Correlation

Dave Lorenz

December 30, 2015

Abstract

These examples demonstrate some of the functions and statistical methods for computing the correlation between two variables that are available in the smwrQW package.

Contents

1 Introduction

The examples in this vignette use the Atra dataset from the NADA package. The examples in this vignette use the function as.lcens to convert those data to a form used by the functions demonstrated. That conversion is not necessary for data of class ”qw” or numeric data. With the exception of the binary and Spearman methods, only censored data techniques are included in this vignette. Techniques that apply to single reporting limits and can require recensoring and simple substitution are not included, as the censored techniques can be used directly by the functions in smwrQW.

```
> # Load the smwrQW package
> library(smwrQW)
> # And the psych package, required for the phi function
> library(psych)
> # And the data
> data(Atra, package="NADA")
> # Convert the data
> Atra <- with(Atra, data.frame(June=as.lcens(June, censor.codes=JuneCen),
+   Sept=as.lcens(Sept, censor.codes=SeptCen)))
```

2 Binary Method

The binary method simply recodes values as 0 or 1 depending on whether the value is less than or greater than or equal to a specified criterion. The recoded values can then be tabulated and the phi coefficient is computed from the table.

The example below illustrates the recoding of values and uses the `phi` function in the `psych` package to compute the phi coefficient of association. The `code01` function returns a data frame of values with missing values removed from the input arguments. The `table` function creates a table of matched 0-0, 0-1, 1-0, and 1-1 pairs. The printed output from `phi` shows only the phi coefficient value; it is not constructed as a null-hypothesis test function.

```
> # Create data frame of 0/1 data for June and September
> Atra.01 <- with(Atra, code01(June, Sept))
> # Tabulate the 0/1 data and print it
> Atra.tbl <- table(Atra.01)
> print(Atra.tbl)
```

```
    Sept
June  0  1
   0  4  5
   1  1 14
```

```
> # And compute the phi coefficient
> phi(Atra.tbl, digits=4)
```

```
[1] 0.4503
```

3 Maximum Likelihood Estimation Method

Helsel (2012) only describes the square root of the likelihood r-squared as a measure of the Pearson correlation however, it is only appropriate for censored y and uncensored x and it only asymptotically approaches Pearson's correlation coefficient as the proportion of censoring decreases. Lyles and others (2001) describe a maximum likelihood estimation method that computes the Pearson correlation coefficient for censored x and y. This method is coded as `censCor` in the `smwrQW` package.

An important first step in any parametric statistical analysis is to plot the data. Figure 1 show the June and September atrazine concentrations on a log-log axis. In figure 1, uncensored data are shown by solid circles, and censored data are shown by open circles, horizontal lines are drawn between the y-axis and the symbol if x-axis data values (June) are censored and vertical lines are drawn between the x-axis and the symbol if the y-axis data values (Sept) are censored. Except for a single outlier, the data appear fairly linear.

```
> setSweave("graph01", 6 ,6)
> # Create the graph
> with(Atra, xyPlot(June, Sept, yaxis.log=TRUE, xaxis.log=TRUE))
> graphics.off()
```

