The GSqwsr R package

Laura De Cicco[1], Steve Corsi[1], and Austin Baldwin[1]

[1] *United States Geological Survey*

April 30, 2014

Contents

1 Introduction to GSqwsr package **2**

2 General Workflow **2**

3 Workflow Details **3**

3.1 Data Retrieval 3

3.2 Data Merging 5

3.3 Data Investigation 6

3.3.1 Narrow down investigation 6

3.3.2 Plot variables 7

3.4 Stepwise Regression 7

3.5 Stepwise Regression Analysis 9

3.6 Model Adjustments 10

3.7 Model Analysis 11

A Getting Started in R **13**

A.1 New to R? 13

A.2 R User: Installing QWSR 13

1 Introduction to GSqwsr package

The GSqwsr (USGS water quality surrogate regressions) package was designed to simplify the process of gathering water quality sample data and unit surrogate data, running a stepwise regression using the USGSwsQW censReg regression function, and analyzing those results. This vignette will first show a general overview workflow (2), then a more detailed description of the workflow with working examples (3).

2 General Workflow

```
library("GSqwsr")

#Sample data included with package:
DTComplete <- StLouisDT
UV <- StLouisUV
QWcodes <- StLouisQWcodes
siteINFO <- StLouisInfo

investigateResponse <- "Ammonia.N"
transformResponse <- "lognormal"

DT <- DTComplete[c(investigateResponse,
                   getPredictVariables(names(UV)),
                   "decYear","sinDY","cosDY","datetime")]
DT <- na.omit(DT)

predictVariables <- names(DT)[-which(names(DT)
                  %in% c(investigateResponse,"datetime","decYear"))]


#Check predictor variables
predictVariableScatterPlots(DT,investigateResponse)

# Create 'kitchen sink' formula:
kitchenSink <- createFullFormula(DT,investigateResponse)

#Run stepwise regression with "kitchen sink" as upper bound:
returnPrelim <- prelimModelDev(DT,investigateResponse,kitchenSink,
                               "BIC", #Other option is "AIC"
                               transformResponse)

steps <- returnPrelim$steps
```

```
modelResult <- returnPrelim$modelInformation
modelReturn <- returnPrelim$DT.mod

# Analyze steps found:
plotSteps(steps,DT,transformResponse)
analyzeSteps(steps, investigateResponse,siteINFO)

# Analyze model produced from stepwise regression:
resultPlots(DT,modelReturn,siteINFO)
resultResidPlots(DT,modelReturn,siteINFO)

# Create prediction plots
predictionPlot(UV,DT,modelReturn,siteINFO=siteINFO)
```

3 Workflow Details

In this section, we will step through the basic workflow.

3.1 Data Retrieval

Data retrieval is currently supported by web service calls to the National Water Information Service (NWIS). The first step is to get the discrete sample data that the regressions are modeling. In this example, we will look at the St Louis River at Scanlon (USGS site ID 04024000). If we don't know the sample data that is available, we can use the whatQW function to discover that information.

```
library(GSqwsr)
```

```
Warning:  package 'lattice' was built under R version 3.0.2
```

```
site <- "04024000"
QWcodes <- whatQW(site, minCount=20)

Warning:  unable to resolve 'waterservices.usgs.gov'
Error:  cannot open the connection

head(QWcodes)

Error:  object 'QWcodes' not found
```

Most likely, there will be a known set of parameters that are to be modeled. If the parameter codes for these analytes are known, the data from NWIS can be accessed directly with the function importNWISqw. The following example shows the process, and then lists the column names returned in the QW dataframe.

```
pCodeQW <- c("00608","00613","00618")
startDate <- "2011-04-22"
endDate <- ""
QW <- importNWISqw(site, params=pCodeQW,
                   begin.date=startDate, end.date=endDate)
```

```
names(QW)
```

```
 [1] "site_no"              "sample_dt"
 [3] "sample_tm"            "tzone_cd"
 [5] "medium_cd"            "Ammonia.N"
 [7] "Nitrite.N"            "Nitrate.N"
 [9] "NO2PlusNO3.N"         "Phosphorus_WW.P"
[11] "OrthoPhosphate.P"     "Chloride"
[13] "NitrogenTotal_WW.sum" "SuspSed"
[15] "datetime"
```

This brings the data in automatically as a 'qw' object. This means that censoring information is embedded within each data point. If any processing needs to be done to the data, it might be easier to import the raw data first, then convert to 'qw' objects with the makeQWObjects function.

```
QWRaw <- retrieveNWISqwData(site,pCodeQW,startDate,
                            endDate,expanded=TRUE)
QW <- makeQWObjects(QWRaw)
```

