# HW 6

*Paul Harmon*

- **Due on Monday 10/16 at the end of the day.**
- **You can work in groups of up to 2 as long as the person is not someone you worked with on the last homework.**
- The following questions pertain to the post at http://ucfagls.wordpress.com/2011/06/12/additive-modelling-and-the-hadcrut3v-global-mean-temperature-series/
- Read the post (ignore the comments) and then answer the following questions:

1) **Describe the data set he is analyzing. What is the sample size, what does it represent, and what is the scale of the response variable?**

The author is analyzing a climate data set called the Hadcrut Temperature Anomaly dataset. It measures temperature anomalies from 1950 to 2017, meaning that it measures the

The response variable refers to the difference between the mean monthly temperature over the span from 1961 to 1990 and the average monthly temperature in the given month/year. Looking at the plot, the temperature anomaly tends to be pretty small during the 1961 to 1990 period (as would be expected, given the construction of the response); further, it appears that the temperatures tended to be less during the roughly 100 years prior to that period and slightly higher in the period after 1990.

There are 12 observations per year, with the annual average being the response. Since there are 168 years in the dataset, the sample size ends up being just 168 since each year has only a single observation. With these data, they could have fit models with 12 anomaly observations per year (2016 total observations) but this would have been more sensitive to short-term monthly temperature trends than would the yearly-aggregated method they chose to use.

2) **What research question(s) is he trying to address in the analysis? What question does the derivative estimation and blue highlighting using "sizer" address?**

The author is interested in assessing statistical evidence of whether or not global temperatures are rising. In particular, the author is interested in temperature trends - were there periods when the temperatures were clearly rising or falling? The derivative estimation and the blue highlighting get at this second question; the periods of time where there is strong evidence that the derivative of the trend differs from 0, the plot is highlighted either red (decreasing) or blue (increasing). Interestingly, no places are highlighted red, indicating that no period of time experience strong evidence for a decreasing temperature trend (at least globally), and two periods experienced strong statistical evidence of an increasing trend.

The author's methodology is interesting; however, he repeatedly refers to statistical 'significance' as his benchmark for determining whether something is out of the ordinary. In the plot, a 'significant' derivative is a period of time for which 0 lies outside a 99-percent confidence interval, indicating that, given the uncertainty in the trend estimates, only periods with fairly large deviations in slope from 0 are highlighted blue.

3) **What is the impact of accounting for autocorrelation on the estimated trend in his analysis?**

The impact of accounting for autocorrelation in the estimated trend helps to keep the smoothing line from overfitting the data. By estimating a lag-1 autocorrelation structure, the fitted line is not pulled towards more extreme values at the bottom of valleys or tops of peaks; the uncorrelated-errors model ends up overfitting the data by getting pulled towards those extreme values. Also, it appears that the AR(1) structure seems to do better near the edges - it is not as affected by the extreme temperature values in 1850 or 2017 and seems more linear.

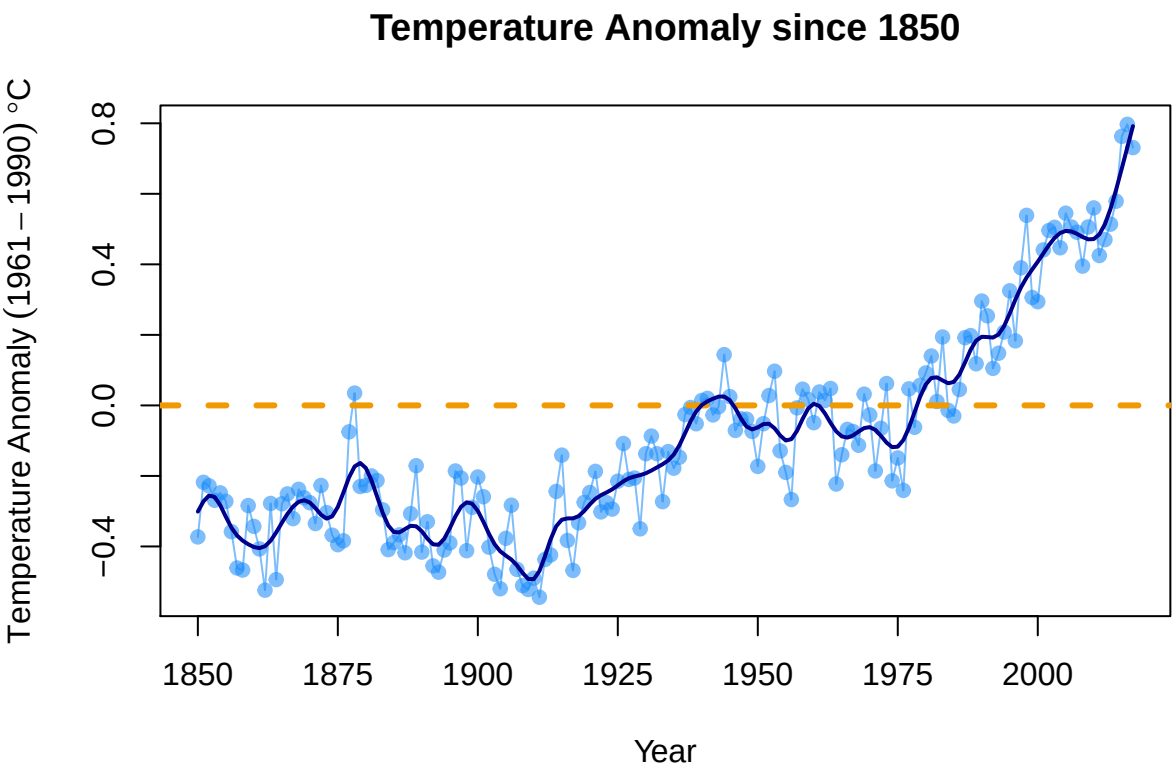
4) **The provided code will get you the updated global yearly temperature data set that goes to 2017 (you may want to save the data set as sometimes they update the data without**