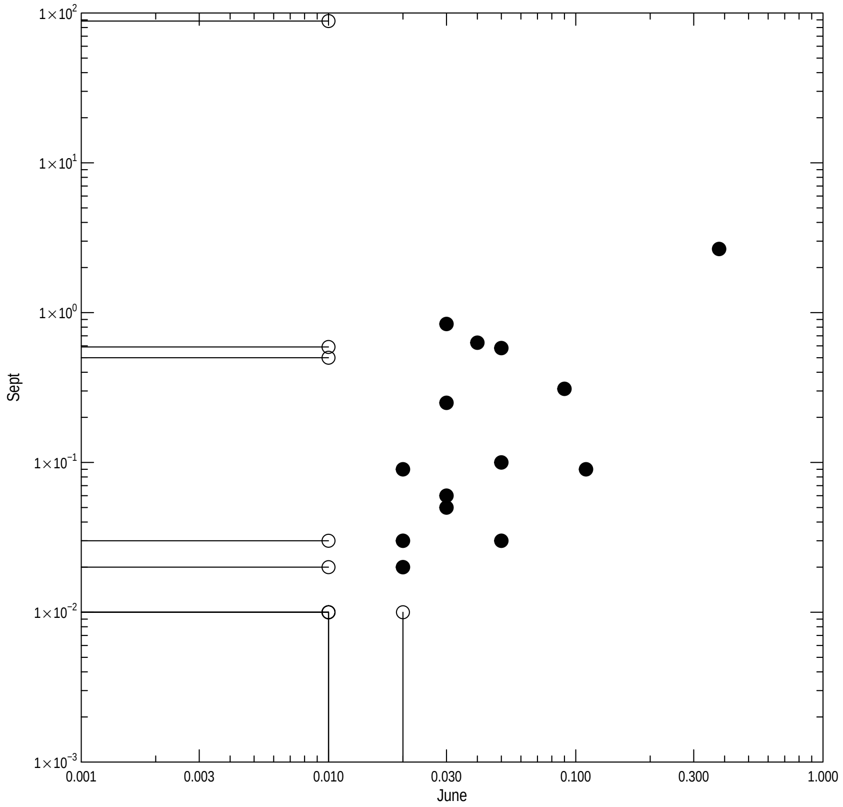

Figure 1. The scatter plot for the June and September atrazine data.

The `censCor` function is used to compute the Pearson correlation between two possibly left-censored variables. The output is a named vector: cor is the Pearson correlation coefficient; mnx and mny are the mean of the `x` and `y` arguments, log(June) and log(Sept) in this example; sdx and sdy are the standard deviations of the `x` and `y` arguments; cx and cy are the proportions of censored values of the `x` and `y` arguments; cxy is the proportion of the data where both `x` and `y` are censored; n is the number of observations; and ll0 and llcor are the log likelihoods of the null model (0 correlation) and the correlation model. For these that the correlation is about 0.38.

```
> # The Pearson correlation:
> with(Atra, censCor(log(June), log(Sept)))

       cor         mnx         mny         sdx         sdy          cx          cy 
 0.3783949  -4.0424356  -2.5851495   1.3662149   2.6696473   0.3750000   0.2083333 
       cxy           n         ll0       llcor 
 0.1666667  24.0000000 -83.9329289 -82.4925242 
```

4 Nonparametric Methods

Helsel (2012) only discusses the computation of Spearman's rho for single detection limit using simple substitution for the censored values. However later in his book, he discuses U coding that computes the equivalent of the rank for multiple detection limit data. Those U-coded values can be used to compute Spearman's rho. The R code following this paragraph illustrates its use on the Atra dataset. The U coded values the equivalent of the rank times 2 minus the number of observations plus 1 and thus are interchangeable with rank and can be either used directly or converted to ranks.

```
> # The Spearman correlation:
> # Step 1: U code the data, and print part
> AtraU <- with(Atra, codeU(June, Sept))
> head(AtraU)

  June Sept
1   23   21
2   11   17
3  -15   15
4    6   -3
5    6   19
6   15   13

> # Compute rho and the test
> with(AtraU, cor.test(June, Sept, method='spearman'))

        Spearman's rank correlation rho

data:  June and Sept
S = 1291.4, p-value = 0.03207
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.4385303
```

Helsel (2012) does discuss the computation of Kendall's tau for multiple detection limits. The function `kendallATS.test` computes Kendall's tau and also computes a slope estimate for y as a function of x. Its use is illustrated in the R code following this paragraph. Note that taub is the value of tau corrected for ties.

```
> # Compute tau and the test
> with(Atra, kendallATS.test(June, Sept))

        Kendall's tau with the ATS slope estimator

data:  June and Sept
taub = 0.36314, p-value = 0.02275
alternative hypothesis: true slope is not equal to 0
sample estimates:
   slope median.x median.y
4.019469 0.020000 0.075000
```

References

[1] Helsel, D.R. 2012, Statistics for Censored Environmental Data Using Minitab and R: New York, Wiley, 324 p.

[2] Helsel, D.R. and Cohn, T.A., 1988, Estimation of descriptive statistics for multiply censored water quality data: Water Resources Research v. 24, n. 12, p.1997–2004

[3] Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 522 p.

[4] Lyles, R.H., Williams, J.K., and Chuachoowong R., 2001, Correlating two viral load assays with known detection limits: Biometrics, v. 57 no. 4, p. 1238–1244.

[5] Lorenz, D.L., 2016, smwrQW–an R package for managing and analyzing water-quality data, version 1.0.0: U.S. Geological Survey Open File Report 2016-XXXX.

[6] Lorenz, D.L., Ahearn, E.A., Carter, J.M., Cohn, T.A., Danchuk, W.J., Frey, J.W., Helsel, D.R., Lee, K.E., Leeth, D.C., Martin, J.D., McGuire, V.L., Neitzert, K.M., Robertson, D.M., Slack, J.R., Starn, J., Vecchia, A.V.,Wilkison, D.H., and Williamson, J.E., 2011, USGS library for S-PLUS for Windows—Release 4.0: U.S. Geological Survey Open-File Report 2011-1130.