Application 5: User-Defined Model with an Additional Variable

Dave Lorenz

December 3, 2015

This example illustrates the development of a user-defined model that includes and extra explanatory variable that is not related to flow or time. In this case, the extra explanatory variable is specific conductance.

The advent of water-quality monitors has facilitated the inclusion of surrogate variables such as specific conductance, pH, water temperature, dissolved oxygen, and turbidity that can be useful in the formulation of rating curve models for load estimation. Following the work of Christiansen and others (2000), this example uses 103 observations of streamflow and specific conductance collected from 1995 to 1998 to build a regression model for alkalinity in the Little Arkansas River, Kansas.

Part 2 delves into the diagnostic plots and builds a "better" model. It then reproduces figure 22 from Runkel and others (2004).

```
> # Load the rloadest package and the data
> library(rloadest)
> data(app5.calib)
> head(app5.calib)
```

```
       DATES TIMES   FLOW     SC Alkalinity
1 1995-02-28  1231   10.0 1425.0        248
2 1995-03-24  1301   11.3 1010.0        205
3 1995-03-28  0801  190.0  519.0         78
4 1995-04-12  0931   13.1  784.0        204
5 1995-04-24  1301   22.4 1750.0        231
6 1995-05-08  1301 2700.0   98.9         27
```

1 Build the Model

The `loadReg` function is used to build the rating-curve model for constituent load estimation. The basic form of the call to `loadReg` is similar to the call to `lm` in that it requires a formula and data source. The response variable in the formula is the constituent concentration, which is converted to load per day (flux) based on the units of concentration and the units of flow. The `conc.units`, `flow.units`, and `load.units` arguments to `loadReg` define the conversion. For these data, the concentration units (`conc.units`) are "mg/L", the flow units are "cfs" (the default), and the load units for the model are "kg" (also the default). If `conc.units` is not set, they are assumed to be "mg/L" and a warning is issued. Two additional pieces of information are required for `loadReg`—the names of the flow column and the dates column. A final option, the station identifier, can also be specified.

For any load model, the centered log flow is not necessary. It is used throughout LOADEST, but is optional for the construction of the rating-curve model. This example application does not use centered log flow in the initial model formulation.

```
> # Create the and print load model.
> app5.lr <- loadReg(Alkalinity ~ log(FLOW) + log(SC), data = app5.calib,
+                    flow = "FLOW", dates = "DATES", conc.units="mg/L",
+                    station="Arkansas River at Halstead, Ks.")
> app5.lr

*** Load Estimation ***

Station: Arkansas River at Halstead, Ks.
Constituent: Alkalinity

           Number of Observations: 103
Number of Uncensored Observations: 103
           Center of Decimal Time: 1996.979
                  Center of ln(Q): 5.126
                Period of record: 1995-02-28 to 1998-10-22

Selected Load Model:
--------------------

Alkalinity ~ log(FLOW) + log(SC)

Model coefficients:
            Estimate Std. Error z-score p-value
(Intercept)   2.2770    0.30151   7.552       0
log(FLOW)     0.8874    0.01559  56.942       0
log(SC)       0.6198    0.03605  17.190       0
```

```
AMLE Regression Statistics
Residual variance: 0.02494
R-squared: 98.33 percent
G-squared: 421.3 on 2 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9951
  p-value = 0.5743
Serial Correlation of Residuals: 0.2043

Variance Inflation Factors:
            VIF
log(FLOW) 3.038
log(SC)   3.038

Comparison of Observed and Estimated Loads
------------------------------------------
      Summary Stats: Loads in kg/d
------------------------------------------
     Min   25%   50%   75%    90%    95%    Max
Est 3440 10900 23400 80200 150000 273000 671000
Obs 3730 10400 25200 63100 170000 281000 947000

Bias Diagnostics
----------------
 Bp: -6.124 percent
PLR: 0.9388
  E: 0.9226
```

2 Diagnostic Plots

The rloadest package contains a `plot` function that creates diagnostic plots of the load model. Most often the user will just enter `plot(app5.lr)` (for this example) in the R Console window to generate the full suite of plots, but this example application will generate each plot individually. And, in general, the user will not need to set up a graphics device. But for this vignette, the graphics device must be set up for each graph.

