Nondetects And Data Analysis:
Multiple Regression with NDs

Dennis R. Helsel, Ph.D
Practical Stats

1

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
```

Explanatory (X) variables

- There are 4 possible explanatory variables: LandUse category (not the best choice to make this a categorical variable), Population Density, Percent Industrial Landuse, and Depth to the water table.
- Which combination of these 4 explanatory variables best predicts TCE Concentration?

2

2

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.

3

3

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

4

4

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:

$X_1 = b_0 + b_2 \bullet X_2 + b_3 X_3 + b_4 * X_4$ 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
```

5

Input format for Censored Regression

To use the cencorreg function to do multiple regression you'll need to input the x variables as a single data frame.
To create the data frame for all 4 X variables:

```
> xvar4 <- data.frame(LandUse, PopDensity, PctIndLU, Depth)
```

Then save the regression to an object name and print the results using summary().

```
> reg4 <- cencorreg(TCEConc, TCECen, xvar4)
 Likelihood R2 = 0.1075                    AIC = 395.8513
 Rescaled Likelihood R2 = 0.1326           BIC = 415.9318
 McFaddens R2 = 0.06833

> summary(reg4)
              Value Std. Error     z      p
(Intercept) -5.38940    2.61512 -2.06  0.039
LandUse      0.32205    0.31035  1.04  0.299
PopDensity   0.21991    0.07829  2.81  0.005
PctIndLU     0.03644    0.05274  0.69  0.490
Depth       -0.00374    0.00238 -1.57  0.117
Log(scale)   1.02763    0.11058  9.29 <2e-16
```

Default Y scale in cencorreg is log(Y). Check this with the QQ plot that is produced. (see next slide)

6

6

Step 1. Choose the best units of Y

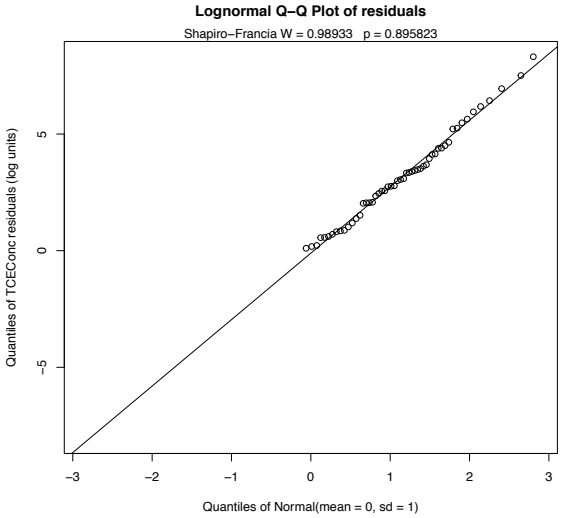
(Do regression residuals follow a normal distribution?)

If you've taken our Applied Environmental Statistics course you'll remember that with regression, it is not the distributional shape of Y or the Xs that matter. Its all about the residuals.

Check whether the residuals from cencorreg (NOT the Y variable itself) follow a normal distribution using the Shapiro Francia W statistic and test. If residuals appear linear on the QQ plot and W is close to 1, you've got a good scale to use for the Y variable. Remember that the default scale with cencorreg is log(Y).

Here W is very close to 1 so residuals from the logY regression look similar to a normal distribution, and we should use the chosen logY units for everything that follows.

(Run cencorreg with the LOG=FALSE option and you'll see that for the original Y scale the Shapiro-Wilk W is lower, and so less like a normal distribution.)



7

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
```

Only PopDensity is significant. Does that mean to drop the other 3 variables and you're done?
NO! There's much more work to do (Steps 2 and 3) to build a good multiple regression equation.

8

Step 2: Transform X variables?

Use partial plots to view the relationship between Y and one X variable. The goal: a linear relationship between each X and Y.

Use the `partplots` function. A simple plot of Y vs one X will NOT tell you whether to transform due to nonlinearity. That plot includes effects from all the other X variables, hiding behind the one X variable on the plot. Partial plots look specifically at the relationship between Y and one X in the context of multiple regression.

9

9

How Are Partial Plots Computed?

The partial plot of Y vs j=4 (Density) is:

$Y_{partial}$ = residuals ⟵ from regression of

Y ~ LandUse + PopDensity + PctIndLU (the other 3 X variables)

or $Y|X_{(j)}$

$X_{partial}$ = residuals ⟵ from regression of

Density ~ LandUse + PopDensity + PctIndLU

or $X_j|X_{(j)}$

$X_{(j)}$ means "the set of X variables excluding j"

10

10

Partial Plots for X_j

Sometimes called “crPlots” or “adjusted-variable plots”

Plots of Y vs. $X_{j = 1, \ldots, k}$ do not show the relationship (curved vs linear) between Y and X_j in the MLR, as they don’t consider effects of the other $X_{(j)}$s

Partial plots graph the partial residual for $Y|X_{(j)}$ against the partial residual of $X_j|X_{(j)}$ for each variable. Censored regression is used for $Y|X_{(j)}$

The slope of the line on the plot is the same as the slope in the regression output for that X_j. This is not true for a simple plot of Y vs X_j

If the relationship between Y and X_j is not linear, the partial residuals will not follow a linear pattern when plotted versus X_j

11

The partplots command

Choose the variable most needing transforming, transform and run partplots again. A sequential process, not all at once.

