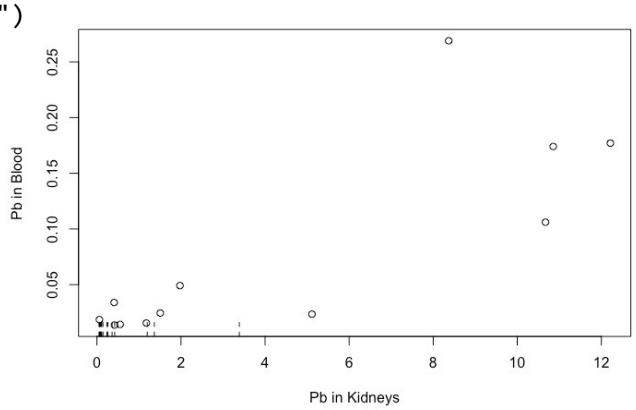
Nondetects And Data Analysis:
Correlation & Regression with NDs

Dennis R. Helsel, Ph.D
Practical Stats

1

Plot the Data First!

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



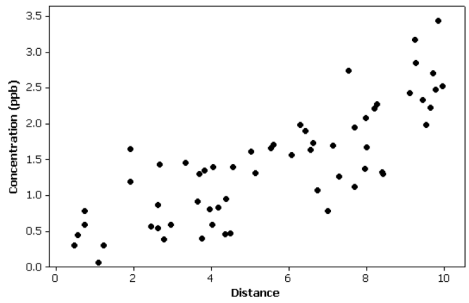
Is there a correlation between Pb in Blood and Kidneys?

What equation best describes the relation?

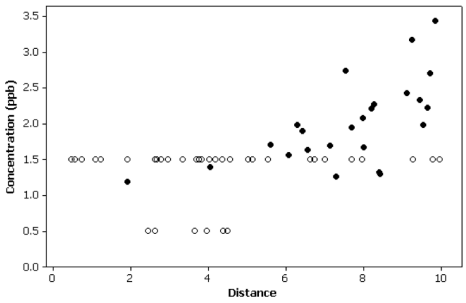
2

2

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).

3

3

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 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!

4

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
Linear Regression	
Regression	Censored MLE regression
Theil-Sen line	Akritas-Theil-Sen line

5

5

Background: MLE for Correlation and Regression

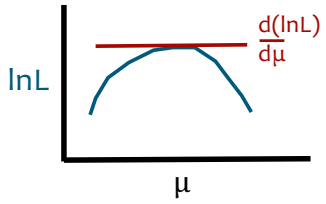
- Starts with the observed data
- Given the observed data, what values for parameters (mean, standard deviation, correlation coefficient) are most likely to have given rise to these data?
- For censored data, we match the observed data by matching 2 types of information: the values for detected observations and the observed proportions of data below each detection limit.

6

6

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 L where L is negative.
- Do this by maximizing lnL
- Do this by setting the derivative of lnL = 0, solve for slope and intercept



7

7

How MLE Regression Works

The likelihood function for estimating the regression coefficients for a censored Y variable is a function of each observed value Y_i, and an indicator δ_i of whether this value is a DL or detected value.

$$L = \prod_{i=1}^{n} [p(Y_i)^{1-\delta i}] [cdf(Y_i)]^{\delta i}$$

detects $\delta_i=0$

nondetects $\delta_i=1$

8

8

Likelihood Function L

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

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 lnL = 0.

What is important to remember?

1. The intercept and slopes are iteratively solved for by maximizing 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. As a prophet once said to Indiana Jones, "Choose wisely".

