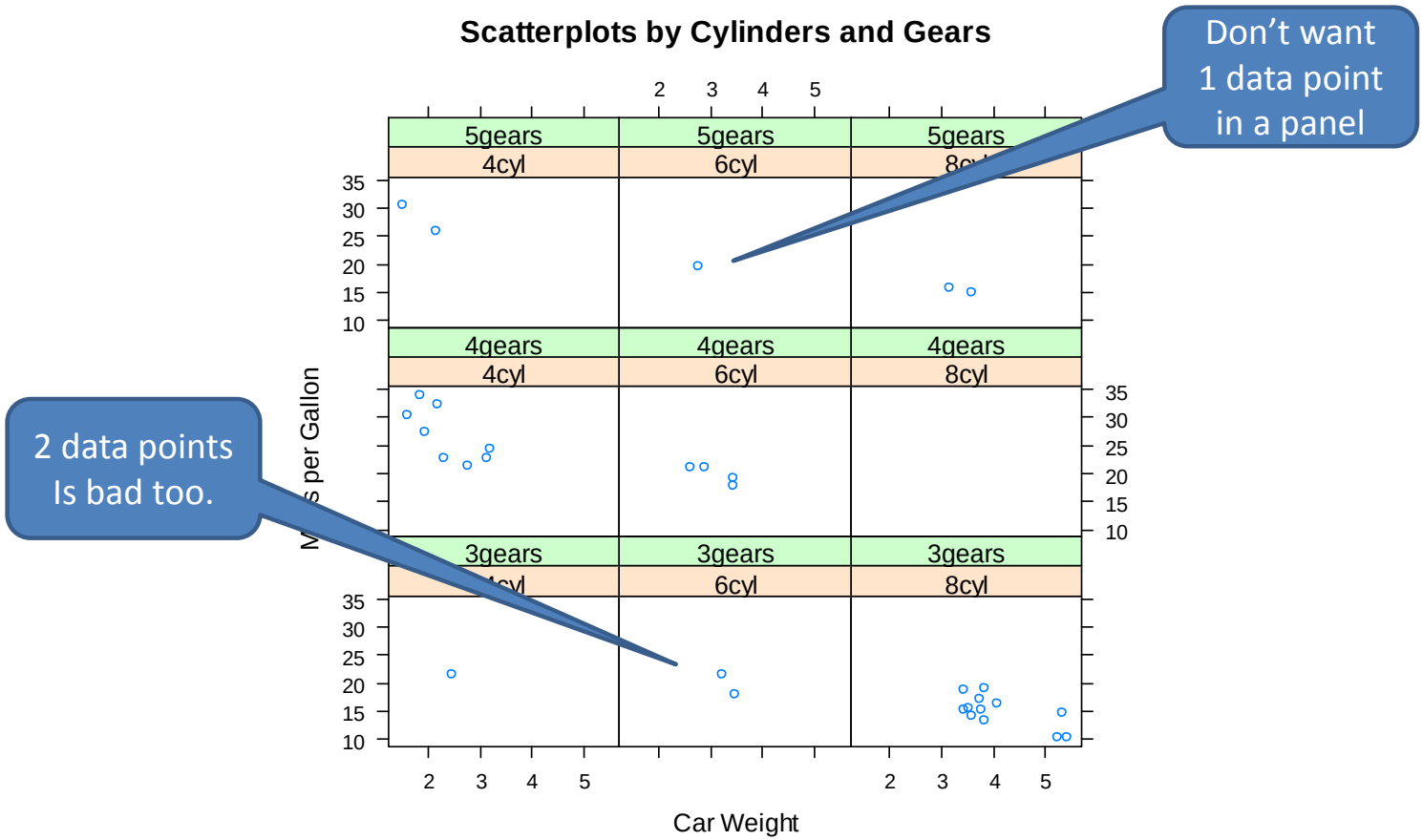
Often have

(y ~ x | categorical factor) Levels of categorical factor determine which panel the data in scatter plot falls in

Can we have lattice plots conditioned on a continuous variable?

(y ~ x | continuous variable)

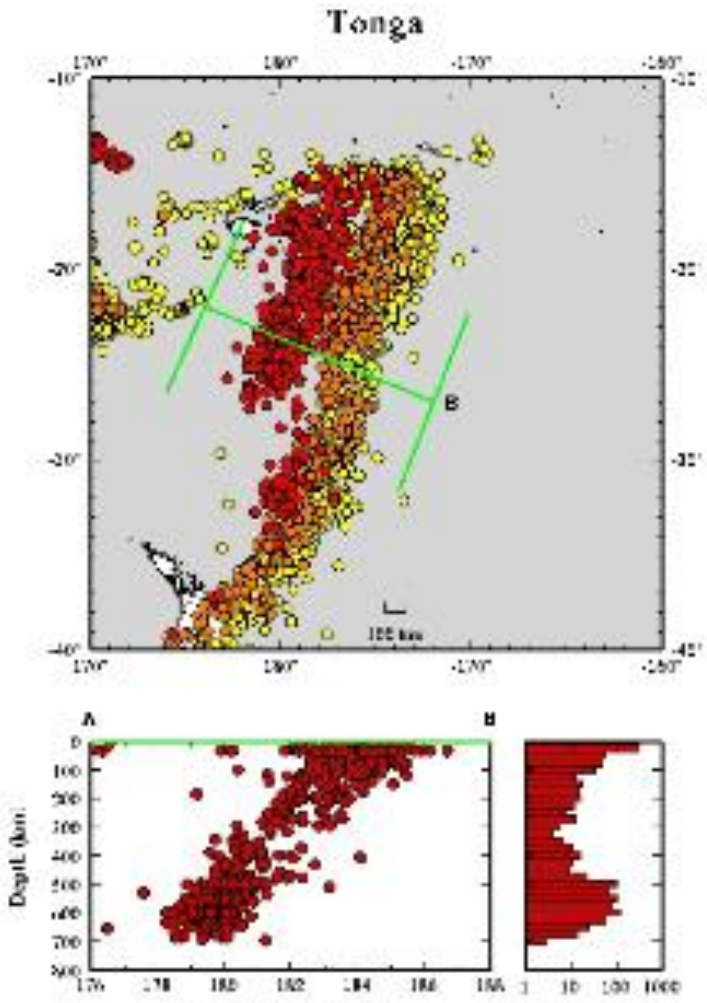
Visualization Methods in R



Would like to have a scatter plot conditioned on a continuous Variable where we have **equal number of data points in each panel**

Visualization Methods in R

Example of Conditioning on Continuous Variable In Trellis Plots



The **quakes** dataset in R

Tonga Trench Earthquakes

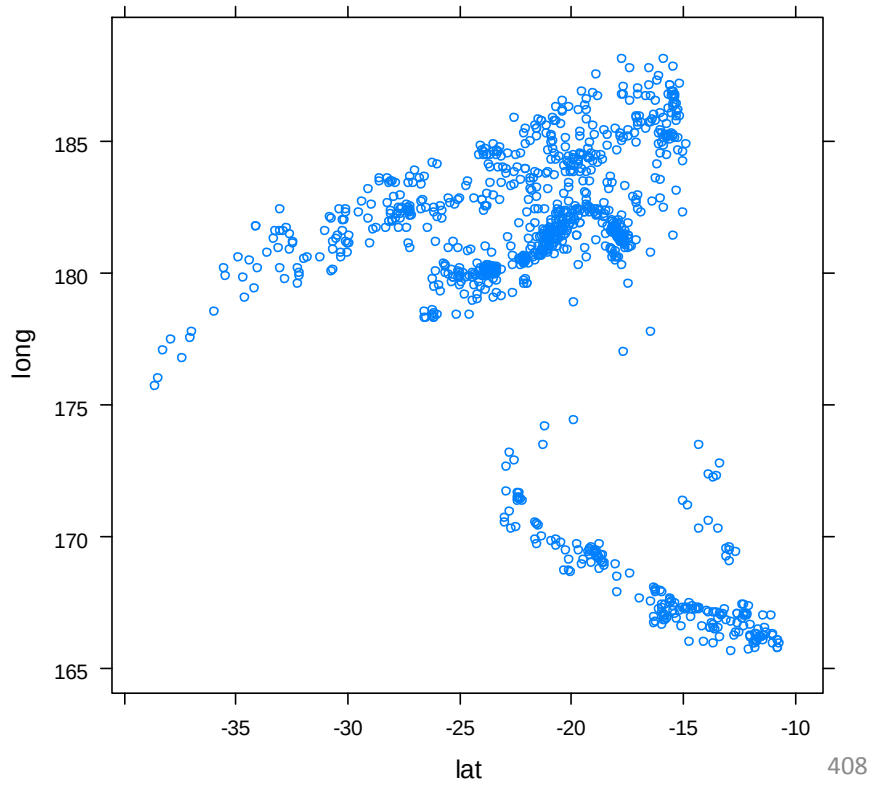
Yellow:	0 — 70 km
Orange:	71 — 300 km
Red:	300 — 800 km.

Visualization Methods in R

```
> head(quakes)
    lat   long depth mag stations
1 -20.42 181.62   562 4.8       41
2 -20.62 181.03   650 4.2       15
3 -26.00 184.10    42 5.4       43
4 -17.97 181.66   626 4.1       19
5 -20.42 181.96   649 4.0       11
6 -19.68 184.31   195 4.0       12
> xyplot(long~lat,data=quakes)
```

Where do the quakes happen by ocean depth?

Scatter plot of earthquakes along the Tonga trench subduction zone.



Visualization Methods in R

The Shingle

A shingle categorizes a continuous variable, but allows some overlap between adjacent shingles

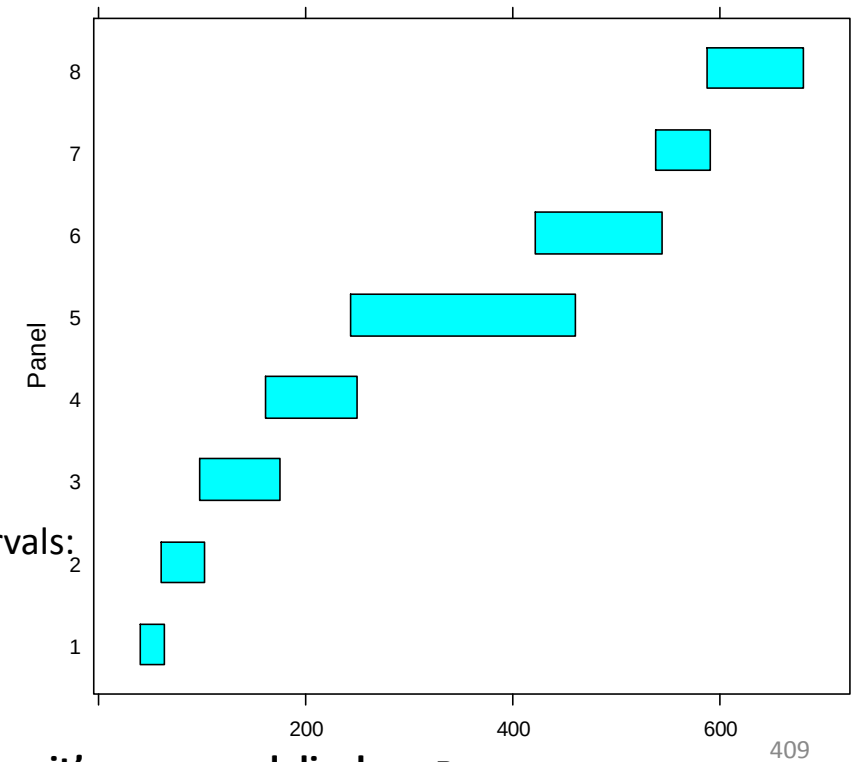
> depthgroup=equal.count(quakes$depth,number=8,overlap=0.1)

> plot(depthgroup)

Intervals:

	min	max	count
1	39.5	63.5	138
2	60.5	102.5	138
3	97.5	175.5	138
4	161.5	249.5	142
5	242.5	460.5	138
6	421.5	543.5	137
7	537.5	590.5	140
8	586.5	680.5	137

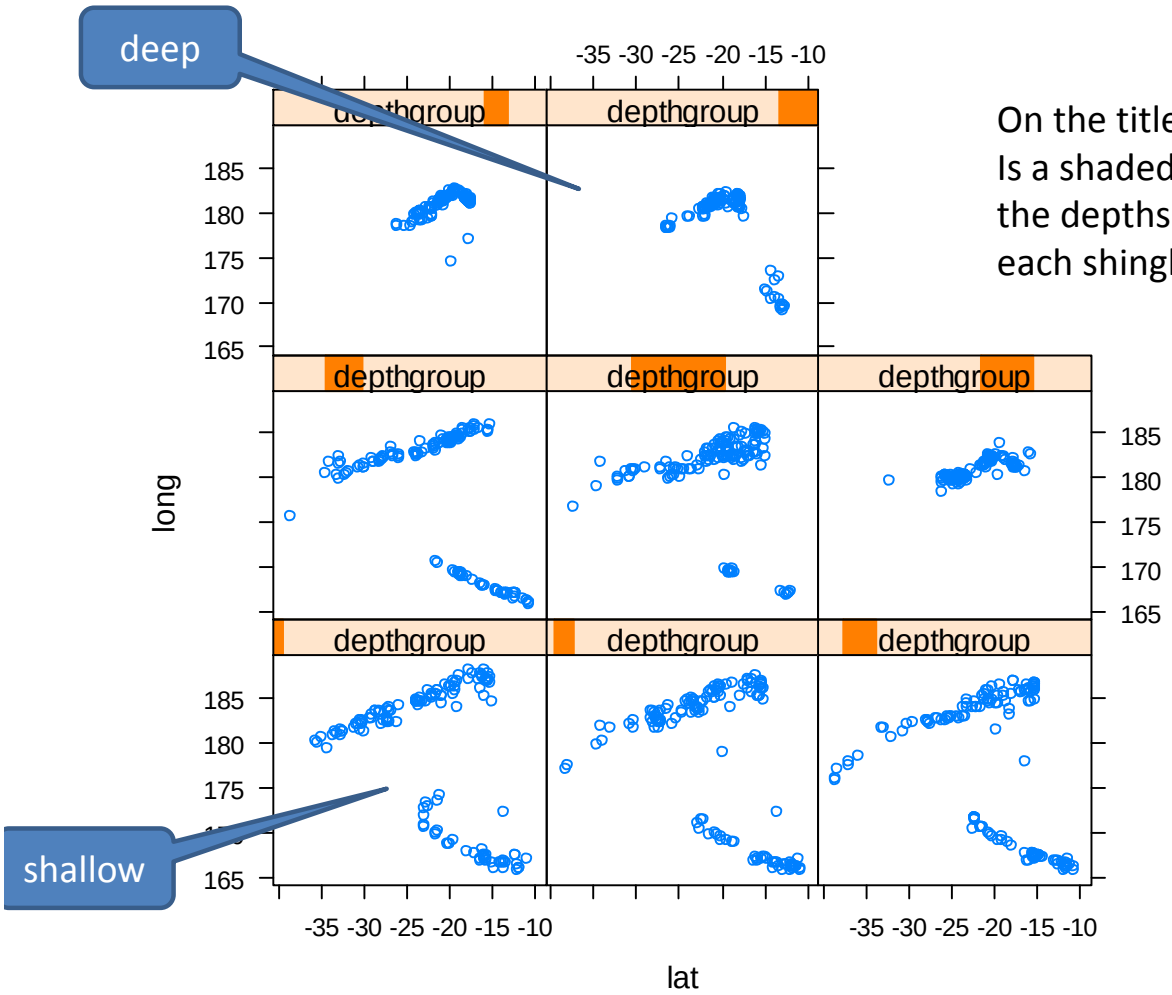
Overlap between adjacent intervals:
[1] 16 14 19 15 14 15 15



Each interval will be used to form it's own panel display

Visualization Methods in R

> xyplot(long~lat|depthgroup,data=quakes)

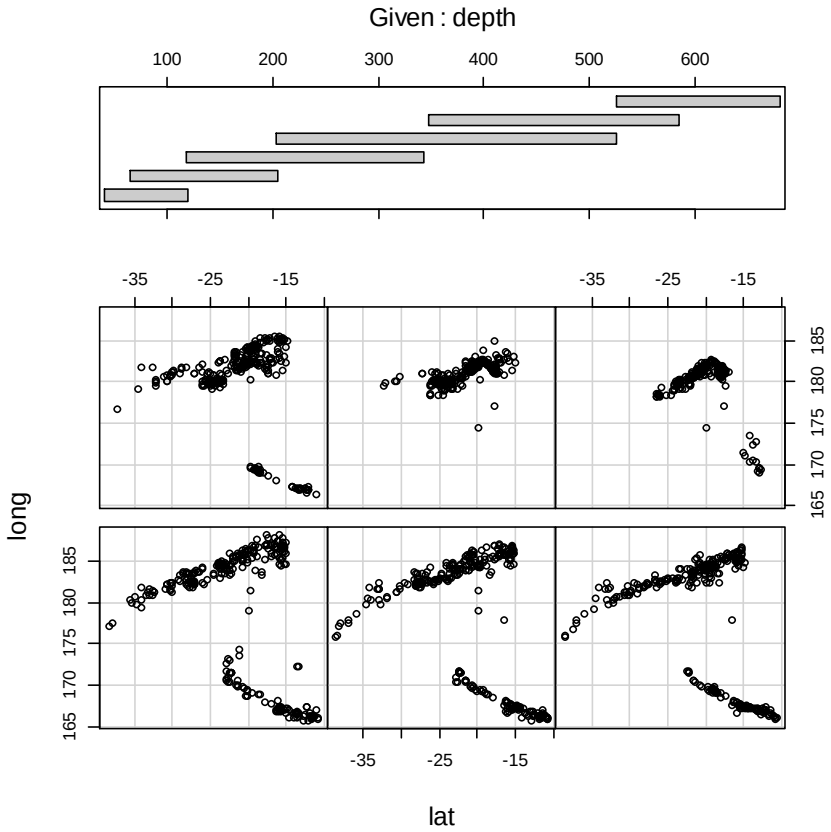


On the title for each pane
Is a shaded zone which depicts
the depths covered for
each shingle

Visualization Methods in R

The Coplot

```
> coplot(long~lat|depth,data=quakes)
```



The coplot combines the shingle plot and xyplot into one.

Visualization Methods in R

The Panel Function

The data region of a panel on a Trellis display is the rectangular region where the data are plotted. A *panel function* has the sole responsibility for drawing in the data regions produced by a general display function. The panel function is given as an argument of the general display function. The other arguments of the general display function manage the superstructure of the graph—scales, labels, boxes around the data region, and keys. The panel function manages the symbols, lines and so forth that encode the data in the data region.

Panel function which plots "M" when the Y value is biggest.

```
panel.special <- function(x,y){
biggest <- y == max(y)
panel.points(x[!biggest], y[!biggest], pch = "+",cex=2)
panel.points(x[biggest], y[biggest], pch = "M",cex=2)
}
```

xyplot(mpg~wt|cyl.f,data=mtcars,panel=panel.special)

Visualization Methods in R

panel.abline
panel.refline
panel.curve
panel.rug
panel.average
panel.linejoin
panel.fill
panel.grid
panel.lmline
panel.mathdensity
panel.identify.qqmath
panel.levelplot
panel.levelplot.raster
panel.linejoin
panel.lines
panel.link.splom
panel.lmline
panel.loess
panel.mathdensity
panel.qqmathline

Some Useful Panel Functions

panel.abline
panel.arrows
panel.average
panel.axis
panel.barchart
panel.brush.splom
panel.bwplot
panel.cloud
panel.contourplot
panel.curve
panel.densityplot
panel.dotplot
panel.error
panel.fill
panel.functions
panel.grid
panel.histogram
panel.identify
panel.identify.cloud
panel.rect
panel.refline
panel.rug
panel.segments
panel.smoothScatter
panel.spline
panel.splom
panel.stripplot
panel.superpose
panel.superpose.2
panel.superpose.plain
panel.text
panel.tmd.default
panel.tmd.qqmath
panel.violin
panel.wireframe
panel.xyplot
panel.number
panel.pairs
panel.parallel
panel.points
panel.polygon
panel.qq
panel.qqmath

Visualization Methods in R

Some Useful Panel Functions

Drawing a loess smoother through a scatter plot is very useful:

```
panel = function(x, y){
panel.xyplot(x, y)  # plots the points
panel.loess(x, y, span = 1) #plots a loess
smoother
}
```

Another Example of using panel functions

```
bwplot(yield ~ site, barley, groups = year,
panel = function(x, y, groups, subscripts, ...) {
panel.grid(h = -1, v = 0)
panel.stripplot(x, y, ..., jitter.data = TRUE, groups = groups, subscripts = subscripts)
panel.superpose(x, y, ..., panel.groups = panel.average, groups = groups, subscripts = subscripts)
 },
auto.key = list(points = FALSE, lines = TRUE, columns = 2))
```

Visualization Methods in R

The ggplot2 Package

Ggplot2 started in 2005 started in 2005 to take the good aspects about The base and Lattice packages and improve upon them by employing A linguistic model for constructing plots.

