Nondetects And Data Analysis:
Trend Analysis with NDs

Dennis R. Helsel, Ph.D
Practical Stats

1

Trend Analysis Methods

	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 GAM 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

2

2

R packages required

If you haven't already installed and loaded these three R packages:

cenGAM, mgcv, nlme

you will need to install and load them before you can perform the 3 nonparametric methods in this Section.

3

3

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

4

4

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 the cencorreg script from the regression section:

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

5

5

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
```

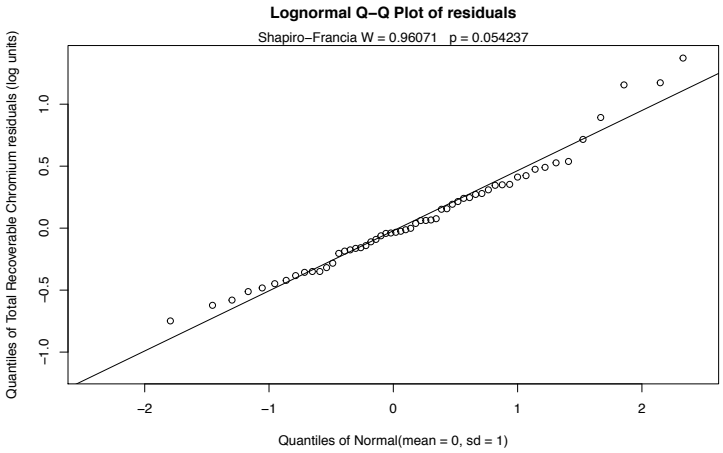
The slope is significant (p = 0.005) showing a decrease of 0.059 log units per year.

6

6

Check Normality of Residuals

Do not reject normality using log(Cr)



7

7

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)

8

8

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

9

9

