Correlation and Regression
for Data with Nondetects

Dennis R. Helsel
PracticalStats.com

Objectives of the 'Correlation Regression' webinar

1. To demonstrate how correlation coefficients and regression models can be computed without substituting a value for nondetects
2. To motivate you to use censored data methods by increasing your understanding of them, and trust in them
3. To highlight one of the many aspects of the new online course Nondetects And Data Analysis now available at http://practicalstats.teachable.com

 2

Outline: Correlation and Regression with NDs

1. What methods are available?
2. Why not substitute $\frac{1}{2}$DL or similar values and run the usual procedures?
3. How does maximum likelihood work for (parametric) correlation and regression?
4. Correlation coefficients for censored data, with an example
5. Regression for censored data, with an example
6. Multiple regression for censored data, with an example

3

Plot the Data First!

```
> attach (Golden)
> cenxyplot(Kidney, KidneyCen, Blood, BloodCen, xlab = "Pb in
Kidneys", ylab = "Pb in Blood")
```

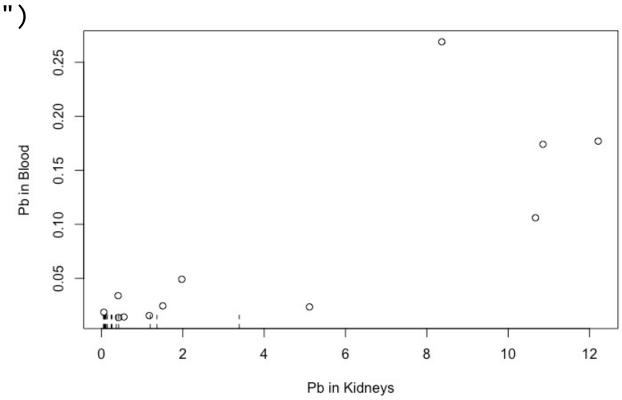
cenxyplot is in the NADA package

can plot data censored for both the X and Y variables

Is there a correlation between Pb in heron blood and kidneys?

55% of blood lead concs are censored.

What equation best describes their relationship?

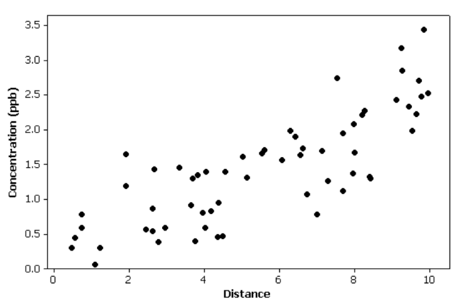


4

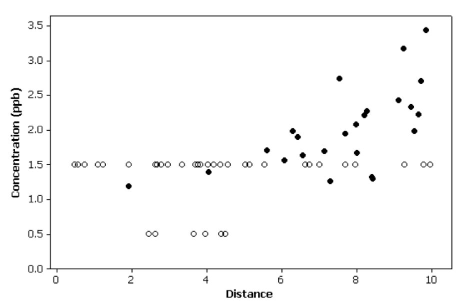
Parallels between standard methods and survival analysis methods

Standard Methods	Methods for Censored Data
Correlation	
Pearson's r	Likelihood r
Kendall's tau	Kendall's tau-b
Linear Regression	
Regression	Censored MLE regression
Theil-Sen line	Akritas-Theil-Sen line

Effects of Substitution



Original data



After censoring and substitution

Effects of substitution are corporate, not just "1/2 DL should be OK" for one observation. It produces an invasive pattern -- flat lines, artificially lower correlation and r2, lowering slope estimates, and usually lowering the standard error (all the fabricated values are identical).

Evaluation of Substitution for regression models

Thompson and Nelson (2003) found that for censored response (y) variables, substituting one-half the DL for nondetects produced

1. biased-low parameter estimates (slopes) and
2. artificially small standard error estimates (explanatory variables who shouldn't be in the regression appear to be significant)

There are better ways!

Background: MLE for Correlation and 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.