The **ggplot2** package, created by Hadley Wickham, offers a powerful graphics language for creating elegant and complex plots. Its popularity in the R community has exploded in recent years. Originally based on Leland Wilkinson's The Grammar of Graphics, ggplot2 allows you to create graphs that represent both univariate and multivariate numerical and categorical data in a straightforward manner. Grouping can be represented by color, symbol, size, and transparency.

The Basic Function is qplot():

qplot(*x*, *y*, data=, color=, shape=, size=, alpha=, geom=, method=, formula=, facets=, xlim=, ylim= xlab=, ylab=, main=, sub=)

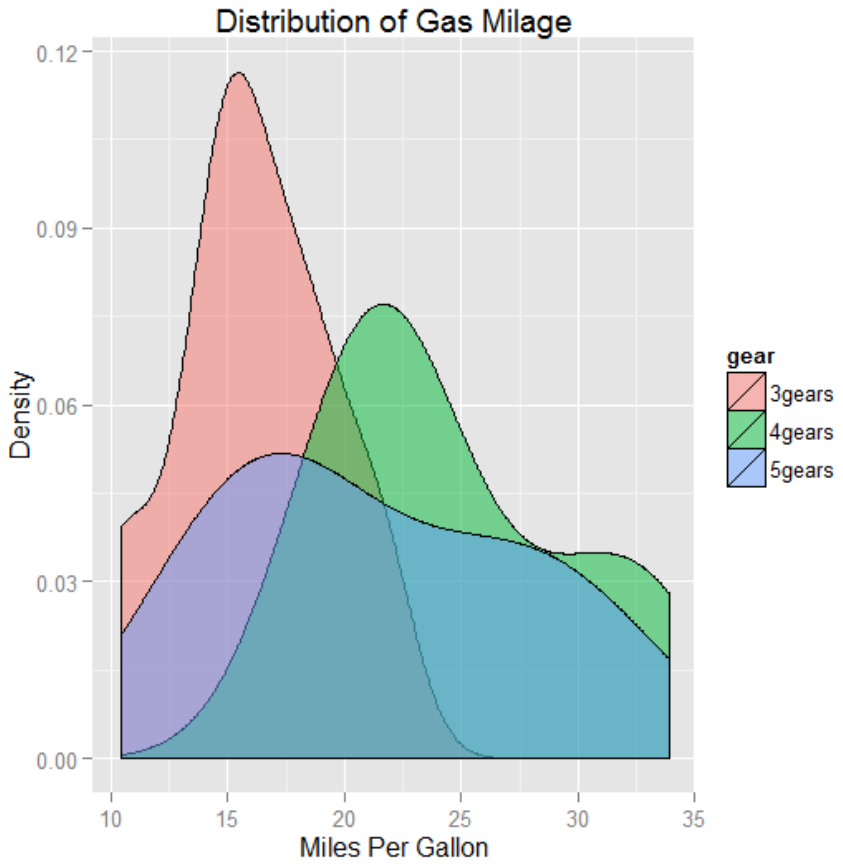
Visualization Methods in R

option	description
alpha	Alpha transparency for overlapping elements expressed as a fraction between 0 (complete transparency) and 1 (complete opacity)
color, shape, size, fill	Associates the levels of variable with symbol color, shape, or size. For line plots, color associates levels of a variable with line color. For density and box plots, fill associates fill colors with a variable. Legends are drawn automatically.
data	Specifies a data frame
facets	Creates a trellis graph by specifying conditioning variables. Its value is expressed as *rowvar* ~ *colvar*. To create trellis graphs based on a single conditioning variable, use *rowvar*~. or .~*colvar*)
geom	Specifies the geometric objects that define the graph type. The geom option is expressed as a character vector with one or more entries. geom values include "point", "smooth", "boxplot", "line", "histogram", "density", "bar", and "jitter".

option	description
main, sub	Character vectors specifying the title and subtitle
method, formula	If geom="smooth", a loess fit line and confidence limits are added by default. Methods include "lm" for regression, "gam" for generalized additive models, and "rlm" for robust regression. The formula parameter gives the form of the fit. For example, to add simple linear regression lines, you'd specify geom="smooth", method="lm", formula=y~x. Changing the formula to y~poly(x,2) would produce a quadratic fit.
x, *y*	Specifies the variables placed on the horizontal and vertical axis. For univariate plots (for example, histograms), omit *y*
xlab, ylab	Character vectors specifying horizontal and vertical axis labels
xlim,ylim	Two-element numeric vectors giving the minimum and maximum values for the horizontal and vertical axes, respectively

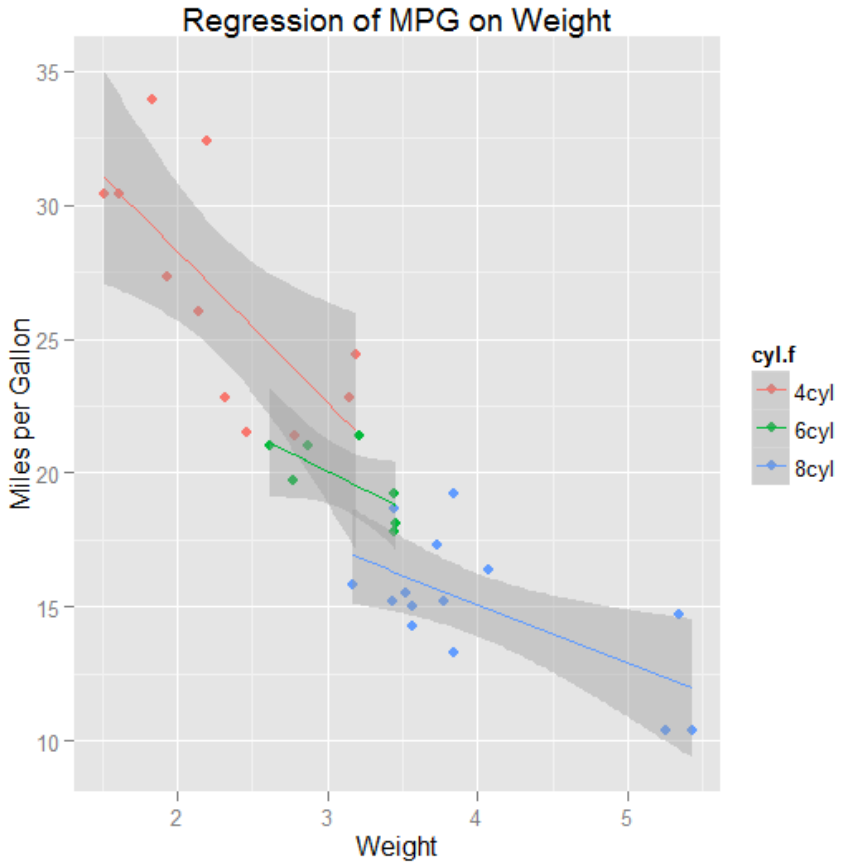
Visualization Methods in R

- qplot(mpg, data=mtcars, geom="density", fill=gear.f, alpha=I(.5), main="Distribution of Gas Milage", xlab="Miles Per Gallon",ylab="Density")



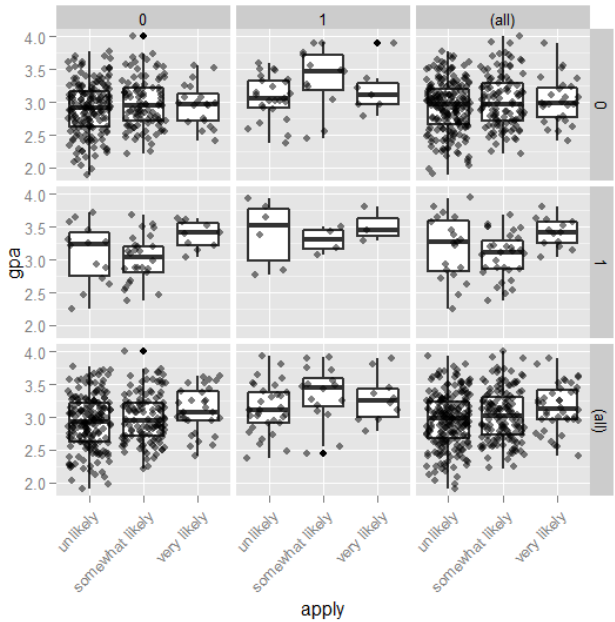
Visualization Methods in R

> qplot(wt, mpg, data=mtcars,geom=c("point", "smooth"),method="lm", formula=y~x, color=cyl.f,main="Regression of MPG on Weight",xlab="Weight", ylab="Miles per Gallon")



Visualization Methods in R

```
> library(foreign)
> dat <- read.dta("http://www.ats.ucla.edu/stat/data/ologit.dta")
> head(dat)
> ggplot(dat, aes(x = apply, y = gpa)) +
 geom_boxplot(size = .75) +
 geom_jitter(alpha = .5) +
facet_grid(pared ~ public, margins = TRUE) +
theme(axis.text.x = element_text(angle = 45, hjust = 1, vjust = 1))
```



Visualization Methods in R

Leland Wilconsin's Idea of the Grammer of Graphics

Forming a graph is like **constructing a sentence or paragraph**.

We put together a graph by joining the "geoms" which is short for geometric objects together. We can think of the geoms like the nouns of a sentence.

When utilizing the geoms (nouns) it is important to generate aesthetic mappings that describe how variables in the data are mapped to visual properties (aesthetics) of geoms. We generate aesthetic mappings by using the Function aes().

```
p <- ggplot(mtcars, aes(factor(cyl), mpg))
```

This says:
Data = mtcars
X = factor(cyl)
Y = mpg

Once the above sentence is written R doesn't know anything but what the X and Y variables are and what dataset you will be using.

```
p + geom_boxplot()
```

This says what "geom" you want to use when you plot x vs y

Visualization Methods in R

If you want to add or modify another geom onto a given aesthetic mapping you can just "add: it onto an existing geom plot

```
p + geom_boxplot() + geom_jitter()
p + geom_boxplot() + coord_flip()
p + geom_boxplot() + geom_jitter()+coord_flip()
```

Available Geoms:
"point", "smooth", "boxplot", "path", "histogram", "freqpoly", "density","bar"

```
p=ggplot(data=diamonds,aes(carat,price))

p+geom_point()+geom_smooth()
```

Can add "faceting" which is the same as "trellising" or breaking apart an existing plot Into subplots (panels or facets) by one or several variables.

```
p+geom_point()+geom_smooth()+facets_grid(cut~color,margins=TRUE)
```

Get the ggplot2 book to learn more

Categorical Data Analysis

What is Categorical Data Analysis?

Categorical data analysis is concerned with the analysis of data where the Y response is a category.

Examples:
Y = Liberal, Moderate, Conservative
Y = Disease Present, Disease Absent

Zoology Example (Alligator food source) : Y = fish, invertebrate, reptile
Medical Example: Y = Initial, advanced, remission, other
Behavior Example: Y = Depressed, neurotic, schizophrenic
Marketing Example: Y = Brand A, Brand B, Brand C, …..

Two kinds of categories:
Ordinal Category (A category with an order to it): Ex: Small, Medium, Large
Nominal Category (A category with no apparent order); Ex: Country, Folk, Rock, Jazz, Pop

Categorical Data Analysis

What is Categorical Data Analysis?

Categorical data analysis is concerned with the analysis of data where the Y response is a category.

Examples:
Y = Liberal, Moderate, Conservative
Y = Disease Present, Disease Absent

Zoology Example (Alligator food source) : Y = fish, invertebrate, reptile
Medical Example: Y = Initial, advanced, remission, other
Behavior Example: Y = Depressed, neurotic, schizophrenic
Marketing Example: Y = Brand A, Brand B, Brand C, …..

Two kinds of categories:
Ordinal Category (A category with an order to it): Ex: Small, Medium, Large
Nominal Category (A category with no apparent order); Ex: Country, Folk, Rock, Jazz, Pop

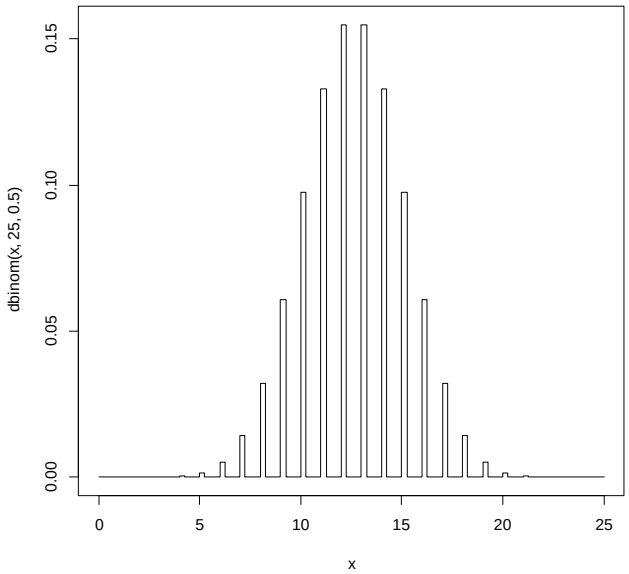
Categorical Data Analysis

Binomial Distribution

$X \sim Bin(n, p)$

Binomial experiment records the number of binary type events called "successes" which occur in n trials, where the probability of each event is constant.

$$\hat{p} = \frac{x}{n} = \frac{successes}{trials}$$



$$E[X] = np, \quad \mathrm{Var}[X] = np(1-p) = npq$$

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = p, \quad \mathrm{Var}[\hat{p}] = \mathrm{Var}\left[\frac{X}{n}\right] = \frac{npq}{n^2} = \frac{p(1-p)}{n}$$

Normal Approximation to Binomial:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Categorical Data Analysis

Two- Way Contingency Tables

	Heart Attack		
Treatment	Yes	No	Total
Placebo	189	10845	11034
Asprin	104	10933	11037
Total	293	21778	22071

	Belief in Afterlife		
Gender	Yes	No	Total
Females	509	116	625
Males	398	104	502
Total	907	220	1127

X and Y are two categorical variables
X has N levels (rows); Y has M levels (Columns)

M*N cells or M x N contingency tables.
Extends to three-way tables, etc….

Categorical Data Analysis

Let's read in some data into R from the internet

```
students = read.table(url("http://www.kkuniyuk.com/RFiles/RTutor.txt"),header=TRUE)
```

```
> students
      Name Level Grade
1   Alberto   Jun     B
2    Beryl   Sen     A
3    Chris  Soph     C
4    Debby   Sen     C
5   Ernesto   Jun     C
6  Florence   Jun     B
7   Gordon   Sen     B
8    Helen  Soph     A
9    Isaac   Jun     A
10   Joyce   Jun     B
11    Kirk   Sen     C
12   Leslie   Jun     A
13 Michael   Sen     A
14  Nadine   Jun     B
15   Oscar   Sen     B
16   Patty  Soph     C
17  Rafael   Jun     B
18   Sandy   Jun     B
19    Tony   Jun     A
20 Valerie   Jun     A
21 William   Jun     B
22  Xavier   Sen     C
23   Yancy   Jun     B
24  Yvette   Jun     A
25     Yul   Sen     B
26   Zorro   Sen     B
```

Categorical Data Analysis

Constructing Table Objects in R

```
> attach(students)
> T=table(Level,Grade)
> T
      Grade
Level  A B C
  Jun  5 8 1
  Sen  2 4 3
  Soph 1 0 2
```

```
> G=xtabs(~Level+Grade)
> G
      Grade
Level  A B C
  Jun  5 8 1
  Sen  2 4 3
  Soph 1 0 2
```

Re-ordering the Levels in a Contingency Table.

```
> G=G[c(3,1,2),]
> G
      Grade
Level  A B C
  Soph 1 0 2
  Jun  5 8 1
  Sen  2 4 3
```

Categorical Data Analysis

Data(Titanic)

4-Way Table

X = Class (1^{st},2^{nd},3^{rd},Crew)
Y = Sex (Male, Female)
Z = Age (Child,Adult)
W = Survived (Yes,No)

See also
data(UCBAdmissions)

Some other Famous Tables in R

```
> Titanic
, , Age = Child, Survived = No

     Sex
Class  Male Female
  1st    0      0
  2nd    0      0
  3rd   35     17
  Crew   0      0

, , Age = Adult, Survived = No

     Sex
Class  Male Female
  1st  118      4
  2nd  154     13
  3rd  387     89
  Crew 670      3

, , Age = Child, Survived = Yes

     Sex
Class  Male Female
  1st    5      1
  2nd   11     13
  3rd   13     14
  Crew   0      0

, , Age = Adult, Survived = Yes

     Sex
Class  Male Female
  1st   57    140
  2nd   14     80
  3rd   75     76
  Crew 192     20
```

Categorical Data Analysis

ftable() takes a many way contingency table and “flattens it” so it looks l
Like an excel file which could be printed

			Survived	
Class	Sex	Age	No	Yes
1st	Male	Child	0	5
		Adult	118	57
	Female	Child	0	1
		Adult	4	140
2nd	Male	Child	0	11
		Adult	154	14
	Female	Child	0	13
		Adult	13	80
3rd	Male	Child	35	13
		Adult	387	75
	Female	Child	17	14
		Adult	89	76
Crew	Male	Child	0	0
		Adult	670	192
	Female	Child	0	0
		Adult	3	20

Categorical Data Analysis

```
> library(gmodels)
> ?CrossTable
> CrossTable(G)

  Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  26

             | Grade
       Level |         A |         B |         C | Row Total |
-------------|-----------|-----------|-----------|-----------|
        Soph |         1 |         0 |         2 |         3 |
             |     0.006 |     1.385 |     2.470 |           |
             |     0.333 |     0.000 |     0.667 |     0.115 |
             |     0.125 |     0.000 |     0.333 |           |
             |     0.038 |     0.000 |     0.077 |           |
-------------|-----------|-----------|-----------|-----------|
         Jun |         5 |         8 |         1 |        14 |
             |     0.111 |     0.366 |     1.540 |           |
             |     0.357 |     0.571 |     0.071 |     0.538 |
             |     0.625 |     0.667 |     0.167 |           |
             |     0.192 |     0.308 |     0.038 |           |
-------------|-----------|-----------|-----------|-----------|
         Sen |         2 |         4 |         3 |         9 |
             |     0.214 |     0.006 |     0.410 |           |
             |     0.222 |     0.444 |     0.333 |     0.346 |
             |     0.250 |     0.333 |     0.500 |           |
             |     0.077 |     0.154 |     0.115 |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |         8 |        12 |         6 |        26 |
             |     0.308 |     0.462 |     0.231 |           |
-------------|-----------|-----------|-----------|-----------|
```

The **CrossTable( )** function in the **gmodels** package produces crosstabulations modeled after PROC FREQ in **SAS** or CROSSTABS in **SPSS**. It has a wealth of options.

