Stat 407 Lab 4 Missing Values
Due: Wednesday, September 12, 2012, in class. Hand in one solution per group.

Purpose: The gain some experience in handling missing values. We will use both R and ggobi to explore missing value distributions, and impute missing values.
Instructions:

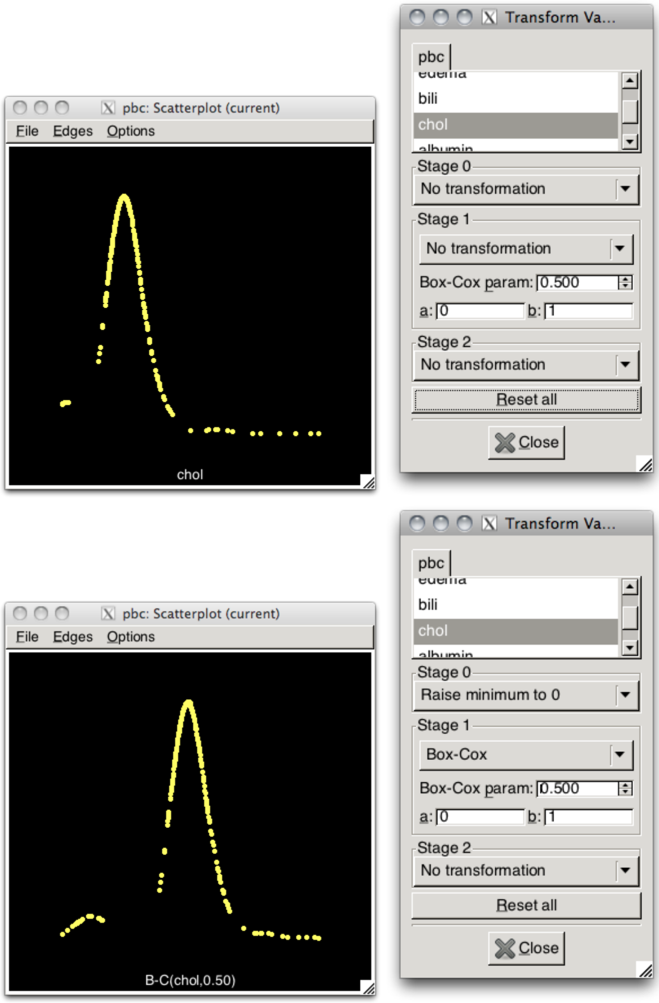
- The `pbc` data can be found on the Data link for the missing values chapter of the `http://www.ggobi.org/book` web page. It is a csv file.
- You can use the ggplot2 package to explore univariate distributions:

```
> library(ggplot2)
> qplot(chol, data=pbc, geom="histogram")
> qplot(log10(chol), data=pbc, geom="histogram")
> pbc.tf <- pbc # Create a new data frame to hold the transformed data
> pbc.tf$chol <- log10(pbc$chol) # Change the values to log values
> qplot(copper, data=pbc, geom="histogram")
> qplot(log10(copper), data=pbc, geom="histogram")
> pbc.tf$copper <- log10(pbc$copper) # Change the values to log values
> qplot(trig, data=pbc, geom="histogram")
> qplot(log10(trig), data=pbc, geom="histogram")
> pbc.tf$trig <- log10(pbc$trig) # Change the values to log values
> qplot(platelet, data=pbc, geom="histogram")
```

- To ese ggobi to impute means and random values, to examine how these are. Once you have the data loaded into ggobi, go to the `Tools` menu to find the Missing values handling methods. Check the ggobi manual for more information on working with missing values. Load the data into ggobi using code similar to:

```
> library(rggobi)
> g <- ggobi(pbc.tf)[1]
```

- You can also use ggobi to explore the trsansformations to normality. Open the `Tools` menu, select `Transform Variables`. Select a variable that you want to make a transformation of, and also plot this variable as a density plot.