Next, the unit value data that will be used as surrogates for the analytes should be retrieved. If the parameters are not known, they can be discovered using the getDataAvailability function, filtering just the 'uv' (unit value) data:

```
UVcodes <- getDataAvailability(site)
```

```
Warning:  unable to resolve 'waterservices.usgs.gov'
```

```
Error:  cannot open the connection
```

```
UVcodes <- UVcodes[UVcodes$service == "uv",]
```

```
Error:  object 'UVcodes' not found
```

```
names(UVcodes)
```

```
Error:  object 'UVcodes' not found
```

```
UVcodes$parameter_cd
```

```
Error:  object 'UVcodes' not found
```

Finally, the unit value data can be retrieved with the getMultipleUV function. Because of the potentially large amount of data being returned, the web service call is automatically split into individual parameter codes.

```
UVpCodes <- c("00010","00060","00095","00300","00400","63680")
UV <- getMultipleUV(site, startDate, endDate, UVpCodes)
```

```
names(UV)
```

```
 [1] "agency_cd"    "site_no"      "datetime"     "tz_cd"
 [5] "Wtemp"        "Wtemp_cd"     "Flow"         "Flow_cd"
 [9] "SpecCond"     "SpecCond_cd" "DO"           "DO_cd"
[13] "pH"           "pH_cd"        "Turb"         "Turb_cd"
```

3.2 Data Merging

We now need to merge the sample and continuous data into one dataframe. This is accomplished using the mergeDatasets function. Both QW and UV dataframes need a column called 'datetime' that has the date and time in an POSIXct object. This may need to be done as shown below.

```
QW$datetime <- as.POSIXct(paste(QW$sample_dt," ",QW$sample_tm, ":00",sep=""))

# Make sure they are in consistant time zones:
QW$datetime <- setTZ(QW$datetime, QW$tzone_cd)
UV$datetime <- setTZ(UV$datetime, UV$tz_cd)

mergeReturn <- mergeDatasets(QW, UV, QWcodes)
```

```
Error:  object 'QWcodes' not found
```

```
DTComplete <- mergeReturn$DTComplete

Error:  object 'mergeReturn' not found

QWcodes <- mergeReturn$QWcodes

Error:  object 'mergeReturn' not found
```

The dataframe DTComplete contains a column of each of the discrete samples, and a column of the nearest (temporally) unit value data. The function mergeDatasets has an argument called 'max.diff'. The default is set to '2 hours', meaning that if the sample and continuous data timestamps do not match, the merge will take the closest continuous data within 2 hours. This value can be changed, see ?mergeNearest for more options.

3.3 Data Investigation

3.3.1 Narrow down investigation

We now want to narrow our investigation down to one analyte. Let us look at nitrate. First we will want a dataframe DT with just nitrate and the unit values. We will call these the 'prediction values' because they will eventually be used to predict nitrate.

```
investigateResponse <- "Nitrate.N"
predictionVariables <- getPredictVariables(names(UV))

DT <- DTComplete[c(investigateResponse,
                   predictionVariables,
                   "decYear","sinDY","cosDY","datetime")]

Error:  object 'DTComplete' not found


names(DT)

Error:  object 'DT' not found
```

For the regression, there can be no NA values in any of the columns. There are many ways in R to deal with this requirement. The easiest way to do it is remove any row that has any NA. This can be done as follows:

```
DT <- na.omit(DT)
```

```
Error:  object 'DT' not found
```

There may be other situations where you want to remove a column that contains the majority of the missing data.

3.3.2 Plot variables

There are a few tools included in this package to explore the data before performing the regression.

```
plotQQTransforms(DT,investigateResponse)
```

```
Error:  object 'DT' not found
```

```
predictVariableScatterPlots(DT,investigateResponse)
```

```
Error:  object 'DT' not found
```

3.4 Stepwise Regression

We are ready to perform a stepwise regression of the data to find the most significant variables to use in the model. This is accomplished with the prelimModelDev function. There are several inputs to this function. DT is the dataframe with all the predictor variables as well as the response we are investigating. We also need to define an upper bound for the stepwise regression to test. This is an equation with all the possible predictor variables, along with their transforms that we are interested in testing. If we want to use all possible variables, and all available transforms, we can use the equation createFullFormula (continuing with our Chloride example):

```
upperBoundFormula <- createFullFormula(DT,investigateResponse)
```

```
Error:  object 'DT' not found
```

```
Error:  object 'upperBoundFormula' not found
```

```
Error:  object 'upperBoundFormula' not found
```

The function will check if any data in DT has less than or equal to zero values. If so, a log transform is not included.

Now to use the stepwise regression within the prelimModelDev function. In this function, the DT dataframe is required, the column name of the response variable (in this example, investigateResponse), and the upper bound formula. The user can then choose a value for k which can be 'AIC' (akaike information criterion), 'BIC' (Bayesian information criterion), or a value of the multiple of the number of degrees of freedom used for the penalty. BIC has a harsher penalty for overfitting the model, which is typically seen as a benefit. The default is BIC. Finally, transformResponse can either be 'lognormal' or 'normal', which will define the transformation of the response variable.