Categorical Data Analysis

Some More Useful R Commands

```
margin.table(mytable, 1) # row frequencies (summed over rows)
margin.table(mytable, 2) # column frequencies (summed over A)

prop.table(mytable) # cell percentages
prop.table(mytable, 1) # row percentages
prop.table(mytable, 2) # column percentages
```

```
> margin.table(G,1)
Level
Soph  Jun  Sen
   3   14    9
> margin.table(G,2)
Grade
 A  B  C
 8 12  6
```

```
> prop.table(G)
      Grade
Level         A         B         C
  Soph 0.03846154 0.0000000 0.07692308
  Jun  0.19230769 0.3076923 0.03846154
  Sen  0.07692308 0.1538462 0.11538462
```

Categorical Data Analysis

Contingency Table Notation

Frequencies:

Y	A	X B	Total
1	n_{11}	n_{21}	n_{+1}
2	n_{12}	n_{22}	n_{+2}
Total	n_{1+}	n_{2+}	n

Row & Column Marginal Totals

Proportions:

Y	A	X B	Total
1	π_{11}	π_{21}	π_{+1}
2	π_{12}	π_{22}	π_{+2}
Total	π_{1+}	π_{2+}	1

Some Texts will use O_{ij} for Observed cell counts:

Y	A	X B	Total
1	O_{11}	O_{21}	O_{+1}
2	O_{12}	O_{22}	O_{+2}
Total	O_{1+}	O_{2+}	n

Categorical Data Analysis

Tests for Independence

H0: X and Y are independent; HA: X and Y are not independent

IDEA: IF X and Y are independent then

$$\pi_{ij} = \pi_{i+} \bullet \pi_{+j} \quad \text{for all i and j}$$

Under assumption that H0 is true:

$\mu_{ij} = n\pi_{ij} = n\pi_{i+} \bullet \pi_{+j}$

So the estimated EXPECTED cell count (cell frequency) in cell i, j is:

$$\hat{\mu}_{ij} = E_{ij} = n\hat{\pi}_{i+} \bullet \hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \left(\frac{n_{i+}n_{+j}}{n}\right)$$

Based on normal Approximation to Binomial

To test independence in a M x N table use

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$ Pearson's Chi-Square Statistic or

Based upon Multinomial Likelihood

$$G^2 = -2\left(\frac{\text{max likelihood when parameters satisfy H0}}{\text{max likelihood when parameters are unrestricted}}\right) = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$$

Under H0 both tests follow $\chi^2_{(M-1)(N-1)}$

Categorical Data Analysis

The Expected Table:

		X	
Y	A	B	Total
1	E_{11}	E_{21}	n_{+1}
2	E_{12}	E_{22}	n_{+2}
Total	n_{1+}	n_{2+}	n

$$E_{ij} = \frac{O_{i+}O_{+j}}{n}$$

Trick to Get the Expected Table:

```
> a=margin.table(G,1)  #  a  is a vector of row marginal totals
> b=margin.table(G,2) #  b  is a vector of column marginal totals
> a
Level
Soph  Jun  Sen
  3   14   9
> b
Grade
 A  B  C
 8 12  6
> Expected=a%o%b/sum(a)    #  Expected is OUTER PRODUCT  of  a  and b
> Expected
     Grade
Level        A        B        C
 Soph 0.9230769 1.384615 0.6923077
 Jun  4.3076923 6.461538 3.2307692
 Sen  2.7692308 4.153846 2.0769231
```

Categorical Data Analysis

Test for Independence in R

Can use chisq.test() or CrossTable() in the gmodels package.

```
chisq.test(x, y = NULL, correct = TRUE,
p = rep(1/length(x), length(x)), rescale.p = FALSE, simulate.p.value = FALSE, B = 2000)
```

```
CrossTable(x, y, digits=3, max.width = 5, expected=FALSE, prop.r=TRUE, prop.c=TRUE,
prop.t=TRUE, prop.chisq=TRUE, chisq = FALSE, fisher=FALSE, mcnemar=FALSE,
resid=FALSE, sresid=FALSE, asresid=FALSE, missing.include=FALSE, format=c("SAS","SPSS"), dnn = NULL, ...)
```

```
> chisq.test(G)

        Pearson's Chi-squared test

data:  G
X-squared = 6.5086, df = 4, p-value = 0.1642

Warning message:
In chisq.test(G) : Chi-squared approximation may be incorrect
```

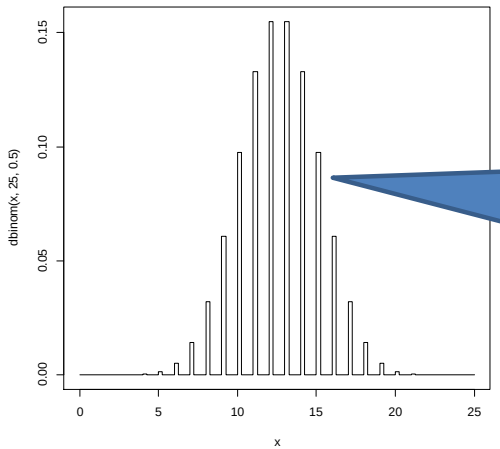
Beware of Chi-Square Test if Expected Cell Counts < 5

Also try:

```
> P=CrossTable(G,expected=TRUE,chisq=TRUE)
```

Categorical Data Analysis

Pearson's Continuity Corrected Chi-Square Test



IDEA: Binomial dist has heights centered at the integers. Tis better to offset them So they are centered at integers + ½

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Peason's Chi-Sq

$$X^2 = \sum \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}$$

Peason's Chi-Sq With Yates Continuity Correction.

Yate's Continuity Correction better approximates $(\text{Normal})^2$ Distribution
Is a little better for small sample sizes but still beware if $E_{ij} < 5$

Categorical Data Analysis

```
> T[3,2]=1
> T
     Grade
Level  A B C
 Jun  5 1 1
 Sen  2 4 3
 Soph 1 1 2
> E=margin.table(T,1)%o%margin.table(T,2)/margin.table(T)
> E
     Grade
Level   A   B   C
 Jun  2.8 2.1 2.1
 Sen  3.6 2.7 2.7
 Soph 1.6 1.2 1.2
```

Observed Table

Expected Table

```
> summary(T)
Number of cases in table: 20
Number of factors: 2
Test for independence of factors:
        Chisq = 5.043, df = 4, p-value = 0.2829
        Chi-squared approximation may be incorrect
```

Pearson's Chi-Square

```
> LR=2*T*log(T/E)
> LR
     Grade
Level         A         B         C
 Jun   5.7981850 -1.4838747 -1.4838747
 Sen  -2.3511467  3.1443407  0.6321631
 Soph -0.9400073 -0.3646431  2.0433025
```

$$2O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right)$$

$$G^2 = 2\sum O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right)$$

```
> G2=sum(LR)
> G2
[1] 4.994445
```

Likelihood Ratio or deviance Chi-Square

G^2 and X^2 are approx the same for large sample Sizes. The sampling Distributions get closer to Chi-square distribution as N → Infinity. The convergence of X^2 is typically quicker than that of G^2

Categorical Data Analysis

Residuals

- Need a cell by cell comparison, for evidence against null hypothesis.
- Observed minus expected (O_{ij} - E_{ij}) is not the best choice since it tends to be large with large E_{ij}
- It is useful to calculate the adjusted (standardized) residuals

$$resid_{ij} = \frac{(n_{ij} - \hat{\mu}_{ij})}{\sqrt{\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})}}$$

- Under H0 the adjusted residuals ~ N(0,1)

Can get adjusted residuals from chisq.test() function or CrossTable() function in R

Illustrate this in R

Categorical Data Analysis

Test for Equality in Proportions

The Chi Square Test Can also be used to test Equality in Proportions!!

```
## Data from Fleiss (1981), p. 139.
## H0: The null hypothesis is that the four populations from which
## the patients were drawn have the same true proportion of smokers.
## A: The alternative is that this proportion is different in at
## least one of the populations.

smokers <- c( 83, 90, 129, 70 )
patients <- c( 86, 93, 136, 82 )
prop.test(smokers, patients)
```

4-sample test for equality of proportions without continuity correction

data: smokers out of patients
X-squared = 12.6004, df = 3, p-value = 0.005585
alternative hypothesis: two.sided
sample estimates:
prop 1 prop 2 prop 3 prop 4
0.9651163 0.9677419 0.9485294 0.8536585

H0: $\pi_i = \pi_j$ for all row indicies i, j ; HA: $\pi_i \neq \pi_j$ for at least one pair of row indexes with $i \neq j$.

Categorical Data Analysis

Fishers Exact Test

```
> T=matrix(c(2,23,5,30),ncol=2,byrow=TRUE)
> rownames(T)=c("non-CVD", "CVD")
> colnames(T)=c("High salt", "Low salt")

> R=as.table(T)
> R
      High salt Low salt
non-CVD         2       23
CVD             5       30
```

Do high salt diets cause more cardiovascular disease than low salt diets?

H0: $\pi_1 = \pi_2$

HA: $\pi_1 \neq \pi_2$

Prop.test() is based upon large sample approx
probably not good for small samples

Categorical Data Analysis

Fisher's Test

	X		
Y	A	B	Total
1	n_{11}	n_{21}	n_{+1}
2	n_{12}	n_{22}	n_{+2}
Total	n_{1+}	n_{2+}	n

	High salt	Low salt	Total
non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

Pretend that the Row Marginal Totals and Column Marginal Totals are fixed.

Row Margins = M1, M2 Column Margins = N1, N2

If Margins are constant the table is uniquely determined by n_{11}

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{1+}-n_{11}}}{\binom{n}{n_{1+}}} = dhyper(n_{11}, n_{+1}, n_{+2}, n_{1+})$$

Categorical Data Analysis

Consider all possible contingency tables
Arranged in assending order of n_{11}

$n_{11} = 0$

	High salt	Low salt	Total
non-CVD	0	25	25
CVD	7	28	35
Total	7	53	60

> dhyper(0,7,53,25)
[1] 0.0174117

$n_{11} = 1$

	High salt	Low salt	Total
non-CVD	1	24	25
CVD	6	29	35
Total	7	53	60

> dhyper(1,7,53,25)
[1] 0.105

$n_{11} = 2$

	High salt	Low salt	Total
non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

> dhyper(3,7,53,25)
[1] 0.2521695

$n_{11} = 3$

	High salt	Low salt	Total
non-CVD	3	22	25
CVD	4	31	35
Total	7	53	60

> dhyper(3,7,53,25)
[1] 0.312

$n_{11} = 4$

	High salt	Low salt	Total
non-CVD	4	21	25
CVD	3	32	35
Total	7	53	60

> dhyper(4,7,53,25)
[1] 0.214

$n_{11} = 5$

	High salt	Low salt	Total
non-CVD	5	20	25
CVD	2	33	35
Total	7	53	60

> dhyper(5,7,53,25)
[1] 0.082

$n_{11} = 6$

	High salt	Low salt	Total
non-CVD	6	19	25
CVD	1	34	35
Total	7	53	60

> dhyper(6,7,53,25)
[1] 0.016

$n_{11} = 7$

	High salt	Low salt	Total
non-CVD	7	18	25
CVD	0	35	35
Total	7	53	60

> dhyper(7,7,53,25)
[1] 0.001

Categorical Data Analysis

P-value is probability of observing a Contingency table
More extreme than the one that was actually observed

$n_{11} = 0$

	High salt	Low salt	Total
non-CVD	0	25	25
CVD	7	28	35
Total	7	53	60

> dhyper(0,7,53,25)
[1] 0.0174117

$n_{11} = 1$

	High salt	Low salt	Total
non-CVD	1	24	25
CVD	6	29	35
Total	7	53	60

> dhyper(1,7,53,25)
[1] 0.105

$n_{11} = 2$

	High salt	Low salt	Total
non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

> dhyper(2,7,53,25)
[1] 0.2521695

$n_{11} = 3$

	High salt	Low salt	Total
non-CVD	3	22	25
CVD	4	31	35
Total	7	53	60

> dhyper(3,7,53,25)
[1] 0.312

$n_{11} = 4$

	High salt	Low salt	Total
non-CVD	4	21	25
CVD	3	32	35
Total	7	53	60

> dhyper(4,7,53,25)
[1] 0.214

$n_{11} = 5$

	High salt	Low salt	Total
non-CVD	5	20	25
CVD	2	33	35
Total	7	53	60

> dhyper(5,7,53,25)
[1] 0.082

$n_{11} = 6$

	High salt	Low salt	Total
non-CVD	6	19	25
CVD	1	34	35
Total	7	53	60

> dhyper(6,7,53,25)
[1] 0.016

$n_{11} = 7$

	High salt	Low salt	Total
non-CVD	7	18	25
CVD	0	35	35
Total	7	53	60

> dhyper(7,7,53,25)
[1] 0.001

Circled ones are More extreme Than observed one

Categorical Data Analysis

Circled ones are less extreme than observed one

$n_{11} = 0$

	High salt	Low salt	Total
non-CVD	0	25	25
CVD	7	28	35
Total	7	53	60

> dhyper(0,7,53,25)
[1] 0.0174117

$n_{11} = 1$

	High salt	Low salt	Total
non-CVD	1	24	25
CVD	6	29	35
Total	7	53	60

> dhyper(1,7,53,25)
[1] 0.105

$n_{11} = 2$

	High salt	Low salt	Total
non-CVD	2	23	25
CVD	5	30	35
Total	7	53	60

> dhyper(2,7,53,25)
[1] 0.2521695

$n_{11} = 3$

	High salt	Low salt	Total
non-CVD	3	22	25
CVD	4	31	35
Total	7	53	60

> dhyper(3,7,53,25)
[1] 0.312

$n_{11} = 4$

	High salt	Low salt	Total
non-CVD	4	21	25
CVD	3	32	35
Total	7	53	60

> dhyper(4,7,53,25)
[1] 0.214

$n_{11} = 5$

	High salt	Low salt	Total
non-CVD	5	20	25
CVD	2	33	35
Total	7	53	60

> dhyper(5,7,53,25)
[1] 0.082

$n_{11} = 6$

	High salt	Low salt	Total
non-CVD	6	19	25
CVD	1	34	35
Total	7	53	60

> dhyper(6,7,53,25)
[1] 0.016

$n_{11} = 7$

	High salt	Low salt	Total
non-CVD	7	18	25
CVD	0	35	35
Total	7	53	60

> dhyper(7,7,53,25)
[1] 0.001

Because the sum of these contingency table probabilities > 0.5

Categorical Data Analysis

Two-Sided Hypothesis Test: Ha $\pi_1 \neq \pi_2$

Pval = 2* sum of probabilities for contingency tables more extreme the observed one

One-Sided Hypothesis Test: Ha $\pi_1 < \pi_2$ or $\pi_1 > \pi_2$

Pval = Sum the probabilities for contingency tables in direction of alternate hypothesis

```
> pval=2*(dhyper(2,7,53,25)+dhyper(1,7,53,25)+dhyper(0,7,53,25))
> pval
[1] 0.6882
```

OR...

```
> pval=2*phyper(2,7,53,25)
> pval
[1] 0.6882
```

OR...

```
> fisher.test(R)

        Fisher's Exact Test for Count Data

data:  R
p-value = 0.6882
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.04625243 3.58478157
sample estimates:
odds ratio
 0.527113
```

Categorical Data Analysis

Visualizing Multi-Way Contingency Tables

```
> library(vcd)
> HairEyeColor
, , Sex = Male

       Eye
Hair    Brown Blue Hazel Green
  Black   32   11    10     3
  Brown   53   50    25    15
  Red     10   10     7     7
  Blond    3   30     5     8

, , Sex = Female

       Eye
Hair    Brown Blue Hazel Green
  Black   36    9     5     2
  Brown   66   34    29    14
  Red     16    7     7     7
  Blond    4   64     5     8
```

The VCD Package

The VCD Package

Graphics in VCD package are based off the structable which is similar to the falt table produced by ftable() Except it explicitly allows for one to select the split direction. One can think of the horizontal dimension In these flat table displays much like defining the Y-variable.

```
HEC = structable(Sex ~ Eye + Hair,data=HairEyeColor)
HEC
        Sex Male Female
Eye   Hair
Brown Black      32    36
    Brown        53    66
    Red          10    16
    Blond         3     4
Blue  Black      11     9
    Brown        50    34
    Red          10     7
    Blond        30    64
Hazel Black      10     5
    Brown        25    29
    Red           7     7
    Blond         5     5
Green Black       3     2
    Brown        15    14
    Red           7     7
    Blond         8     8
```

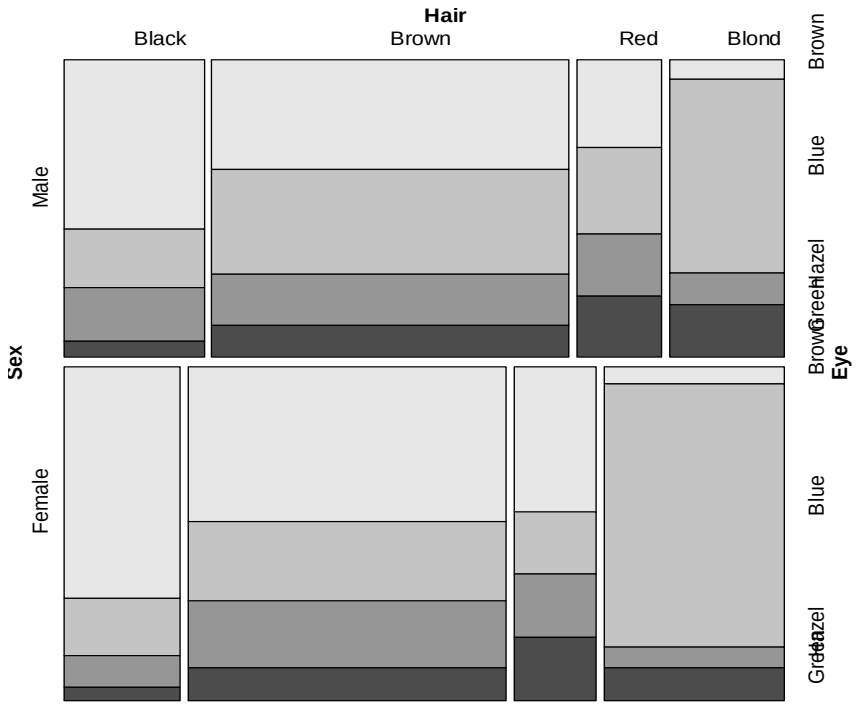
```
HEC = structable(Eye ~ Sex + Hair,data=HairEyeColor)
HEC
        Eye Brown Blue Hazel Green
Sex    Hair
Male   Black      32   11    10    3
    Brown         53   50    25   15
    Red           10   10     7    7
    Blond          3   30     5    8
Female Black      36    9     5    2
    Brown         66   34    29   14
    Red           16    7     7    7
    Blond          4   64     5    8
```

Can read the two vignettes on the package by typing in vignette("strucplot")

The VCD Package

The Mosaic Plot

mosaic(Eye ~ Sex + Hair, data = HairEyeColor)




```
HEC = structable(Eye ~ Sex + Hair,data=HairEyeColor)
HEC
        Eye Brown Blue Hazel Green
Sex    Hair
Male   Black    32  11    10    3
       Brown    53  50    25   15
       Red      10  10     7    7
       Blond     3  30     5    8
Female Black    36   9     5    2
       Brown    66  34    29   14
       Red      16   7     7    7
       Blond     4  64     5    8
```