**too much warning). The code reads in the data set and makes a nice looking (but not as nice-looking as if it were made in GGplot ;) ) plot, adding the fitted values from a potential GAM to the plot. Discuss the estimate from the GAM that was fit. How is it similar to or different from his results? Make sure to address differences in options for fitting the GAMs, the shape of the estimated trend as well as evidence for the trend and complexity of the estimated curves.**

```
URL <- url("https://crudata.uea.ac.uk/cru/data/temperature/HadCRUT4-gl.dat")
gtemp <- read.table(URL, fill = TRUE)
## Don't need the even rows
gtemp <- gtemp[-seq(2, nrow(gtemp), by = 2), ]
## set the Year as rownames
rownames(gtemp) <- gtemp[,1]
## Add colnames
colnames(gtemp) <- c("Year", month.abb, "Annual")

## Plot the data
library(grDevices)
ylab <- expression(Temperature~Anomaly~(1961-1990)~degree*C)
plot(Annual ~ Year, data = gtemp, type = "o",pch = 20, cex = 1.4,col = rgb(30,144,255,alpha = 150,maxCo
abline(h=0,col="orange2",lwd=3, lty = 2)
title("Temperature Anomaly since 1850")
axis(1,at = seq(1850,2010, by = 25))

library(mgcv)
m1<-gam(Annual~s(Year,k=round(168/3),bs="tp"),data=gtemp,method="GCV.Cp")

summary(m1)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Annual ~ s(Year, k = round(168/3), bs = "tp")
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.094375   0.006694   -14.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Year) 35.87  42.94 42.61  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.916   Deviance explained = 93.4%
## GCV = 0.0096453  Scale est. = 0.0075285  n = 168
```

```
lines(gtemp$Year,fitted(m1),col="blue4",lwd=2)
```