9

Using the Likelihood Function L

- lnL is reported by some software. It is a negative number.
- Most software reports -2lnL, called the residual deviance G, a measure of error of the regression model. This is a positive number. The lower the G the less error for a given set of data.
- Because lnL and -2lnL are sums of values for each observation (L was a product of each observation's effect), there is no scale for a 'good' or 'bad' -2lnL = G. It's value depends on how many observations there are -- how many small errors are added together.
- For the same data set, the model with the lower G has less error. The G for differing models larger: ($y = b_0 + X_1 + X_2$) vs smaller: ($y = b_0 + X_1$) on the same data are compared to determine which is best.
- The decision of which model is best is whether the decrease in error when adding additional variables ($G_{smaller} - G_{larger}$) is large enough to offset the decrease in degrees of freedom (the additional number of variables in the larger model).
- An important comparison is when a regression model is compared to the smaller model that has zero X variables, called the 'null' or 'intercept only' model, which has error G_0.

10

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 I present three of the most common.

11

11

1. Likelihood Correlation Coefficient

Parametric approach: MLE

The Likelihood Ratio correlation coefficient:

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

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

12

12

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).

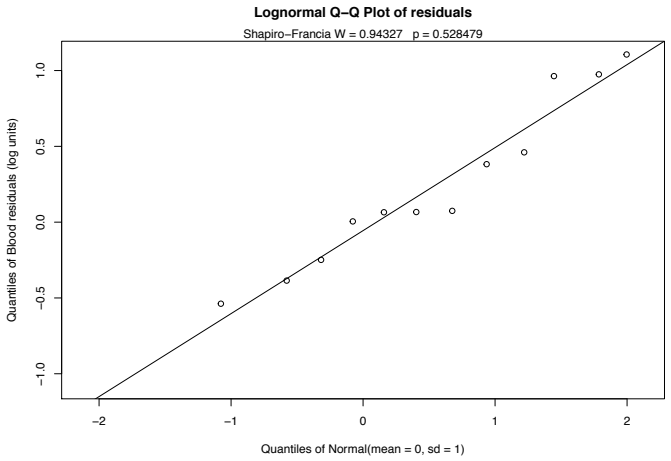
Censoring only allowed for the Y variable

```
> 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



13

13

Do Not Use the Correlation Coefficient to Decide Which Model is Better

Parametric method: coefficients fit by MLE

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

Use the cencorreg option LOG=FALSE to use Y instead of log(Y).

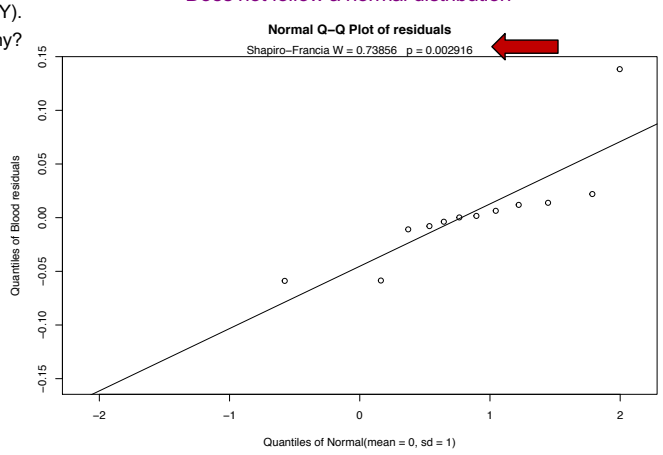
Here, the Likelihood R is higher but the model is far worse. Why?

```
> cencorreg(Blood, BloodCen, Kidney, LOG=FALSE)