```
> partplots(TCEConc, TCECen, xvar4)
LandUse
 untransformed
 Likelihood R2 = 0.1075            AIC = 395.8513
cube root
 Likelihood R2 = 0.1074            AIC = 395.8607
log transform
 Likelihood R2 = 0.1074            AIC = 395.8689
 Decrease in AIC from transformation of
PopDensity = 0

PopDensity
 untransformed
 Likelihood R2 = 0.1075            AIC = 395.8513
cube root
 Likelihood R2 = 0.1255            AIC = 390.8004
log transform
 Likelihood R2 = 0.1332            AIC = 388.6177
Decrease in AIC from transformation of PopDensity
= 7.233565
```

largest drop in AIC due to transformation

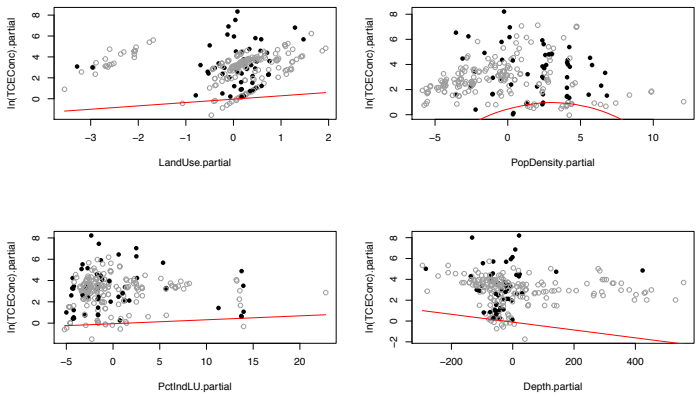
```
PctIndLU
 untransformed
 Likelihood R2 = 0.1075      AIC = 395.8513
cube root
 Likelihood R2 = 0.1108      AIC = 394.914
log transform
Cannot take logs of zero or negative values.
Decrease in AIC from transformation of PctIndLU =
0.9372712

Depth
 untransformed
 Likelihood R2 = 0.1075      AIC = 395.8513
cube root
 Likelihood R2 = 0.1033      AIC = 396.9991
log transform
 Likelihood R2 = 0.1006      AIC = 397.7427
Decrease in AIC from transformation of Depth = 0
```

12

Partial plots with gam smooth

Gray circles are residuals for nondetects using the DL value



Outliers less important than a curve.

Lines are low because a high % of data are nondetects.

Only curve is for PopDensity.

Take logs of PopDensity to lower AIC.

13

13

Four variable model with lnPopDen: AIC = 388.6

```
> xvar4b <- data.frame(LandUse, lnPopDen, PctIndLU, Depth)
> x4b.reg <- cencorreg(TCEConc, TCECen, xvar4b)
 Likelihood R2 = 0.1332                    AIC = 388.6177
 Rescaled Likelihood R2 = 0.1644           BIC = 408.6983
 McFaddens R2 = 0.08593

> summary(x4b.reg)
               Value Std. Error     z       p
(Intercept) -5.85579    2.63740 -2.22 0.02640
LandUse      0.23787    0.30875  0.77 0.44105
lnPopDen     1.59733    0.46190  3.46 0.00054
PctIndLU     0.02515    0.05229  0.48 0.63045
Depth       -0.00379    0.00233 -1.63 0.10403
Log(scale)   1.01132    0.11009  9.19 < 2e-16
```

A better regression than with original units for PopDen.

Keep variables in these units during the next step.

None of the other three variables is helped by transformation.

14

14

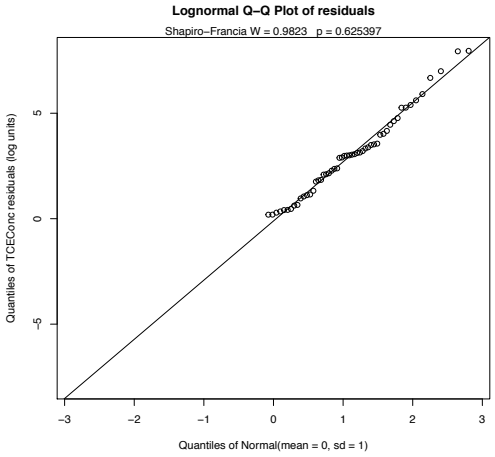
Step 3: Find the best set of X variables. Goal is a lower AIC

3a. Can do a stepwise deletion of variables, dropping 1 at a time:

```
> xvar3 <- data.frame(LandUse, lnPopDen, Depth)
> reg3 <- cencorreg(TCEConc, TCECen, xvar3)
 Likelihood R2 = 0.1324          AIC = 386.8478
 Rescaled Likelihood R2 = 0.1634
                                 BIC = 403.415
```

AIC has gone from 388.6 down to 386.8, so a better model without PctIndLU.

Don't drop 2 or more Xs at a time. One by one! Each X variable interacts with the others, so see the effect of dropping 1 first, it will change the pvalues of the others. Otherwise you may drop a significant and important variable by mistake!



15

15

Three variable model AIC = 386.8