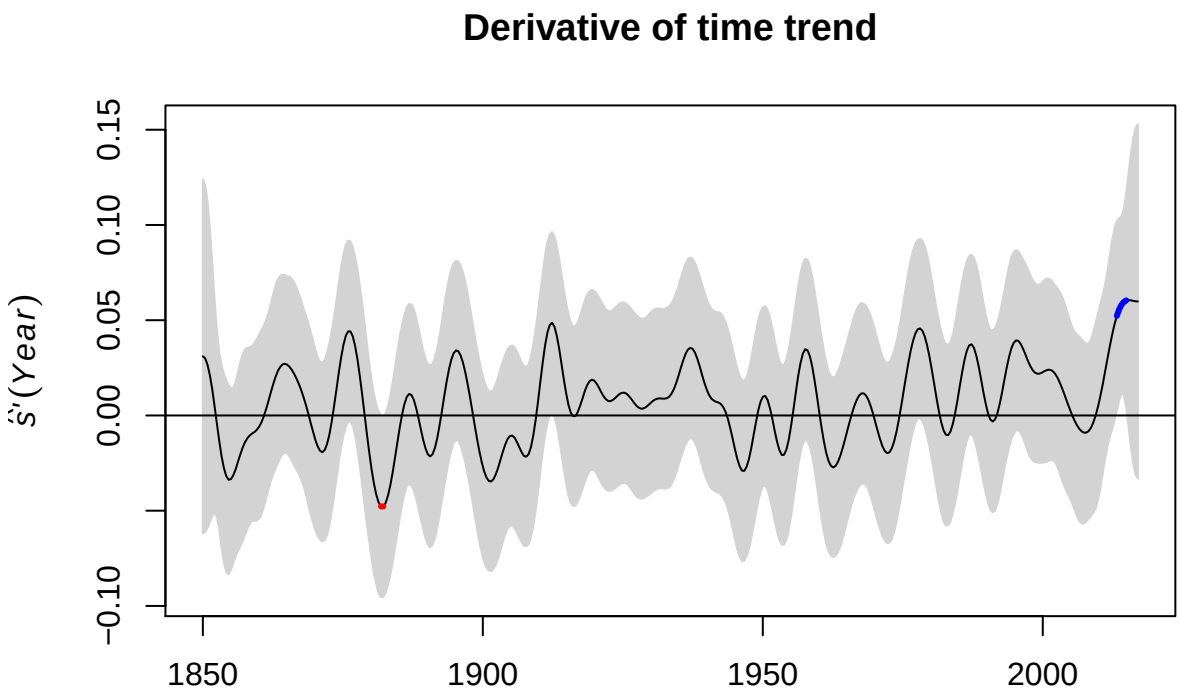



5) **Using the following code, you can obtain the estimated first derivative of the time trend. Smoothing with GCV and autocorrelated data can lead to overly aggressive fits. One option (now the default in gam) is to use "REML" estimation to get the optimal amount of smoothness. This approach exploits a connection to mixed models that I will explain later this semester. The model m2 is fit using this approach. Compare the edf used by the two models and the estimated derivatives.**

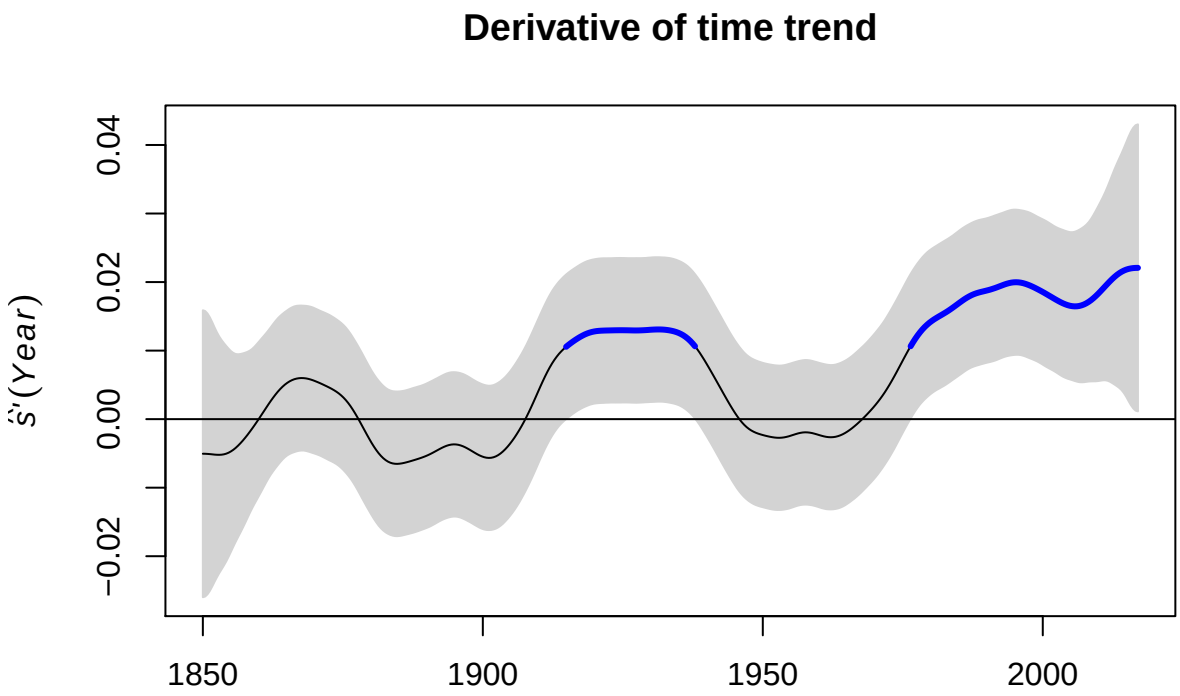
The REML-estimated model allows for residual-based maximum likelihood estimation. The estimated derivative for the REML model is not as wiggly as the one that usess maximum likelihood. Based on this, it is likely that the ML model is overfitting the data.