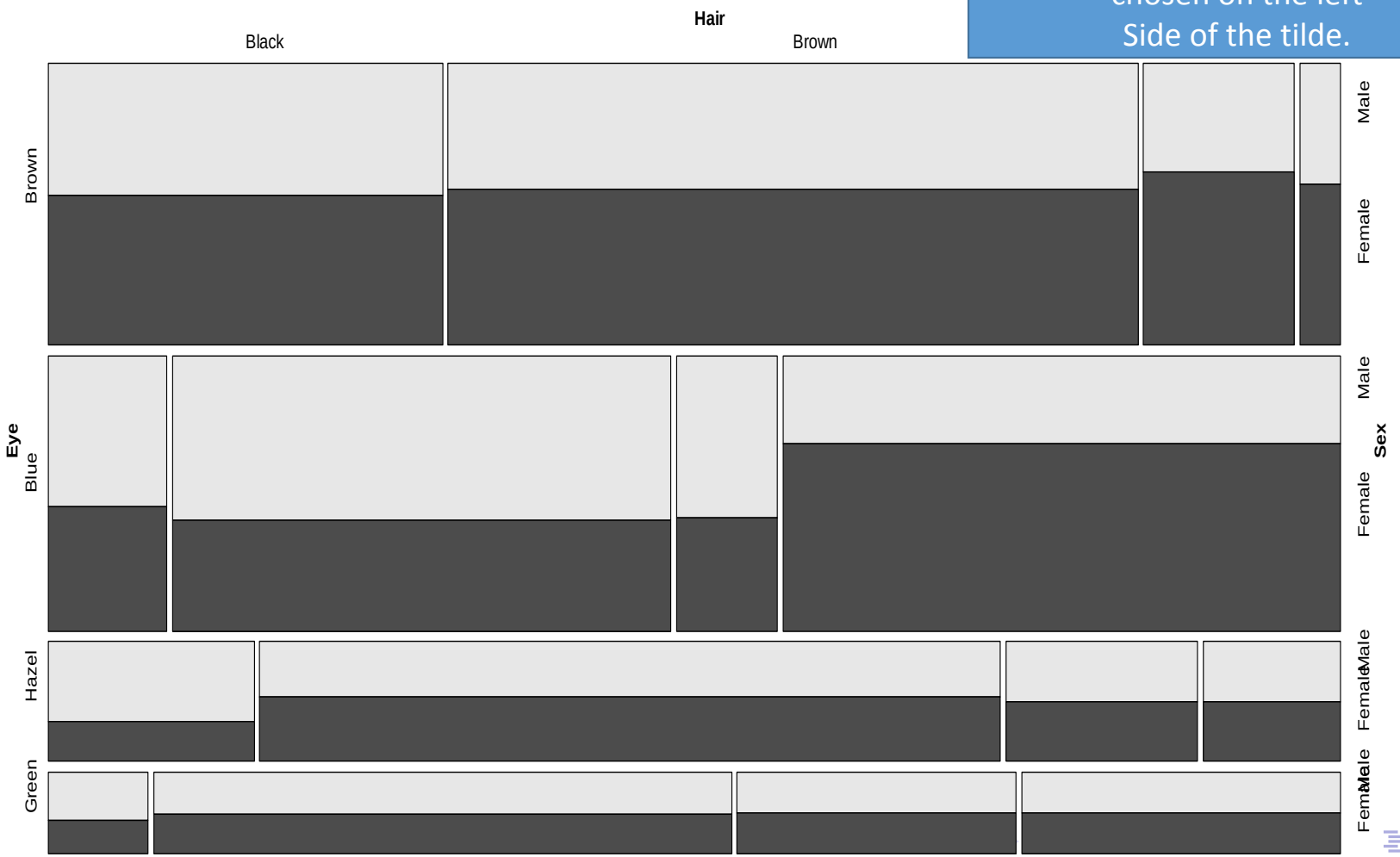
Is better than stacked bar charts because the width of the plot is controlled by the count in the column $n_{+i.}$ The height is controlled by the count in the row n_{j+}

Area of each rectangle is proportional to count

The VCD Package

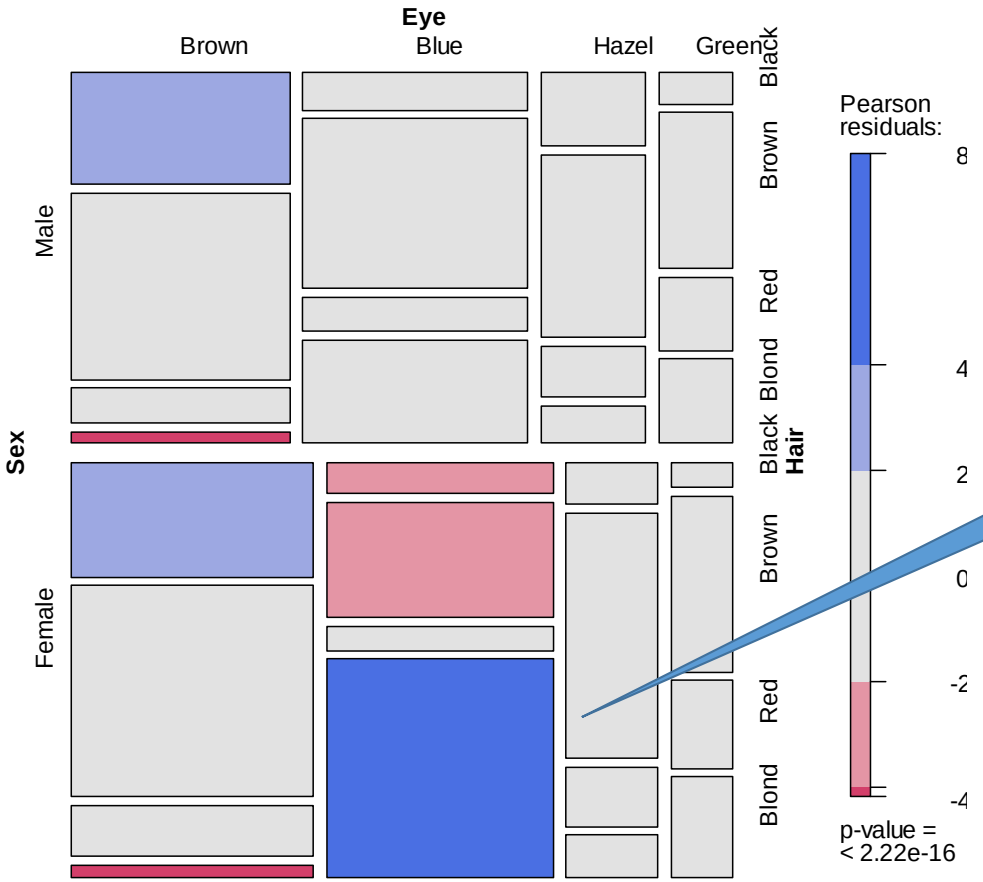
mosaic(Sex ~ Eye + Hair, data = HairEyeColor)

What you dimension you shade is chosen on the left Side of the tilde.



The VCD Package

mosaic(~ Sex + Eye + Hair, data = HairEyeColor,shade=TRUE,legend=TRUE)

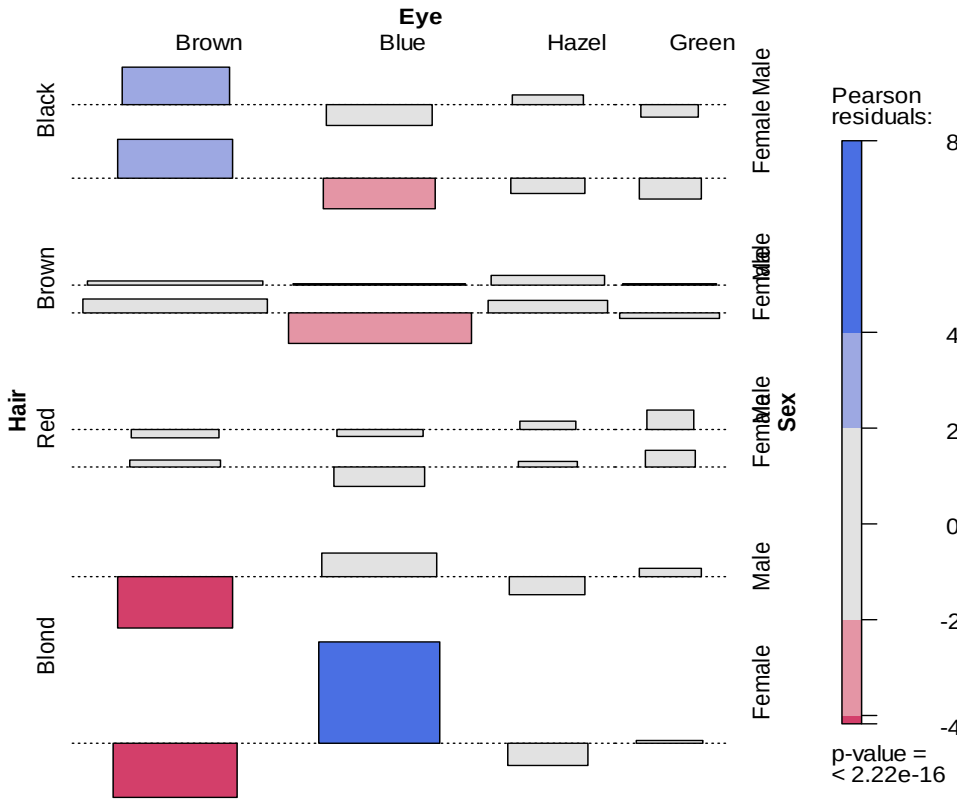


Shading is determined by Pearson residuals when shade=TRUE

Area of rectangle is proportional to the count in each cell.

The VCD Package

assoc(HairEyeColor,shade=TRUE,legend=TRUE)

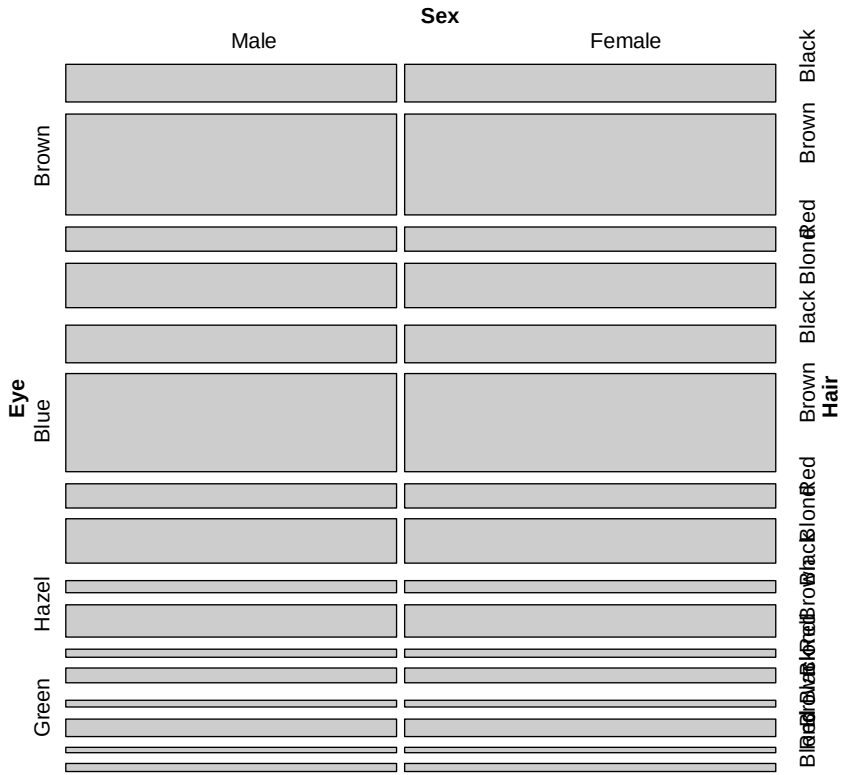


Association plot is plot of residuals

Width of plot is controlled by count in each cell

The VCD Package

mosaic(HEC, type = "expected")



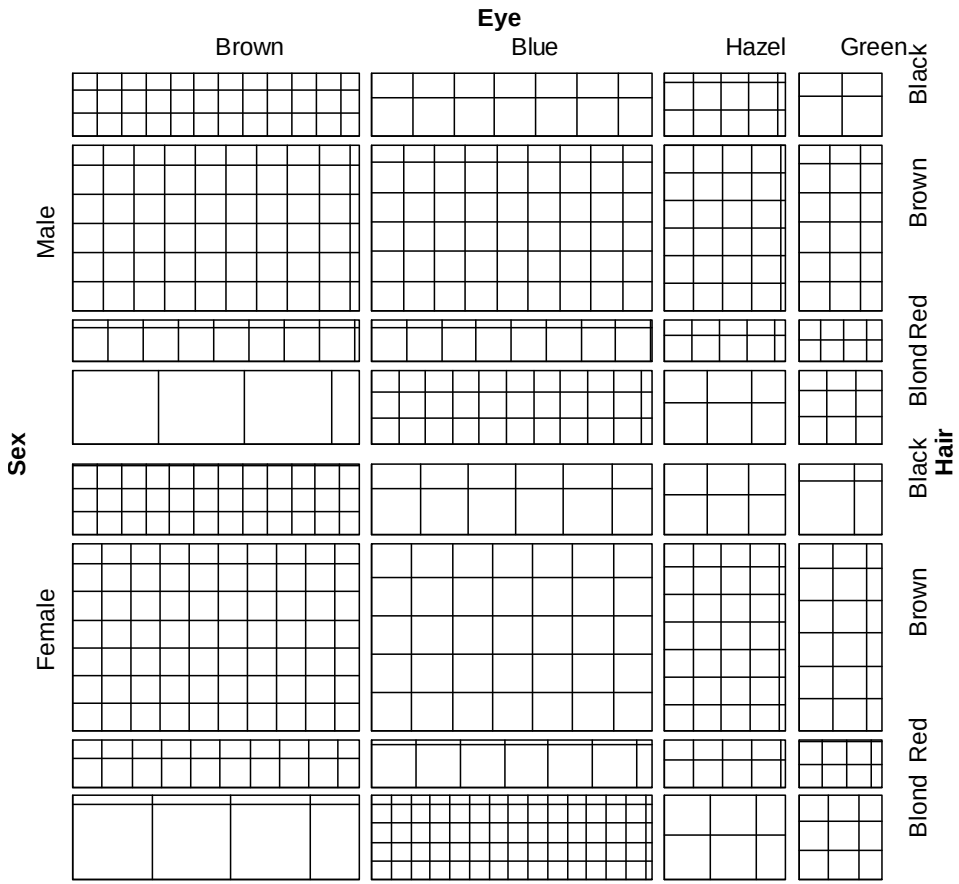
If we say type = “expected” the width of rectangle = n_{i+}

The height of each rectangle is proportional to n_{j+}

So the area of each rectangle is proportional to $E_{ij} = (n_{+}$

The VCD Package

sieve(~Sex + Eye + Hair, data = HEC)



The Sieve Plot

When the two variables are independent, then the expected frequency is:

$$E_{ij} = n_{i+}n_{+j}/n_{++}$$

In a sieve plot, each E_{ij} is represented by a rectangle. The width of the rectangle is proportional to the total frequency in each column, n_{+j} and the height is proportional to the total frequency in each row, n_{i+}. The area of the rectangle is then proportional to E_{ij}.

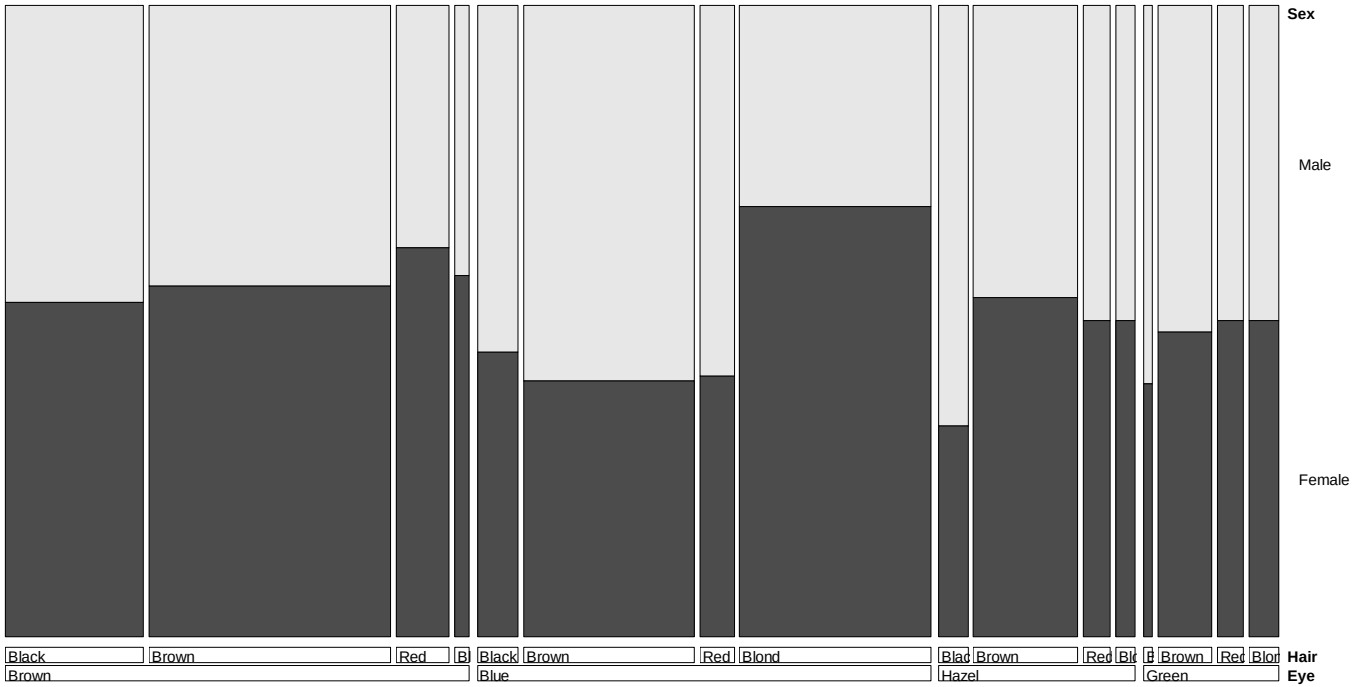
Each rectangle is then cross-ruled based on the c
frequency. The deviations from independence are
reflected in the density of the shading. Denser sha
indicates the observed frequency is greater than e
and sparse shading indicates the observed freque
less than expected. As an additional cue, positive
negative departures from independence can be c
with different colors.

The VCD Package

Double Decker Plot

Similar to stacked bar chart only that area is proportional to observed count

```
doubledecker(Survived ~ Class + Sex + Age, data = Titanic)
HEC = structable(Sex ~ Eye + Hair,data=HairEyeColor)
doubledecker(HEC)
```



The VCD Package

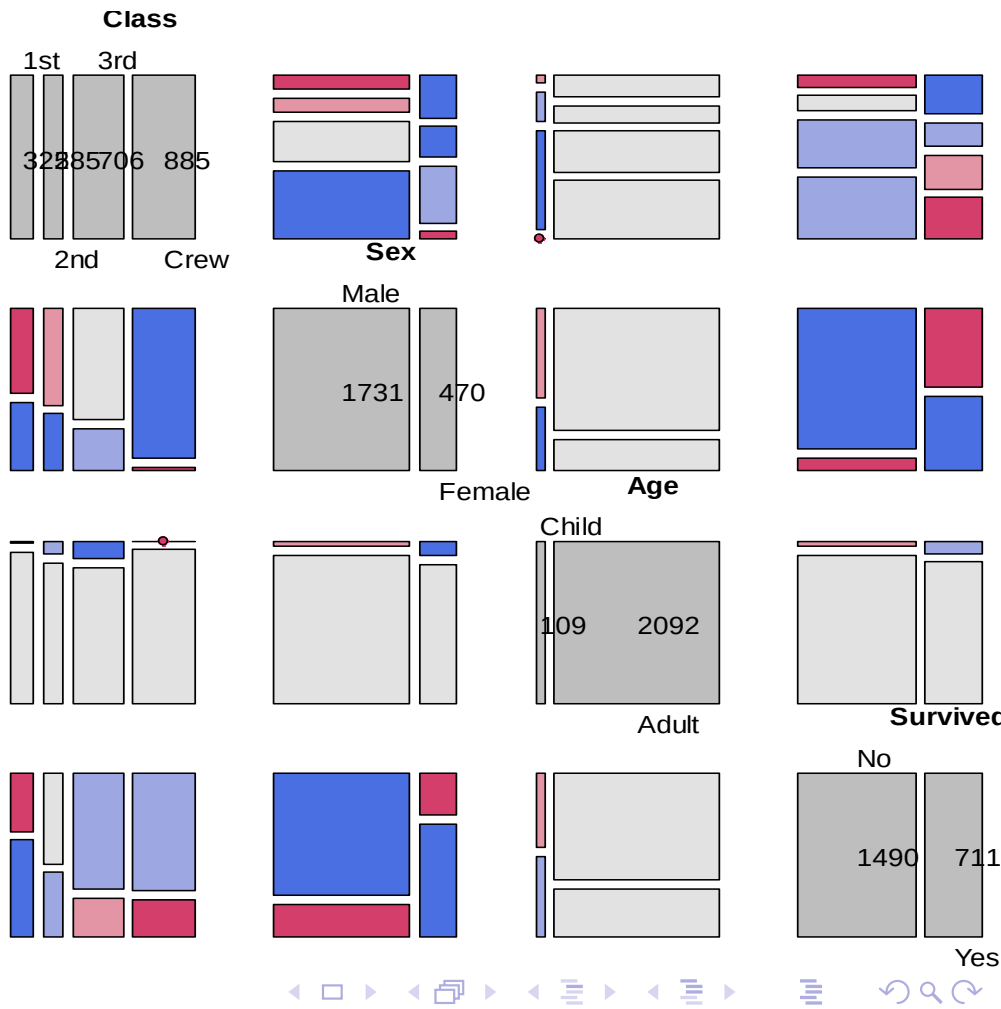
The Pairs Plot

The pairs plot is similar to a scatter plot matrix.