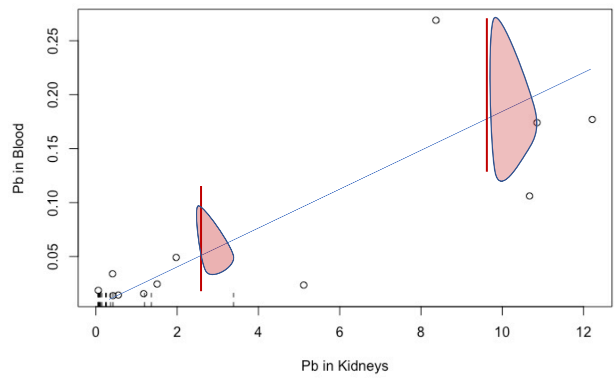
Distribution of the Residuals

For Pearson's r and regression, residuals should have a normal distribution and constant variance. This often is not the case with environmental data, with or without censoring

Colored areas on the plot are the density of residuals projecting out from the plane of the slide (pardon by lack of artistry).

Note skewness and increasing variance of residuals.

If residuals are skewed and the variance changes, consider taking the logs of the Y variable (equation then assumes a lognormal distribution of residuals)

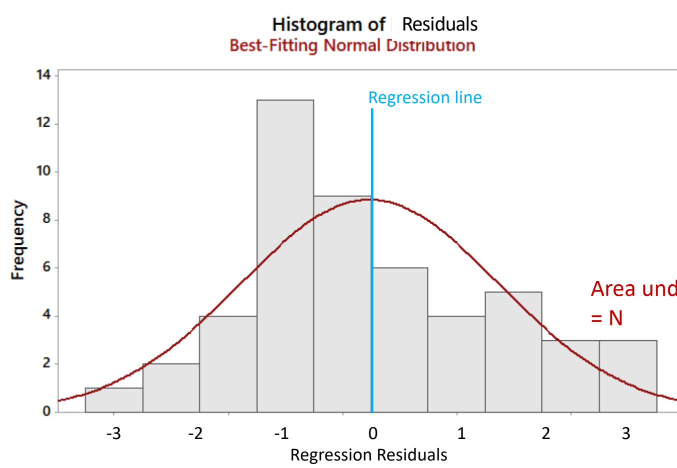


 9

Probability Density Function of Residuals -- log Y equation

The familiar "bell shaped curve" of the normal distribution

Frequency scale: Total = N. Or divide by N to get "density", the % of the observations. Total percentage (sum of area of the bars) =1



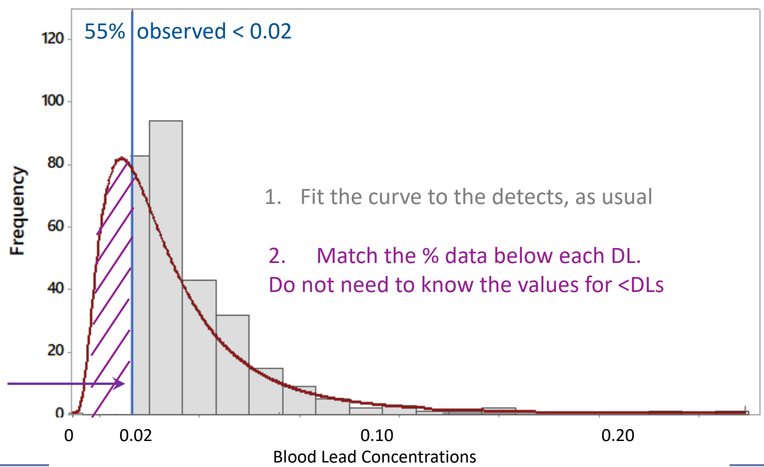
 10

Probability Density Function for Censored Data

Minimize the log-likelihood. For censored data it 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%



 11

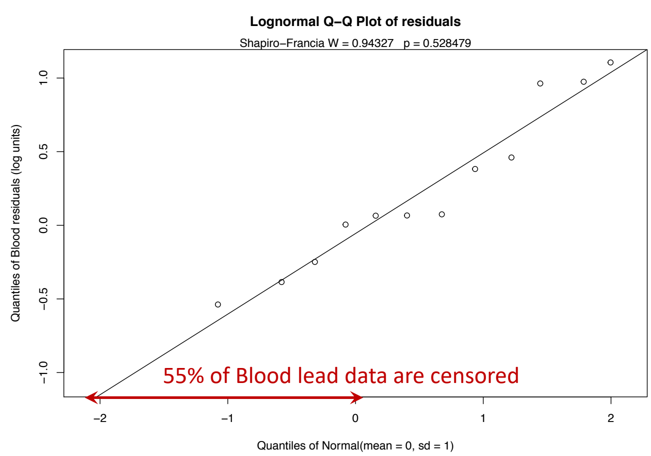
How Best to Visualize? Q-Q plot of residuals

The assumed distribution of residuals is represented on the x axis by the linear Normal Quantiles scale, where the mean = 0 and the scale is in standard deviations above and below the mean.

