Generalized additive model example in R

This is a study of the relationship between atmospheric ozone concentration, O_3 and other meteorological variables in the Los Angeles Basin in 1976. To simplify matters, let's only focus on three predictors: temperature measured at El Monte, `temp`, inversion base height at LAX, `ibh`, and inversion top temperature at LAX, `ibt`. The cases with missing values were removed for simplicity. The data were first presented by Breiman and Friedman (1985). First, let's fit a linear model for reference purpose.

```
> library(faraway)
> data(ozone)
> olm <- lm(O3~temp+ibh+ibt,ozone)
> summary(olm)

Call:
lm(formula = O3 ~ temp + ibh + ibt, data = ozone)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3224  -3.1913  -0.2591   2.9635  13.2860

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.7279822  1.6216623  -4.765 2.84e-06 ***
temp         0.3804408  0.0401582   9.474  < 2e-16 ***
ibh         -0.0011862  0.0002567  -4.621 5.52e-06 ***
ibt         -0.0058215  0.0101793  -0.572    0.568
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 4.748 on 326 degrees of freedom
Multiple R-squared:  0.652,        Adjusted R-squared:  0.6488
F-statistic: 203.6 on 3 and 326 DF,  p-value: < 2.2e-16
```

Note that `ibt` is not significant in this model. One task among others in a regression analysis is to find the right transformation on the predictors. Additive models can help here. Let's fit an additive model using the Gaussian response as the default. We used the loess smoother here by specifying `lo` in the model formula for all three predictors.

```
> library(gam)
> amgam <- gam(O3 ~ lo(temp) + lo(ibh) + lo(ibt), data=ozone)
> summary(amgam)

Call: gam(formula = O3 ~ lo(temp) + lo(ibh) + lo(ibt), data = ozone)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
```

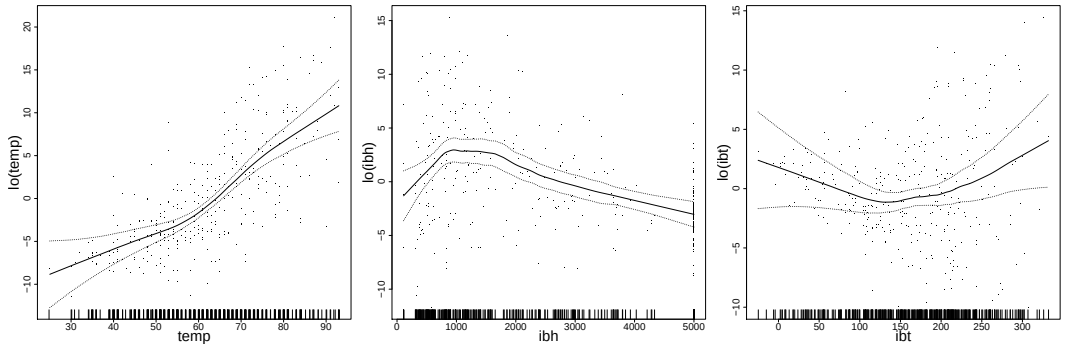
```
-13.1146  -2.3624  -0.2092   2.1732  12.4447

(Dispersion Parameter for gaussian family taken to be 18.6638)

    Null Deviance: 21115.41 on 329 degrees of freedom
Residual Deviance: 5935.096 on 318.0005 degrees of freedom
AIC: 1916.049

Number of Local Scoring Iterations: 2

Anova for Parametric Effects
           Df  Sum Sq Mean Sq  F value    Pr(>F)
lo(temp)    1 11958.2 11958.2 640.7153 < 2.2e-16 ***
lo(ibh)     1  1117.3  1117.3  59.8646 1.358e-13 ***
lo(ibt)     1     3.5     3.5   0.1898    0.6634
Residuals 318  5935.1    18.7
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Anova for Nonparametric Effects
            Npar Df Npar F     Pr(F)
(Intercept)
lo(temp)        2.5 7.4550 0.0002456 ***
lo(ibh)         2.9 7.6205 8.243e-05 ***
lo(ibt)         2.7 7.8434 9.917e-05 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

> 1-5935.1/21115

[1] 0.7189155
```