Figure 1 shows the response versus the fitted values. The LOWESS curve agrees very well with the regression line and the scatter is very small. There does appear to be a small amount of curvature in the fit suggested by the response values at the ends are above the fit and a bump below the line between the fitted values of 11 and 12. Subsequent diagnostic plots can be used to assess whether the nonlinearity is a concern or not.

```
> # setSweave is required for the vignette.
> setSweave("app5_01", 5, 5)
> plot(app5.lr, which=1, set.up=FALSE)
> graphics.off()
```

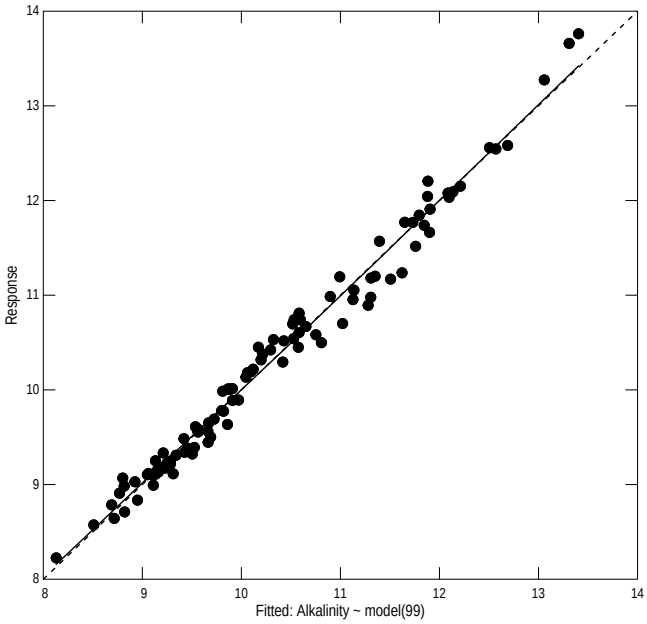

Figure 1. The rating-curve regression model.

Figure 2 is a scale-location (S-L) graph that is a useful graph for assessing heteroscedasticity of the residuals. The horizontal dashed line is the expected value of the square root of the absolute value of the residuals and the solid line is the LOWESS smooth. The slope of the LOWESS line indicates some cause concern for unequal variance in the estimates.

```
> # setSweave is required for the vignette.
> setSweave("app5_02", 5, 5)
> plot(app5.lr, which=3, set.up=FALSE)
> graphics.off()
```

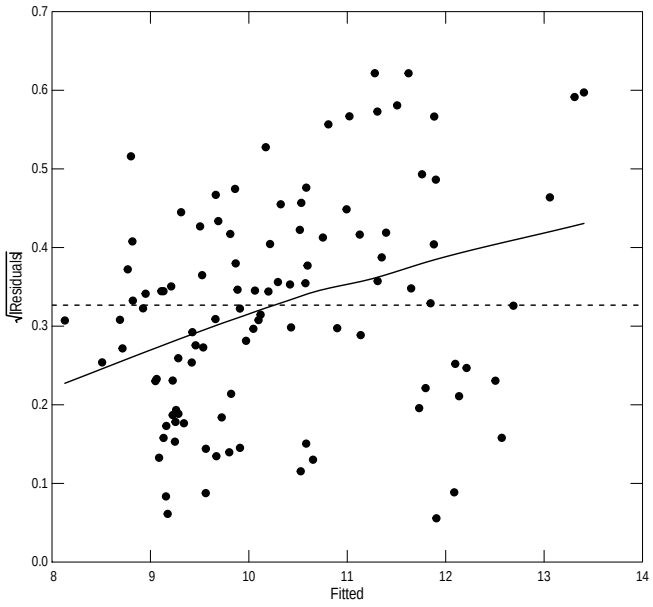

Figure 2. The scale-location graph for the regression model.