Stage 0
No transformation

Stage 1
No transformation

Box-Cox param: 0.500

a: 0 b: 1

Stage 2
No transformation

Stage 0
Raise minimum to 0

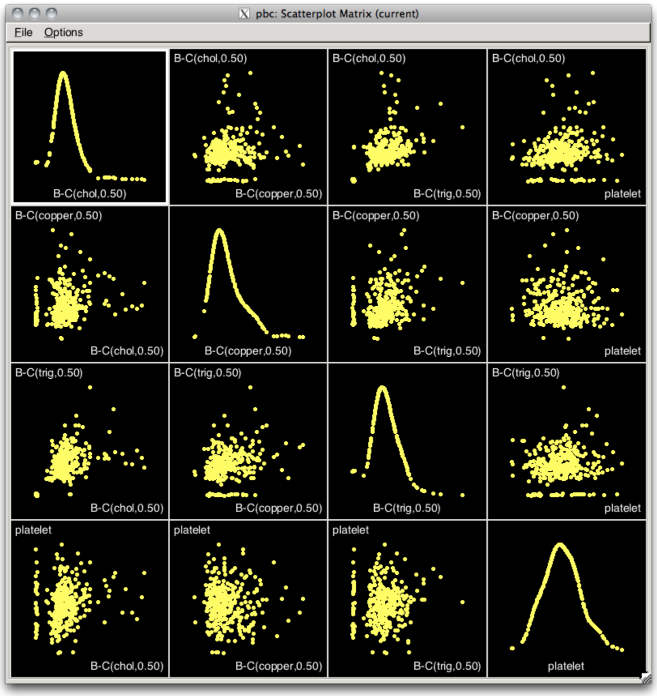
Stage 1
Box-Cox

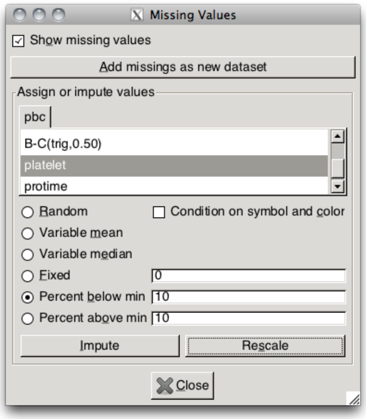
Box-Cox param: 0.500

a: 0 b: 1

Stage 2
No transformation

- Click on `Tools` menu and select the `Missing Values` item. This will bring up a control panel that will allow you to impute missing values in different ways, and generate a shadow matrix.




Missing Values

☑ Show missing values

Add missings as new dataset

Assign or impute values

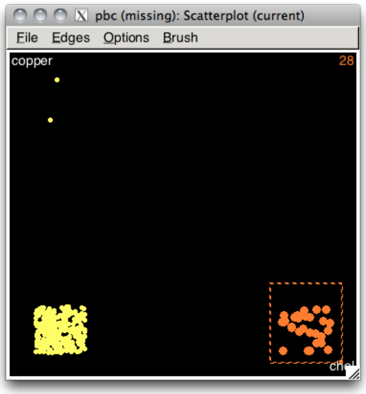
pbc

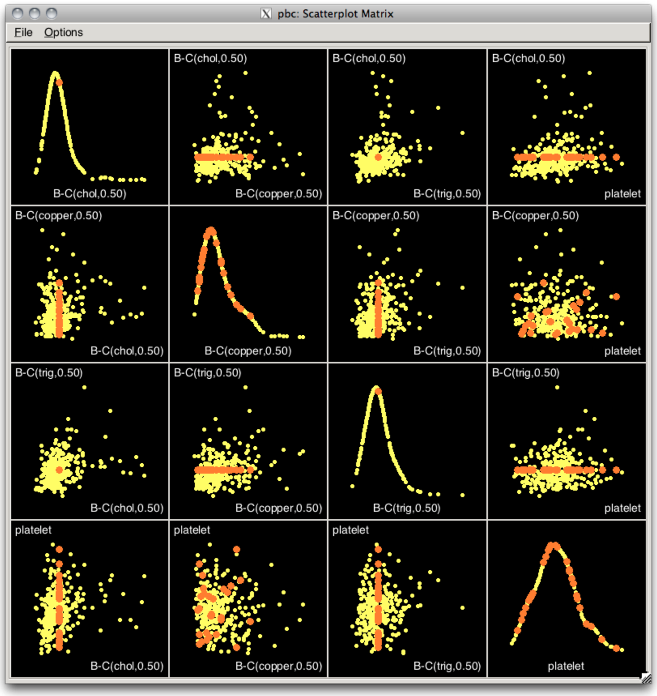
B-C(trig,0.50)
platelet
protime

○ Random ☐ Condition on symbol and color
○ Variable mean
○ Variable median
○ Fixed 0
● Percent below min 10
○ Percent above min 10