Compared to the linear model, the R^2 of GAM is improved by more than 10%. However, the loess fit does use more degrees of freedom (i.e. the effective number of parameters is estimated by the trace of the projection matrix). The `gam` package uses a score test for testing the significance for each predictor. But the p-values are only approximate at best and should be viewed with some skepticism. It is generally better to fit the model without the predictor of interest and then construct the F-test.

```
> amgamr <- gam(O3 ~ lo(temp) + lo(ibh) , data=ozone)
> anova(amgam,amgamr,test="F")

Analysis of Deviance Table

Model 1: O3 ~ lo(temp) + lo(ibh)
Model 2: O3 ~ lo(temp) + lo(ibh) + lo(ibt)
  Resid. Df Resid. Dev     Df Deviance      F Pr(>F)
1    321.67     6044.6
2    318.00     5935.1 3.6648   109.47 1.6005  0.179
```

Although the p-value from F-test is still an approximation, we can see some evidence that `ibt` is not significant. Let's examine the fit for all three variables.

```
> par(mfrow=c(1,3),mar=c(5,5,2,2),cex.lab=3,cex.axis=2)
> plot(amgam,residuals=TRUE,se=TRUE,pch=".")
```




```
> detach(package:gam, unload = TRUE) #close the gam library
```

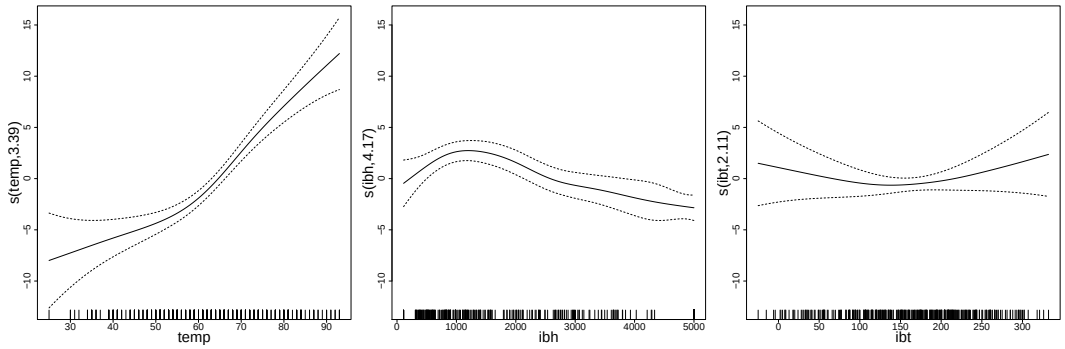
We can see for `ibt`, the confidence band can hold a constant function, which reinforces the conclusion that it is not significant. For variable `temp`, we can clearly see an 'elbow' around 60^o, while for `ibh` it reaches maximum around 1000. The partial residuals allow us to identify the outliers and influential observations. Although it seems no problem in this data, loess smoother is recommended where such problem arises. Allowing more choice of smoother is the favorite feature of `gam` package.

Another method of fitting GAM is provided by `mgcv` package of Wood (2000). Although splines are the only choice of smoother in the `mgcv` package, it has an automatic choice in the amount of smoothing as well as wider functionality.

```
> library(mgcv)
> ammgcv <- gam(O3 ~ s(temp)+s(ibh)+s(ibt),data=ozone)
```

We see that the R^2 is about the same as the `gam` fit. We can also examine the transformation used for each variable. We can see that the fitted transformations are again similar to `gam` fit. Variable `ibt` does not appear to be significant.

```
> par(mfrow=c(1,3),mar=c(5,5,2,2),cex.lab=3,cex.axis=2)
> plot(ammgcv)
```