The correlogram in figure 4 is a adaptation of the correlogram from time-series analysis, which deals with regular samples. The horizontal dashed line is the zero value and the solid line is a kernel smooth rather than a LOWESS line. The kernel smooth gives a better fit in this case. The solid line should be very close to the horizontal line. In this case, there is a slight regular pattern with a higher line at 0 and 1 and a low line at about 0.5. This might suggest a seasonal lack of fit.

```
> # setSweave is required for the vignette.
> setSweave("app5_03", 5, 5)
> plot(app5.lr, which=4, set.up=FALSE)
> graphics.off()
```

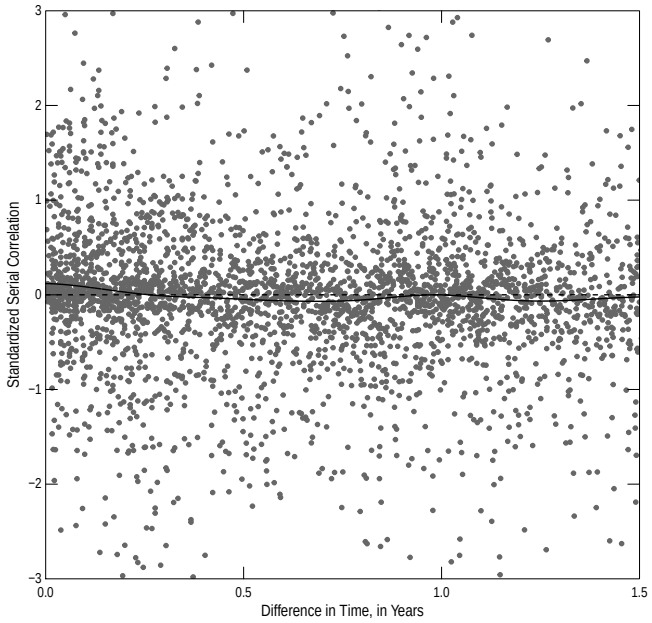

Figure 3. The correlogram from the regression model.

Figure 4 is a q-normal plots that shows the standardized residuals, which are assumed to have a standard deviation of 1. The solid line is the theoretical fit of mean of 0 and standard deviation of 1. The visual appearance of figure 5 confirms the results of the PPCC test in the printed output—the residuals are reasonably normal in distribution.

```
> # setSweave is required for the vignette.
> setSweave("app5_04", 5, 5)
> plot(app5.lr, which=5, set.up=FALSE)
> graphics.off()
```

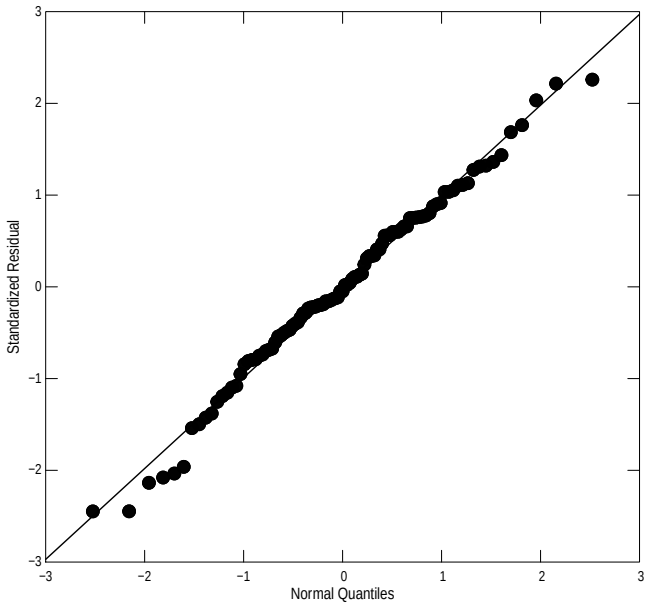

Figure 4. The Q-normal plot of the residuals.

