Appendix 2.5A: Percent Change Function

Original document:
Notes on computing estimates of percent change based on GAMs trend analysis.
Elgin Perry
eperry@chesapeake.net
Aug 17, 2015.

This document gives a conceptual explanation for the chosen method of estimating percent change for the period of record, the mathematical details of implementing these concepts with gams, and details for obtaining an estimate of standard error for statistical inference about the degree of change. Finally the methods for implementing the mathematics in the r-programming language are addressed.

Concept

It has become clear after working with smoothing estimation methods such as Weighted Regression on Time Discharge and Space (WRTDS) and Generalized Additive Models (GAMs) that estimates near the edge of the independent variables space can exhibit high uncertainty. An example of particular interest is estimating mean levels of a response variable at the beginning and end of the Period of Record (POR). It is these estimates that form the basis of percent change during the POR which informs us of progress toward established goals. To compensate for uncertainty at the very beginning and end of the POR, it seems prudent to estimate the beginning and end values by averaging over a defined period at the beginning and end of the POR.

As an initial procedure, that may be modified after some experience, the GAMs approach will be to average the once a month estimates over the first two years or baseline period and last two years or current period of the POR. Estimates of percent change are computed based on the difference of these baseline and current period estimates relative to the baseline estimate. The discussion below is largely concerned with computing the estimate of the mean difference between baseline and current periods in a way that simplifies obtaining the standard error of this estimate.

Mathematics

For the GAMs implementation, the mathematics of obtaining the estimate and precision of change during the POR is based on a simple idea from linear models theory.

We make the usual linear model assumptions such that the model parameter vector and its variance-covariance matrix are estimated as

$$\hat{\boldsymbol{\beta}} = \left(Z^{\mathrm{T}}Z\right)^{-1}Z^{\mathrm{T}}Y \text{ and } \hat{\sum}_{\beta} = s^2\left(X^{\mathrm{T}}X\right)^{-1}.$$

Appendix 2.5A

Here Z is the matrix of linear predictors in the cubic spline basis for the GAM model which is computed from the matrix X of independent variables. The matrix Z is computed by differencing the columns of X (see Wood 2006 for details) and here we rely on the mgcv package of r to compute Z from X and thus do not concern ourselves with these details. The vector $\hat{\boldsymbol{\beta}}$ is the estimated parameter vector for the GAM and $\hat{\sum}_{\beta}$ is the estimated variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. Both of these estimates can be obtained from mgcv. The goal here will be to define a matrix Z_d such the $Z_d\hat{\boldsymbol{\beta}}$ is an estimate the difference between the baseline and current periods. Then we can estimate the standard error of the difference as $\sqrt{Z_d \hat{\sum}_{\beta} Z_d{}^{\mathrm{T}}}$ (Rao 1973). Given the estimate and standard error, we can obtain tests of significance and confidence intervals in the usual way.

For this example, let us look at percent change by averaging quarterly data for the first two and last two years of the POR rather than monthly data in order to keep the size of the matrices manageable. Also assume that the GAM is fairly simple with just a smooth for time (year) and a smooth for season (doy), although the methods easily extend to more complex GAMs. Assume that the period of record is 1999 to 2014. The first task is define X_p with columns for year and doy such that is will form the bases of computing the POR difference.

$$X_p = \begin{bmatrix} 1999 & 46 \\ 1999 & 136 \\ 1999 & 228 \\ 1999 & 320 \\ 2000 & 46 \\ 2000 & 136 \\ 2000 & 228 \\ 2000 & 320 \\ 2013 & 46 \\ 2013 & 136 \\ 2013 & 228 \\ 2013 & 320 \\ 2014 & 46 \\ 2014 & 136 \\ 2014 & 228 \\ 2014 & 320 \end{bmatrix}$$

The mgcv package is used to convert X_p to Z_p such that $Z_p\hat{\boldsymbol{\beta}}$ is a vector of length 16 corresponding to the predicted values of the 16 quarters defined in X_p. Next is to formulate a matrix so when it pre-

multiplies $Z_p\hat{\boldsymbol{\beta}}$, it will result in the averages of the first 8 and last 8 quarters in $Z_p\hat{\boldsymbol{\beta}}$. For that we use the matrix

$$A=\begin{bmatrix} 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 & 1/8 \end{bmatrix}$$