 Likelihood R = 0.8525                AIC = 17.48548
 Rescaled Likelihood R = 0.9445       BIC = 20.48209
 McFaddens R = 0.8773
```

Just as with ordinary regression, you cannot compare the correlation coefficient, AIC, BIC, or r2 between models to decide which model is better after taking a transformation of the Y variable. Y and logY, and so also these statistics, are on different scales. Instead, use the Shapiro-Francia test of normality to decide which distribution best fits the residuals.

Does not follow a normal distribution



14

14

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 of the 3

15

15

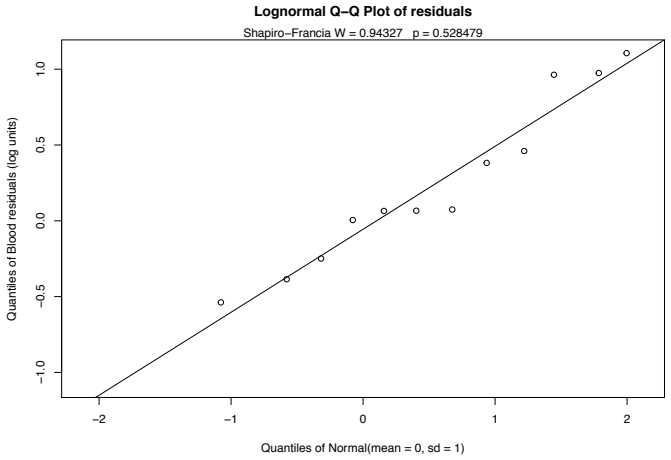
2. Rescaled likelihood ratio: compute using cencorreg

Parametric method: coefficients fit by MLE
Check the Q-Q plot to see if the residuals appear close to a normal 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
```



16

16

3. McFadden's Correlation Coefficient

Parametric approach: MLE

McFadden's (or 'deviance') correlation coefficient:

$$r_{McF} = \pm\sqrt{1 - \frac{lnL_R}{lnL_0}}$$

where

- the algebraic sign of the correlation coefficient (+ or -) is the sign of the regression slope, and
- lnL_R is the log likelihood of the regression model and lnL_0 is the log-likelihood of the null model (one with no explanatory variables)
- simple in concept -- a decrease in error produces a higher McFadden's r
- close in concept to Pearson's r, but on a scale generally below it and other MLE correlation coefficients

17

17

3. McFadden's r : compute using cencorreg

Parametric method: coefficients fit by MLE

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

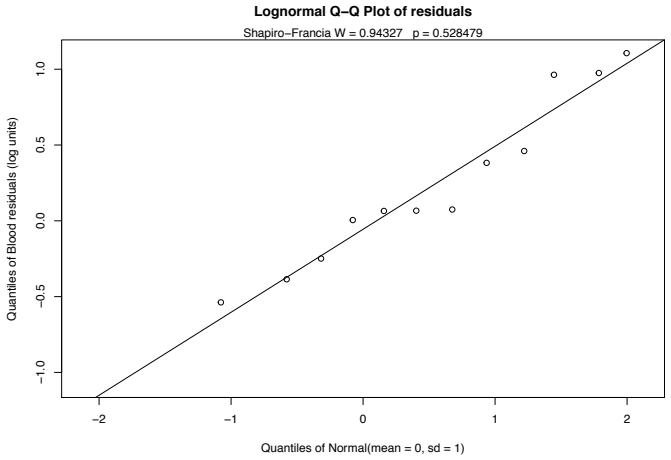
cencorreg default is to take the log(Y).

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
```



18

18

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

19

19

Nonparametric Correlation with censored data

Nonparametric approach: Kendall's tau

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

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.

20

20

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	

21

21

Kendall's tau with censored data

Use the ATS script. 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 = 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

22

22

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                  Loglik(model)= -14.7   Loglik(intercept only)= -30
Kidney       0.2436     0.0302   8.07 7.1e-16                  Chisq= 30.62 on 1 degrees of freedom, p= 3.1e-08
Log(scale)  -0.6737     0.2036  -3.31 0.00094
```

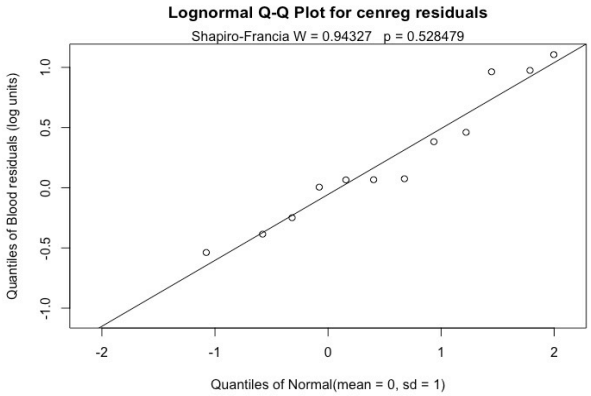
ln(blood Pb) = -4.457 + 0.244*kidney Pb or blood Pb = $e^{-4.457} \bullet kidneyPb^{0.244}$