Multiple detection limits are no problem in this process.

Only residuals from detected obs are plotted as points, but the residuals from nondetects affect the spacing of the points on the plot.

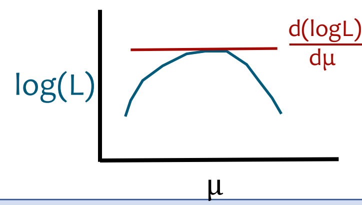
The straight line of points and Shapiro-Francia normality test (accounting for nondetects) show that residuals from a logY model support the lognormal assumption



 12

How MLE Regression Works

- Write a likelihood function L = function (slope, intercept).
  This evaluates the match between Y_i and $(b_0 + b_jX_j)$
  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



Likelihood Function L

$$L = \prod_{i=1}^{n} \left(\frac{1}{\sigma}\, p\left(\frac{Y_i - X_i\beta_j}{\sigma}\right)\right)^{1-\delta_i} \left(cdf\left(\frac{X_i\beta_j - Y_{DL}}{\sigma}\right)\right)^{\delta_i}$$

detects δ_i=0 nondetects δ_i=1

where p = the pdf (probability distribution function) for a normal distribution,

and cdf = cumulative distribution function of a normal distribution

The values for the β_j (intercept plus slopes) can be solved for by setting the derivative of log(L) = 0.

What is important to remember?

1. The intercept and slopes are iteratively solved for by maximizing log(L). That finds estimates most likely to have produced both the observed detected data, and the observed proportions of data below each detection limit.
2. The pdf and cdf are mathematical formulae specifically for a particular distribution. If you choose a different distribution you change the results.

Correlation Coefficients for Censored Data

- Parametric correlation coefficients using maximum likelihood estimation (MLE) are not necessarily on the same scale as Pearson's r.
- However, they should be used in the same context as Pearson's r – they measure linear correlation (not curved) with normal residuals.
- They are based either on the log-likelihood -- the measure of error for MLE methods -- or on the likelihood ratio test (G_0 - G_{model}), which determines whether the regression equation explains a significant amount of variation as compared to a null intercept-only model.
- There are several suggested "pseudo r^2" statistics whose square roots serve as correlation coefficients. Here are two of the most common.

1. Likelihood Correlation Coefficient

Parametric approach: MLE

The Likelihood Ratio correlation coefficient:

$$r_{LR} = \pm\sqrt{1-\exp\left(\frac{-G2}{n}\right)} = \pm\sqrt{1-\exp\left(\frac{-7.35}{9}\right)} = -0.747$$

(not for the blood and kidney Pb data)

where

- the algebraic sign of the correlation coefficient ($\pm$) is the sign of the regression slope, and
- G2 = (G_0 - G_{model}) = 2($\ln L_{model}$ - $\ln L_0$), or "the -2 log likelihood", 2 times the difference in log likelihoods between this model and one with no explanatory variables (the null model)
- Is perhaps the most-reported correlation coefficient for MLE
- However, in theory it can be greater than 1

1. Likelihood R: How to compute from MLE regression output

Parametric method: coefficients fit by MLE

Check the Q-Q plot to see if the residuals appear close to a normal distribution

- cencorreg default is to take the log(Y).
- Censoring only allowed for the Y variable

```
> cencorreg(Blood, BloodCen, Kidney)
 Likelihood R = 0.8236
 Rescaled Likelihood R = 0.8721
 McFaddens R = 0.714
Loglik(model)= -14.7
Loglik(intercept only)= -30
Chisq= 30.62 on 1 degrees of freedom, p= 3.14e-08
n= 27
```

not different from a lognormal distribution

2. Rescaled Likelihood Ratio Correlation Coefficient

Parametric approach: MLE

Rescaled likelihood ratio (or Nagelkerke) correlation coefficient:

$$r_N = \pm \sqrt{\frac{1-\exp(\frac{-G2}{n})}{1-\exp(D_0/n)}}$$

where

- the algebraic sign of the correlation coefficient (+ or -) is the sign of the regression slope
- values are between 0 and 1
- has values more similar to Pearson's r than other coefficients
- is generally my choice for censored correlation and regression