We can also test whether there is a nonlinear trend for variables `temp` by fitting a model with a linear term of `temp` and then make the F-test. The test result confirms that there is really a change in the trend of temperature.

```
> am1 <- gam(O3 ~ s(temp)+s(ibh),data=ozone)
> am2 <- gam(O3 ~ temp+s(ibh),data=ozone)
> anova(am2,am1,test="F")
```

```
Analysis of Deviance Table

Model 1: O3 ~ temp + s(ibh)
Model 2: O3 ~ s(temp) + s(ibh)
  Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
1    323.67       6950
2    320.97       6054 2.7031   895.98 17.573 7.943e-10 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
```

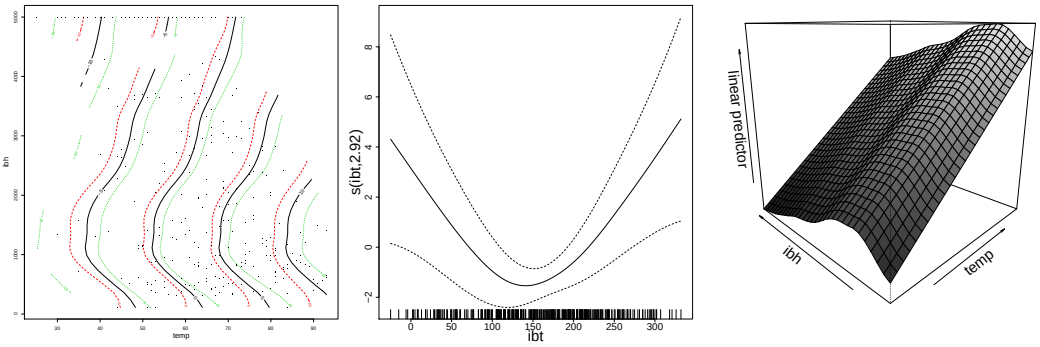
We can also do the bivariate transformations in `mgcv`. Suppose we suspect there is an interaction between temperature and IBH. We can fit a model with this interaction, and compare to previous additive model. We can see that in fact fewer d.f. was used to fit the bivariate model. And the results show the additive model fits better than the interaction model. Hence, in spite of the significant p-value, we suspect there is no interaction between temperature and IBH. A side-effect of the interaction model is that variable `ibt` becomes significant.

```
> amint <- gam(O3 ~ s(temp,ibh)+s(ibt),data=ozone)
> summary(amint)

Family: gaussian
Link function: identity

Formula:
O3 ~ s(temp, ibh) + s(ibt)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.7758     0.2409   48.88   <2e-16 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Approximate significance of smooth terms:
              edf Ref.df      F  p-value
s(temp,ibh) 6.346  8.040 14.881  < 2e-16 ***
s(ibt)      2.917  3.679  9.805 6.16e-07 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

R-sq.(adj) =  0.702   Deviance explained =   71%
GCV = 19.767  Scale est. = 19.152    n = 330

> anova(ammgcv,amint,test="F")

Analysis of Deviance Table

Model 1: O3 ~ s(temp) + s(ibh) + s(ibt)
Model 2: O3 ~ s(temp, ibh) + s(ibt)
  Resid. Df Resid. Dev       Df Deviance      F   Pr(>F)
1    319.33     5977.9
2    319.74     6123.6 -0.40915  -145.66 19.017 0.001411 **
```

```
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1
```

We can graphically examine the bivariate contour plot as the following. The left plot shows a perspective view of the contour plot on the left. Based on both the left and right plots, we conclude there is no obvious interaction between temperature and IBH.