Figure 5 is one of the partial-residual plots showing the relation to the log of flow. For this example, the span is reduced from the default of 1.0 to 0.5 to emphasize the curvature in the relation. The smooth line is does better represent the curvature of the relation and the second order polynomial test for linearity clearly indicates the strength of the curvature because the p-value is much less than 0.05.

```
> # setSweave is required for the vignette.
> setSweave("app5_05", 5, 5)
> plot(app5.lr, which="log(FLOW)", span=0.5, set.up=FALSE)
> graphics.off()
```

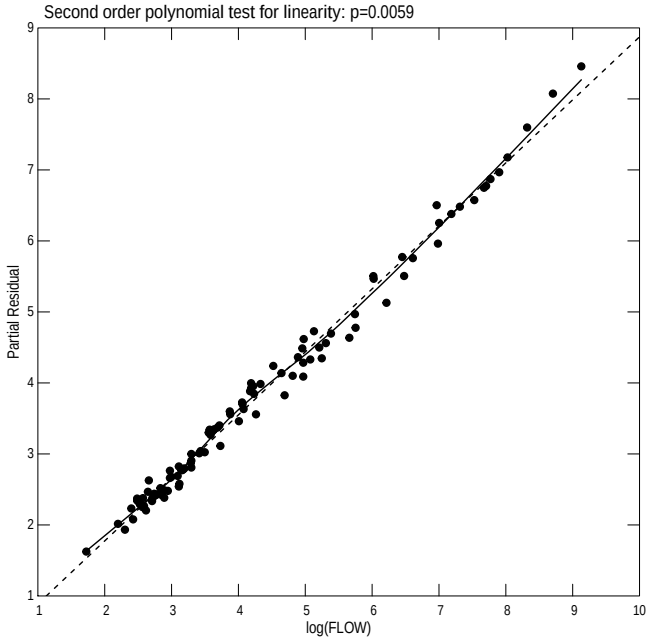

Figure 5. Partial residual plot for the log flow.

3 Part 2 Revise the Model

Figure 5 clearly indicated that the relation with log flow was not linear. The most straight forward option for these data is to use a second order polynomial. The correlogram (fig. 3) suggested that seasonality might be a concern and adding quadratic terms can account for some seasonality when the largest or smallest values are not being estimated correctly, so the revised model will include only quadratic flow and specific conductance. The residuals will be looked at to explain and resolve the structure of the correlogram.

```
> # Create the revised load model and plot the correlogram.
> app5.lrR1 <- loadReg(Alkalinity ~ quadratic(log(FLOW)) + log(SC),
+                      data = app5.calib,
+                      flow = "FLOW", dates = "DATES", conc.units="mg/L",
+                      station="Arkansas River at Halstead, Ks.")
> # setSweave is required for the vignette.
> setSweave("app5_06", 5, 5)
> plot(app5.lrR1, which=4, set.up=FALSE)
> graphics.off()
```

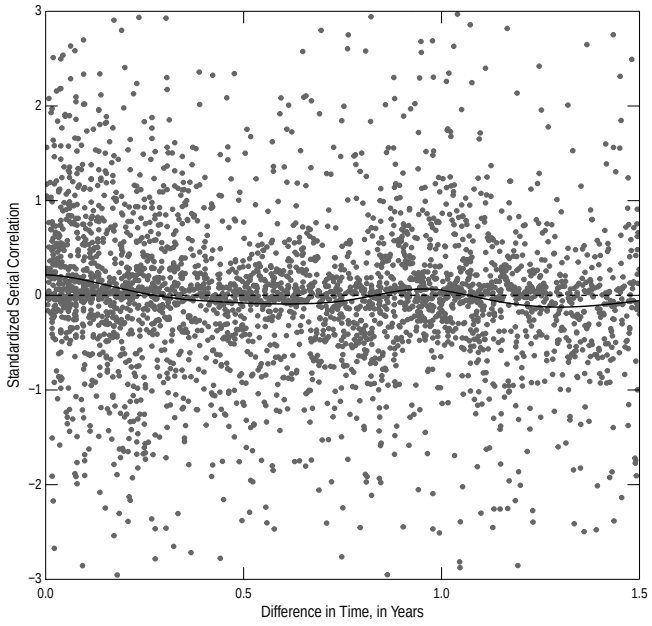

Figure 6. The correlogram from the revised regression model.

Normally, the correlogram shown if figure 6 would suggest that seasonal terms should be added and that would be the logical next step. However, for these data, that would not show a substantial difference in the correlogram and the sine and cosine terms would not significantly improve the model.