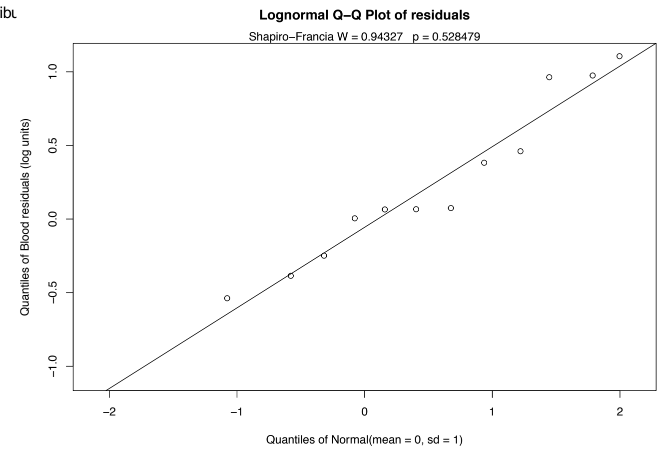
2. Rescaled likelihood ratio: compute using cencorreg

Parametric method: coefficients fit by MLE

Check the Q-Q plot to see if the residuals appear close to a normal distribu

- cencorreg default is to take the log(Y).
- Censoring only allowed for the Y variable

```
> cencorreg(Blood, BloodCen, Kidney)

 Likelihood R = 0.8236
 Rescaled Likelihood R = 0.8721
 McFaddens R = 0.714

Loglik(model)= -14.7
Loglik(intercept only)= -30
Chisq= 30.62 on 1 degrees of freedom, p= 3.14e-08
n= 27
```



 19

Nonparametric Correlation: Kendall's tau

- Does not measure only linear relationships, but all monotonic relationships
- Does not require normality of residuals
- Allows for censoring in both the X and Y variables
- Has an associated 'regression' line with it for one X variable: the Akritas-Theil-Sen line

 20

Nonparametric Correlation with censored data

Nonparametric approach: Kendall's tau

Kendall's tau -a $$\tau = \frac{N_c - N_d}{\frac{N(N-1)}{2}}$$

where N_c = # concordant pairs (+) [x going same direction as y]

and N_d = # discordant pairs (-) [x going opposite direction than y]

N = number of (x,y) pairs

With Kendall's tau, ties count as evidence for the null hypothesis. So many <1 vs <1 or <3 vs 1 or <1 vs <3, all of which are ties, will result in a high p-value.

Kendall's tau with censored data

Computing tau:

With data ordered by increasing x, does y consistently increase (+) or decrease (–) ?

For some example data:

X	Y	result
1980	20	– – – –
1981	<10	0 0 0
1982	7	– –
1983	3	–
1984	< 3	

Kendall's tau with censored data

The ATS command is similar to the cenken command in NADA for R. The ATS plot will take logs of Y by default, but a power transformation doesn't change the Kendall's tau correlation. Same Kendall's tau for both Y and log(Y)

```
> ATS(Blood, BloodCen, Kidney, KidneyCen)
Akritas-Theil-Sen line for censored data


ln(Blood) = -4.5128 + 0.295 * Kidney
Kendall's tau-b = 0.4217        p-value = 0.00043
```

Without censoring, tau is on a scale of about 0.2 lower than Pearson's r. Due to increased ties with censoring, tau often is even lower than the MLE correlation coefficients. Compare tau with tau for other models/data

Summary: Correlation with censored data

Parametric approach. Likelihood Ratio r or Rescaled Likelihood Ratio r.

- Make sure residuals follow the assumed (normal, lognormal) distribution
- X-Y relationship must be linear.
- Only the Y variable may be censored. (there is now an R package that can compute the likelihood correlation coef. when both X and Y are censored. No regression yet, though).