With this defined we have the result that $AZ_p\hat{\boldsymbol{\beta}}$ is a column vector with two entries: the first is the mean for the first 8 quarters and the second is the mean for the last 8 quarters. All that remains to get the desired difference is to pre-multiply by the row vector $\mathbf{d}=\begin{bmatrix}-1 & 1\end{bmatrix}$ such that $\mathbf{d}AZ_p\hat{\boldsymbol{\beta}}$ is a point estimate for the difference of the last two years minus the first two years. If we define

$Z_d=\mathbf{d}AZ_p$ then we have the desired end product such that the difference is estimated by $Z_d\hat{\boldsymbol{\beta}}$ and the standard error of the difference is estimated by $\sqrt{Z_d\hat{\sum}_\beta Z_d{}^{\mathrm{T}}}$.

Citations:

Rao, C. R. (1973). Linear Statistical Inference and Its Applications, 2nd Ed. John Wiley & Sons, New York. pp625.

Wood, Simon N. (2006). Generalized Additive Models (An Introduction with R). Chapman & Hall/CRC, Boca Raton, Florida. p392.

Appendix 2.5A

Implementing in R

Code to implement this method appears in the r-function gam.por.diff() at the end of this file. In what follows, we relate key elements of the mathematics to parts of the code.

The function has the data frame for fitting the GAM (tsdat) and the GAM (gam1) as arguments

```
gam.por.diff <- function(gam1,tsdat)

  {
```

First obtain the period of record range from the data frame, it is assumed that year is a column in tsdat.

```
  por.rng <- range(tsdat$year)
```

Make a 4 element vector for the first two and last two years

```
  yr.set <- c(por.rng[1],por.rng[1]+1,por.rng[2]-1,por.rng[2])  # first two and last two years
```

Make a 12 element vector of days of year (doy) that roughly spaced once a month

```
  doy.set <- seq(15,365,30)          # roughly once a month
```

Create a prediction data frame that pairs every element of yr.set with each element of doy.set. This corresponds to X_p above

```
  pct.chg.dta <- expand.grid(doy.set,yr.set)

  names(pct.chg.dta) <- c('doy','year')
```

Center year to create cyear as was done for fitting GAM.

```
  pct.chg.dta$cyear <- pct.chg.dta$year - mean(tsdat$year)

  # the following commented code is a check

#  pct.chg.dta$pdep <- predict(gam1,newdata=pct.chg.dta) # this is for testing

#  por.mns <- tapply(pct.chg.dta$pdep,pct.chg.dta$bl,mean,na.rm=TRUE) # this is for testing
```

Extract the coefficients vector and its variance-covariance estimate from the GAM fit. These correspond to $\hat{\boldsymbol{\beta}}$ and $\hat{\sum}_{\beta}$ above.

```
  beta <- gam1$coefficients # extract coefficients vector

  VCmat <- gam1$Vp    # extract variance-covariance matrix of coefficents
```

Use the predict function to compute the linear predictors that correspond to the data in pct.chg.dta. Xpc corresponds to Z_p.

```
  Xpc <- predict(gam1,newdata=pct.chg.dta,type="lpmatrix")  # extract matrix of linear predicters
```

```
  pdep <- Xpc%*%beta # check that is same predicted values as gotten from predict()
```

Create a matrix to average the current and baseline time period predictions. The matrix avg.per.mat corresponds to A above.

```
  xa <- c(rep(1/24,24),rep(0,24),rep(0,24),rep(1/24,24)) # construct a matrix to average baseline
and current periods

  avg.per.mat <- matrix(xa,nrow=2,ncol=48, byrow=TRUE)

  period.avg <- avg.per.mat %*% pdep # pre-multiply by averaging matrix
```

Create a vector to compute the difference of the time periods. The vector diff.mat corresponds to $\mathbf{d} = \begin{bmatrix} -1 & 1 \end{bmatrix}$ above.

```
  diff.mat <- c(-1,1)           # contruct matrix to get difference of current minus baseline

  diff.avg <- diff.mat %*% period.avg # pre-multiply by differencing matrix to check results
```

