Trend Analysis for Data with Nondetects

Dennis R. Helsel

PracticalStats.com

1

Objectives of the 'Trend Analysis' webinar

1. To see why DL/2, DL/sqrt(2) and other fractions should never be substituted for nondetects when doing trend analysis
2. To see what is possible when testing for trends with censored data
3. To give an overview of how trend analysis methods work
4. To highlight one of the many new sections in our online course Nondetects And Data Analysis available at https://practicalstats.teachable.com

 2

2

Why Not Substitute DL/2 etc. for nondetects?

What's wrong with substitution?

- Most commonly substituted values are DL/2 and DL/sqrt(2)
- Produces invasive data alien to the concentrations actually in samples
- Substituting a constant always results in a poor estimate of std dev, and with multiple censored data, sneaks in flat zero-slope lines
- Results in poor estimates and incorrect statistical tests

There are better ways

 3

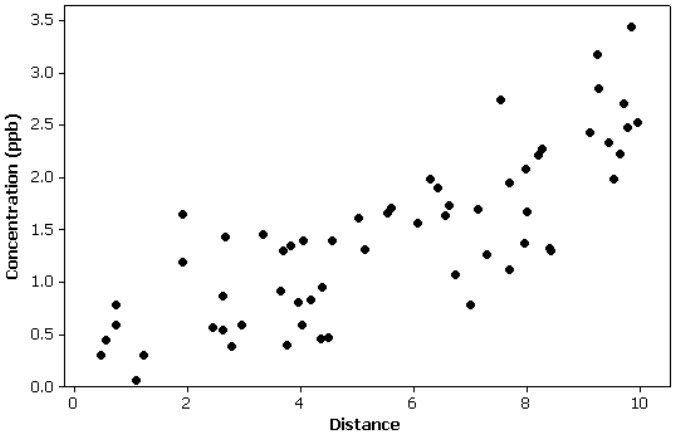
3

What's wrong with substitution?

Correlation and Regression

Before censoring, the true correlation is r=0.81

What if 50% of the previous data were instead found to be nondetects?



 4

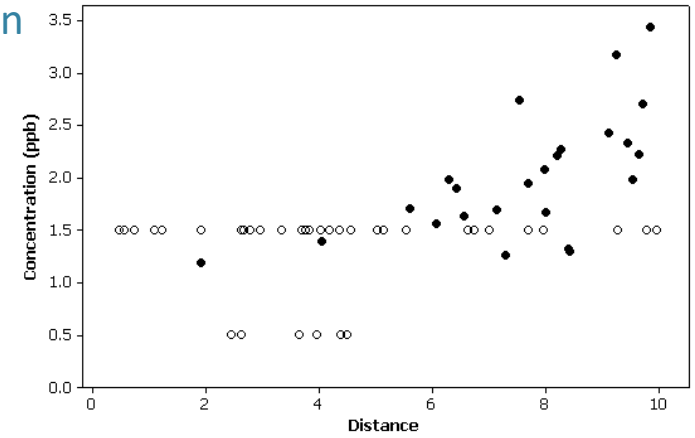
4

Not finding a trend that is there

Correlation and Regression

2 DLs. DL/2 substituted for each.

After substitution. invasive data form flat lines, lowering correlation to r=0.55



5

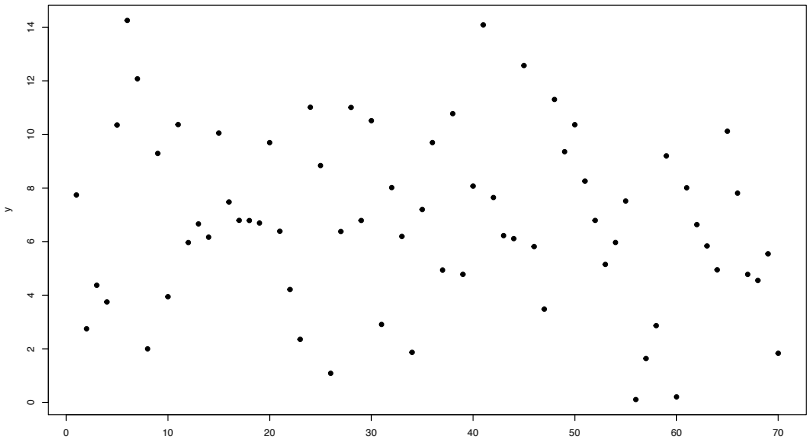
5

Finding a trend that isn't there

Before Substitution: the true situation is No correlation over time (tau = 0.11, p = 0.18)

Suppose 23% of samples had higher background salinity, were diluted and resulted in nondetects.

Because of better lab equipment, detection limits decreased over time, from 10 to 7 to 5.