Down the diagonal are univariate mosaic plots Of each variable.

On off diagonals are bivariate mosaic plots of pairwise combinations of variables.

pairs(Titanic,shade=TRUE)



The VCD Package

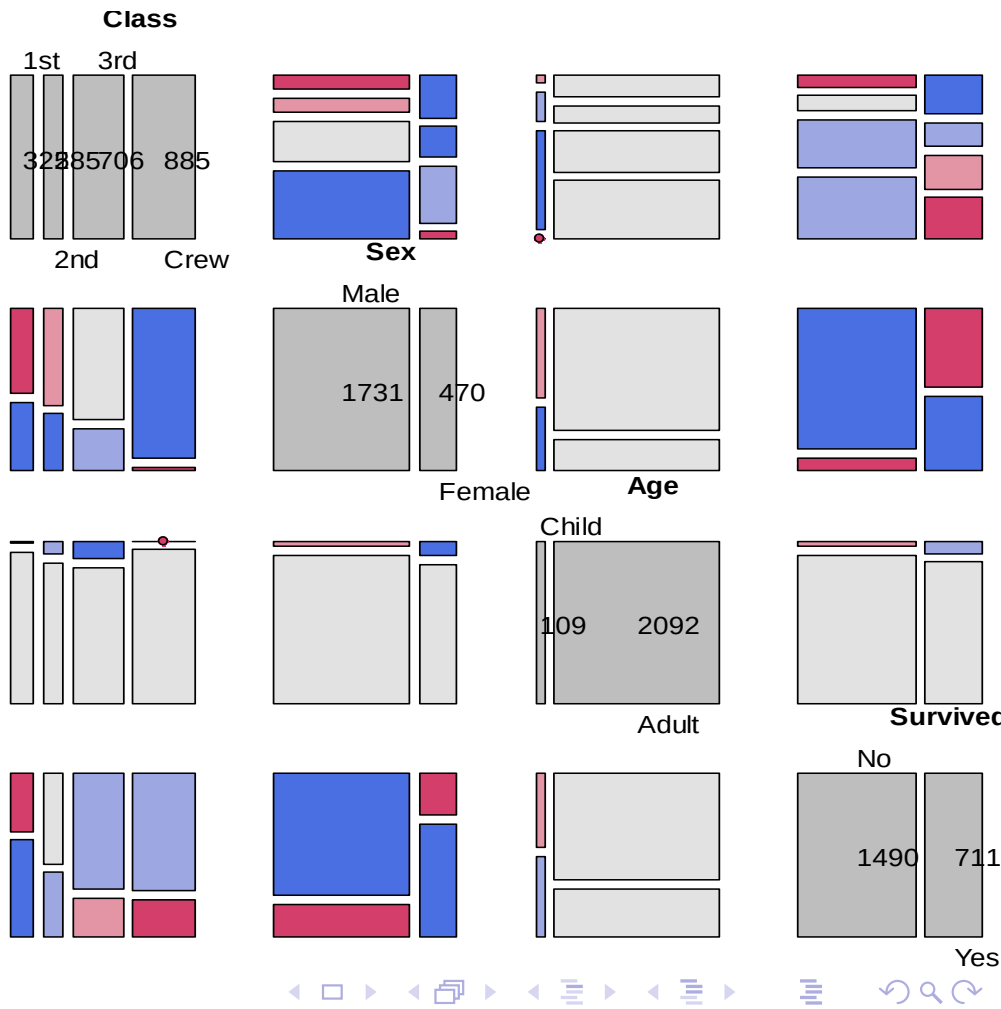
The Pairs Plot

The pairs plot is similar to a scatter plot matrix.

Down the diagonal are univariate mosaic plots Of each variable.

On off diagonals are bivariate mosaic plots of pairwise combinations of variables.

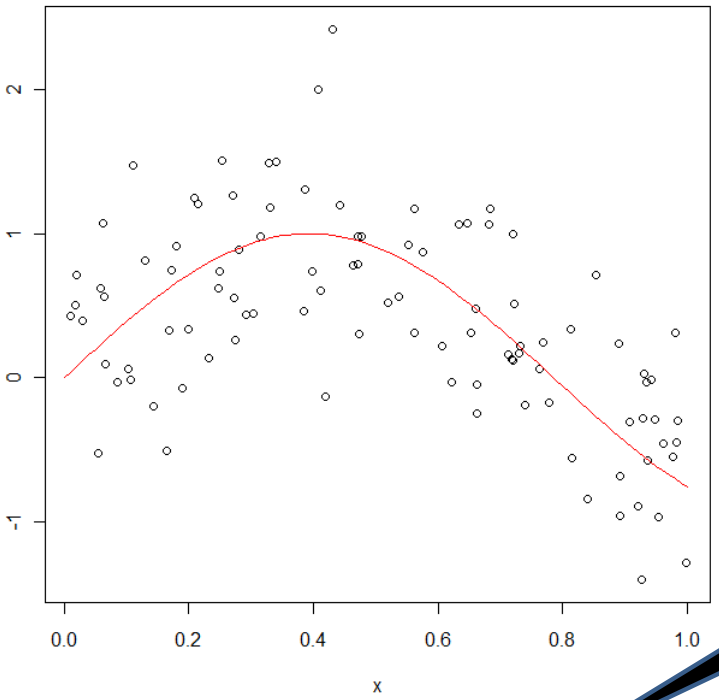
pairs(Titanic,shade=TRUE)



Bootstrapping and Crossvalidation

Model Assessment, Bootstrapping and Crossvalidation

Bootstrapping and Crossvalidation



- x=runif(100)
- truey=sin(4*x)
- errors=rnorm(100,mean=0,sd=0.5)
- y=truey+errors
- plot(x,y)
- curve(sin(4*x),from=0,to=1,col="red")

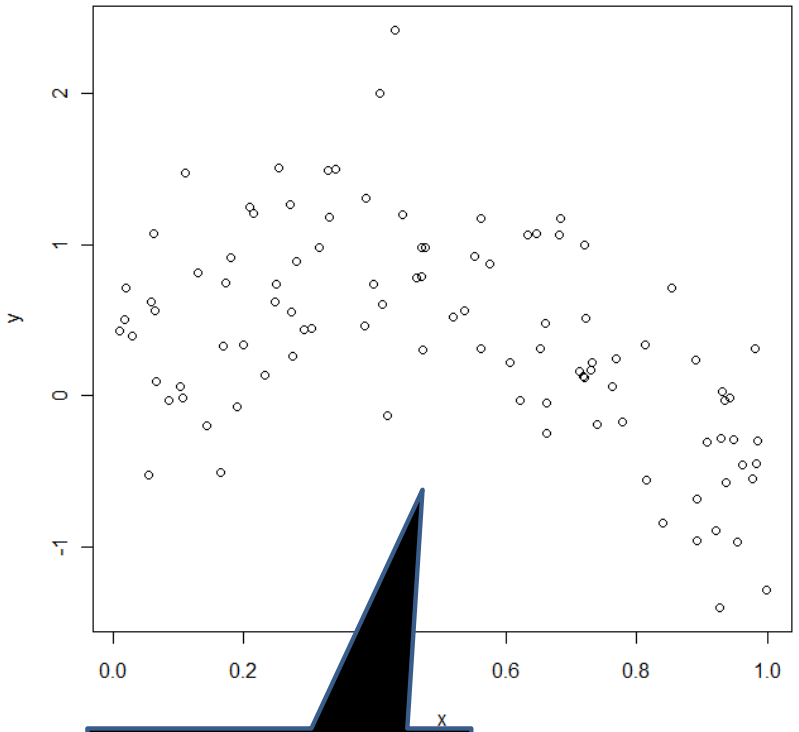
True Functional Relationship between X and Y

$Y = f(X) + e$

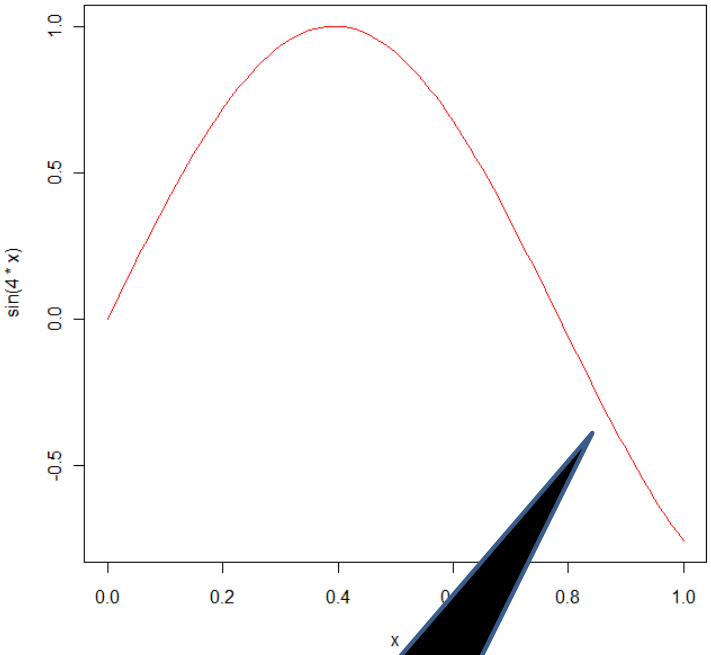
Experimental Errors

Game: Estimate the true relationship between X and Y

Bootstrapping and Crossvalidation



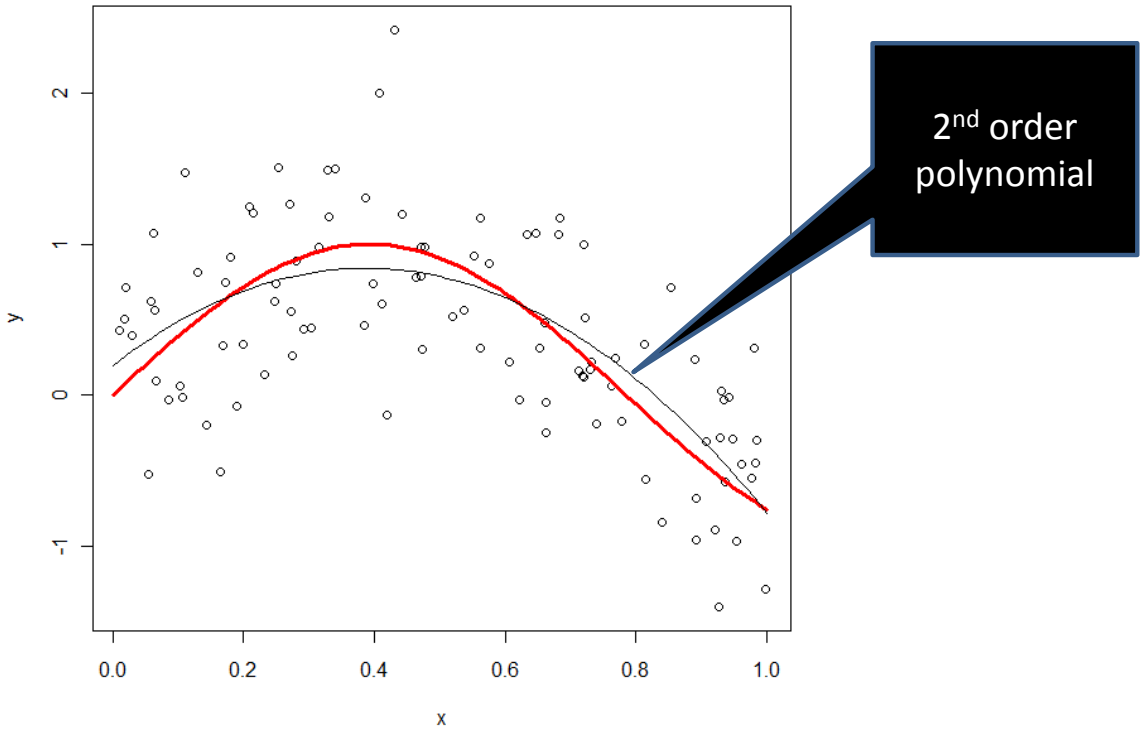
Mortals get to "see" this – the data = signal + noise



Mortals will never know the true signal

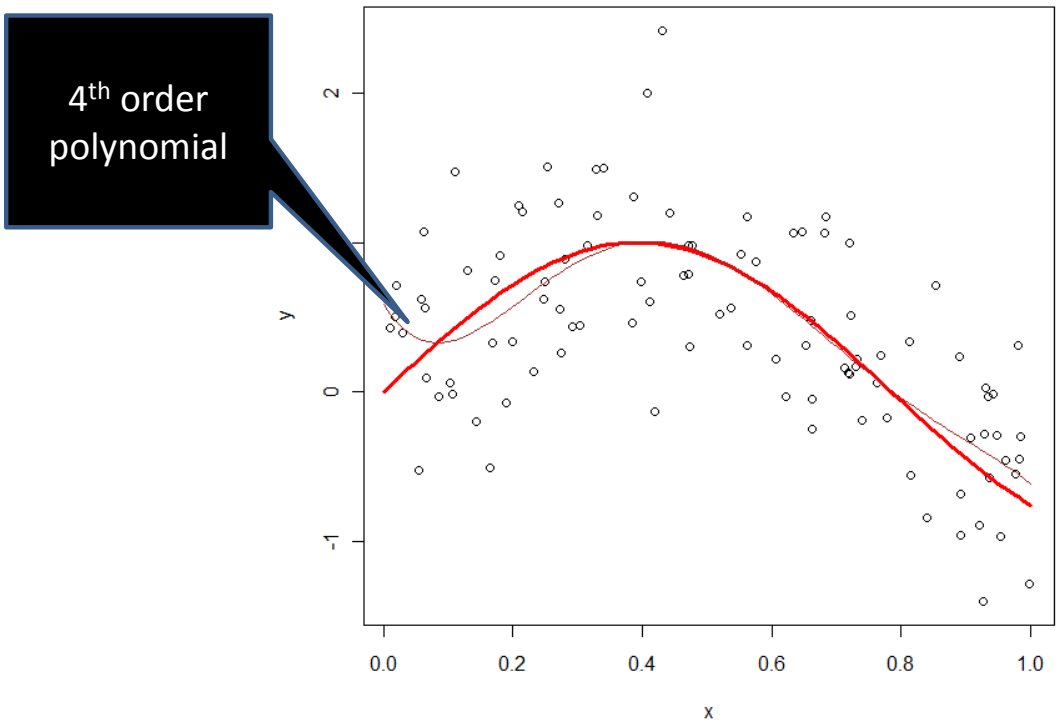
Bootstrapping and Crossvalidation

```
> fit1=lm(y~poly(x,degree=2),data=data)   # fit quadratic to data
> fit2=lm(y~poly(x,degree=3),data=data)   # fit 3rd order polynomial
> fit3=lm(y~poly(x,degree=4),data=data)
> fit4=lm(y~poly(x,degree=5),data=data)
```

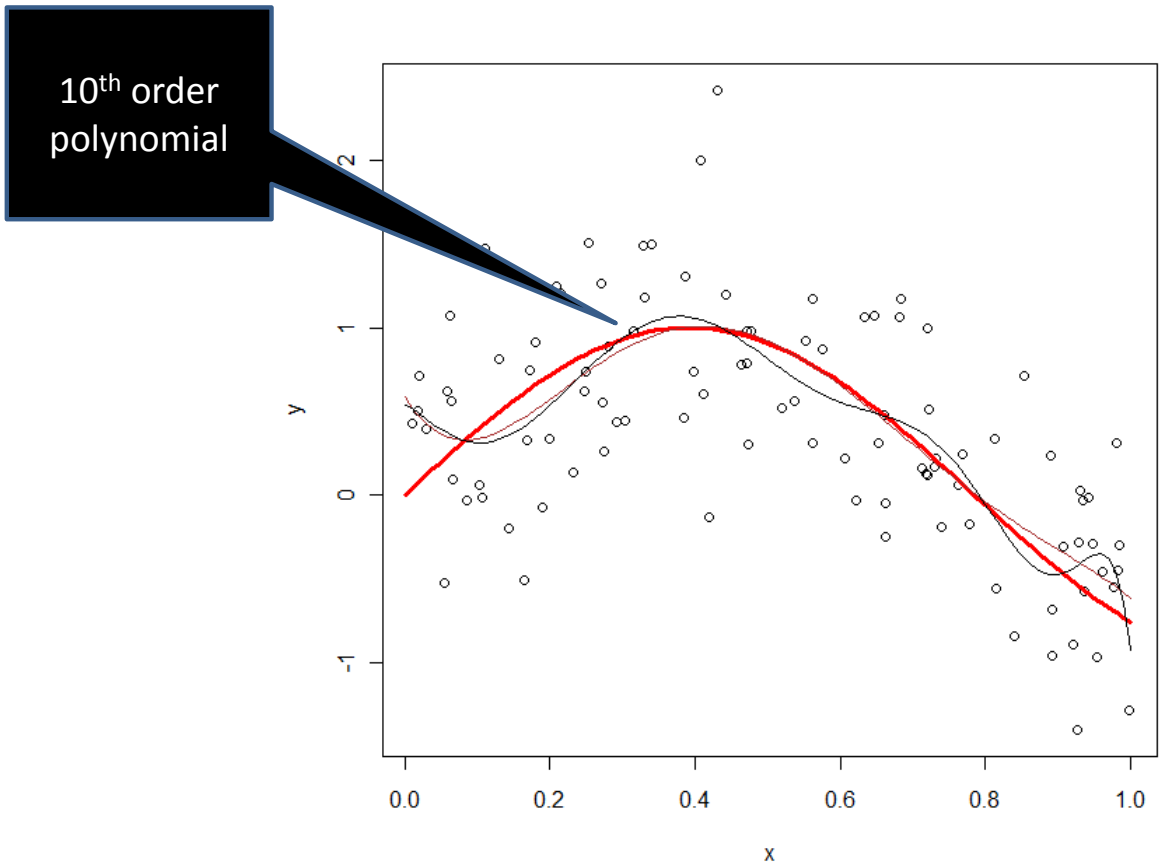


Bootstrapping and Crossvalidation

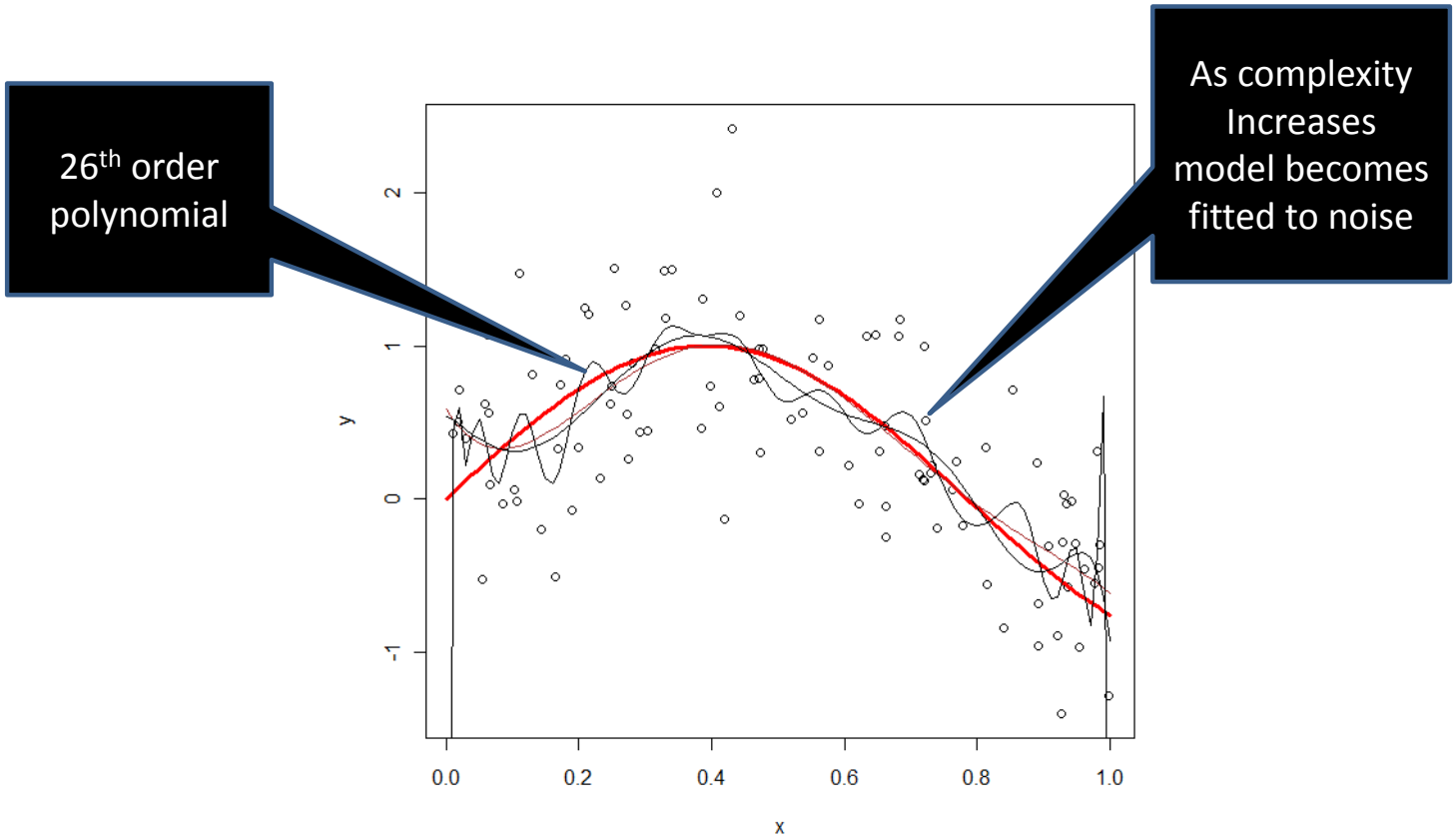
```
> fit1=lm(y~poly(x,degree=2),data=data)   # fit quadratic to data
> fit2=lm(y~poly(x,degree=3),data=data)   # fit 3rd order polynomial
> fit3=lm(y~poly(x,degree=4),data=data)
> fit4=lm(y~poly(x,degree=5),data=data)
```