The estimated degrees of freedom are 10.61 for the REML model and 35.87 for the ML smoother, indicating that the ML model is using more complexity to get to its estimates. The REML estimate is more parsimonious in that it needs fewer knots to estimate an average and thus will be less wiggly than the ML version.

```
source("http://www.math.montana.edu/courses/s217/documents/derivFun.R.txt")

m1.d <- Deriv(m1, n = 400)
par(mfrow=c(2,1))
plot(m1.d, sizer = TRUE, alpha = 0.01,ylab=expression(italic(hat(s)*"'"*(Year))),main="Derivative of ti
```




```
m2<-gam(Annual~s(Year,k=round(168/3),bs="tp"),data=gtemp,method="REML")
m2.d <- Deriv(m2, n = 400)
plot(m2.d, sizer = TRUE, alpha = 0.01,ylab=expression(italic(hat(s)*"'"*(Year))),main="Derivative of ti
```




```
summary(m2)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Annual ~ s(Year, k = round(168/3), bs = "tp")
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.094375   0.007544  -12.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df     F p-value
## s(Year) 10.61  13.24 104.9  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.893   Deviance explained =   90%
## -REML = -130.64  Scale est. = 0.0095623  n = 168
```

6) **In this problem, you will explore the role of the smoothing parameter in GAMs. I do not recommend this option, but it is possible to fix the smoothing parameter and explore the impact on the estimated smooth curves. Describe the impacts of increasing the smoothing**

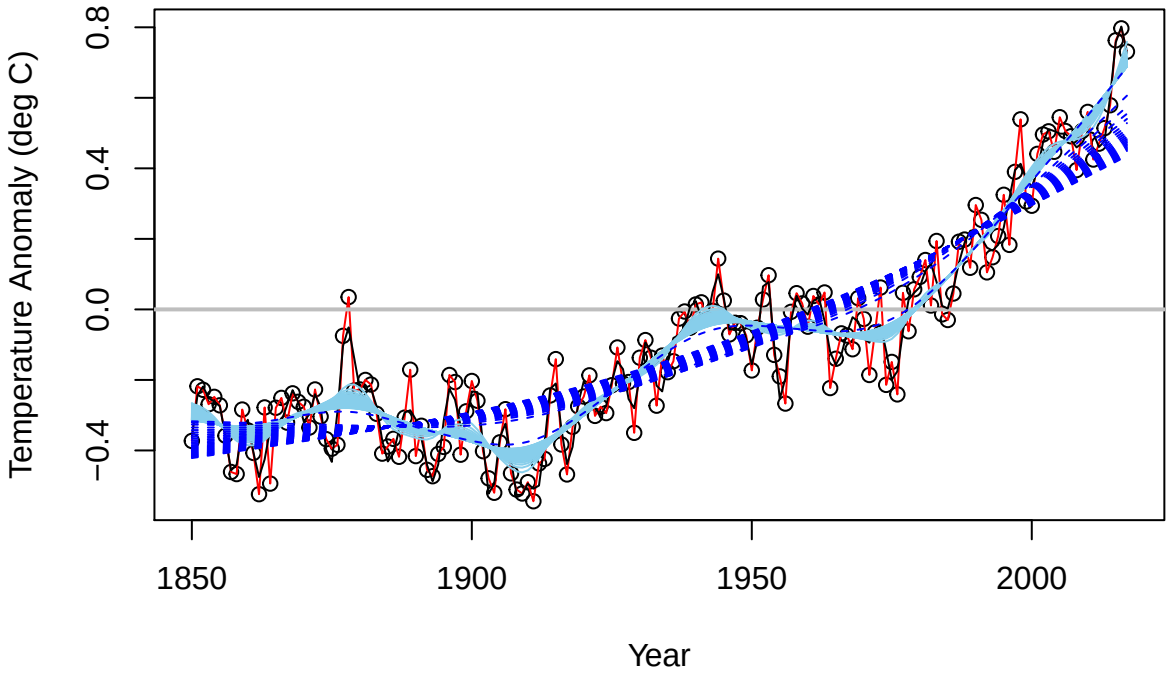
**parameter and what the optimal amount of smoothing was based on the GCV measure.**

There are several plots below. The first one shows the fitted values from the two GAM fits. The GAMs with the smaller set of smoothing parameters tend to start with no smoothing - basically the line fits the points - and then tends to add in minor additional smoothing from there. The black line shows the line with no smoothing and the light blue lines show the smoother estimates. Then, for larger smoothing parameter values, the estimated model may be overly smooth - it looks almost like a quadratic fit.