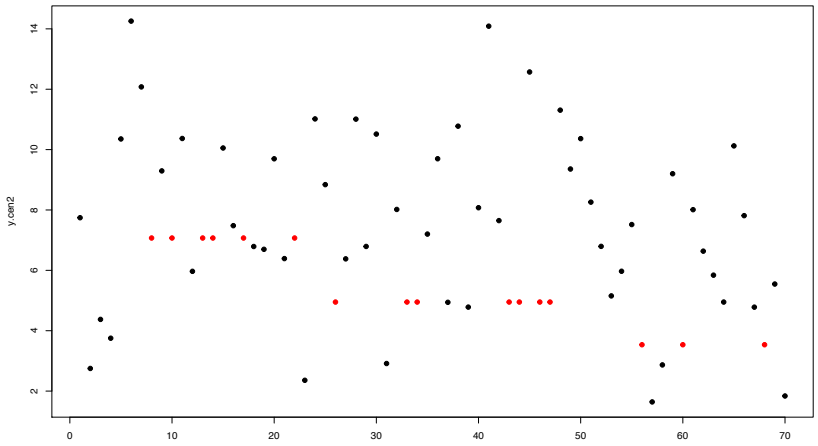
6

6

Finding a trend that isn't there

Invasive pattern

- Detection limits decrease over time
- Censored data shown in red
- By substituting DL/sqrt(2) you put in an invasive downtrend that wasn't there in the original data
- A false correlation with time (false trend) becomes significant
  p = 0.04



7

False positives and negatives

- Substitution with no change in DL puts in a flat, zero-slope "no change" line.
- Can turn a dataset with trend into one without trend.
- Substitution with decreasing detection limits over time puts in a false decreasing "trend" pattern.
- Can turn a dataset with no trend into one with a "significant" trend.

8

Substitution → The Man Who Wasn't There

Yesterday upon the stair
I saw a man who wasn't there
He wasn't there again today
Oh how I wish he'd go away.

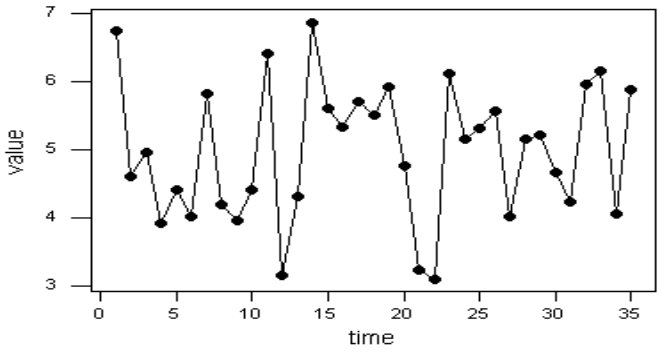
-- Hughes Mearns (1875-1965)

9

What is trend analysis?

Any test where time is an explanatory variable

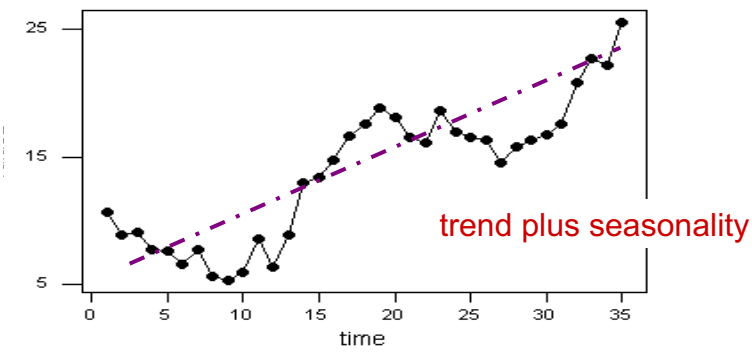
Often measured as correlation with time



Null hypothesis for all trend tests:
No signal, only random noise over time

10

Trend Signal



Correlation of Y with time

- Often Y is affected by other covariates than time. This is noise obscuring trend and often must be accounted for.
- Seasonal variation often found in natural world. This also obscures trend and must be accounted for in order to see a trend.

11

Trend Analysis Methods for Censored Data

	Time only X var	Time + Covariate	Seasonal
Parametric	1 MLE Simple Regression cencorreg (y, ycen, x)	2 MLE Multiple Regression cencorreg (y, ycen, x.frame)	3 MLE Regression with sin and cos terms cencorreg (y, ycen, x.frame)
Nonparametric	4 Akritas-Theil-Sen ATS (y, ycen, time)	5 ATS on residuals from a covariate smooth centrend (y, ycens, x, time)	6 Censored Seasonal-Kendall test censeaken (y, ycen, time, season)

None of these methods substitute a number like DL/2 for nondetects

12

Parametric Trend Analysis: MLE Regression

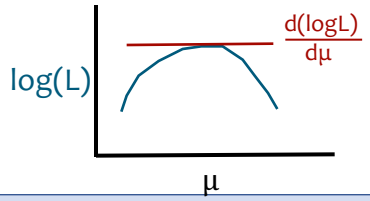
- Starts with the observed data
- Given the observed data, what values for parameters (slope, intercept) are most likely to have given rise to these data?
- For censored data, 2 types of information are utilized: the values for detected observations and the observed proportions of data below each detection limit (how is the proportion of <DL data changing with increasing X?)
- We must assume that the residuals from the regression (bivariate residuals for correlation) follow a chosen distribution.