Bootstrapping and Crossvalidation



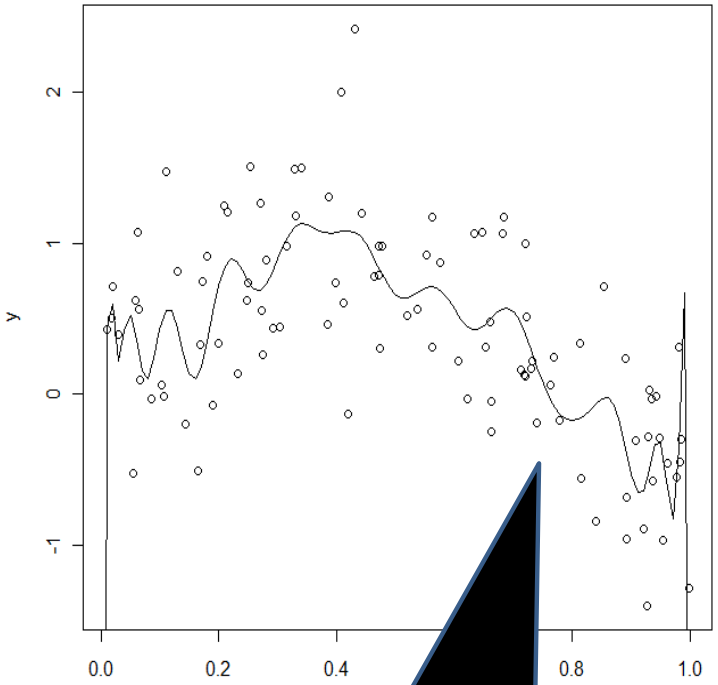
Bootstrapping and Crossvalidation



Bootstrapping and Crossvalidation

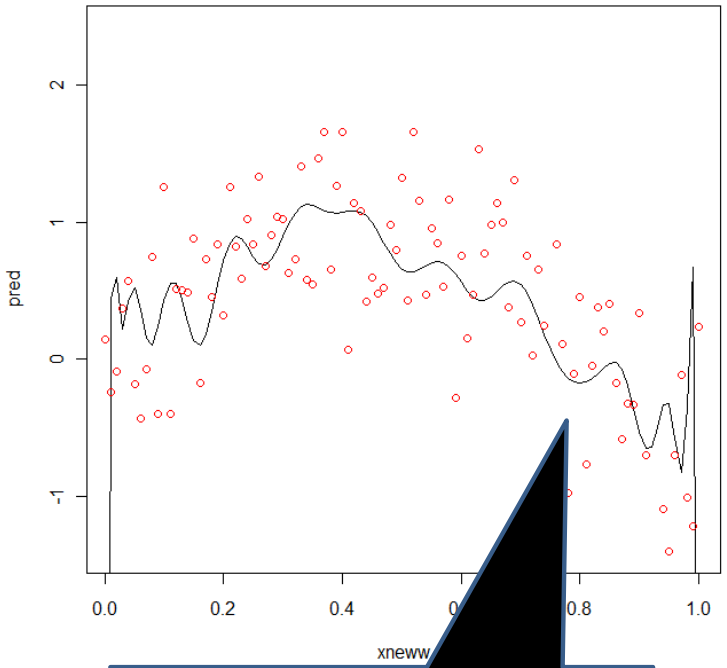
The Dangers of Over-fitting a Model

Training Data



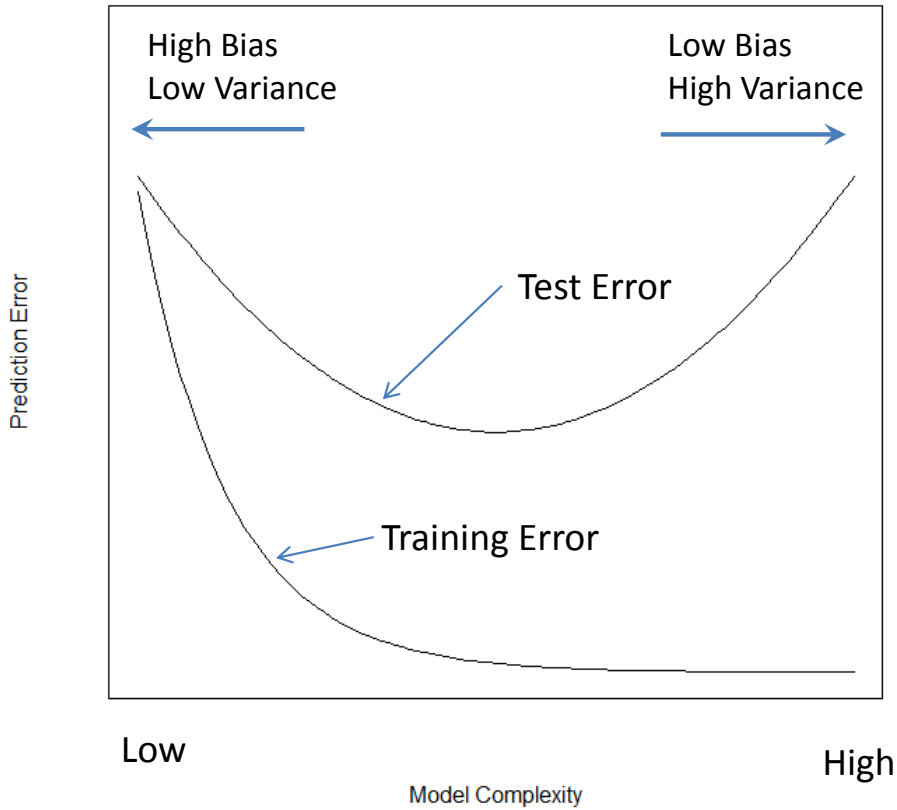
Because the complex Model was fit to every Every little quirky feature of the training data…

New Data



The complex model does a poor job of predicting new – as yet unseen – data coming from the process.

Bootstrapping and Crossvalidation



Training Err = MSE will always go down as model complexity increases

Bootstrapping and Crossvalidation

Ways of measuring error, using a "loss" function:

$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{Squared Error} \\ \left|Y - \hat{f}(X)\right| & \text{Absolute Error} \end{cases}$$

$$\text{Test Err} = Err = E\left[L(Y, \hat{f}(X)) \right]$$

Expectation taken over everything random, including X and Y as well as randomness In the training sample which produced $\hat{f}(x)$

$$\text{Training Err} = err = \frac{1}{N} \sum_{i=1}^{N} \left[L(y_i, \hat{f}(x_i)) \right]$$

Empirical average of loss observed in the training data

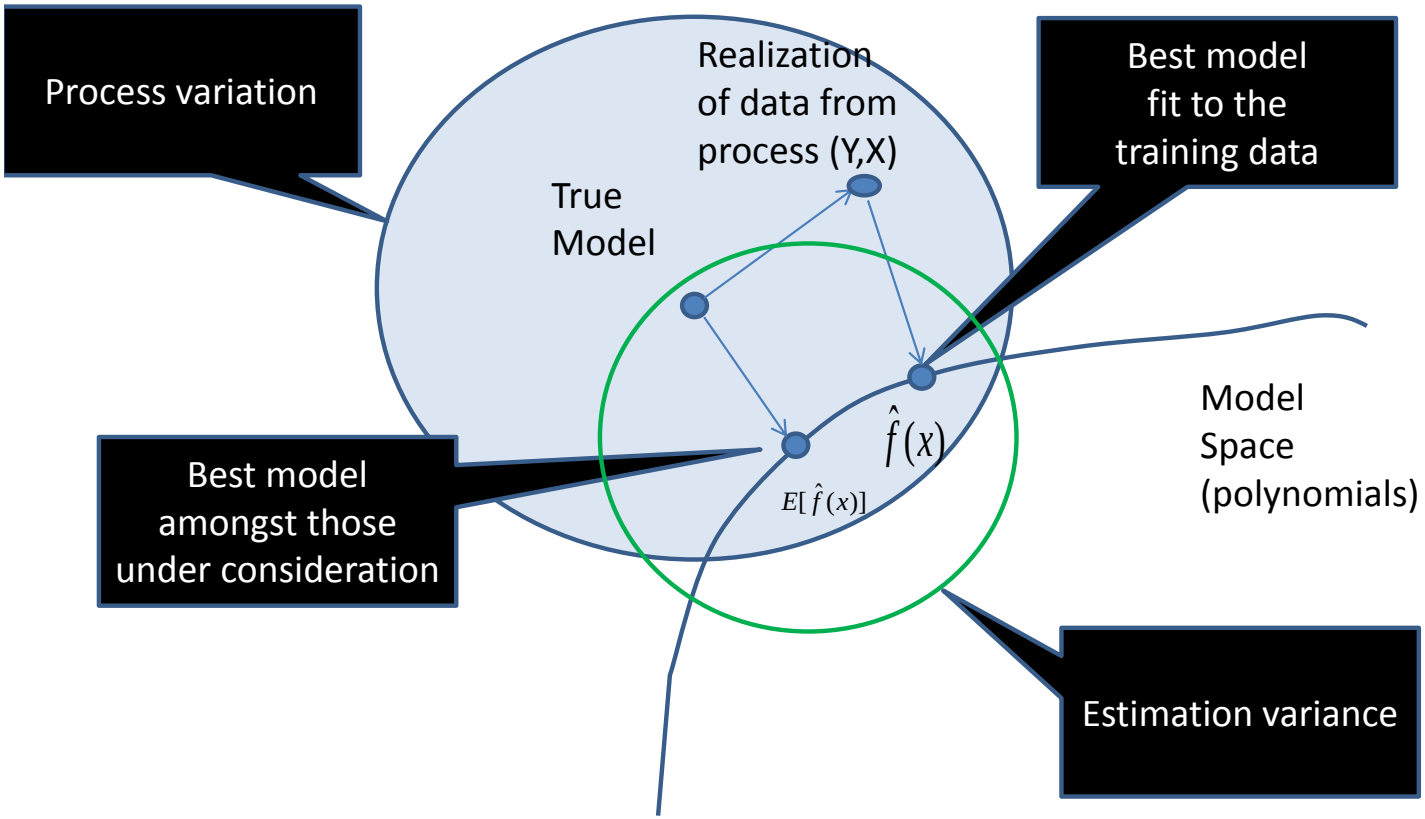
$$\text{Training Err} < \text{Test Err}$$

Training error underestimates test error and does worse as model complexity grows

Bootstrapping and Crossvalidation

Schematic of the Modeling Process

$$Y = f(X) + \varepsilon \qquad \mathrm{Var}(\varepsilon) = \sigma_{\varepsilon}^{2}$$



Bootstrapping and Crossvalidation

The Bias – Variance Decomposition

The Process: $Y = f(X) + \varepsilon$

Prediction error at X = x_0 =

$$Err(x_0) = E\left[(Y - \hat{f}(x_0))^2 \mid X = x_0\right]$$
$$= E\left[(Y - f(x_0))^2 \mid X = x_0\right] + \left[(f(x_0) - E(\hat{f}(x_0)))^2\right] + E\left[(\hat{f}(x_0) - E(\hat{f}(x_0)))^2 \mid X = x_0\right]$$
$$= \sigma_\varepsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))$$
$$= \text{Irreducible Error} + \text{Bias}^2 + Variance$$

- First term is the variation of the target $f(x_0)$ around its true mean

And never goes away no matter how well we estimate f.

- Second term measures deviation from the true function $f(x_0)$

to best estimable function $E[\hat{f}(x_0)]$

- Third term is the variation present in estimation

Model Complexity ↑ Bias ↓ Variance ↑

Bootstrapping and Crossvalidation

Extra-Sample Error, In-Sample Error, and Optimism

$$\text{Test Err} = Err = E\left[L(Y, \hat{f}(X)) \right]$$

Test Err (or generalization error) is a kind of extra-sample error since the Test features occur at different X values than the samples in the training data.

$$Err_{in} = \frac{1}{N} \sum_{i=1}^{N} E_y E_{Y^{New}} \left[L(Y_i^{New}, \hat{f}(X_i)) \right]$$

In sample error measures the expected difference between N new responses Y_i^{New} At each of the training points x_i, i = 1, …, N

$$Op = Err_{in} - E_y[err]$$

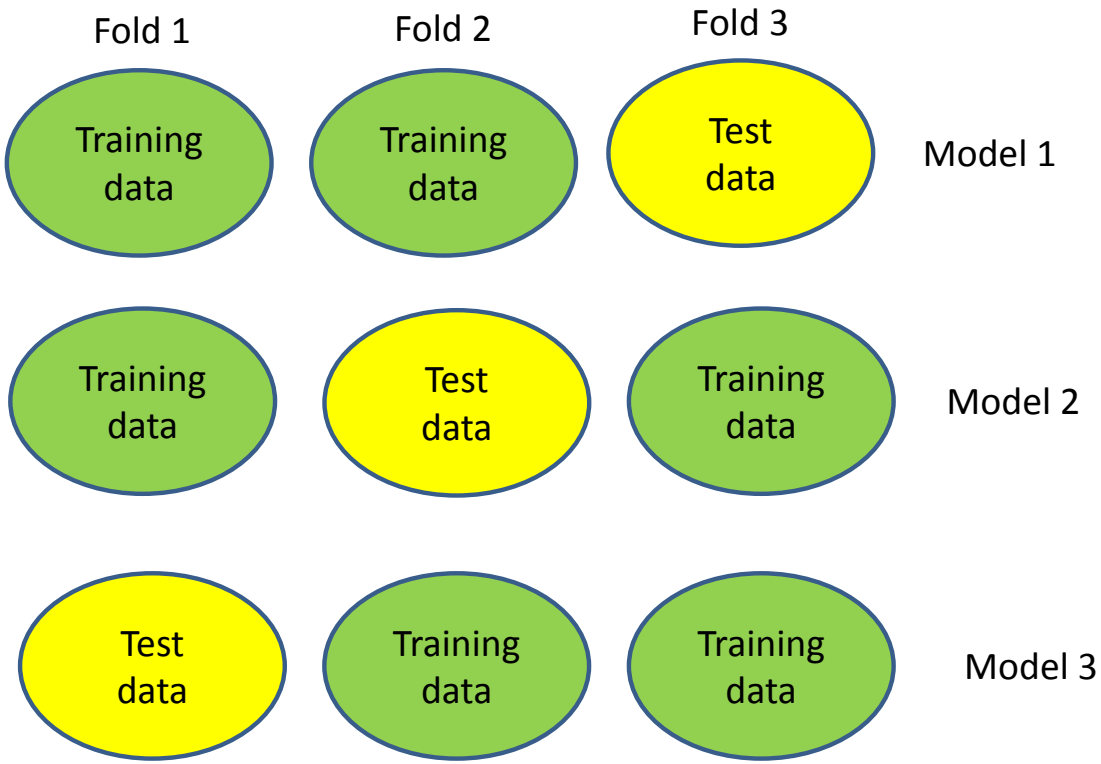
Optimism is the expected difference between In sample error and training error.