Nonparametric approach. Kendall's tau.

- Data may be straight or curved for tau, whose value and p-value will not change when using a power transformation (log, cube root, etc.).
- Both Y and X may be censored.

- Don't use substitution.

Regression with censored data

Regression by Maximum Likelihood Estimation – a Parametric method: the cencorreg command. Only the Y variable can be censored.

```
> Pbreg <- cencorreg(Blood, BloodCen, Kidney)
 Likelihood R = 0.8236
 Rescaled Likelihood R = 0.8721
 McFaddens R = 0.714
> summary(Pbreg)
Call:
survreg(formula = "log(Blood)", data = "Kidney", dist = "gaussian")
              Value Std. Error      z       p
(Intercept) -4.4573     0.1733 -25.72 < 2e-16
Kidney       0.2436     0.0302   8.07 7.1e-16
Log(scale)  -0.6737     0.2036  -3.31 0.00094

Loglik(model)= -14.7   Loglik(intercept only)= -30
Chisq= 30.62 on 1 degrees of freedom, p= 3.1e-08
```

ln(blood Pb) = -4.457 + 0.244*kidney Pb or blood Pb = $e^{-4.457} \bullet kidneyPb^{0.244}$

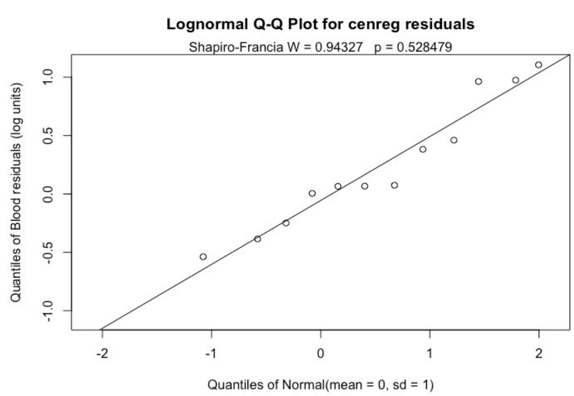
 25

Check Assumptions with QQ Plot

MLE regression is a parametric method:

Check the assumption of a normal distribution with a Q-Q plot of residuals.

- Here using the default of log(Y) fits well.
- It often does for data with NDs (because they are close to zero)



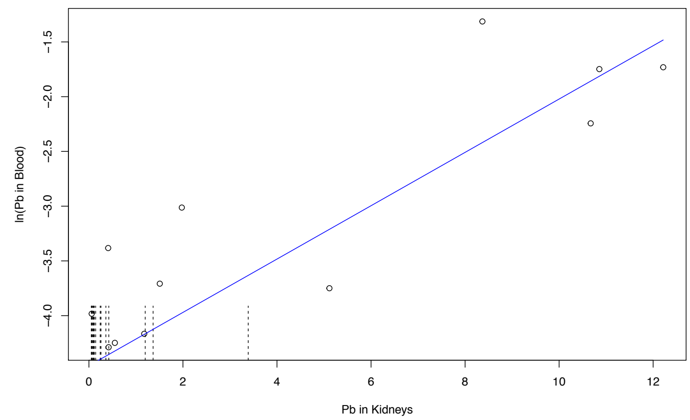
 26

Plotting the regression line

Plot of the logY regression in the units the regression was run:

Blue line is the MLE regression line

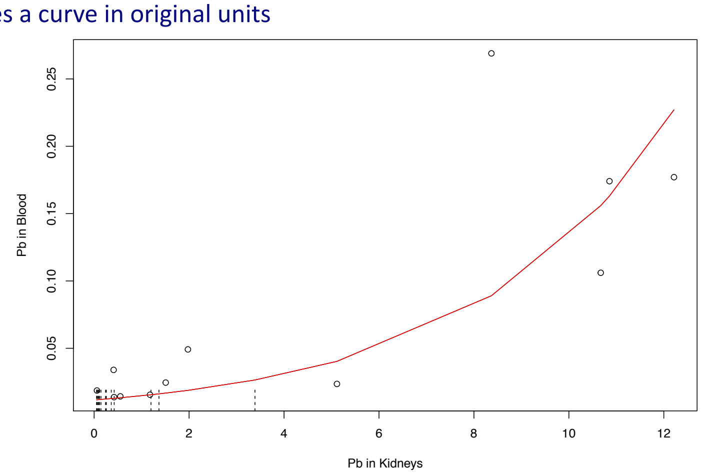
Dashed lines are the censored Y variable data



 27

Plotting the regression line

Regression straight line in log units becomes a curve in original units



 28

Nonparametric Regression with Censored Data

Nonparametric method: the Akritas-Theil-Sen line.