 13

13

How MLE Regression Works

- Write a likelihood function L = function (slope, intercept).
  This evaluates the match between observed Y_i and the model $(b_0 + b_jX_j)$
  i = 1…n observations j = 1...k X variables
- Want to maximize log(L) where log(L) is negative.
- Do this by setting the derivative of log(L) = 0, and solve for slope and intercept

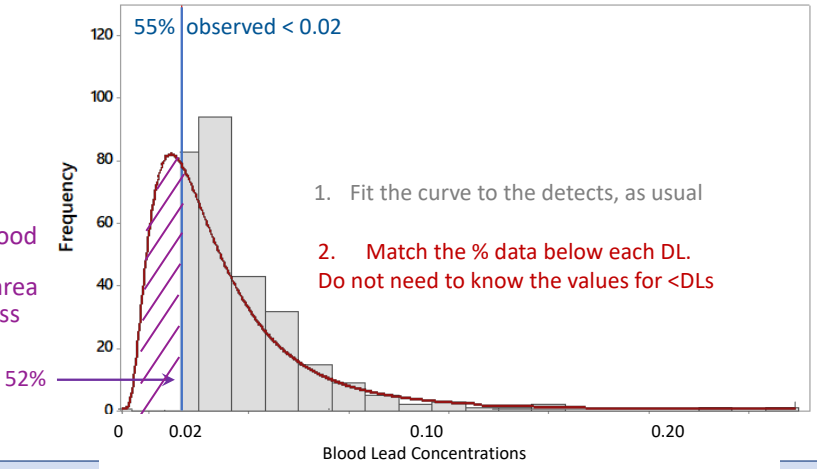


 14

14

Probability Density Function for Censored Data

Find the best-fitting mean and sd for a chosen (here, lognormal) distribution. The fit has two parts, one for detects and one for nondetects.



Maximum Likelihood (MLE) best fit has 52% of the total area under its curve less than 0.02.

52%

1. Fit the curve to the detects, as usual

2. Match the % data below each DL.
Do not need to know the values for <DLs

 15

15

Parametric Trend Analysis

Censored regression solved by Maximum Likelihood Estimation (MLE) -- cencorreg function

1. Check for multicollinearity between X variables
2. Use the cencorreg script to compute the regression equation
3. Check that residuals follow the assumed distribution
4. When comparing models, choose the one with the lowest AIC

 16

16

Example Data

DairyCreekCr.Rdata includes Total Recoverable Chromium concentrations (some nondetects) and dectime (decimal time) for the day of sampling. Provided by a colleague.

Note: Data have been altered from the original (I filled in some flow data so fewer were missing).

Censoring indicator variable (here CrND distinguishes 1 = a detection limit in the Y column from 0 = detected concentration in the Y column.

To perform a simple regression (only dectime as the X variable), use:

```
> cencorreg(`Total Recoverable Chromium`, CrND, dectime)
```

17

1 Simple Regression
(one X variable -- time)

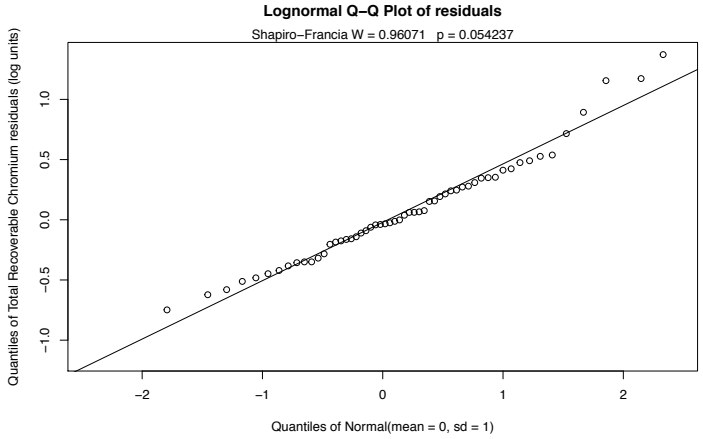
```
> cencorreg(`Total Recoverable Chromium`, CrND, dectime)
 Likelihood R = -0.339                    AIC = 96.39843
 Rescaled Likelihood R = -0.3824          BIC = 101.8751
 McFaddens R = -0.2815
Call: survreg(formula = "log(Total Recoverable Chromium)", data = "dectime",
    dist = "gaussian")          NOTE: default is to use log(Y)

Coefficients:
(Intercept)    dectime
119.7497387 -0.0596987
Scale= 0.4767561
Loglik(model)= -44.7   Loglik(intercept only)= -48.5
        Chisq= 7.69 on 1 degrees of freedom, p= 0.00555
n= 63
```

The slope is significant (p = 0.005) showing a decrease of 0.059 log units per year.