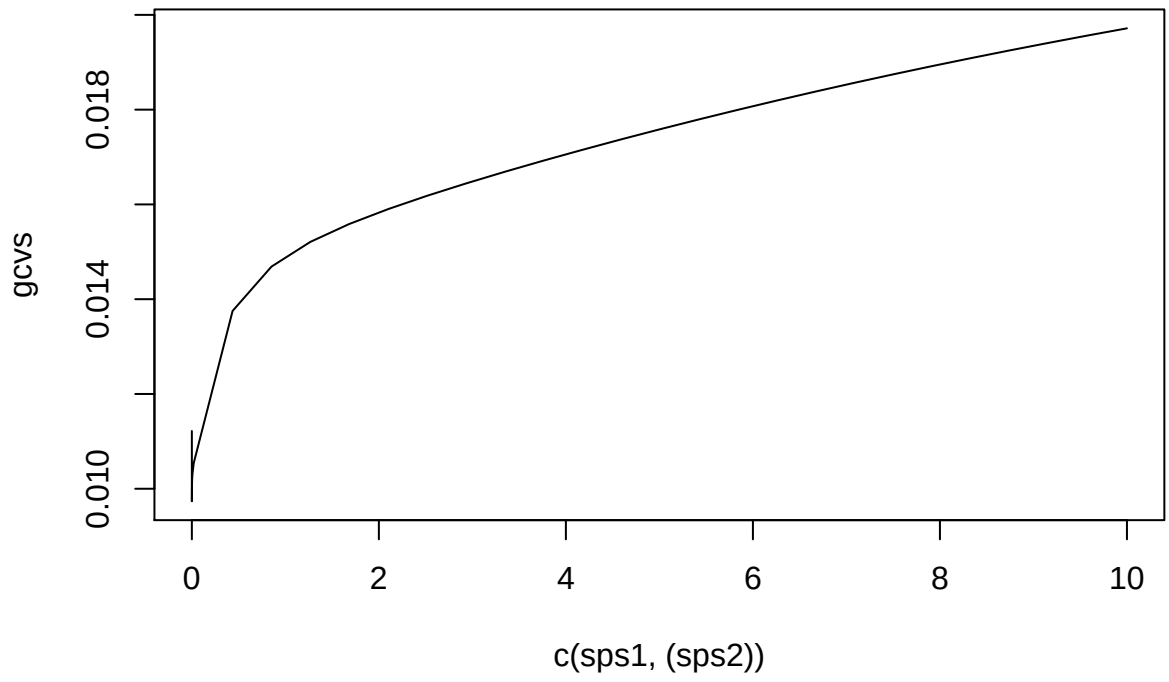
I didn't really like the way that the GCV plot looked in base R, but thankfully GGplot allows for faceting. While not a perfect solution (because it distorts the x-axis), the plot below shows the behavior of the GCV measures near where it is minimized.

Increasing the smoothing parameter has an initial effect of lowering the GCV measure; however, after it is minimized at $x = 0.0002$ and increases rather rapidly from there.