Multiply these matrices to get a single premultiplying matrix for estimating the difference. The matrix xpd corresponds to $Z_d = \mathbf{d}AZ_p$ above.

```
  xpd <- diff.mat%*%avg.per.mat%*%Xpc # premultiply linear predictors by averaging and
differencing matrices.
```

Use that pre-multiplying matrix to get the point estimate of the difference and the corresponding standard error.

```
  diff.est <- xpd%*%beta; diff.est  # compute estimate of difference

  diff.se <-  sqrt(xpd%*%VCmat%*%t(xpd)); diff.se  # compute Std. Err. by usual rules

  diff.t <- diff.est / diff.se; diff.t

  diff.pval <- 2*pt(abs(diff.t), gam1$df.null, 0, lower.tail = FALSE)

  diff.ci <- c(diff.est - 1.96*diff.se,diff.est + 1.96*diff.se)

  gam.por.diff.return <- list(por.rng = por.rng,
                              per.mns = as.vector(period.avg),
                              diff.est=diff.est,
                              diff.se=diff.se,
                              diff.ci=diff.ci,
                              diff.t=diff.t,
                              diff.pval=diff.pval)

  }  # end gam.por.diff
```

Appendix 2.5A

```
gam.por.diff <- function(gam1,tsdat)
  {
  # compute estimate of period of record difference with std. err. and confidence interval
  por.rng <- range(tsdat$year)
  yr.set <- c(por.rng[1],por.rng[1]+1,por.rng[2]-1,por.rng[2])  # first two and last two years
  doy.set <- seq(15,365,30)         # roughly once a month
  pct.chg.dta <- expand.grid(doy.set,yr.set)
  names(pct.chg.dta) <- c('doy','year')
  pct.chg.dta$bl <- pct.chg.dta$year <= por.rng[1]+1
  pct.chg.dta$cyear <- pct.chg.dta$year - mean(tsdat$year)
  # the following commented code is a check
#  pct.chg.dta$pdep <- predict(gam1,newdata=pct.chg.dta)
#  por.mns <- tapply(pct.chg.dta$pdep,pct.chg.dta$bl,mean,na.rm=TRUE)

  beta <- gam1$coefficients # extract coefficients vector
  VCmat <- gam1$Vp    # extract variance-covariance matrix of coefficents
  Xpc <- predict(gam1,newdata=pct.chg.dta,type="lpmatrix")  # extract matrix of linear predicters
  pdep <- Xpc%*%beta # check that is same predicted values as gotten from predict()
  xa <- c(rep(1/24,24),rep(0,24),rep(0,24),rep(1/24,24)) # construct a matrix to average baseline
and current periods
  avg.per.mat <- matrix(xa,nrow=2,ncol=48, byrow=TRUE)
  period.avg <- avg.per.mat %*% pdep # pre-multiply by averaging matrix
  diff.mat <- c(-1,1)           # contruct matrix to get difference of current minus baseline
  diff.avg <- diff.mat %*% period.avg # pre-multiply by differencing matrix to check results

  xpd <- diff.mat%*%avg.per.mat%*%Xpc # premultiply linear predictors by averaging and
differencing matrices.
  diff.est <- xpd%*%beta; diff.est  # compute estimate of difference
  diff.se <-  sqrt(xpd%*%VCmat%*%t(xpd)); diff.se  # compute Std. Err. by usual rules
  diff.t <- diff.est / diff.se; diff.t
  diff.pval <- 2*pt(abs(diff.t), gam1$df.null, 0, lower.tail = FALSE)
  diff.ci <- c(diff.est - 1.96*diff.se,diff.est + 1.96*diff.se)
  gam.por.diff.return <- list(por.rng = por.rng,
                              per.mns = as.vector(period.avg),
                              diff.est=diff.est,
                              diff.se=diff.se,
                              diff.ci=diff.ci,
                              diff.t=diff.t,
                              diff.pval=diff.pval)
  }  # end gam.por.diff
```