Use the cenken command in the NADA package, or my ATS command (script): ATS (Y, Ycen, X, Xcen)

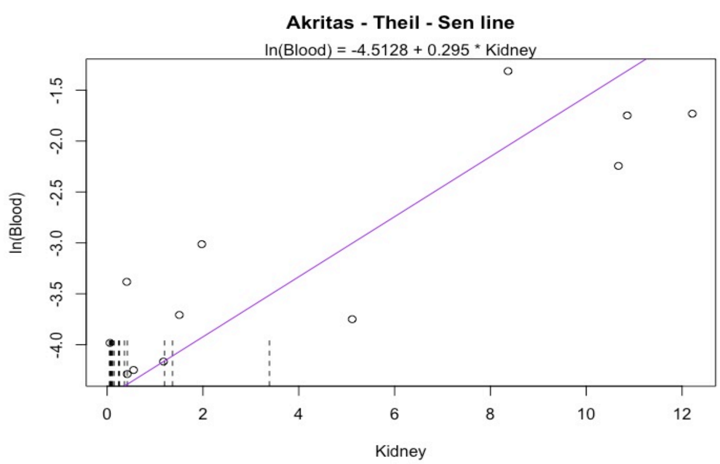
```
> Pbk<- ATS(Blood, BloodCen, Kidney, KidneyCen)

Akritas-Theil-Sen line for censored data
ln(Blood) = -4.5128 + 0.295 * Kidney
Kendall's tau = 0.4217   p-value = 0.00043
```

 29

The ATS line

Kendall's tau and its test for significance are identical whether or not the log or original scales were used.

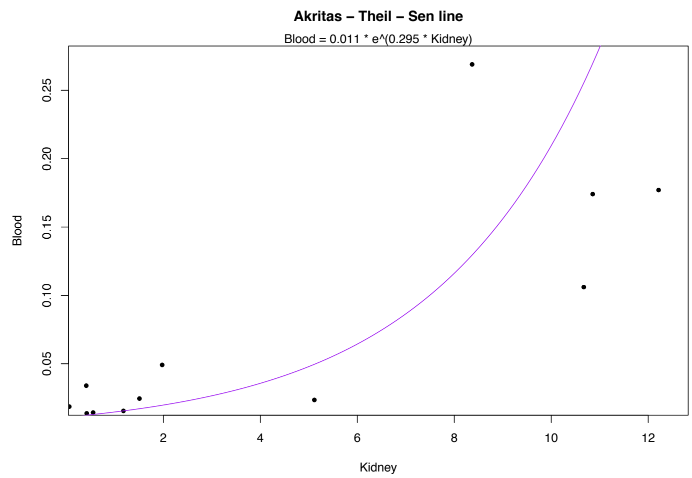
Log(Y) was used because the data were judged to be more straight than on the original scale.



 30

The ATS line

A straight line regression using logY is curved when transformed back to the original scale:



How does ATS get its slope?

- Without censoring, the (Sen) slope of the Theil-Sen line is the slope that produces a tau = 0 when subtracted from the data
- This "inverse solution" is used by ATS. Initial estimates of slope and intercept are computed, and the Kendall's tau of residuals is computed. Slope and intercept are iteratively adjusted until the residuals have a zero slope.

Summary: Regression with censored data

Parametric approach. Fit coefficients using MLE.

- Only the Y variable may be censored. (cannot yet compute MLE regression slopes and intercept when both X and Y are censored).
- Make sure residuals follow a normal distribution
- The relationship must be linear (in order to summarize it using a single slope).
- Estimates a linear mean.

Nonparametric approach. Fit using Akritas-Theil-Sen line

Estimates a linear median.