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

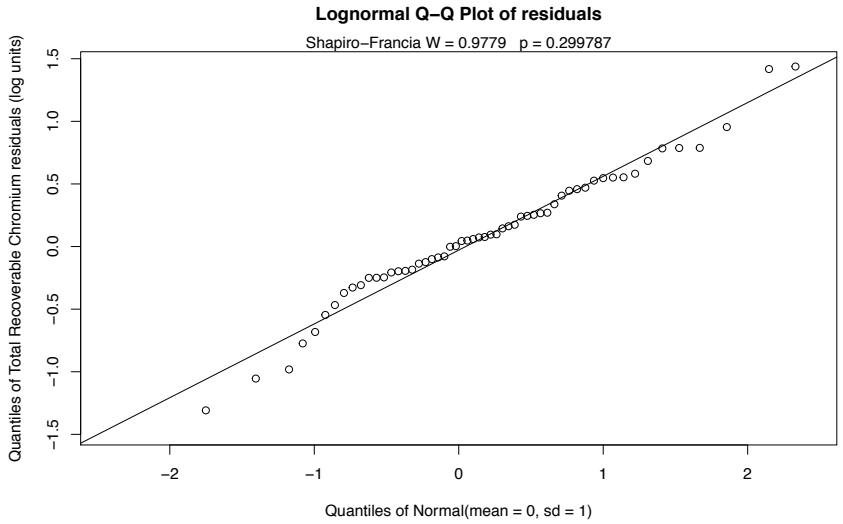
You may use the lm command for uncensored regression because vifs do not have anything to do with the Y variable. They just measure the multicollinearity (multiple correlations) between the X variables.

10

10

Always Check Normality of Residuals

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



11

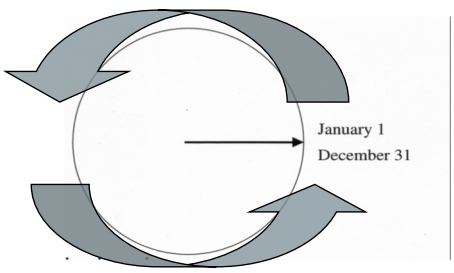
11

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 … $2\pi T$



12

12

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.

13

13

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

14

14

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.

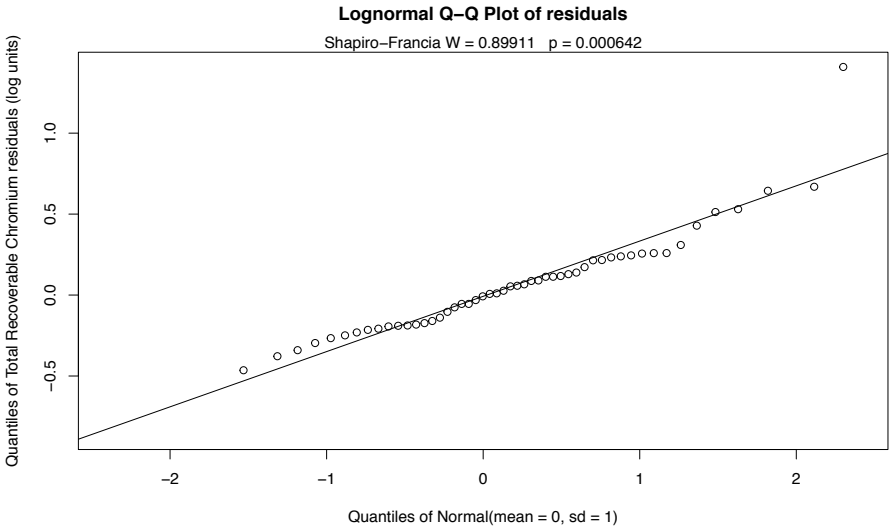
15

15

Always Check Normality of Residuals

Reject normality of residuals from the multiple regression. But there's not much else you can do -- its caused by the one large outlier.

Check the value for that outlier but use these units because of the straight line for all but one point.



16

16

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