Impute Rescale

Close





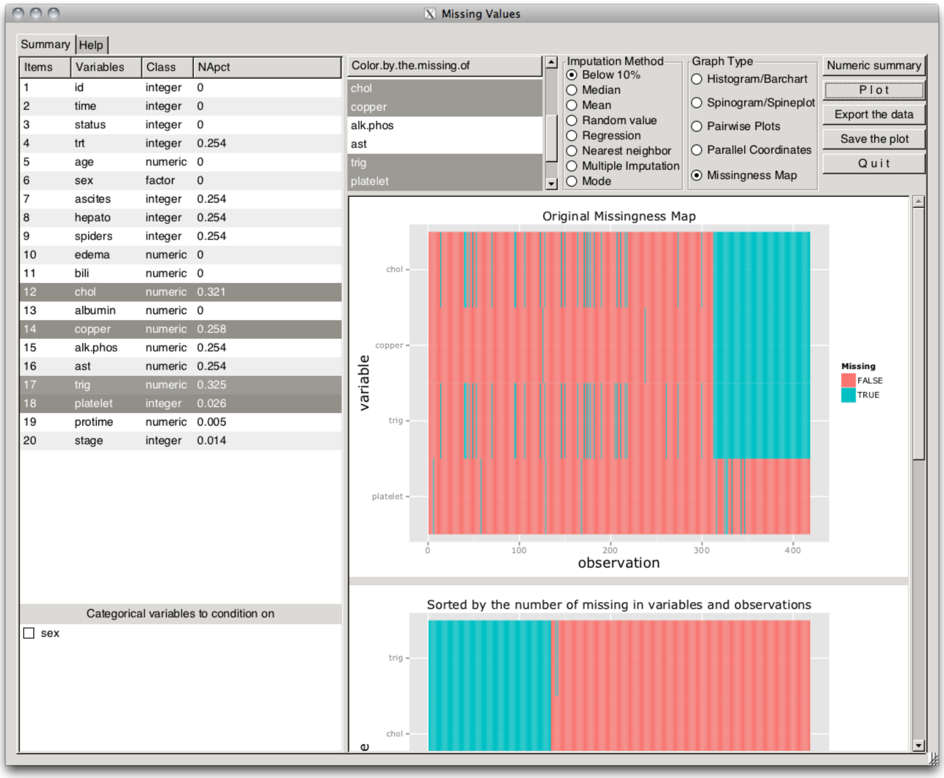
- Use the `norm` package for doing multiple imputation:

```
> library(norm)
> pbc.nm <- prelim.norm(as.matrix(pbc.tf))
> pbc.nm$nmis
> pbc.nm$r
> rngseed(1234567)
> theta <- em.norm(pbc.nm,showits=TRUE)
> for (i in 1:10) {
  pbc.impute <- imp.norm(pbc.nm, theta)
  colnames(pbc.impute) <- colnames(pbc)
  g[,"chol"] = pbc.impute[,"chol"]
  g[,"copper"] = pbc.impute[,"copper"]
  g[,"trig"] = pbc.impute[,"trig"]
  g[,"platelet"] = pbc.impute[,"platelet"]

  readline("Ready to continue? Press return ")
}
```

- An alternative to ggobi for exploring missing values is the missing data GUI package. It doesn't have a tour, but it has got a scatterplot matrix. Using the menu items, you can the % of cases missing on each variable, get a numerical summary of the missing values by case and variable, make plots and impute missings using a variety of methods.

```
> library(MissingDataGUI)
> MissingDataGUI(pbc.tf)
```



Items	Variables	Class	NApct
1	id	integer	0
2	time	integer	0
3	status	integer	0
4	trt	integer	0.254
5	age	numeric	0
6	sex	factor	0
7	ascites	integer	0.254
8	hepato	integer	0.254
9	spiders	integer	0.254
10	edema	numeric	0
11	bili	numeric	0
12	chol	numeric	0.321
13	albumin	numeric	0
14	copper	numeric	0.258
15	alk.phos	numeric	0.254
16	ast	numeric	0.254
17	trig	numeric	0.325
18	platelet	integer	0.026
19	protime	numeric	0.005
20	stage	integer	0.014

Categorical variables to condition on

□ sex

Imputation Method
● Below 10%
○ Median
○ Mean
○ Random value
○ Regression
○ Nearest neighbor
○ Multiple Imputation
○ Mode

Graph Type
○ Histogram/Barchart
○ Spinogram/Spineplot
○ Pairwise Plots
○ Parallel Coordinates
● Missingness Map

Exercises:

1. In the last lab, working with the olympics data, what were the missing values in the data? How did it happen that there were missings? How did you handle the missings? What value did you use to impute the missing values? Why?

2. For the Primary Biliary Cirrhosis (PBC) data:

   (a) Describe the univariate distributions of complete cases for chol, copper, trig, and platelet. What transformations might be used to make the distributions more bell-shaped? Make these transformations, and use the transformed data for the rest of this exercise.

   (b) Examine a scatterplot matrix of chol, copper, trig, platelet with missing values plotted in the margins.

      i. Describe the pairwise relationships among the four variables.

      ii. Describe the distribution between missings and non-missings for trig and platelet.

   (c) Generate the shadow matrix, and brush the missing values a different color.

   (d) Substitute means for the missing values, and examine the result in the scatterplot matrix (and a tour is using ggobi). What pattern is obvious among the imputed values?

   (e) Substitute random values for the missings, and examine the result in the scatterplot matrix (and a tour is using ggobi). What pattern is obvious among the imputed values?

   (f) **Extra credit:** In R, generate imputed values using multiple imputation. Examine different sets of imputed values in the scatterplot matrix. Do these sets of values look consistent with the data distribution?