18

Check Normality of Residuals

Do not reject normality using log(Cr)



 19

19

2 Multiple Regression, X and Time

Regression using both flow and dectime as explanatory variables. To do multiple regression using cencorreg, input the x variables as a single data frame. Create the data frame for both variables, then run the model:

```
> xvar2 <- data.frame(dectime, mean_daily_flow_cfs)
> reg.cr <- cencorreg(`Total Recoverable Chromium`, CrND, xvar2)
 Likelihood R2 = 0.4617                    AIC = 63.5293
 Rescaled Likelihood R2 = 0.5846           BIC = 70.83945
 McFaddens R2 = 0.3971
> summary(reg.cr)
```

smaller than the 1-variable model, so this is better

(see next slide)

 20

20

2 MLE Regression Results

```
> summary(reg.cr)
Call:
survreg(formula = "log(Total Recoverable Chromium)", data = "dectime+mean_daily_flow_cfs",
    dist = "gaussian")
                       Value Std. Error      z       p
(Intercept)         1.02e+02   3.31e+01   3.09  0.0020.
dectime            -5.11e-02   1.64e-02  -3.11  0.0019
mean_daily_flow_cfs 6.19e-04   9.89e-05   6.26 3.9e-10
Log(scale)         -1.01e+00   1.01e-01 -10.03 < 2e-16
Scale= 0.362

Gaussian distribution
Loglik(model)= -27.3   Loglik(intercept only)= -45.2
         Chisq= 35.92 on 2 degrees of freedom, p= 1.6e-08
n=58 (5 observations deleted due to missingness)
```

Downtrend of 0.051 log units per year. Adj for flow

Significant increase in log(Cr) with flow

Overall significant model

Always check that VIFs < 10

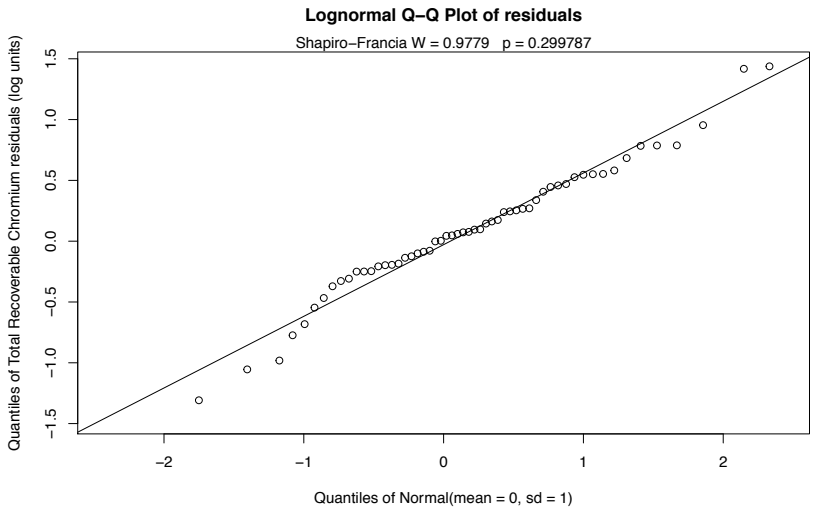
```
> vif(lm(`Total Recoverable Chromium`~ dectime + mean_daily_flow_cfs))
            dectime      mean_daily_flow_cfs
           1.000662            1.000662
```

No multicollinearity present

- See the webinar "Correlation and Regression for Nondetects" on our Training Site for details on vifs.
- vifs do not have anything to do with the Y variable. They just measure the multicollinearity (multiple correlations) between the X variables.

Always Check Normality of Residuals

Do not reject normality of residuals from the multiple regression using log(Cr)



 23

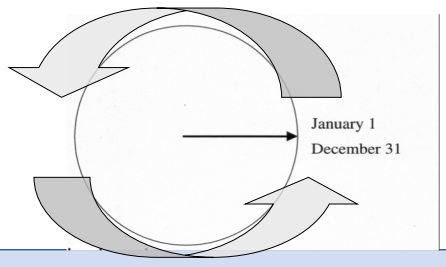
23

3 Seasonal Regression with sine and cosine

Two new explanatory variables are created, and added to the regression equation

These are the sine and cosine of $2\pi T$, where T is time in decimal years (1997.5)

Resulting in one revolution every year … $2\pi T$



 24

24

3 Regression with sine and cosine

$$Y = b_0 + b_1*T + b_2*X + b_3*\sin(2\pi T) + b_4*\cos(2\pi T)$$

Keep both sin and cos seasonal terms, or keep neither.