- This should sound familiar. See the ATS portion of the regression section.

17

17

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)

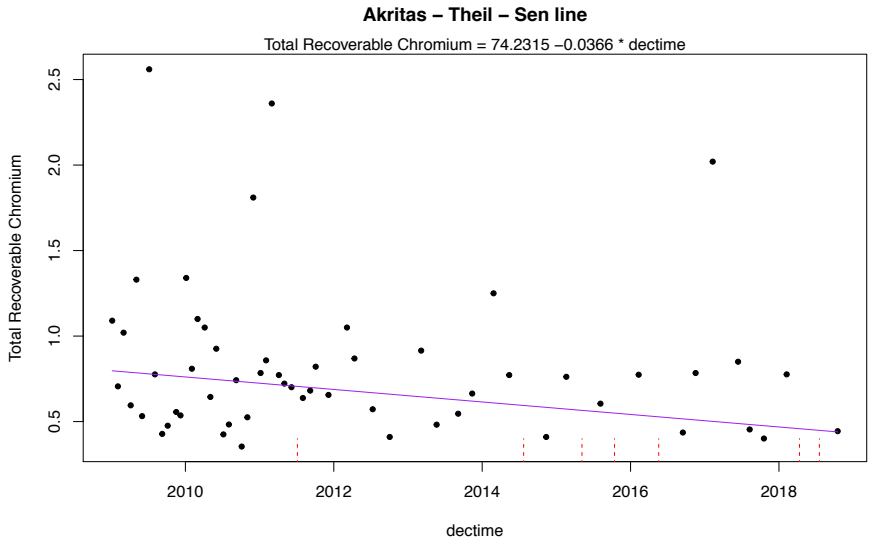
There is a significant downtrend. The model is linear over time. So there is a median decrease of 0.0366 ug/L of Chromium per year.

18

18

4. ATS line

This test is not strongly affected by outliers. Logs of Cr could be taken instead, producing a curved relationship. Your call. Same p-value.



19

19

5. Nonparametric Trend with a Covariate "multiple regression"

1. Compute a smooth of censored Y vs X, where X is not time, using Generalized Additive Models (GAM)
2. 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)

20

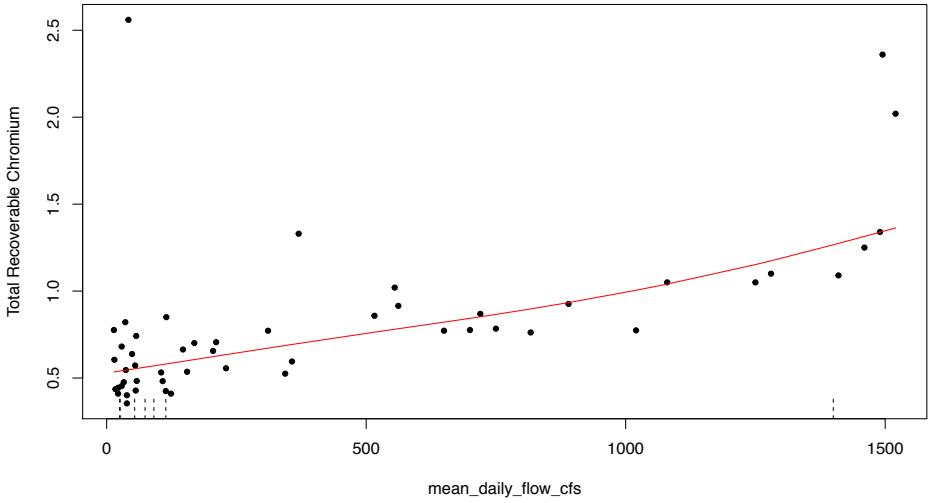
20

centrend function

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

Smooth of Cr concentrations vs. mean daily flow by censored GAM

dashed lines are the censored data

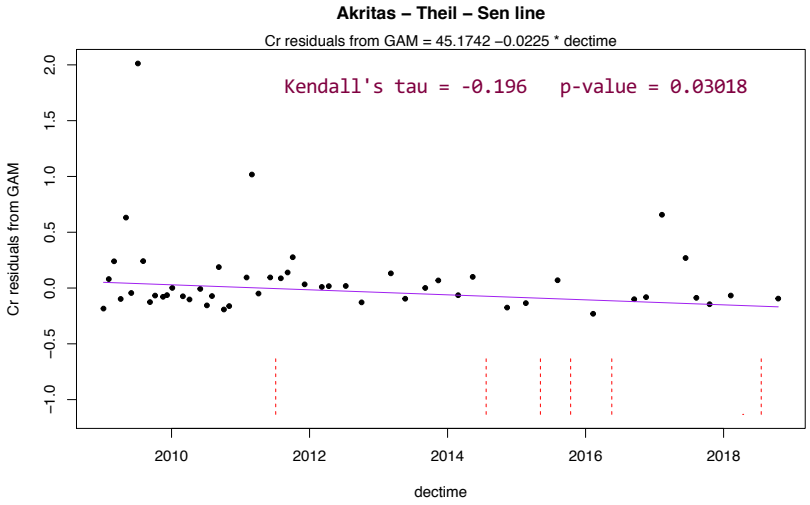


21

21

ATS (Mann-Kendall) test on censored GAM residuals

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



