

# Bivariate EDA in R Handout

## 1 Background

Measurements of drinking water and toenail levels of arsenic, as well as related covariates, were measured on 21 individuals with private wells in a New Hampshire community. The variables below were recorded in the **Arsenic.txt** file located on the R Resources web page.

- *age* : Age (yrs) of person
- *sex* : Sex of person
- *usedrink* : Household well used for drinking (1=<1/4 2=1/4 3=1/2 4=3/4 5=>3/4)
- *usecook* : Household well used for cooking (1=<1/4 2=1/4 3=1/2 4=3/4 5=>3/4)
- *arswater* : Arsenic in water (ppm)
- *arsnails* : Arsenic in toenails (ppm)

## 2 Initialization

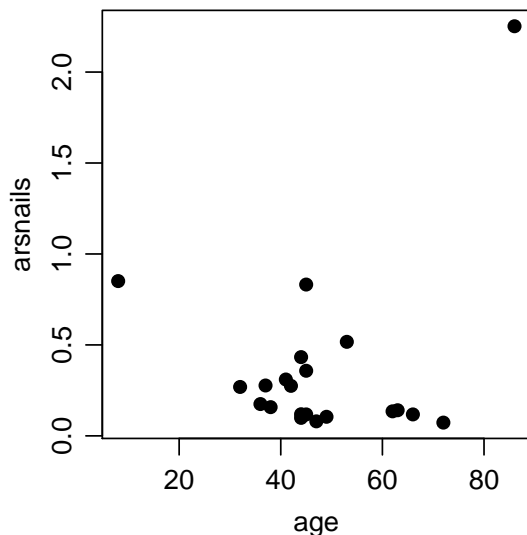
You must change the directory with the File...Change Dir menu or `setwd()` (as below, but for your directory) before the following command

```
> setwd("C://aaaWork//Class Materials//MTH107//F08//Lecture//H0s//")
> library(NCStats)

> Ars <- read.table("Arsenic.txt",header=TRUE)
> Ars$f.usedrink <- factor(Ars$usedrink)
> Ars$f.usecook <- factor(Ars$usecook)
> attach(Ars)
```

## 3 Bivariate EDA – Quantitative

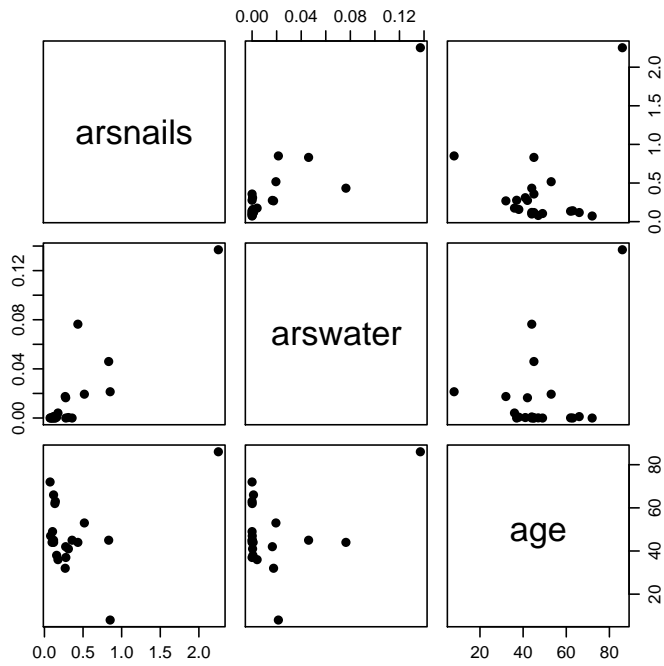
```
> plot(arsnails~age,pch=19)
```



```
> cor(arsnails,age)
```

```
[1] 0.2807416
```

```
> pairs(~arsnails+arswater+age,pch=19)
```



```
> cor(cbind(arsnails,arswater,age))
```

```
      arsnails  arswater   age
arsnails 1.000000 0.8964172 0.2807416
arswater 0.8964172 1.0000000 0.3452031
age       0.2807416 0.3452031 1.0000000
```

## 4 Bivariate EDA – Categorical

```
> freq.tbl <- table(f.usecook,f.usedrink)
> freq.tbl
```

```
      f.usedrink
f.usecook 1  2  3  4  5
      2  0  0  1  0  0
      5  1  1  1  3 14
```

```
> row.tbl <- prop.table(freq.tbl,margin=1)
> row.tbl
```

```

      f.usedrink
f.usecook  1  2  3  4  5
2 0.00 0.00 1.00 0.00 0.00
5 0.05 0.05 0.05 0.15 0.70

> col.tbl <- prop.table(freq.tbl,margin=2)
> col.tbl

```

```

      f.usedrink
f.usecook  1  2  3  4  5
2 0.0 0.0 0.5 0.0 0.0
5 1.0 1.0 0.5 1.0 1.0

> ttl.tbl <- prop.table(freq.tbl)
> ttl.tbl

```

```

      f.usedrink
f.usecook      1      2      3      4      5
2 0.00000000 0.00000000 0.04761905 0.00000000 0.00000000
5 0.04761905 0.04761905 0.04761905 0.14285714 0.66666667

```

## 5 Good “Ending” Commands

```
> detach(Ars)
```

## 6 Class Exercise

- Urbanization poses a major threat to stream and watershed ecosystems. One aspect of urbanization is the conversion of natural areas to land with impervious surfaces, thus increasing runoff of rain and, likely, pollutants. A University of Washington researcher recorded the percent of impervious land and the benthic index of biotic integrity (IBI) for 14 areas in the state of Washington. The IBI has been described as a measure of “the capability of supporting and maintaining a balanced, integrated, adaptive community of organisms having a species composition and functional organization comparable to that of natural habitat in the region.” The data for this study are below but much more information can be obtained at the [QELP site](#). Enter the data into Excel, save as a tab-delimited text file, and read the data into R. Describe the relationship between IBI and the percent of impervious area. Use a scatterplot [put IBI on the y-axis] and summary statistics to support your description.

```
% imperv 60 43 43 34 27 25 21 18 11  8  8  8  7  7  5  4  2
IBI      9 11 13 23 31 31 21 23 27 37 39 29 31 43 33 35 37
```

- Researchers surveyed people in a Minneapolis neighborhood and focused on two questions – “How big a problem is crime in your neighborhood (No problem, Small problem, or Big problem)?” and “How often do you walk at night in your neighborhood (Never, Sometimes, Always)?” The results of their survey can be found in [Walkcrime.txt](#). Use these data to describe the relationship between walking at night and the respondent’s perception of how big a problem crime is in their neighborhood. Provide tables to support your answers (and make sure that the levels of the variables are ordered appropriately).