Base the decision on significance of b_3, b_4.

If either are significantly different than zero, keep both terms.

You can instead compare the AIC for models with and without the sin and cos terms. The model with the lowest AIC is better.

25

3 Regression with sine and cosine

```
> cosT <- cos(2*pi*dectime
> sinT <- sin(2*pi*dectime)
> xvar4 <- data.frame(dectime, mean_daily_flow_cfs, sinT, cosT)

> reg4 <- cencorreg(`Total Recoverable Chromium`, CrND, xvar4)
 Likelihood R2 = 0.4645                    AIC = 67.22336
 Rescaled Likelihood R2 = 0.5882           BIC = 78.68859
 McFaddens R2 = 0.4005
```

AIC was 63.53 without sine and cosine, so the 2 variable model was better.

No significant seasonal variation

continued on next slide:

```
> vif(lm(`Total Recoverable Chromium`~ dectime + mean_daily_flow_cfs + sinT + cosT))
         dectime    mean_daily_flow_cfs                sint                cost
          1.008583               2.555278            1.883205            1.596627
```

No multicollinearity problems

26

3 Regression with sine and cosine

```
survreg(formula = "log(Total Recoverable Chromium)", data = "dectime+mean_daily_flow_cfs+sinT+cosT",
    dist = "gaussian")

                       Value Std. Error      z       p
(Intercept)         1.00e+02   3.31e+01   3.03 0.00241
dectime            -5.02e-02   1.64e-02  -3.05 0.00226
mean_daily_flow_cfs 5.73e-04   1.57e-04   3.64 0.00027
sinT                4.73e-02   8.98e-02   0.53 0.59848
cosT                2.59e-03   8.98e-02   0.03 0.97700
Log(scale)         -1.02e+00   1.01e-01 -10.06 < 2e-16
Scale= 0.361

Gaussian distribution
Loglik(model)= -27.1   Loglik(intercept only)= -45.2
Chisq= 36.23 on 4 degrees of freedom, p= 2.6e-07
n=58 (5 observations deleted due to missingness)
```

Significant down trend in log(Cr)

Significant relation to flow

Not Significant

Not Significant

Conclusion: No seasonal variation.
Use the 2 variable model.

 27

27

Nonparametric Trend Tests with Censored Data

Based on ATS: The Akritas-Theil-Sen line

- Slope is the one that produces a Kendall's tau of 0 for the residuals from the line.
- Test for slope = 0 is the test for Kendall's tau of data vs. time – the Trend Test
- See the textbook *Statistics for Censored Environmental Data Using Minitab and R* (Helsel, 2012) for more detail on the ATS method.

 28

28

4. Simple Nonparametric Regression

```
> ATS(`Total Recoverable Chromium`, CrND, dectime, LOG = FALSE)
Akritas-Theil-Sen line for censored data

Total Recoverable Chromium = 74.2315 -0.0366 * dectime
Kendall's tau = -0.2232   p-value = 0.00979
```

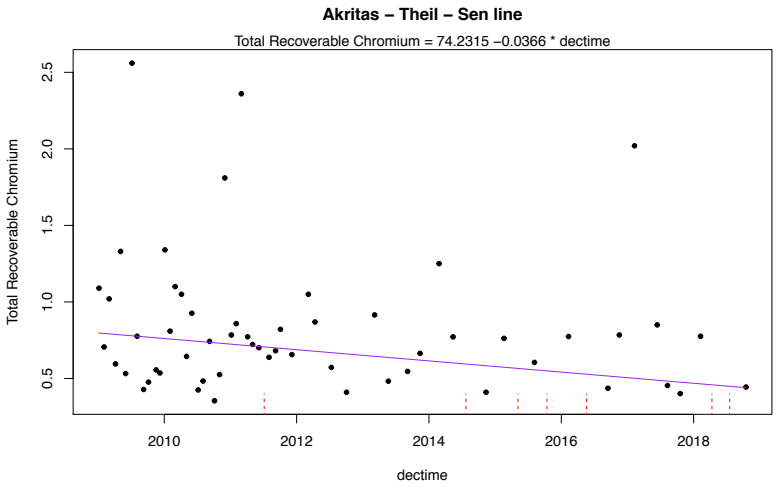
(tau = -0.22 is something like -0.4 for Pearson's r correlation)

There is a significant downtrend. The model is linear over time. So there is a median decrease of 0.0366 ug/L of Chromium per year.