22

22

What's a GAM?

Generalized Additive Models

- relate Y to Xs using smooth curves instead of just a linear model
- smooths are a weighted combination of multiple functions
- a smoothing parameter determines how the functions are combined, so that the smoothness can be customized to the need
- Cubic regression splines (the default in centrend) use a cubic regression form (x, x^2, x^3 terms). Multiple cubic regressions are run and combined to maximize smoothness, but not so smooth that a straight line results.

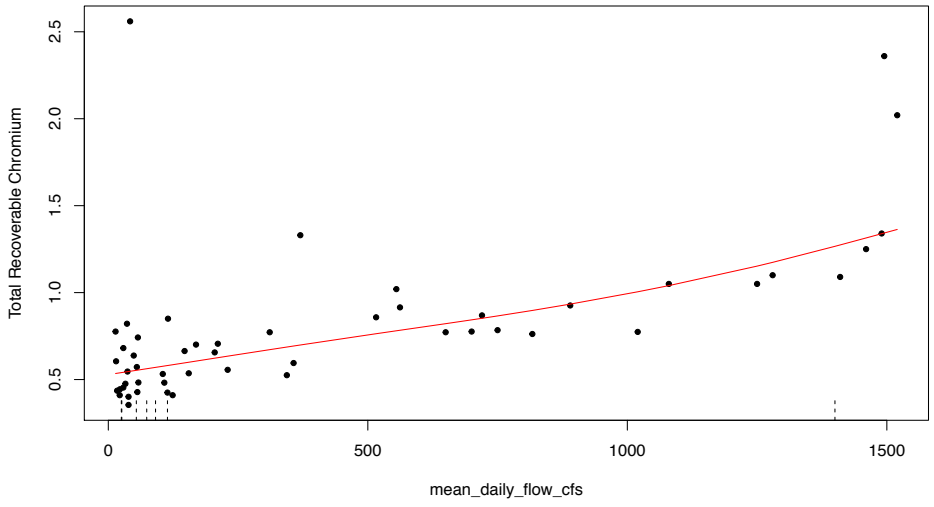
23

23

Note that the relationship is not linear

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

Smooth of Cr concentrations vs. mean daily flow using a cubic regression spine function



24

24

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

25

25

Seasonal Kendall test

- Compare all data within the same season to one another
- DO NOT compare data across different seasons
- Akritas-Theil-Sen slope is the slope that produces a zero Kendall's tau correlation coeff. after the line has been subtracted from the data (residuals).