To improve the model, one must understand the setting and the hydrologic conditions. There is a low-head dam at this site, which can affect the relation between flow, specific conductance and alkalinity. Flows were much higher in January through April of 1998 when groundwater would normally be expected to dominate and the presence of the low-head dam could affect those relations during periods of abnormal flows during a season.

Figure 7 shows the relation between alkalinity and flow with the data for winter and early spring 1998 highlighted in red. The red points are all to the right of the cloud of other points, suggesting a change in the structural relation between alkalinity and flow for that time period. For the purposes of estimating loads for 1996, we will simply subset the data to exclude all of 1998.

```
> # setSweave is required for the vignette.
> setSweave("app5_07", 5, 5)
> AA.pl <- with(app5.calib, xyPlot(FLOW, Alkalinity, yaxis.log=TRUE,
+                                  xaxis.log=TRUE))
> with(subset(app5.calib, DATES > "1998-01-01" & DATES < "1998-05-01"),
+      addXY(FLOW, Alkalinity,Plot=list(what="points", color="red"),
+            current=AA.pl))
> graphics.off()
```

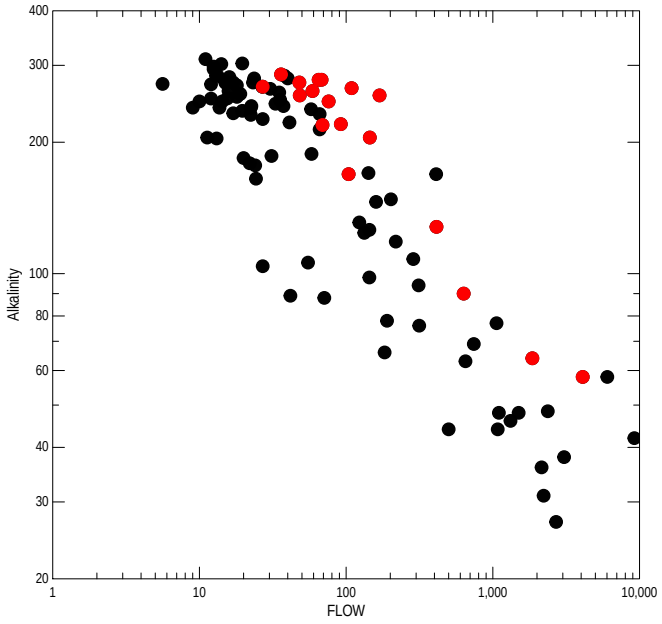

Figure 7. The relation between alkalinity and flow highlighting winter and early spring 1998 (in red).

The correlogram in figure 6 does indicate seasonality, so the second revision of the load model will include the sine and cosine terms (using the function `fourier`). The correlogram for the second revised model (fig. 8) is much improved.