29

4. ATS line

A "linear median". The test is not strongly affected by outliers. Logs of Cr could be taken instead, producing a curved relationship. Your call. Same tau and p-value.



30

5. Nonparametric Trend with a Covariate "multiple regression"

1. Compute a smooth of censored Y vs X, where X is not time. This models the relationship between Y and X.
2. Subtract off that relationship by taking the residuals from the smooth.
3. Compute an ATS on the residuals -- Kendall's tau test of change in residuals over time. Slope is still in Y units per time.

R function centrend (Y, Y.cen, X, time)

time is often as decimal time (i.e. 2013.5 for halfway through the year)

 31

31

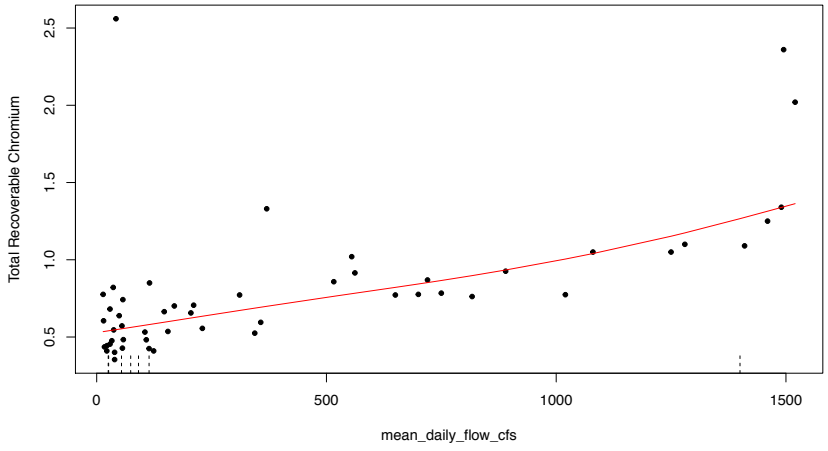
centrend function

```
> resid.trend <- centrend(`Total Recoverable Chromium`, CrND,
mean_daily_flow_cfs, dectime)
```

Smooth of Cr concentrations vs. mean daily flow

Note that its not necessarily linear

dashed lines are the censored data



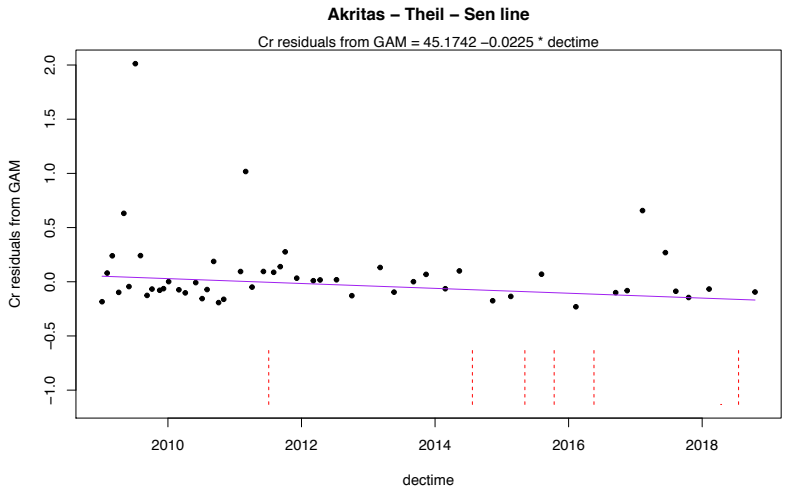
 32

32

ATS (Mann-Kendall) test on censored GAM residuals

```
Kendall's tau = -0.196
p-value = 0.03018
```

Low censored data (dashed lines) contribute to the test results. Note that they appear mostly at later times, increasing evidence for a downtrend.



33

6. Seasonal Kendall test on censored data

- Computes an ATS line and test for each season separately
- Combines them to produce an overall SK test
- Is a test of 'consistent trend' -- if one season shows a significant increasing trend and a 2nd a significant decreasing trend, these can cancel each other out so that there is no overall significant Seasonal Kendall trend

34

Seasonal Kendall test

- Compare all data within the same season to one another
- DOES NOT compare data across different seasons
- Sum up the individual season's S test statistic to get an overall Seasonal Kendall test