26

26

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:

$$S_i = P_i - M_i$$

27

27

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

28

28

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

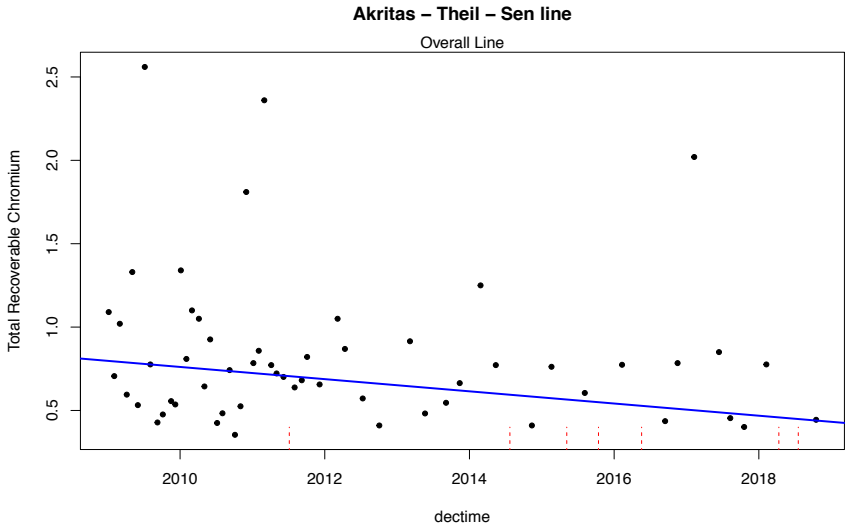
29

29

6. censeaken function

Nondetects influence the line and test.

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



30

30

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

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

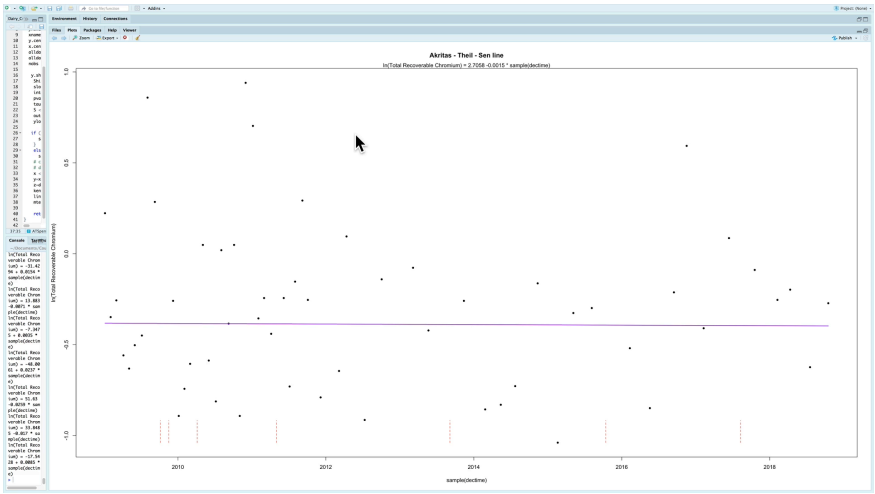
31

31

Permutations by shuffling time

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

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



Then S is computed for each shuffle.

32

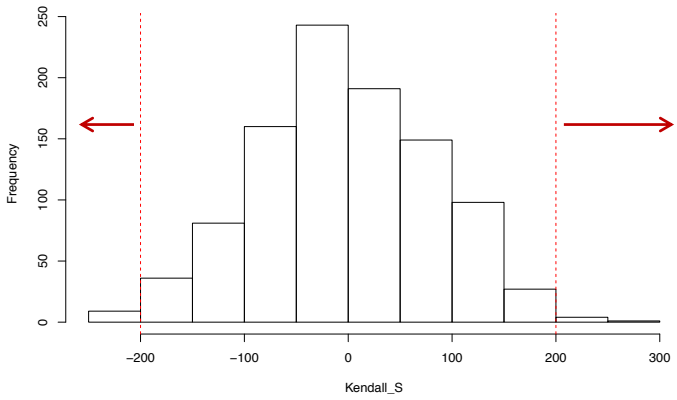
32

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.

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)

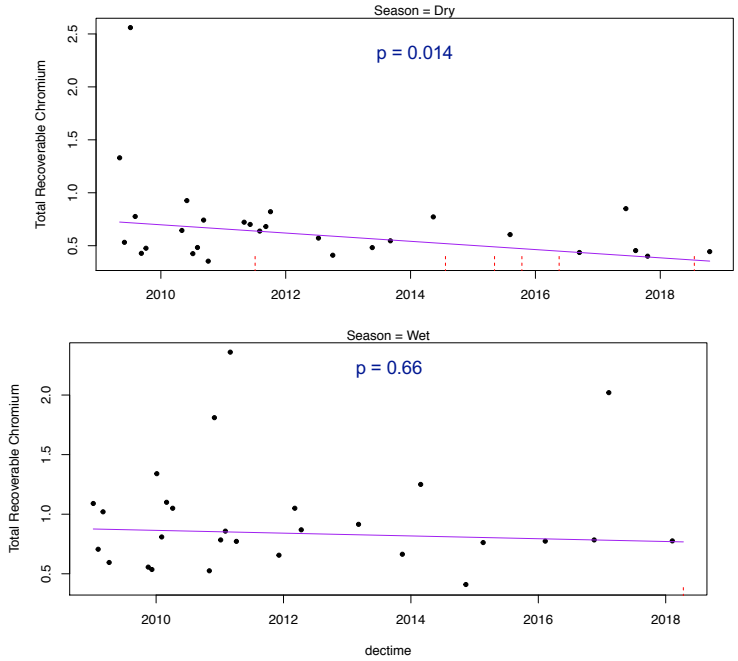


33

33

Optional graphs for each season

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



34

34