The report for the new model does indicate a poor goodness-of-fit for the normality of the residuals and the variance inflation factors for linear flow and SC are relatively large, greater than 5. The normality of the residuals is assessed later in this section. For this model, the relatively large variance inflation factors should not be a concern because the correlation structure is similar between flow and specific conductance in the calibration and estimation data, but a bit stronger in the calibration (-0.893) than in the estimation data (-0.811).

```
> # Create, print the revised load model and plot the correlogram.
> app5.lrR2 <- loadReg(Alkalinity ~ quadratic(log(FLOW)) + log(SC) +
+                        fourier(DATES),
+                      data = app5.calib, subset=DATES < "1998-01-01",
+                      flow = "FLOW", dates = "DATES", conc.units="mg/L",
+                      station="Arkansas River at Halstead, Ks.")
> app5.lrR2

*** Load Estimation ***

Station: Arkansas River at Halstead, Ks.
Constituent: Alkalinity

           Number of Observations: 74
Number of Uncensored Observations: 74
           Center of Decimal Time: 1996.566
                  Center of ln(Q): 5.1571
                Period of record: 1995-02-28 to 1997-12-29

Selected Load Model:
--------------------

Alkalinity ~ quadratic(log(FLOW)) + log(SC) + fourier(DATES)

Model coefficients:
                               Estimate Std. Error z-score p-value
(Intercept)                     5.95099    0.31973  18.612   0e+00
quadratic(log(FLOW))(5.12598)1  0.91116    0.01875  48.605   0e+00
quadratic(log(FLOW))(5.12598)2  0.02903    0.00516   5.626   0e+00
log(SC)                         0.73392    0.04857  15.111   0e+00
fourier(DATES)sin(k=1)         -0.10864    0.02358  -4.608   0e+00
fourier(DATES)cos(k=1)         -0.09143    0.02670  -3.425   6e-04
```

```
AMLE Regression Statistics
Residual variance: 0.01496
R-squared: 99.06 percent
G-squared: 345.6 on 5 degrees of freedom
P-value: <0.0001
Prob. Plot Corr. Coeff. (PPCC):
  r = 0.9789
  p-value = 0.0172
Serial Correlation of Residuals: 0.1002

Variance Inflation Factors:
                                  VIF
quadratic(log(FLOW))(5.12598)1 5.856
quadratic(log(FLOW))(5.12598)2 1.181
log(SC)                        7.342
fourier(DATES)sin(k=1)         1.284
fourier(DATES)cos(k=1)         1.441

Comparison of Observed and Estimated Loads
------------------------------------------
      Summary Stats: Loads in kg/d
-------------------------------------------
     Min   25%   50%   75%    90%    95%    Max
Est 5540 10300 16400 66400 169000 283000 865000
Obs 5290 10100 16900 58600 176000 281000 947000

Bias Diagnostics
----------------
 Bp: -2.08 percent
PLR: 0.9792
  E: 0.9867

> # setSweave is required for the vignette.
> setSweave("app5_08", 5, 5)
> plot(app5.lrR2, which=4, set.up=FALSE)
> graphics.off()
```

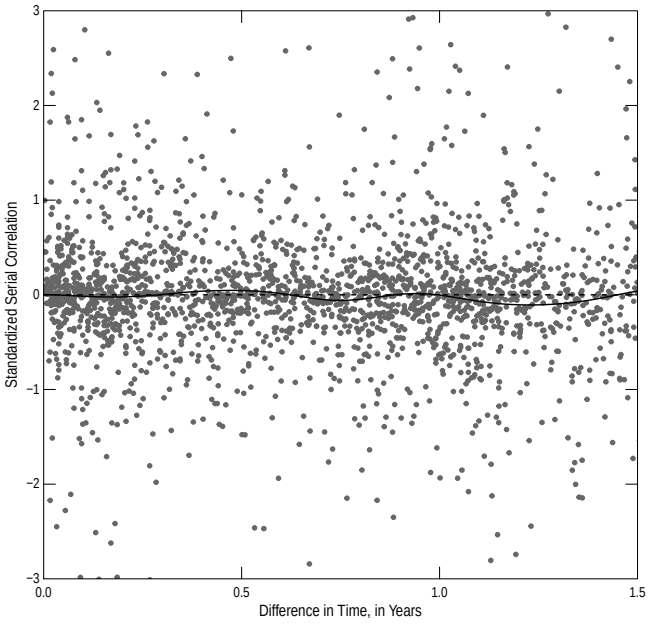

Figure 8. The correlogram from the second revised regression model.

Figure 9 shows the response versus the fitted values for the second revised model. The curvature is reduced from figure 1.

```
> # setSweave is required for the vignette.
> setSweave("app5_09", 5, 5)
> plot(app5.lrR2, which=1, set.up=FALSE)
> graphics.off()
```