- The pattern of observed data should be linear to fit a straight line that can be summarized by a single slope.

Both Y and X may be censored.

 33

33

Multiple Regression with Censored Data

Load the TCEReg.rda dataset. The Y variable is TCE Concentration. There are 4 detection limits, indicated by the TCECen variable.

```
> attach(TCEReg)
> head(TCEReg)
  TCECen TCEConc LandUse PopDensity PctIndLU Depth PopAbv1
1   TRUE       1       9          9       10   103       1
2   TRUE       1       8          3        4   142       1
3   TRUE       1       8          3        4   209       1
4   TRUE       1       5          1        3   140       1
5   TRUE       1       5          2        1   218       1
6   TRUE       1       9         13        5    98       1
```

There are 4 possible explanatory variables: LandUse category (not very precise), Population Density, Percent Industrial Landuse, and Depth to the water table.

Which combination of these 4 explanatory variables best predicts TCE Concentration?

 34

First, Check for Multicollinearity

- Multicollinearity is the biggest problem in multiple regression
- Cause -- Redundant variables. More than one X variable is explaining same effect. X variables are correlated (not always pairwise) with one another
- Symptoms:
  1. Slope coefficients with signs that make no sense.
  2. Two variables describing same effect with opposite signs.
  3. p-values are inflated so variables that should be in the model are tossed out.

 35

Measure Multicollinearity with VIFs

Variance Inflation Factor (VIF)

- Measures the correlation (not just pairwise) among the j >1 X variables
- Has nothing to do with the response variable Y, so censoring of the Y data aren't an issue
- One VIF is computed for each X variable
- Want all VIFs < 10

 36

How is the VIF computed?

Variance Inflation Factor (VIF)

$$VIF_j = \frac{1}{1 - r_j^2}$$

where r^2_j is the r^2 between X_j and all the other X variables. So for X_1:
$X_1 = b_0 + b_2 \cdot X_2 + b_3 X_3 + b_4 * X_4$ with an r-squared = r_1^2

Want all VIFs < 10 = $r_j^2 < 0.9$

Compute with the command vif(Lreg) from the car package, where Lreg is a linear model created using the lm command. Can do this in one line.

For the TCE data's 4 variables:

```
> vif(lm(TCEConc ~ LandUse + PopDensity + PctIndLU + Depth))
   LandUse PopDensity   PctIndLU      Depth
  1.337049   1.221461   1.040476   1.184310
```

 37

Four variable model: AIC = 395.8

```
> summary(reg4)
Call:
survreg(formula = "log(TCEConc)", data = "LandUse+PopDensity+PctIndLU+Depth",
    dist = "gaussian")
               Value Std. Error     z      p       Loglik(model)= -191.4
(Intercept) -5.38940    2.61512 -2.06  0.039       Loglik(intercept only)= -205.5
LandUse      0.32205    0.31035  1.04  0.299       Chisq=28.08 on 4 degrees of freedom, p=1.2e-05
PopDensity   0.21991    0.07829  2.81  0.005                Number of Newton-Raphson Iterations: 4
PctIndLU     0.03644    0.05274  0.69  0.490                n= 247
Depth       -0.00374    0.00238 -1.57  0.117
Log(scale)   1.02763    0.11058  9.29 <2e-16
Scale= 2.79
```

 38

Three variable model AIC = 394.3

```
> reg3 <- cencorreg(TCEConc, TCECen, xvar3)
 Likelihood R2 = 0.1057                    AIC = 394.3252
 Rescaled Likelihood R2 = 0.1305           BIC = 410.8924
 McFaddens R2 = 0.06718

> summary(reg3)
               Value Std. Error     z      p
(Intercept) -5.44065    2.62890 -2.07 0.0385
LandUse      0.33855    0.31107  1.09 0.2764
PopDensity   0.22621    0.07797  2.90 0.0037
Depth       -0.00367    0.00239 -1.54 0.1239
Log(scale)   1.02852    0.11059  9.30 <2e-16
Scale= 2.8

Loglik(model)= -191.7   Loglik(intercept only)= -205.5
           Chisq= 27.61 on 3 degrees of freedom, p= 4.4e-06
```

- Is better than the 4 variable model due to lower AIC. Adding the additional PctIndLU variable does not explain much variation, and costs one degree of freedom.
- LandUse has a relatively high p-value. What about a 2-variable model?