23

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)



24

Plotting the regression line

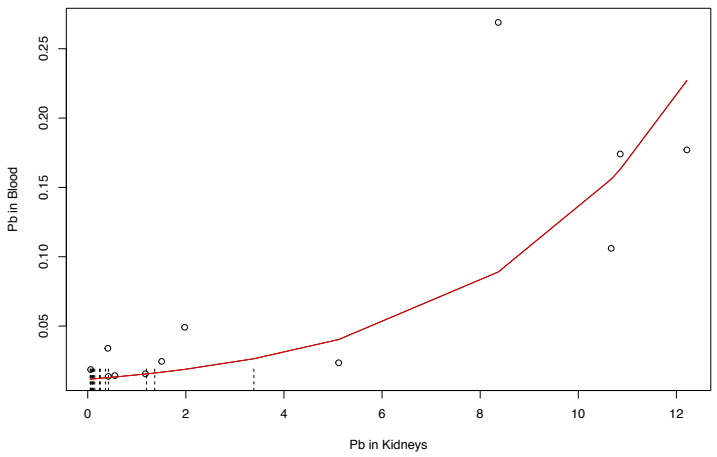
Regression straight line in log units becomes a curve in original units

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

> ik <- order(Kidney)

> lines(Kidney[ik],
    exp(predict(Pbreg)[ik]),
    col = "red")
```

The order function computes the rank of obs in the vector. Use that ordering to 'sort' variables in the following commands



25

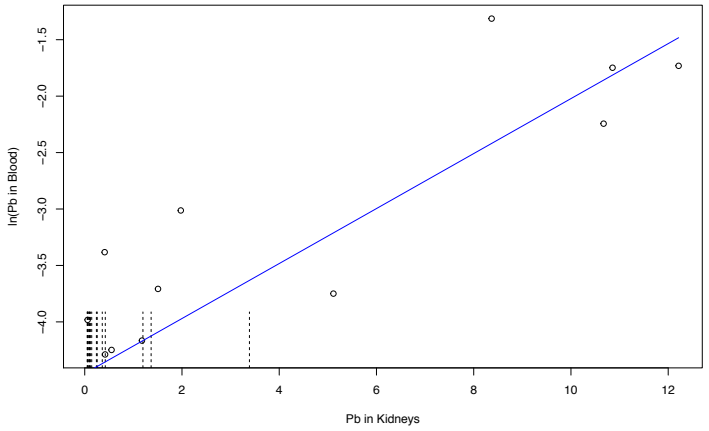
25

Plotting the regression line

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

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

> lines(Kidney[ik],
    predict(Pbreg)[ik],
    col="blue”)
```