Additionally, this function has an argument 'autoSinCos' which is a logical input. The default is set to TRUE, in this case - if the sine of decimal year (sinDY) is picked during the stepwise regression, the next step is forced to be cosDY. Likewise, if cosDY is picked, sinDY is forced on the next step. This feature can be turned off by setting autoSinCos to FALSE.

```
transformResponse <- "lognormal"

returnPrelim <- prelimModelDev(DT,
                 investigateResponse,
                 upperBoundFormula,
                 "BIC", #Other option is "AIC"
                 transformResponse)
```

```
Error:  object 'DT' not found
```

```
steps <- returnPrelim$steps
```

```
Error:  object 'returnPrelim' not found
```

```
modelResult <- returnPrelim$modelInformation
```

```
Error:  object 'returnPrelim' not found
```

```
modelReturn <- returnPrelim$DT.mod
```

```
Error:  object 'returnPrelim' not found
```

The output during the function shows the steps that the stepwise regression determined were ideal, information from the final model (modelInformation) such as the terms, their coefficients, standard error, p value as calculated by the censReg function, and standard coefficient (PARAML/STDDEV), and the raw data returned from censReg (DT.mod).

3.5 Stepwise Regression Analysis

It might be a good idea here to verify that the results from the stepwise regression are indeed what you want. The process can be observed using two functions: plotSteps and analyzeSteps.

analyzeSteps creates a plot with correlation, slope, RMSE, PRESS, and AIC statistics. Correlation and slope are values that should trend towards 1. Correlation is the correlation between observed and predicted, slope is the slope of observed and predicted. RMSE should trend towards zero, RMSE is the root mean squared error of the observed vs. predicted data. PRESS (predicted residual sums of squares) is calculated from the external studentized residuals, and should trend downward (for a better model fit). Residuals are calculated for censored values using the detection limit. AIC (akaike information criterion) is returned from the ANOVA output of the stepwise regression. It is always called 'AIC' whether or not 'AIC' or 'BIC' was specified. AIC will also trend downward for better model fits.

```
siteINFO <- getSiteFileData(site, interactive=FALSE)

Warning:  unable to resolve 'waterservices.usgs.gov'

Error:  cannot open the connection

analyzeSteps(steps, investigateResponse,siteINFO,
             xCorner=0.01,yCorner=0.3)

Error:  object 'steps' not found
```

In this case, it may seem strange that the AIC value goes up then down. This is because the first parameter that was picked was sinDY (sine of decimal year). As mentioned earlier, if sinDY is picked, we automatically force cosine to be the next parameter.

plotSteps shows the observed versus predicted for each step along the way of the stepwise regression.There are two lines included, the blue line is a one-to-one line, the red line is the slope of the observed versus predicted values as calculated with a linear regression (lm). In this simple regression, censored values are taken as their detection limits. Red points indicate potential outliers as calculated based on external studentized residuals greater than 3 or less than -3. Censored values are represented with a line segment from the detection limit to zero (for left-censored data).

```
m <- t(matrix(c(1:6), nrow = 2, ncol = 3))
layout(m)
par(mar=c(2,2,2,2))
plotSteps(steps,DT,transformResponse)

Error:  object 'steps' not found
```

3.6 Model Adjustments

There may be situations in which the user wants to explore alternative models compared to the results of the stepwise regression. The first tool offered in the package for this type of work is the function generateParamChoices. This function will create a dataframe, and optionally save it to a csv file with all of the parameter choices.

```
#Change this to a relavent path:
pathToSave <- "C:/RData/"
choices <- generateParamChoices(predictionVariables,
                                modelReturn,pathToSave,save=TRUE)
```

```
Error:  object 'modelReturn' not found
```

This produces a file that can be opened in Microsoft Excel, or any text editor:

	A	B	C	D	E	F	G	H	I	J	K	L
1	variableNames	Scalar	Wtemp	Flow	SpecCond	DO	pH	Turb	log(Flow)	log(SpecCond)	log(DO)	log(Turb)
2	Wtemp	1	0	0	0	0	0	0	0	0	0	0
3	Flow	0	0	0	0	0	0	0	0	0	0	0
4	SpecCond	0	0	0	0	0	0	0	0	0	0	0
5	DO	0	0	0	0	0	0	0	0	0	0	0
6	pH	1	0	0	0	0	0	0	0	0	0	0
7	Turb	0	0	0	0	0	0	0	0	0	0	0
8	log(Flow)	0	0	0	0	0	0	0	0	0	0	0
9	log(SpecCond)	1	0	0	0	0	0	0	0	0	0	0
10	log(DO)	0	0	0	0	0	0	0	0	0	0	0
11	log(Turb)	1	0	0	0	0	0	0	0	0	0	0