Cross-Validation and Bootstrapping measure Extra-Sample Error

Bootstrapping and Crossvalidation

K-Fold Cross Validation

Depiction of 3-fold cross validation:



Divide your data into K –folds, train a model on (K-1) folds and test model On the remaining 1 fold

Bootstrapping and Crossvalidation

Big Idea in Cross Validation

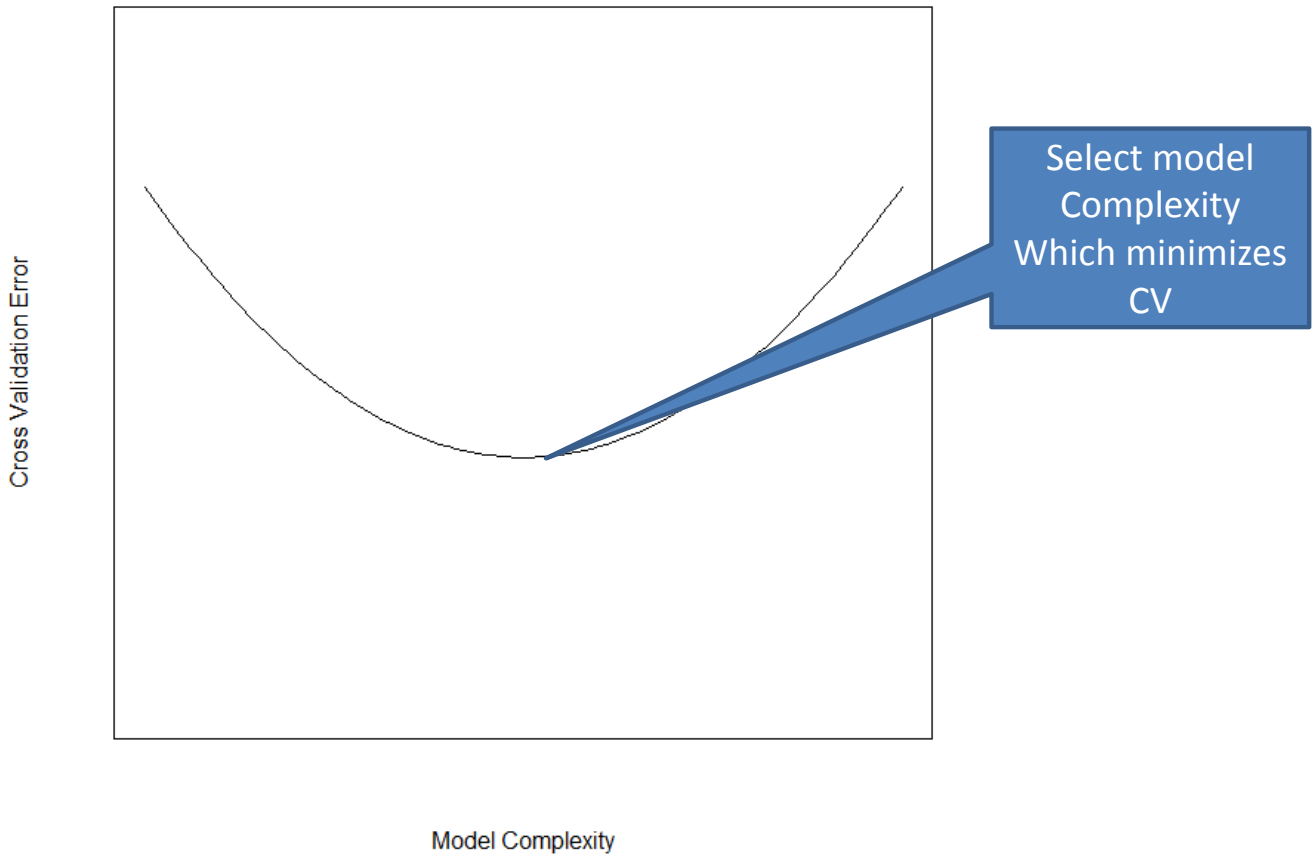
Because the Test data was not used to build each of the models,
The test data acts as NEW INDEPENDENT, YET TO BE OBSERVED DATA FROM THE PROCESS

Let $\hat{f}^{-K}(x)$ denote the fitted function with the Kth part of the data removed

$$CV = \frac{1}{KN} \sum_{i=1}^{K} \sum_{j=1}^{N} (y_i - \hat{f}^{-K}(x_i))^2$$

CV is an estimator of Extra-Sample Err

Bootstrapping and Crossvalidation



Bootstrapping and Crossvalidation

Write an R function which will compute CV

Bootstrapping and Crossvalidation

Bootstrapping

A Dataset:

Row	X	Y
1	A	F
2	B	G
3	C	H
4	D	I
5	E	J

Sample the rows 1:5 WITH REPLACEMENT randomly:

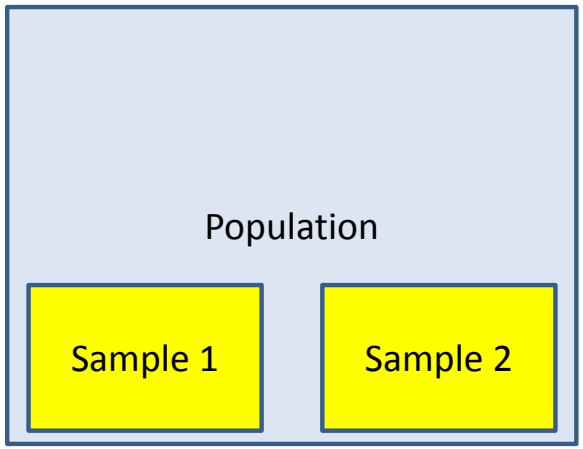
```
> sample(1:5,5,replace=TRUE)
[1] 5 1 2 2 1
```

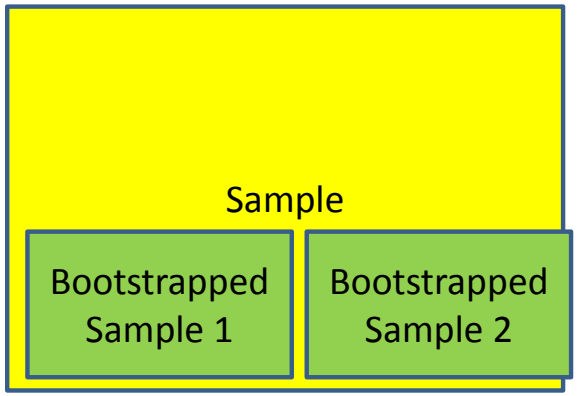
The bootstrapped data is the dataset with the rows in the sample

Rows	X	Y
5	E	J
1	A	F
2	B	G
2	B	G
1	A	F

Bootstrapping and Crossvalidation

Idea behind Bootstrapping:





The process of repeatedly subsampling a sample should mimic the process of repeatedly drawing samples from a population.

Bootstrapping and Crossvalidation

Bootstrapping

Sample the rows 1:5 WITH REPLACEMENT randomly:

```
> sample(1:5,5,replace=TRUE)
[1] 5 1 2 2 1
```

The bootstrapped data is the dataset with the rows in the sample

Rows	X	Y
5	E	J
1	A	F
2	B	G
2	B	G
1	A	F

Rows	X	Y
3	C	H
4	D	I

OOB = out of bag

The rows which were selected are “in the bag”,
the rows not selected are “out of the bag”.

Bootstrapping and Crossvalidation

Uses for Bootstrapping

Bootstrapping can be used to measure the variance of any statistic S(Data)

$$\hat{V}(S(Data)) = \frac{1}{B-1}\sum_{i=1}^{B}(S(Data^{*i}) - \bar{S}^{*})^2$$

Bootstrapping can also be used to measure the bias of any statistic.

OOB data can be used to estimate extra-sample error.

If C_{-i} denotes the set of bootstrap samples which do not contain observation i

$$\hat{Err} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{|C_{-i}|}\sum_{b\in C_{-i}} L(y_i, \hat{f}^{*b}(x_i))$$

Jackknifing

From an historical standpoint leave one out crossvalidation, or n-fold crossvalidation was the first measure of out of sample error which was measured. Consider the general linear model

$$\mathbf{y}_{(i)} = \mathbf{X}_{(i)}\beta_{(i)} + \epsilon_{(i)}$$

formed by **deleting row** i from the model. Here $\mathbf{y}_{(i)}$ and $\epsilon_{(i)}$ are $(n-1) \times 1$ vectors, $\mathbf{X}_{(i)}$ is a $(n-1) \times p$ matrix and $\beta_{(i)}$ is $p \times 1$. What makes leave one out crossvalidation special is we know explicit formulas for how the prediction will change. One formula comes as a consequence of the Sherman-Morrison-Woodbury theorem we have

$$\begin{aligned}(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} &= (\mathbf{X}^T\mathbf{X} - \mathbf{x}_i\mathbf{x}_i^T)^{-1} \\ &= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}\end{aligned}$$

Jackknifing

where $\mathbf{x}_i$ denotes the $p \times 1$ vector corresponding to the i^{th} row of $\mathbf{X}$. Now since $\mathbf{x}_i = \mathbf{X}^T\mathbf{e}_i$ with $\mathbf{e}_i = (0, \ldots, 1, \ldots, 0)^T$ denoting the i^{th} element of the standard basis in $\mathbb{R}^n$ it follows that

$$\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i = \mathbf{e}_i^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{e}_i = h_{ii}$$

hence

$$\begin{aligned}(\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1} &= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - \mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i} \\ &= (\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - h_{ii}}.\end{aligned}$$

It follows that

$$\begin{aligned}\hat{\beta}_{(i)} &= (\mathbf{X}_{(i)}^T\mathbf{X}_{(i)})^{-1}\mathbf{X}_{(i)}^T\mathbf{y}_{(i)} \\ &= \left[(\mathbf{X}^T\mathbf{X})^{-1} + \frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}}{1 - h_{ii}}\right]\left[\mathbf{X}^T\mathbf{y} - \mathbf{x}_i y_i\right]\end{aligned}$$

Jackknifing

Hence,

$$\begin{aligned}\hat{\beta}_{(i)} &= \hat{\beta} + \left[\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}\right]\left[\mathbf{x}_i^T\hat{\beta} - y_i h_{ii}\right] - y_i(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i \\ &= \hat{\beta} + \left[\frac{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i}{1-h_{ii}}\right][\hat{y}_i - y_i h_{ii} - y_i(1-h_{ii})] \\ &= \hat{\beta} + \left[\frac{\hat{\epsilon}_i}{1-h_{ii}}\right]\left[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\right].\end{aligned}$$

Thus,

$$\begin{aligned}\hat{\epsilon}_{i(i)} \equiv y_i - \hat{y}_{i(i)} &= y_i - \mathbf{x}_i^T\hat{\beta}_{(i)} \\ &= y_i - \mathbf{x}_i^T\hat{\beta} + \left[\frac{\hat{\epsilon}_i}{1-h_{ii}}\right]\left[\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i\right] \\ &= \hat{\epsilon}_i + \left[\frac{h_{ii}\hat{\epsilon}_i}{1-h_{ii}}\right] = \frac{\hat{\epsilon}_i}{1-h_{ii}}.\end{aligned}$$

PRESS

The PRESS residuals are defined by $\hat{\epsilon}_{i(i)} = \frac{\hat{\epsilon}_i}{1-h_{ii}}$ and the PRESS statistic is given by

$$PRESS = \sum_{i=1}^{n}(y_i - y_{i(i)})^2 = \sum_{i=1}^{n}(\hat{\epsilon}_{i(i)})^2 = \sum_{i=1}^{n}\left(\frac{\hat{\epsilon}_i}{1-h_{ii}}\right)^2.$$

A related statistic is given by

$$R^2_{PRESS} = 1 - \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Notice that since formulas are given for CV it saves us a lot of computation.

Scenario: We have:

- a sample of observations: $x_1, \ldots, x_n$
- a target parameter in mind: θ
  (e.g., σ, or $\log \sigma^2$, or e^{μ}, or σ/μ, ...)
- an estimate of θ, $\hat{\theta} \equiv y_{all}$
  e.g., s or $\log s^2$, or $e^{\bar{x}}$, or 0.7413· F-spread/median ...

Questions:

1. What is a confidence interval for θ based on $\hat{\theta}$?
2. If $\hat{\theta}$ is biased for θ (i.e., if you repeatedly estimate θ using $\hat{\theta}$ and found that mean of $\hat{\theta}$'s is **not** θ), can we find an estimate of θ, say $\hat{\theta}^*$, which is less biased?

Answers:

1. Yes. Construct CIs using pseudo-values, or bootstrap estimates.
2. Yes. $\hat{\theta}_{JK}$ or $\hat{\theta}_{B}$.

We illustrate Jackknife approach first, then bootstrap.

Question 1: Confidence Interval

Q1: What is a confidnence interval for θ based on $\hat{\theta}$?

Ans: For "95% CI", expect answer is of the form

$$\hat{\theta} \pm t_{n-1}(0.975) \cdot [??]$$

where [??] is probably something like $\sqrt{var(\hat{\theta})}$ and $t_{n-1}(0.975)$ is 97.5%-point of Student's t, $n-1$ d.f.

- If $\hat{\theta} = \bar{x}$, then we use $\widehat{Var}(\hat{\theta}) = s^2/n$, because theory says $Var(\bar{x}) = \sigma^2/n$ and s^2 estimates σ^2:

- But if $\hat{\theta} = \log s^2$ or $e^{\bar{x}}$ or $0.7413 \cdot F - spread/median$?

Question 1: Confidence Interval

Situations where we *know* the answer:

1. Sample mean
2. LS coefficients (Gaussian error)

Cases where we *don't know* the answer: **Everything else**

1. Most nonlinear statistics (e.g., RRline)
2. Non-Gaussian error distributions
3. Long-tails

What can we do?

- If we had many samples, then we would have many $\hat{\theta}$'s, and then we could calculate a standard deviation of them, as a measure of the variability of the $\hat{\theta}$'s.
- Generally we do not have a lot of samples– we have only one, and only one $\hat{\theta}$ (y_{all}).
- We need to generate more $\hat{\theta}$'s.
- We do so by estimating θ on subsamples of original sample.
- Jackknife: subsample $=$ {all x's except one} (size $n-1$)
- Bootstrap: sample (with replacement) from the x's (size n)

Jackknife – notation

Jackknife notation:

- **Pseudo-values** $= ny_{all} - (n-1)y_{(-i)}$
- **Target parameter** $= \theta$
- **Estimate of target parameter** $=$ statistic $S \equiv \hat{\theta}$
- $y_{all} \equiv \hat{\theta}_{all}$
- Mean of pseudo-values $\equiv \hat{\theta}_{JK} =$ jackknife estimate of θ

Jackknife – Example 1

Example: **Sample mean, single batch** $x_1, \ldots, x_{20}$:
Confidence limits on θ using $S = y_{all} = \bar{x}$
($\bar{x} = 31.11826$, SE=SD$/\sqrt{20} = 4.09458$)

```
 5.913934 17.225504 -0.525662  5.838253 16.726366
18.323374 20.972659 31.927836 20.437125 25.978228
26.039943 29.045681 42.689422 45.187752 47.038066
48.324157 53.771513 59.261140 51.370595 56.819336
```

Jackknife – Example 1

- $y_{all} = \text{mean}(y) = \sum_{i=1}^{20} x_i/20$
- $y_{(-i)} = \text{mean}(x \text{ without } x_i) = \sum_{k \neq i} x_k/19$
- Pseudo-values $= 20 \cdot y_{all} - 19 \cdot y_{(-i)}$:

5.913934	17.225504	-0.525662	5.838253	16.726366
18.323374	20.972659	31.927836	20.437125	25.978228
26.039943	29.045681	42.689422	45.187752	47.038066
48.324157	53.771513	59.261140	51.370595	56.819336

Jackknife – Example 1

- average of pseudo-values $= 31.11826 = \bar{x}$
- Standard error of pseudo-values is $s/\sqrt{n} = 4.09458$

So, when statistic S is the sample mean,

- jackknife mean = usual sample mean
- jackknife standard error = usual standard error

i.e., results are expected.

Jackknife

Notes:

1. Mean of pseudo-values = "jackknife estimate of θ" ("*an estimate of much reduced bias*")
2. SE(pseudo-values) = "jackknife SE of $\hat{\theta}$"
3. **HW**: Show: SE(PVs) $= [(n-1)/\sqrt{n}] \cdot SD(y_{(-i)})$

Jackknife – procedure

Procedure: Statistic $S = \hat{\theta}$, $\hat{\theta}_{all} = y_{all}$

1. Calculate $\hat{\theta}_{all} = y_{all}$ using all data
2. Calculate same statistic without x_i : $y_{(-i)}$
3. Calculate pseudo-values $\hat{\theta}_i^* \equiv PV_i \equiv n y_{all} - (n-1) y_{(-i)}$
4. Calculate: $\hat{\theta}_{JK} = \sum \hat{\theta}_i^* / n$

$$\widehat{Var}(\hat{\theta}_{JK}) = (\text{SE(mean PVs)})^2 = \sum_{i=1}^{n} (\hat{\theta}_i^* - \hat{\theta}_{JK})^2 / [n(n-1)]$$

5. Calculate approximate 95% CI for θ using

$$\text{mean (PVs)} \pm t_{n-1}(0.975) \cdot \text{SE(mean PVs)}$$

$$= \hat{\theta}_{JK} \pm t_{n-1}(0.975) \cdot \sqrt{\widehat{Var}(\hat{\theta}_{JK})}$$

Jackknife – Example 2

Example: (Confidence interval on a standard deviation) A sample from a distribution produced the 11 values

$$0.1, 0.1, 0.1, 0.4, 0.5, 1.0, 1.1, 1.3, 1.9, 1.9, 4.7$$

There is no reason to suppose that the distribution is normal and some reason to suppose it is not.

- Sample standard deviation: y_{all}
- Leave-one-out: $y_{(-i)}$
- Pseudo-values: $\hat{\theta}_i^*$
- Average of pseudo-values: $\hat{\theta}_{JK}$
- SE($\hat{\theta}_{JK}$):
- 95% CI for σ:

Question 2: Reduce Bias

Q2: Is $\hat{\theta}_{JK}$ *"an estimate of much reduced bias"?*

Example: Use sample statistic s to estimate σ

- $E(s) \neq \sigma$; $E(s) = a(n) \cdot \sigma$ where $a(n) < 1$ (i.e., s slightly underestimates σ).
- $\lim_{n \to \infty} a(n) = 1$. So, as n gets large, the bias is negligible.
- But when $n = 5$, $E(s) = 0.8812\sigma$; i.e., s is about 12% too small.
- In general, $E(s) - \sigma \neq 0$, i.e., $E(s)$ is biased for σ.

We show that, in general, $\hat{\theta}_{JK}$ is less biased (or has no greater bias) for θ than $\hat{\theta} = y_{all}$ is.

Jackknife – Reduce Bias

- Suppose the bias has this form:

$$bias_n = E(\hat{\theta}) - \theta = a_1/n + a_2/n^2 + a_3/n^3 + \cdots$$

  i.e., bias is of order $1/n$.
- Then bias in $\hat{\theta}_{(-i)}$ ($n-1$ observations) is:

$$bias_n = E(\hat{\theta}_{(-i)}) - \theta = a_1/(n-1) + a_2/(n-1)^2 + a_3/(n-1)^3 + \cdots$$

- So bias in (average of leave-out-ones) $\equiv \hat{\theta}_L$ is

$$bias_n = E(\hat{\theta}_L) - \theta = (1/n) \cdot \sum_{i=1}^{n} E(\hat{\theta}_{(-i)} - \theta) = \sum_{k=1}^{\infty} a_k/(n-1)^k$$

Jackknife – Reduce Bias

- So bias in $\hat{\theta}_{JK}$ is

$$bias(\hat{\theta}_{JK}) = E(\hat{\theta}_{JK}) - \theta$$

$$= E[n(\hat{\theta}_{all} - \theta) - (n-1)(\hat{\theta}_L - \theta)]$$

$$= n(a_1/n + a_2/n^2 + \cdots) - (n-1)(a_1/(n-1) + a_2/(n-1)^2 + \cdots)$$

$$= (a_2/n - a_2/(n-1)) + a_3(1/n^2 - 1/(n-1)^2) + \cdots$$

$$= -a_2/(n(n-1)) - a_3(2n-1)/(n^2(n-1)^2) - \cdots$$

- Leading term is $a_2/(n(n-1)) \approx a_2/n^2$, of order $1/n^2$, less than order $1/n$: "*much reduced bias*".
- If $a_2 = a_3 = \cdots = 0$, then $\hat{\theta}_{JK}$ is **unbiased** for θ.

Jackknife – problem

- We can use the jackknife for any procedure — e.g., intercepts and slopes of RRline, effects from median polish, etc.
- Usually the jackknife over-estimates the standard error.
- But, more worrisome, sometimes it can under-estimate it.
- Efron (1979) asked himeself, "Why does the jackknife work?"
- In developing theory for it, he developed an alternative (And, in many ways, more intuitive) way of generating more $\hat{\theta}$'s
- So we next discuss Efron's bootstrap, then illustrate on
  — LS line and RR line
  — Median polish

About standard errors

Standard errors:

- SE=SD(statistic); e.g., SD(estimate)
- n observations $\rightarrow SD^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$
- For Gaussian$(0, \sigma^2)$: SD $\approx$ F-spread$/1.349$
- What happens when $n = 1$?

Jackknife and Bootstrap: Applications

RRline: We have only

- **one** intercept
- **one** slope

Median polish: We have only

- **one** M,
- **one** a_1, **one** a_2, ...
- **one** b_1, **one** b_2, ...

How to compute a SE when we have only one of each ??

Efron's bootstrap

- Ideally, we'd like to have another sample, so we can calculate another θ
- All we have is the sample at hand, which we hope is representative of the entire population. If the distribution of the quantity in the entire population is F, then we have only an $\hat{F}$, a sample from F
- We can use $\hat{F}$ to generate another sample, say $\hat{\hat{F}}$
- $\hat{\hat{F}}$ is obtained by take a random sample, with replacement, of the original x's.
- Note that some x_i will be duplicated, or triplicated, or ..., while other x_i's will not be represented at all.