```
> summary(reg3)

survreg(formula = "log(TCEConc)", data = "LandUse+lnPopDen+Depth",
    dist = "gaussian")
```

	Value	Std. Error	z	p
(Intercept)	-5.89982	2.64683	-2.23	0.02581
LandUse	0.24852	0.30905	0.80	0.42131
lnPopDen	1.62374	0.45984	3.53	0.00041
Depth	-0.00374	0.00233	-1.60	0.10875
Log(scale)	1.01156	0.11010	9.19	< 2e-16

```
Scale= 2.75
```

LandUse has a relatively high p-value. What about a 2-variable model? Drop LandUse and see if AIC goes down.

16

16

Two variable model AIC = 385.5

```
> xvar2 <- data.frame(lnPopDen, Depth)
> reg2 <- cencorreg(TCEConc, TCECen, xvar2)
 Likelihood R2 = 0.1301                    AIC = 385.5172
 Rescaled Likelihood R2 = 0.1605           BIC = 398.5709
 McFaddens R2 = 0.08375

> summary(reg2)
survreg(formula = "log(TCEConc)", data = "lnPopDen+Depth", dist =
     "gaussian")
               Value Std. Error     z       p
(Intercept) -3.99871    1.05133 -3.80 0.00014
lnPopDen     1.75184    0.44359  3.95 7.8e-05
Depth       -0.00429    0.00226 -1.90 0.05741
Log(scale)   1.01515    0.11015  9.22 < 2e-16
Scale= 2.76
```

- This is better than the 3 variable model due to lower AIC
- Depth is now at p=0.057. This is why you don't just throw away multiple variables at the beginning! Tossing 2 variables allows the important effect of a third, Depth, to be seen.
- I generally keep variables with p < 0.10, as model selection stats like AIC and BIC generally underfit (choose too few explanatory variables)
- Just as in ordinary regression, R2 increases with each added variable, so is no help in choosing a model. Rescaled R2 here is 0.160, while with the 3-variable model it was 0.163. This does NOT mean the 3-variable model is better.
- What about a 1-variable model, with just lnPopDensity?

17

17

One variable model AIC = 387.8

```
>  reg1 <- cencorreg(TCEConc, TCECen, lnPopDen)
 Likelihood R = 0.3388                     AIC = 387.8185
 Rescaled Likelihood R = 0.3763            BIC = 397.3588
 McFaddens R = 0.2707

> summary(reg1)
survreg(formula = "log(TCEConc)", data = "lnPopDen", dist
    = "gaussian")
             Value Std. Error     z      p
(Intercept) -5.10       1.03 -4.93  8e-07
lnPopDen     2.03       0.46  4.41  1e-05
Log(scale)   1.02       0.11  9.26 <2e-16

Scale= 2.78
```

- AIC goes back up for the 1-varaible model. So AIC picks the 2-variable model. AIC for 1 variable is worse than the 3-variable model.
- Summary: Choose the 2-var model. Why?

1. AIC is best
2. the p-value for Depth in 2-var model is 0.057, and since AIC/BIC underfits, keep variables below p = 0.10, especially if AIC is better.

- Should also 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 meaningful to me.

18

18

Or use 3b: bestaic function. Computes AIC for all possible models

Use all possible X vars, possibly transformed as in Step 2.

Model xvar4b used LandUse, lnPopDen, PctIndLU, Depth as X variables.

```
 > bestaic(TCEConc, TCECen, xvar4b)

Evaluating 15 models and printing the 10 lowest AIC models
 n.xvars                    model.xvars      aic
   2                    lnPopDen Depth 385.5172
   3              LandUse lnPopDen Depth 386.8478
   3            lnPopDen PctIndLU Depth 387.2307
   2                  LandUse lnPopDen 387.7902
   1                        lnPopDen 387.8185
   4     LandUse lnPopDen PctIndLU Depth 388.6177
   2                  lnPopDen PctIndLU 389.5974
   3          LandUse lnPopDen PctIndLU 389.6428
   2                     LandUse Depth 401.8533
   3              LandUse PctIndLU Depth 402.7185
```

- Start with the full set of variables, using transformations instead of original variables if found to be better in Step 2.
- There are 2^k - 1 possible combinations of k X variables; all of these censored regression models are run and their AICs reported.
- The "best" (lowest AIC) model is printed at the top. Here it is the 2-variable model that we settled on in slide 17.
- This is a quick way to get to a few good models. Run each of those you consider and inspect their partplots, look at p-values and Q-Q plots before accepting them as "My Model".

19

19

Remember the 3 Steps to a Good Multiple Regression Model

Step 1: Choose the units of Y (maximize W)

Step 2. Use partial plots to choose the units of the X variables (minimize AIC through transformation of X variables)

Step 3. Only then, lower AIC by

a. Dropping X variables one by one from the equation. Begin with the variable with the highest p-value , or

b. Use the bestaic function

20

20

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 (cencorreg)	ATS line (only 1 x variable)

21

21