Figure 1: Output of generateParamChoices shown in Excel

The first column, 'Scalar', is pre-populated with zeros and ones, where ones represent the variables picked in the stepwise regression. Changing the 1's and 0's in this column will allow the user to easily set which parameters should be modeled. So, adding a 1 to the Flow row in the Scalar column will tell the program to include Flow in the model (as well as any other parameters with 1's). The next set of columns are used to allow users to include interaction terms. For example, if the interaction between log(Flow) and Turbidity was thought to be useful, a 1 in row 8, column H (log(Flow):Turb) would be required.

Once the parameter choice file is adjusted, it can be read in using read.csv, and a new formula can be created using the createFormulaFromDF function.

```
choicesNew <- read.csv(pathToSave)
newFormula <-createFormulaFromDF(choicesNew)
```

```
Error:  object 'choices' not found

Error:  object 'choicesNew' not found

Error:  object 'choicesNew' not found

Error:  object 'choicesNew' not found
```

```
newFormula

Error:  object 'newFormula' not found
```

From this formula, the stepwise regression routine can be re-run (if certain parameters were deleted for example), or the model can be created:

```
newUpperFormula <- paste(investigateResponse," ~ ", newFormula, sep="")

Error:  object 'newFormula' not found

modelReturnCustom <- censReg(newUpperFormula,
                        dist=transformResponse, data=DT)

Error:  object 'newUpperFormula' not found

modelReturnCustom

Error:  object 'modelReturnCustom' not found
```

3.7 Model Analysis

Finally, the package offers several ways to analyze and report on the model results. We will look at the results of the original model returned from the stepwise regression (not the custom model we created in the last section). Censored values are plotted as points with segments (from the detection limit towards zero for left-censored data). Also, left-censored residuals are calculated using the detection limit.

The function resultPlots plots a set of plots. All model results include observed vs. predicted (A), residuals vs. predicted (B), residuals vs. time (C), and residual quantiles vs. theoretical quantiles (D). After that, there is a plot for observed vs. each parameter in the model (E...).

```
resultPlots(DT,modelReturn,siteINFO)
```

```
Error:  object 'modelReturn' not found
```

The function resultResidPlots plots a set of plots. All model results include observed vs. predicted (A), residuals vs. predicted (B), residuals vs. time (C), and residual quantiles vs. theoretical quantiles (D). After that, there is a plot for residuals vs. each parameter in the model (E...).

```
resultResidPlots(DT,modelReturn,siteINFO)
```

```
Error:  object 'modelReturn' not found
```

The predictionPlot function plots the predicted values based on all of the unit value data available (from the UV dataframe) in blue. Red dots representing the actual measured data are also included. Left-censored points are shown at their detetction limit, with a segment going towards zero.

```
predictionPlot(UV,DT,modelReturn,siteINFO=siteINFO)
```

```
Error:  object 'modelReturn' not found
```

Finally, a summary printout can be obtained with the function summaryPrintout in either the R console or saved to a file.

```
summaryPrintout(modelReturn, siteINFO)
```

```
Error:  object 'modelReturn' not found
```

A Getting Started in R

This section describes the options for downloading and installing the GSqwsr package.

A.1 New to R?

If you are new to R, you will need to first install the latest version of R, which can be found here: `http://www.r-project.org/`.

There are many options for running and editing R code, one nice environment to learn R is RStudio. RStudio can be downloaded here: `http://rstudio.org/`. Once R and RStudio are installed, the dataRetrieval package needs to be installed as described in the next section.

At any time, you can get information about any function in R by typing a question mark before the functions name. This will open a file (in RStudio, in the Help window) that describes the function, the required arguments, and provides working examples.

```
library(GSqwsr)
?plotSteps
```

To see the raw code for a particular code, type the name of the function:

```
plotSteps
```

A.2 R User: Installing QWSR

Before installing GSqwsr, the dependent packages must be first be installed:

```
install.packages(c("XML", "lubridate", "akima", "KernSmooth",
                   "leaps", "car", "mvtnorm", "digest",
                   "relimp", "BSDA", "RODBC","memoise",
                   "boot","survival","splines","RColorBrewer",
                   "lattice","MASS"), dependencies=TRUE)
install.packages(c("USGSwsBase","USGSwsData","dataRetrieval",
                   "USGSwsGraphs","USGSwsStats",
                   "USGSwsQW","GSqwsr"),
                 repos="http://usgs-r.github.com")
```

After installing the package, you need to open the library each time you re-start R. This is done with the simple command:

```
library(GSqwsr)
```