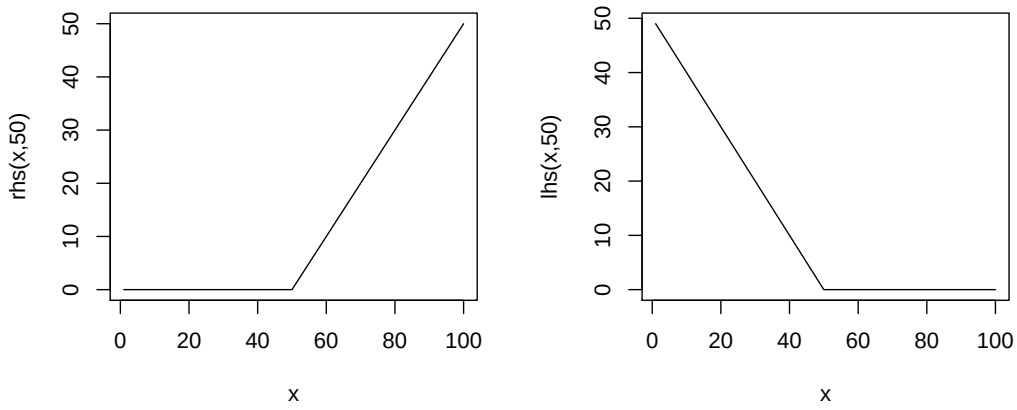
```
> par(mfrow=c(1,3),mar=c(5,5,2,2),cex.lab=3,cex.axis=2)
> plot(amint)
> vis.gam(amint,theta=-45,color="gray")
```



One use for additive models is as an exploratory tool for standard parametric regression modeling. We can use the fitted function to help us find suitable simple transformation on the predictors. In this case, we have found both predictors `temp` and `ibh` can be modeled as piecewise linear regression (also known as segmented regression.) We can define the right and left "hockey-stick" functions as follows, and fit a parametric model using cutoff points of 60 and 1000 for `temp` and `ibh`, respectively. The cutoff points are picked based on the figures above.

```
> rhs <- function(x,c) ifelse(x > c, x-c, 0)
> lhs <- function(x,c) ifelse(x < c, c-x, 0)
> par(mfrow=c(1,2))
> plot(1:100,rhs(1:100,50),type="l",xlab="x",ylab="rhs(x,50)")
> plot(1:100,lhs(1:100,50),type="l",xlab="x",ylab="lhs(x,50)")
> olm2 <- lm(O3 ~ rhs(temp,60)+lhs(temp,60)+rhs(ibh,1000)+lhs(ibh,1000),ozone)
> summary(olm2)

Call:
lm(formula = O3 ~ rhs(temp, 60) + lhs(temp, 60) + rhs(ibh, 1000) +
    lhs(ibh, 1000), data = ozone)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2042  -2.6307  -0.2887   2.3179  12.6720

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.6038321  0.6226512  18.636  < 2e-16 ***
rhs(temp, 60)   0.5364407  0.0331849  16.165  < 2e-16 ***
lhs(temp, 60)  -0.1161735  0.0378660  -3.068  0.00234 **
rhs(ibh, 1000) -0.0014859  0.0001985  -7.486 6.72e-13 ***
lhs(ibh, 1000) -0.0035544  0.0013138  -2.705  0.00718 **
```

```
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 4.342 on 325 degrees of freedom
Multiple R-squared:  0.7098,        Adjusted R-squared:  0.7062
F-statistic: 198.7 on 4 and 325 DF,  p-value: < 2.2e-16
```



Compared this model with the first ordinary linear regression model, the fit is better and about as good as the GAM fit. It is unlikely for us to discover these transformation without the help of the intermediate additive models. Compared to GAMs, this model has compact form and simpler interpretation.

For GAM, we can predict new values with standard error:

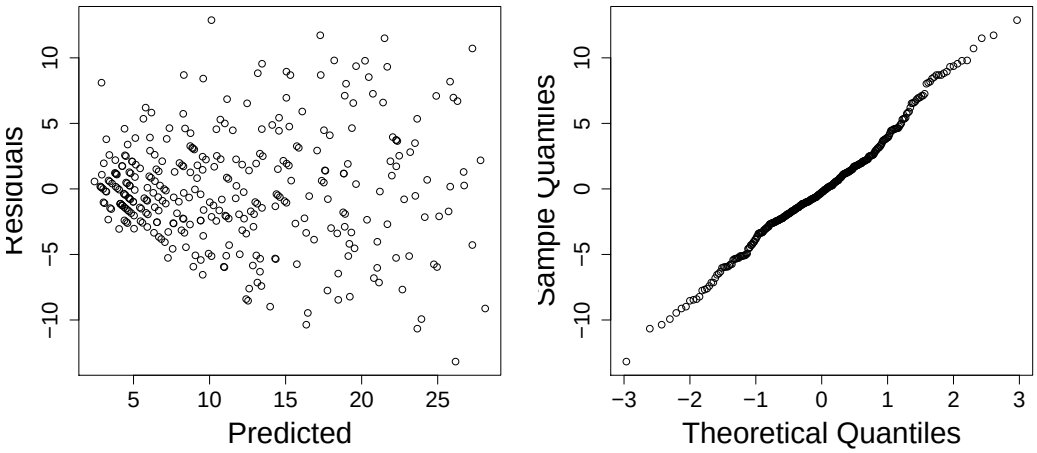
```
> predict(ammgcv,data.frame(temp=60,ibh=2000,ibt=100),se=T)

$fit
       1
11.01278

$se.fit
        1
0.9727755
```