 35

35

Computing the Seasonal Kendall test

The test statistic S_i for each season is the "Mann-Kendall test" -- the number of pluses P_i (increases in Y as time increases) minus the number of minuses M_i (decreases in Y as time increases), comparing data only within that season. For season i we have:

$$S_i = P_i - M_i$$

 36

36

Seasonal Kendall test statistic S

For the i = 1 to m seasons,

$$S = \sum_{i=1}^{m} S_i$$

S becomes significant as it becomes more and more nonzero

 37

37

6. censeaken function

```
> censeaken (dectime, `Total Recoverable Chromium`, CrND, group = Season)

 DATA ANALYZED: Total Recoverable Chromium vs dectime by Season
----------
  Season  N    S    tau      pval intercept    slope
1    Dry 34 -176 -0.314 0.0091337    79.103 -0.03901
----------
  Season  N   S     tau    pval intercept    slope
1    Wet 29 -24 -0.0591 0.66604    24.355 -0.01169
----------
Seasonal Kendall test and Theil-Sen line
   N    S    Tau Pvalue_SK Nreps Intercept    Slope
1 63 -200 -0.207     0.014   999    74.232 -0.03655
```

Significant downtrend in Dry season

No significant trend in Wet season

Significant trend overall. SK slope is -3.6 ug/L per year

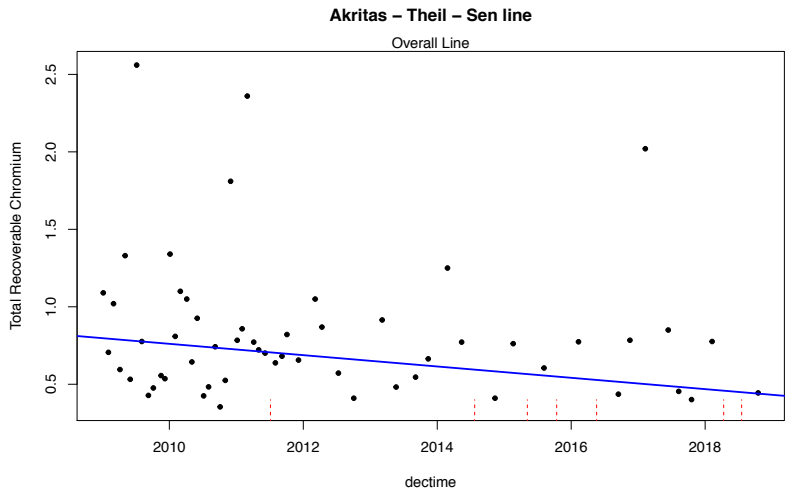
 38

38

6. censeaken function

Nondetects influence the line and test.

They occur more frequently at later times, adding to the evidence of a downtrend.



 39

39

Permutation p-value for the SK test

The SK test without censored data uses a normal approximation to the SK test statistic.
(not a normal assumption for the data, just a smart move by a statistician to form the test statistic)

However the variance of S is not easily computed in a formula when there are censored data. Solution? A permutation test

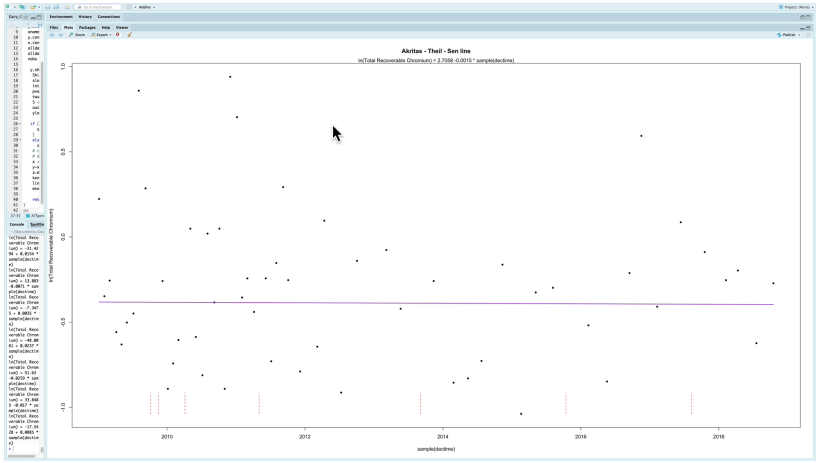
$$Z_S = \begin{cases} \dfrac{S-1}{\sqrt{VAR(S)}} & \text{if } S > 0 \\ 0 & \text{if } S = 0 \\ \dfrac{S+1}{\sqrt{VAR(S)}} & \text{if } S < 0 \end{cases}$$

 40

40

Permutations by shuffling time

The time variable is randomly shuffled 1000s of times and re-assigned to the Y data.



Then S is computed for each shuffle.