 39

Two variable model AIC = 393.6

```
> reg2 <- cencorreg(TCEConc, TCECen, xvar2)
 Likelihood R2 = 0.1012                    AIC = 393.5758
 Rescaled Likelihood R2 = 0.1249           BIC = 406.6296

> summary(reg2)
Value Std. Error     z       p
(Intercept) -2.79067    0.81018 -3.44 0.00057
PopDensity   0.25959    0.07405  3.51 0.00046
Depth       -0.00434    0.00234 -1.85 0.06367
Log(scale)   1.03487    0.11068  9.35 < 2e-16

Scale= 2.81
Gaussian distribution
Loglik(model)= -192.3   Loglik(intercept only)= -205.5
           Chisq= 26.35 on 2 degrees of freedom, p= 1.9e-06
```

- This is better than the 3 variable model due to lower AIC
- Depth is now at p=0.06
- I generally keep variables with p < 0.10, as model selection stats like AIC and BIC generally underfit (too few explanatory variables)
- Just as in ordinary regression, R^2 increases with each added variable, so is no help in choosing a model. Rescaled R^2 here is 0.125, while with the 3-variable model it was 0.130. This does NOT mean the 3-variable model is better.
- What about a 1-variable model, with just PopDensity?

 40

One variable model AIC = 395.7

```
> reg1 <- cencorreg(TCEConc, TCECen, PopDensity)
 Likelihood R = 0.2934                    AIC = 395.6935
 Rescaled Likelihood R = 0.3259           BIC = 405.2338
 McFaddens R = 0.2326

> summary(reg1)
survreg(formula = "log(TCEConc)", data = "PopDensity", dist =
"gaussian")
             Value Std. Error     z       p
(Intercept) -3.7343     0.7493 -4.98 6.2e-07
PopDensity   0.3087     0.0736  4.20 2.7e-05
Log(scale)   1.0418     0.1109  9.39 < 2e-16
Scale= 2.83

Loglik(model)= -194.3   Loglik(intercept only)= -205.5
            Chisq= 22.24 on 1 degrees of freedom, p= 2.4e-06
```

- AIC is higher for the 1-varaible model. So AIC picks the 2-variable model. BIC is lowest for the 1-variable model and is known to underfit.
- AIC here is no better than the 4 variable model.
- Summary: I'd choose the 2-var model:
- AIC is better; the p-value for Depth is 0.06. But also should examine if a decrease of 0.04 ug/L per 10 feet of depth in the 2-variable model is scientifically meaningful or not. Seems reasonable to me.

Conclusions: Correlation and Regression for Censored Data

- Correlation coefficients and linear regression models can be computed for data with nondetects without substituting $\frac{1}{2}$DL or similar values for nondetects
- The methods are available in the 'survival analysis' sections of statistics software. Some of this but not all is available in the 'NADA' R package.
- Maximum likelihood methods for parametric correlation and regression require an assumption of a distribution
- Nonparametric methods based on Kendall's tau are also available
- All of this it shown in more detail, along with scripts to make it easy to use, in our online course Nondetects And Data Analysis (NADA). See http://practicalstats.teachable.com

Our Next Webinar
Tuesday Sept 24th 11 am Mountain time

- A rebroadcast of "The Mystery of Nondetects: How Censored Data Methods Work". Invite people who don't know how these methods work, but need to.
- Sign up for our newsletter/announcement list to get the registration link emailed to you. You can subscribe at http://practicalstats.com/news/
- Or check our webinars page periodically at http://practicalstats.com/training/webinar.html to register for it.

This 'Correlation and Regression' webinar will be available Thursday for streaming

- at our Online Training Site

http://practicalstats.teachable.com/

and click the "Show more courses" button to see the free webinars.

Let colleagues who missed it know about it.

Thank you for attending

Type and Send Your Questions

- Much of the material is based on the book Statistics for Censored Environmental Data Using Minitab and R, 2nd Edition by Dennis Helsel (2012).
- All of the material, plus much more, is now available 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.

Get in touch!

Dennis Helsel ask@practicalstats.com

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