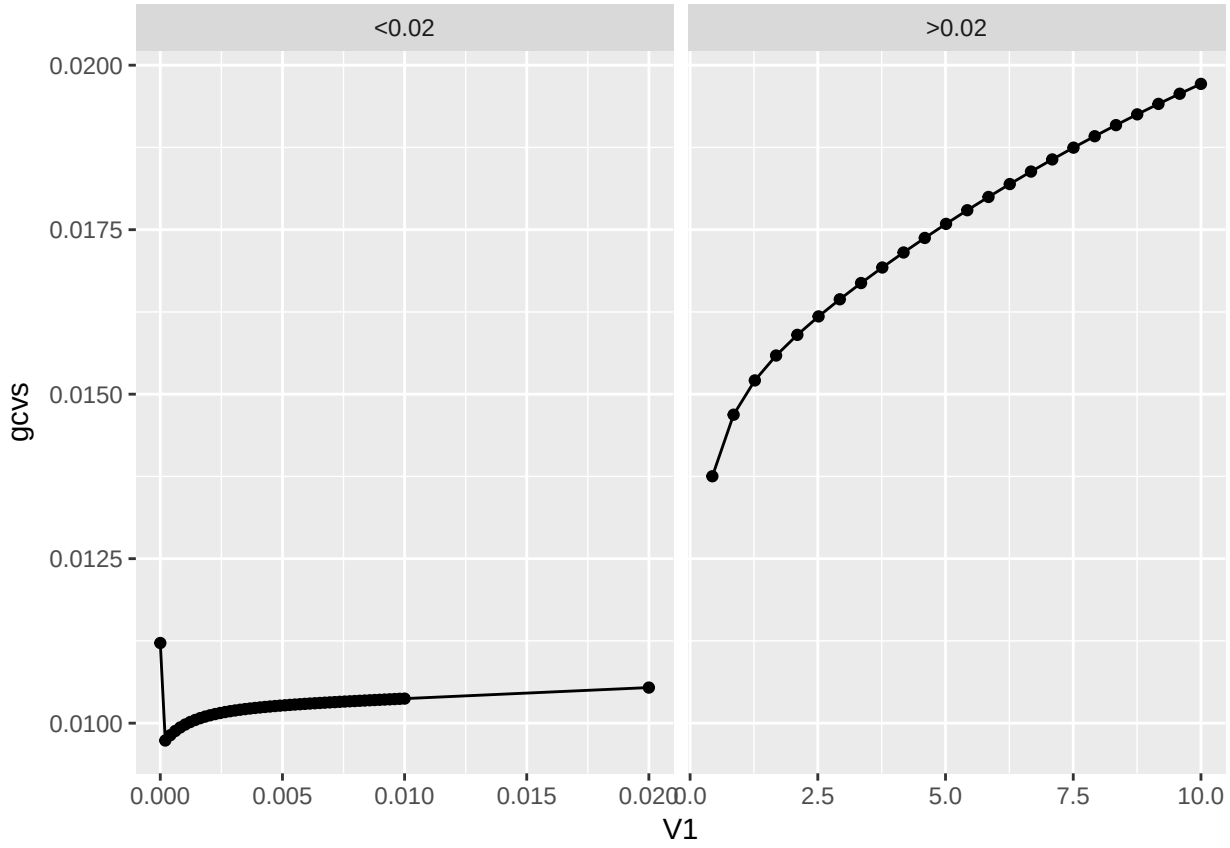
```
plot(gtemp$Year,gtemp$Annual,xlab="Year",ylab="Temperature Anomaly (deg C)")
lines(gtemp$Year,gtemp$Annual,col="red")
abline(h=0,col="grey",lwd=2)

#plots the fitted values
sps1<-c(seq(from=.000000001,to=.01,length.out=50))
gcvs<-numeric(0)
sick.colors <- c("black",rep("skyblue",length(sps1)-1))
for (j in (1:length(sps1))){
  lam1<-sps1[j]
  m2<-gam(Annual~s(Year,k=round(168/2),bs="tp",sp=(lam1)),data=gtemp,method="GCV.Cp")
  gcvs<-c(gcvs,m2$gcv.ubre)
  lines(gtemp$Year,fitted(m2),col=sick.colors[j])
}

sps2<-c(seq(from=.02,to=10,length.out=25))
for (j in (1:length(sps2))){
  lam1<-sps2[j]
  m2<-gam(Annual~s(Year,k=round(168/2),bs="tp",sp=(lam1)),data=gtemp,method="GCV.Cp")
  gcvs<-c(gcvs,m2$gcv.ubre)
  lines(gtemp$Year,fitted(m2),col="blue",lty=2)
}
```




```
#Plot of (Smoothing Parameters) vs GCV scores
plot(c(sps1,(sps2)),gcvs,type="l")
```




```
x <- data.frame(cbind(c(sps1,sps2),gcvs))
x$bigsmall <- ifelse(x$V1 > 0.02,">0.02","<0.02")
library(ggplot2)
ggplot(x,aes(x = V1, y = gcvs)) + geom_point() + geom_line() + facet_grid(.~bigsmall,scales = "free_x")
```


<0.02
>0.02
0.0200
0.0175
0.0150
0.0125
0.0100
gcvs
0.000
0.005
0.010
0.015
0.020
0.0
2.5
5.0
7.5
10.0
V1