26

26

Nonparametric Regression with Censored Data

Nonparametric method: the Akritas-Theil-Sen line.
The 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
```

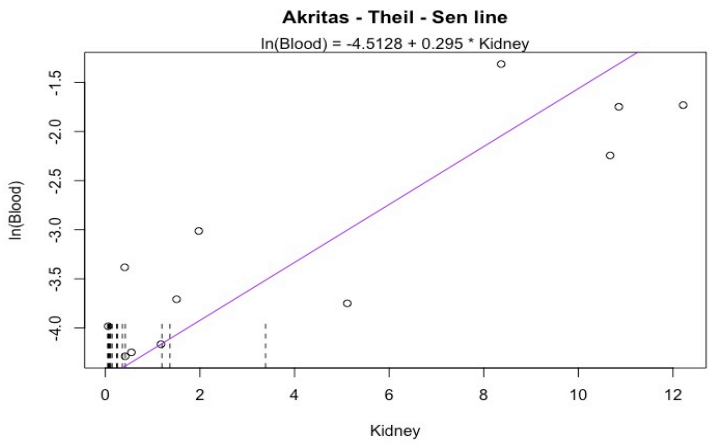
27

27

The ATS line

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

(takes the log of the Y variable to obtain a straight line as its default)



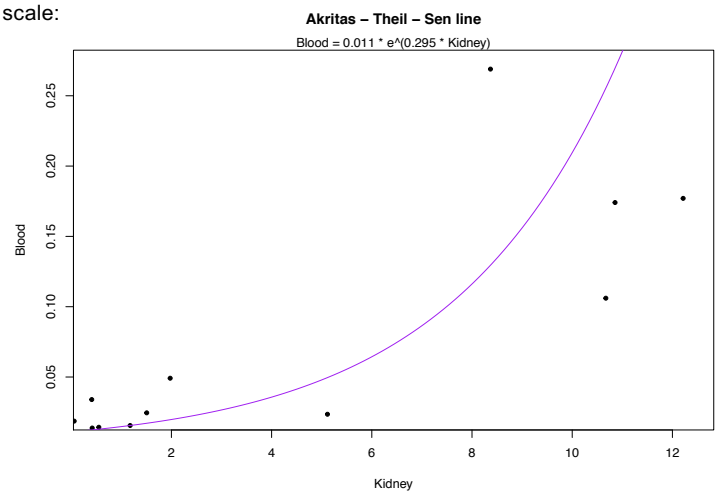
28

28

The ATS line

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

The logY model is curved in the original scale:



29

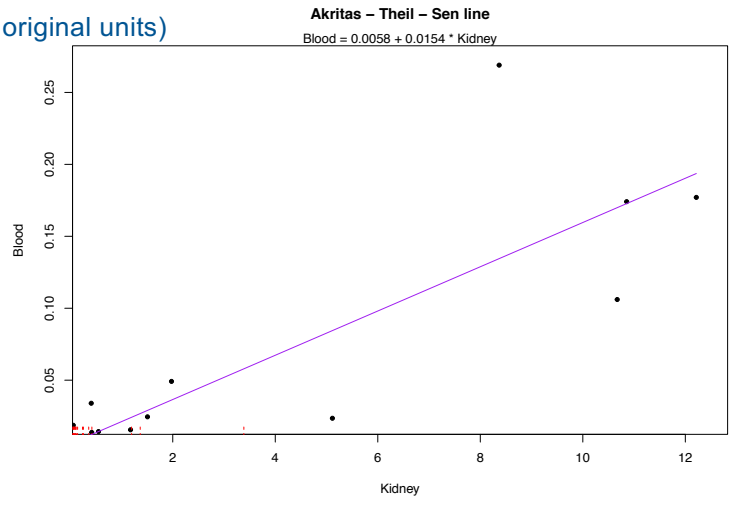
29

To Transform Y or Not?

```
> ATS(Blood, BloodCen, Kidney, KidneyCen, LOG = FALSE)
```

(LOG=F gives straight line in original units)

Choose to transform or not based on the scale that the X-Y relationship is more linear. No normality test or assumption required.



30

30

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 computed. Slope and intercept are iteratively adjusted until the residuals have a zero slope.

31

31

Conclusions: 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. cencorreg

Nonparametric approach. Kendall's tau. ATS

- 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.

32

32

Conclusions: Regression with censored data

Parametric approach. Fit coefficients using MLE.

- Only the Y variable may be 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. cencorreg

Nonparametric approach. Fit line using Akritas-Theil-Sen ATS

- 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

Methods for censored data

Method	Parametric	Nonparametric
Descriptive stats	MLE ROS	Kaplan-Meier
Intervals	Bootstrapping MLE ROS	Bootstrapping K-M
Paired Data	CI on differences by MLE	PPW
2 Indep Groups	MLE Regression on 0/1 variable	Generalized Wilcoxon
3+ Indep Groups	MLE Regression on 0/1 variable	Generalized Wilcoxon
Correlation	Likelihood R by MLE	Kendall's tau
Regression	MLE Regression	ATS line

34

34