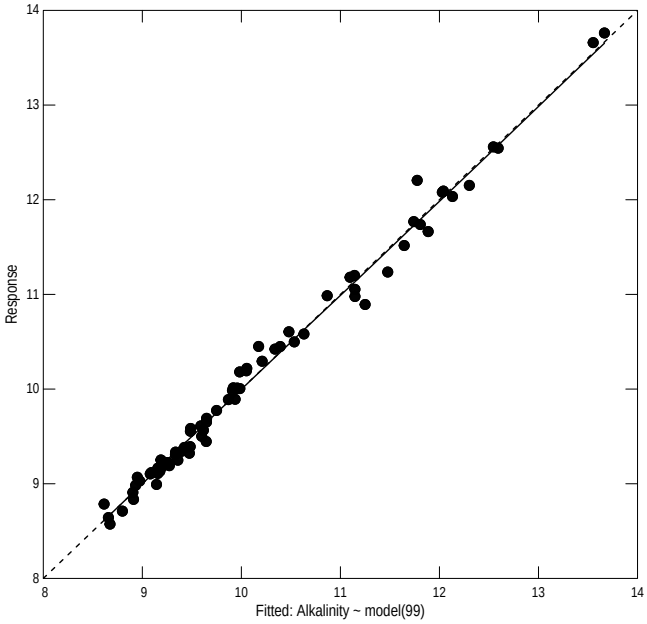

Figure 9. The rating-curve regression model.

Figure 10 is the scale-location (S-L) graph of the second revised model. There is still a bit of heteroscedasticity, particularly lower scatter in the lower fitted values, but constant in the larger values, where most of the load is transported.

```
> # setSweave is required for the vignette.
> setSweave("app5_10", 5, 5)
> plot(app5.lrR2, which=3, set.up=FALSE)
> graphics.off()
```

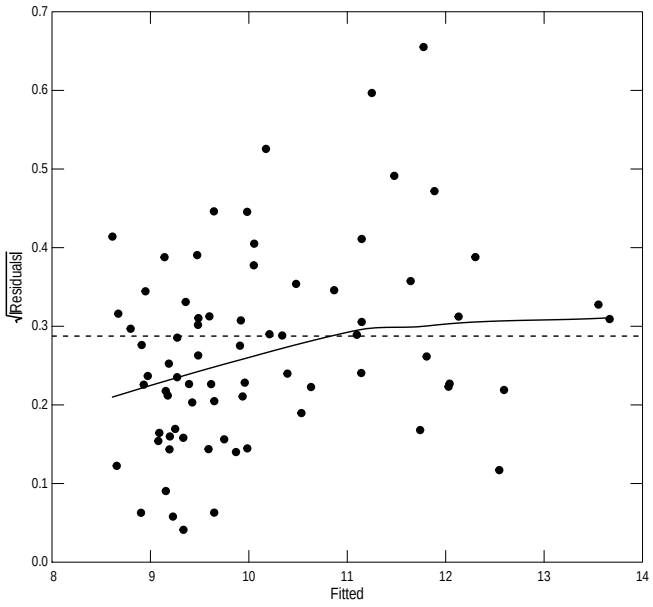

Figure 10. The scale-location graph for the regression model.

Figure 11 is the q-normal plot for the second revised model. In this case, there are a couple of outliers, one high and one low, which contribute to the small p-value recorded in the PPCC test.

```
> # setSweave is required for the vignette.
> setSweave("app5_11", 5, 5)
> plot(app5.lrR2, which=5, set.up=FALSE)
> graphics.off()
```

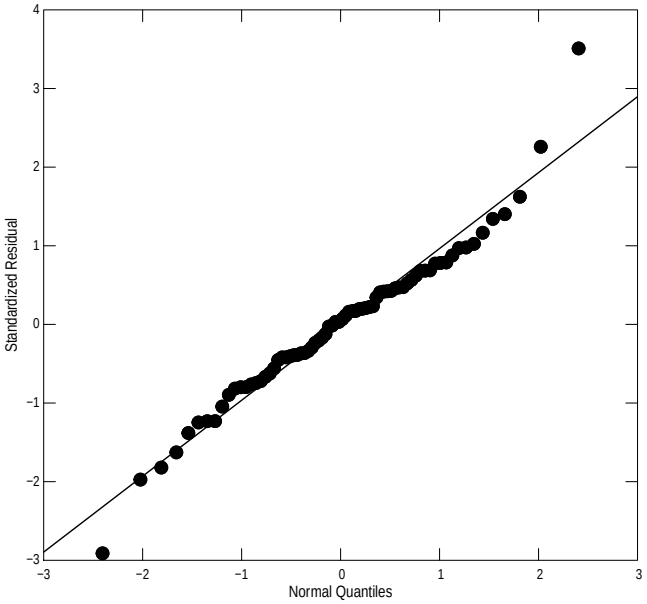

Figure 11. The Q-normal plot of the residuals.