(notice how the times of the nondetect lines change between shuffles)

 41

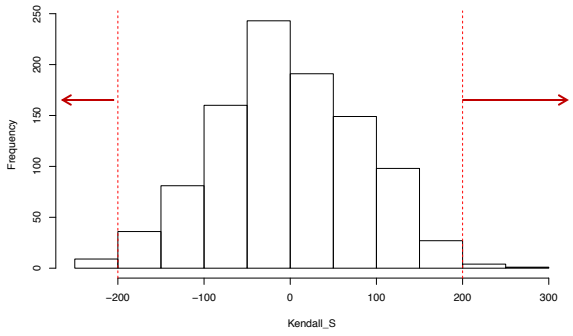
41

Permutation p-value for the SK test

For each shuffle, the S = P - M test statistic is computed for each season and summed to produce the overall SK S statistic. The collection of the 1000s of SK S statistics put together in one histogram is a picture of the null hypothesis. The p-value is the proportion of times that just by chance the same or greater strength of trend (same S observed from your data) occurs when the null hypothesis (no trend) is true.

For the Dairy Creek data, S = -200. The proportion that |S| ≥ 200 is the p-value. Here it was 14/1000, or 0.014

(it may be slightly different the next time)

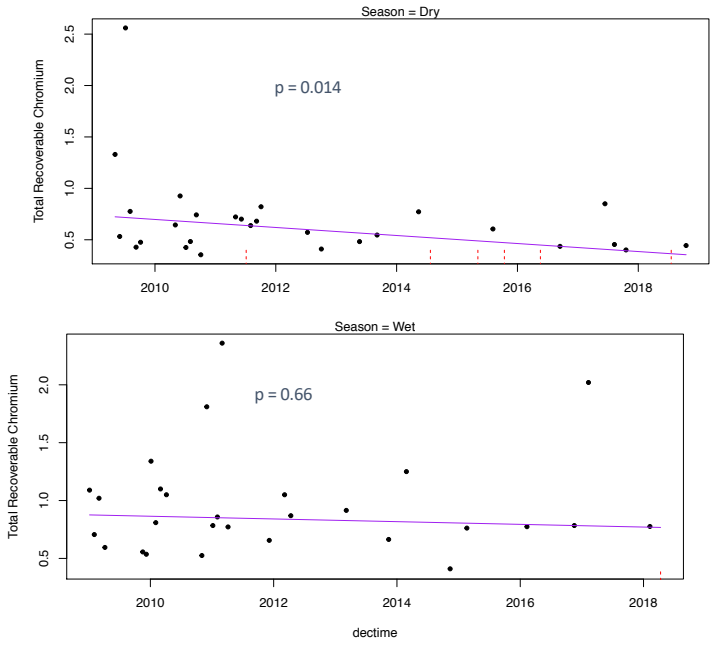


 42

42

Optional graphs for each season

```
censeaken (dectime, `Total
Recoverable Chromium`, CrND,
group = Season, seaplots = TRUE)
```



43

Summary: Trend Tests for Data with Nondetects

1. Substituting values that are a function of the detection limit(s) will cause havoc with trend tests. It's the classic error: DLs go down over time, so DL/2 goes down over time producing a 'trend' that probably isn't in your field data. It just reflects changes in the lab.
2. There are excellent methods for conducting trend tests with censored data that do not substitute values for nondetects
3. Parametric methods are based on censored regression
4. Nonparametric methods are based on Kendall's tau and the Akritas-Theil-Sen line
5. All of these are found in our Nondetects And Data Analysis (NADA) online training course

44

Our Next Webinar

Tuesday January 21st 11 am Mountain time

Topic to be Determined

- Will be a topic from either our *Nondetects And Data Analysis* or *Applied Environmental Statistics* online courses
- Previously nominated topics for upcoming webinars:
  - How to include both nondetects and "greater-thans" in analyses
  - Principal Components Analysis

  (email me at ask@practicalstats.com and add your suggestion)
- Online signup for our newsletter/announcement list to directly receive announcements each month is at http://practicalstats.com/news/
- Or check our webinars page periodically at http://practicalstats.com/training/webinar.html to see the announcement and to register.

45

This 'Trend Analysis for Data with Nondetects' webinar will be available Thursday for streaming

- at our Online Training Site

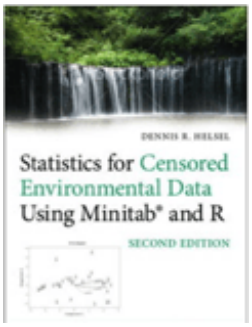
  http://practicalstats.teachable.com/

  (and click the "View all courses" button to see the free webinars)

Let colleagues who missed it know about it.

46

Thank you for attending



- Some of the material is contained in my textbook Statistics for Censored Environmental Data Using Minitab and R, published by Wiley (2012)
- This topic and much more is now covered in our online course Nondetects And Data Analysis, on our Training Site.
- All opinions are my own and do not represent those of anyone else you can think of.

Answers to your questions: Some now, all are answered by Thursday -- the file will be on our Downloads page: http://practicalstats.com/info2use/downloads.html

Get in touch!

Dennis Helsel ask@practicalstats.com

Courses & free webinars at our Training Site: http://practicalstats.teachable.com

 47

47