Bootstrap – procedure

- Calculate $\hat{\theta}$ on this "bootstrap sample" from $\hat{F}$; denote it by $\hat{\theta}_b$, $b = 1, 2, \ldots, B$ (B can be very large).
- $\hat{\theta}_B = \sum_{b=1}^{B} \hat{\theta}_b / B \equiv$ "bootstrap estimate of θ"
- $\widehat{Var}(\hat{\theta}_B) \equiv \sum_{b=1}^{B} (\hat{\theta}_b - \hat{\theta}_B)^2 / (B-1)$
- Approximate 95% CI: $\hat{\theta}_B \pm t_{n-1}(0.975) \cdot \sqrt{\widehat{Var}(\hat{\theta}_B)}$
- If $\hat{F}$ was not a good representation of true F, we are sorely out of luck

Bootstrap: How to obtain multiple samples/estimates?

We need to get more intercepts/slopes (RRline) or more Ms, a_is, b_js
Jackknife: Single sample
Bootstrap: More general

- Collect residuals in a pot
- Shuffle them around
- Put them back in "residual"
- Add back in the effects
- Now you have new set of data $\rightarrow$ recompute
- Get another set of estimates
- Repeat! Calculate standard errors

Jackknife and Bootstrap on LS and RRline

- $x = (1:20)$, $y =$ as before
- "True" line: $y = 1 + 3x + error$ (Gauusian, mean 0, SD=5)
- LS:$(\hat{a}_{LS}, \hat{b}_{LS}, RMS) = (0.1487, 2.9495, 5.412)$
- RR:$(\hat{a}_{RR}, \hat{b}_{RR}, |res|)$=(-1.06, 3.186, 91.9942)
- Jackknife estimate and jackknife SE:

	apv	bpv	rpv	Apv	Bpv	Rpv
mean	0.3389	2.9354	5.9777	-18.9427	5.7761	215.9494
SE	3.4719	0.2558	0.8624	1.6759	0.2102	13.7573

- LS theory: $\hat{SE} = RMS \cdot \sqrt{diag\{(X'X)^{-1}\}}$
  $SE(\hat{a}_{LS}) = 2.514$, $SE(\hat{b}_{LS}) = 0.210$
- Jackknife SEs are generous when $\hat{\theta} = \hat{\theta}_{LS}$

Bootstrap: 3 approaches:

1. Sample indicies: repeat $B(200?)$ times:

```
ii <- sample(1:20, 20, replace=TRUE); xb <- x[ii];
yb <- y[ii]; lm(yb ~ xb); run.rrline(xb,yb)
```

2. Sample residuals, add back to original line:

```
res <- lm(y $\sim$ x)$res
for (j in 1:200) {
   b.res <- sample(res,20,replace=TRUE)
   yb <- 0.1487 + 2.9495*x + b.res
   b.coef <- lm(yb $\sim$ x)$coef
   [ save b.coef in file ]
   }
```

Depends critically on y being linear in x

Theory of Bootstrapping

- Statistics: Construct a statistic $T = f(\text{data})$ whose target of inference is an unknown parameter θ for a distribution function.
- Often, T is known to follow some distribution F.
- We can estimate F by use of the empirical CDF $\hat{F}$ via a function $t(\hat{F})$.
- Suppose, we want to calculate a $(1 - 2\alpha)$ confidence interval for θ.
- Often it is possible to show that $T \sim N(\theta + \beta, v)$ where v is the variance and β is the bias of T. I
- If both β and v are known then

$$P(T \leq t|F) \cong \Phi\left(\frac{t - (\theta + \beta)}{v^{1/2}}\right),$$

where $\Phi(\cdot)$ is the CDF of a standard normal.

Theory of Bootstrapping

If the α quantile of the standard normal distribution is $z_\alpha = \Phi^{-1}(\alpha)$, then an approximate $(1-2\alpha)$ confidence interval for θ has limits

$$(t - \beta - v^{1/2} z_{1-\alpha}, t - \beta - v^{1/2} z_\alpha),$$

where t is the assumed value of T as the above CI follows from

$$P(\beta + v^{1/2} z_\alpha \leq T - \theta \leq \beta + v^{1/2} z_{1-\alpha}) \cong 1 - 2\alpha.$$

The Two Flavors of Bootstrapping

- There are two different flavors of bootstrapping: **Parametric and Non-Parametric Bootstrapping**
- Parametric Bootstrapping: Suppose the CDF $F(\cdot|\theta)$ is known, then estimating the bias and variance of T can be accomplished quite easily.
- In this regard, suppose that $y_1, \ldots y_n$ are i.i.d. and that the CDF and PDF are $F(y|\theta)$ and $f(y|\theta)$, respectively.
- When we estimate θ with $\hat{\theta}$, it's substitution into the model gives the fitted model, with CDF $\hat{F}(y) := F(y|\hat{\theta})$ can be used to estimate the properties of T.
- We shall use Y^* to denote the random variable distributed according to the fitted model $\hat{F}$, and the notation E^* and Var^* will be used when these moments are calculated according to the fitted distribution.
- We may then generate many simulated data sets to estimate the bias and variance of T.

Parametric Bootstrapping

To illustrate the computation of the bias and variance of T from a single data set, we let $Y_1^*, \ldots Y_n^*$ be a independently drawn sample from the fitted distribution $\hat{F}$. When the statistic of interest is calculated from such a simulated data set, we denote it by T^*. From R repetitions of the data simulation we obtain $T_1^*, \ldots T_R^*$. The estimator of the bias $b(F) = \mathrm{E}[T|F] - \theta$ of T is then

$$\hat{B} = b(\hat{F}) = \mathrm{E}[T|\hat{F}] - t = \mathrm{E}^*[T^*] - t,$$

and this in turn is estimated by

$$\hat{B}_R = \frac{1}{R}\sum_{r=1}^{R} T_r^* - t = \bar{T}^* - t.$$

Note that in the simulation, t is the parameter value of the model, so that $T^* - t$ is the simulation analogue of $T - \theta$. The corresponding estimator of the variance of T is then

$$\hat{V}_R = \frac{1}{R-1}\sum_{r=1}^{R}(T_r^* - \bar{T}^*)^2,$$

and estimators for other moments can be made in similar fashion.

Non-Parametric Bootstrapping

- In many cases we have no parametric model, but we can assume $Y_1, \ldots Y_n$ are independently and identically distributed according to an unknown distribution function F. We use the **empirical CDF** $\hat{F}$ to estimate the unknown CDF F. To estimate the properties of T then, we utilize $\hat{F}$ just as we would in the parametric model when drawing simulated samples.
- Because the empirical CDF $\hat{F}$ puts equal probabilities on the original data values $y_1 \ldots, y_n$, each Y^* is independently sampled uniformly from these values.

Non-Parametric Bootstrapping

- Therefore each sample $Y_1^*, \ldots, Y_n^*$ is a random sample taken with replacement from the data.
- We may then repeat such sampling R times and estimate the bias of T by

$$\hat{B}_R = \frac{1}{R}\sum_{r=1}^{R} T_r^* - t = \bar{T}^* - t,$$

and the variance with

$$\hat{V}_R = \frac{1}{R-1}\sum_{r=1}^{R}(T_r^* - \bar{T}^*)^2,$$

which are the same formulae as in the parametric case, with the only exception being that in the non-parametric case the samples are drawn in different fashion.

Non-Parametric Bootstrapping

- Suppose that T estimates θ and we seek a confidence interval on θ with both left- and right-tail errors both equal to α. If the quantiles of $T - \theta$ are denoted a_p, we have
$$P(T - \theta \leq a_\alpha) = \alpha = P(T - \theta \geq a_{1-\alpha}).$$
Rewriting the events $T - \theta \leq a_\alpha$ and $T - \theta \geq a_{1-\alpha}$ as $\theta \geq T - a_\alpha$ and $\theta \leq T - a_{1-\alpha}$, respectively, we see that the $(1 - 2\alpha)$ equi-tailed confidence interval has limits
$$(\hat{\theta}_\alpha := t - a_\alpha, \hat{\theta}_{1-\alpha} := t - a_{1-\alpha}).$$
- This ideal solution to the confidence interval rarely applies because the distribution of $T - \theta$ is usually unknown. This leads us to several approximate methods, most of which are based off approximating the quantiles of $T - \theta$.

Normal Approximation Method

- The simplest approach is to apply a $N(\beta, v)$ approximation for $T - \theta$. This leads to approximate confidence limits given by

$$\hat{\theta}_\alpha, \hat{\theta}_{1-\alpha} = t - \hat{B}_R \mp \hat{V}_R^{1/2} z_{1-\alpha},$$

  where $\hat{B}_R$ and $\hat{V}_R$ are calculated with the bias and variance formulas in the previous slide.
- Whether or not a normal approximation method is appropriate can be assessed through making a Q-Q plot of the simulated estimates $t_1^*, \ldots, t_R^*$. If such a plot suggests that the normal approximation is poor, then we can either try to improve the approximation in some way or replace it completely.

Normal Approximation Method

- If we start again at the general confidence interval formula, we can estimate the quantiles a_α and $a_{1-\alpha}$ by the corresponding quantiles of $T^* - t$. Assuming that the R simulations result in $t^*_{(1)} - t, \ldots, t^*_{(R)} - t$, ordered realizations of $T^* - t$, then the respective quantiles can be approximated by $a_\alpha = t^*_{((R+1)\alpha)} - t$ and $a_\alpha = t^*_{((R+1)(1-\alpha))} - t$, respectively.
- By substituting these values into the confidence interval for θ this results in the lower and upper confidence limits on θ given by

$$\hat{\theta}_\alpha = 2t - t^*_{((R+1)(1-\alpha))}, \quad \text{and} \quad \hat{\theta}_{1-\alpha} = 2t - t^*_{((R+1)\alpha)}.$$

- These are referred to as the basic bootstrap confidence limits for θ.

Studentized Bootstrap Method

- A modification of this is to use the form of the normal approximation confidence limit in (133), by replacing the $N(0,1)$ approximation for $Z = (T-\theta)/V^{1/2}$ by a bootstrap approximation.
- In this method, each simulated sample is used to calculate t^*, the variance estimate $\hat{V}^*$, and hence the bootstrap version $z^* = (t^* - t)/(\hat{V}^*)^{1/2}$ of Z.
- We note that in order to calculate $\hat{V}^*$ for each simulated sample for example, this method requires that a bootstrap be done for each simulated sample, or a bootstrap performed for each bootstrap if you will. Once the R simulated values of z^* are computed, they are ordered, and the p^{th} quantile of Z is estimated by the $(R+1)p^{th}$ ordered value of these.
- Then the confidence limits are replaced by

$$\hat{\theta}_\alpha = t - (\hat{V})^{1/2} z^*_{((R+1)(1-\alpha))}, \quad \text{and} \quad \hat{\theta}_{1-\alpha} = t - (\hat{V})^{1/2} z^*_{((R+1)\alpha)}.$$

- These are referred to as the studentized bootstrap confidence limits for θ.

Example

3. Fit distribution to residuals and sample from it:
Ex: mean(residuals) = 0, SD = RMS = 5.412, N(0, sd=5.412):

```
for ( j in 1:200)  {
     b.res <- 5.412*rnorm(20)
     yb <- 0.1487 + 2.9495*x + b.res
     b.coef <- lm(yb $\sim$ x)$coef
     [ save b.coef in file ]
     }
```

Depends critically on y being linear in x **and** distribution of residuals in Gaussian

Example

U.S. Infant mortality rates, 1964-1966
(# deaths per 1000 live births)
EDTTS p.41 Tbl 2-2; also 2-16 p.55

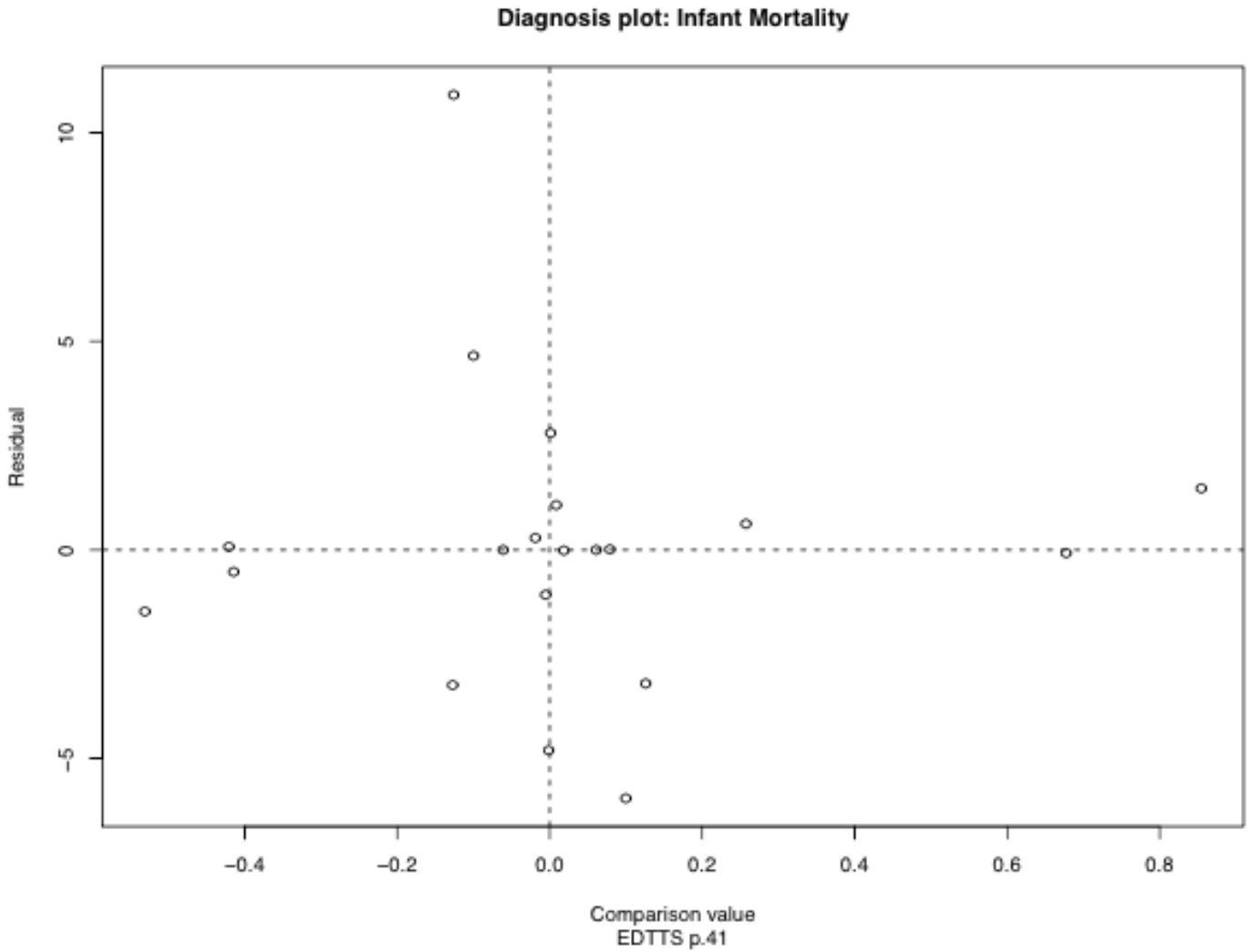
	Father's education				
Region	<=8	9-11	12	13-15	>=16
NE	25.3	25.3	18.2	18.3	16.3
NC	32.1	29.0	18.8	24.3	19.0
South	38.8	31.0	19.3	15.7	16.8
West	25.4	21.1	20.3	24.0	17.5

Example

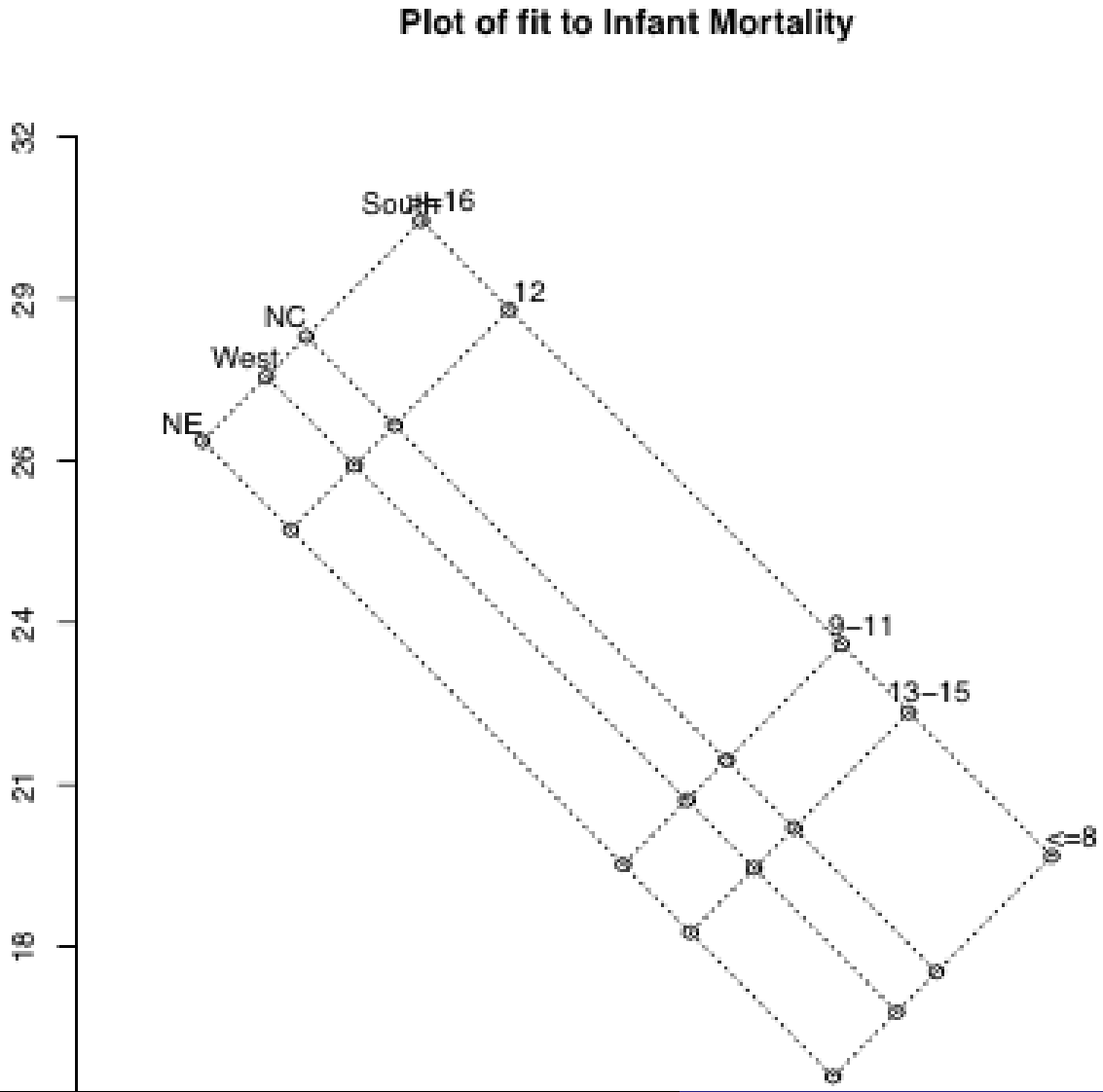
Median polish of infant mortality rates:

	Father's education					
Region	<=8	9-11	12	13-15	>=16	Row
NE	-1.475	0.075	0.012	-1.075	0.625	-1.475
NC	1.475	-0.075	-3.237	1.075	-0.525	2.375
South	10.900	4.650	-0.012	-4.800	0.000	-0.350
West	-3.200	-5.950	0.288	2.800	0.000	0.350
Col	7.475	5.925	-1.113	0.075	-3.625	20.775

Example


Diagnosis plot: Infant Mortality
Residual
10
5
0
−5
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Comparison value
EDTTS p.41

Example


Plot of fit to Infant Mortality
32
29
26
24
21
18
South
NC
West
NE
>=16
12
9–11
13–15
<=8

Example

	M	a1	a2	a3	a4
mean	21.211	-1.629	2.304	-0.515	0.237
SD	1.064	1.328	1.331	1.148	1.158

	b1	b2	b3	b4	b5
mean	6.979	5.442	-1.254	-0.508	-3.891
SD	1.941	1.917	1.403	0.943	1.840