The complete view of the diagnsotic plots suggests that there is some nonlinearity in the sine and cosine terms. Most often that would suggest adding additional, second-order, sine and cosine terms, or using a different seasonal term such as the seasonal wave, which was shown in applications 2 and 4. For this particular example, adding terms does not substantially improve the model, so we stop at the first-order terms.

4 Predict Daily Loads for 1999

Figure 22 in Runkel and others (2004) requires daily load estimates and the calibration load data. The `c2load` function can be used to calculate the daily loads in the 1999 dataset (app5.1999). Note that the daily estimation data set, `app5.est`, has 7 missing days and so has only 358 observations instead of 365 for 1999.

```
> # Get the estimation data
> data(app5.est)
> # Predict daily loads for 1999
> app5.ld <- predLoad(app5.lrR2, app5.est, by="day",
+                     load.units="pounds")
> # Get the 1999 sample data and merge to get daily flows and compute
> # load, note that the units must match what was selected for
> # estimation! The mergeQ function is in the USGSwsBase package;
> # the default names for dates and flow agree with the current datasets.
> data(app5.1999)
> app5.1999 <- mergeQ(app5.1999, Qdata=app5.est, Plot=FALSE)
> app5.1999$Load <- with(app5.1999, c2load(Alkalinity, FLOW,
+                                          conc.units="mg/L",
+                                          load.units="pounds"))
> # Create the graph
> setSweave("app5_12", 5, 5)
> AA.pl <- with(app5.ld, timePlot(Date, Flux,
+   Plot=list(name="Daily load estimate", what="overlaid",
+             size=0.03),
+   ytitle="Alkanity Load, in pounds per day"))
> AA.pl <- with(app5.1999, addXY(DATES, Load,
+   Plot=list(name="Observed load", what="points"),
+   current=AA.pl))
> addExplanation(AA.pl, where="ur", title="")
> graphics.off()
```

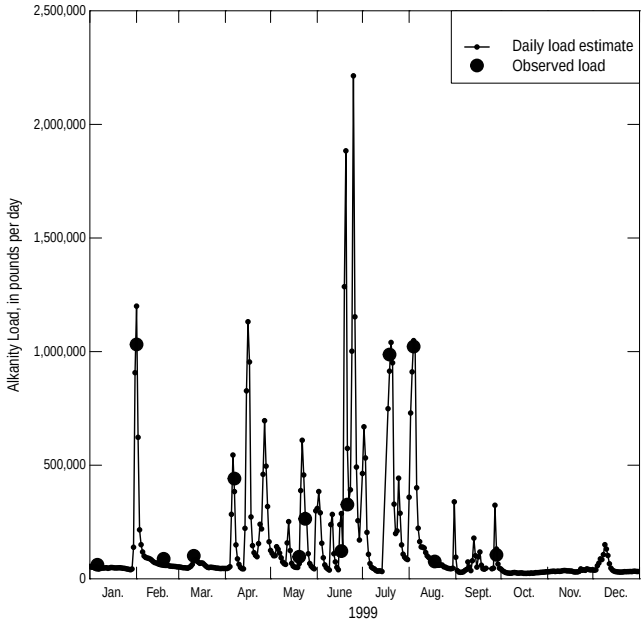

Figure 12. Daily load estimates for 1999.

The estimated loads for 1999 are extrapolated beyond the record used for calibration and as such are reasonable checks for the rating-curve model. In general, the largest loads shown in figure 12 are larger than those shown in figure 22 in Runkel and others (2004). The reason for that general observation can mostly be attributed to including the second order flow term in the second revised model, which in this case increase the estimates of the largest loads becuase the largest loads were underestimated in the original model. The observed loads agree fairly well with the estimated loads, but 2 deserve additional explanation—the observed loads in July and early August agree better with these estimated loads than in Runkel and other (2004) figure 22.