If we try to make predictions for predictor values outside the original range of data, we will need to linearly extrapolate the spline fits, which is highly dangerous. See the SE is much larger although this likely does not fully reflect the uncertainty.

```
> predict(ammgcv,data.frame(temp=120,ibh=2000,ibt=100),se=T)

$fit
       1
35.51078

$se.fit
       1
5.726077
```

We should also examine the usual diagnostics. We can see that although the residuals look normal, there is some nonconstant variance.

```
> par(mfrow=c(1,2),cex.lab=2,cex.axis=1.5)
> plot(predict(ammgcv),residuals(ammgcv),xlab="Predicted",ylab="Residuals")
> qqnorm(residuals(ammgcv),main="")
```



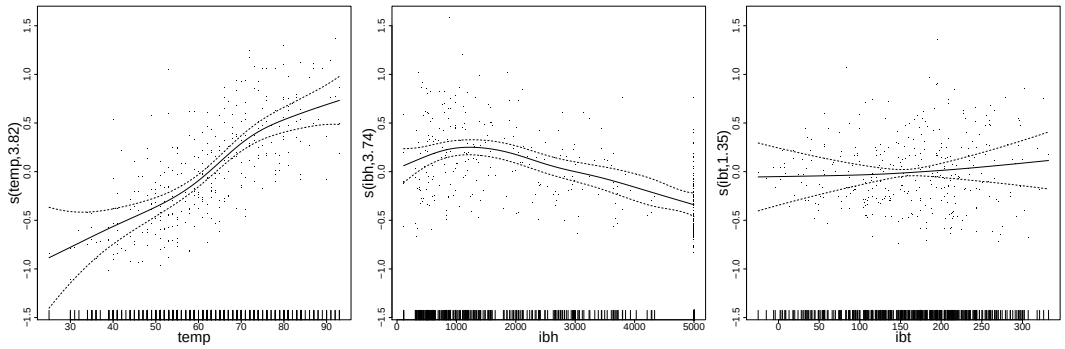
The ozone data has a response with relatively small integer values. Furthermore, the diagnostic plot above shows nonconstant variance. This suggests that a Poisson response might be suitable.

```
> table(ozone$O3)

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 3
 2  9 29 27 25 21 22 10 18 13  9 18 10 14  8 10  9  7 11  6  2  7  6  6  2  9  3  4  4  2  1  1

> gammgcv <- gam(O3 ~ s(temp)+s(ibh)+s(ibt),family=poisson,scale=-1,data=ozone)
> summary(gammgcv)

Family: poisson
Link function: log

Formula:
O3 ~ s(temp) + s(ibh) + s(ibt)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.29269    0.02304   99.49   <2e-16 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Approximate significance of smooth terms:
          edf Ref.df      F  p-value
s(temp) 3.816  4.736 16.817 4.20e-14 ***
s(ibh)  3.737  4.568 10.603 1.04e-08 ***
s(ibt)  1.348  1.623  0.574    0.529
---
```

```
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

R-sq.(adj) =  0.712   Deviance explained = 72.9%
GCV = 1.5062  Scale est. = 1.4573    n = 330
```

We have set `scale=-1` because negative values for this parameter indicate that the dispersion should be estimated rather than fixed at one. Since we do not truly believe the response is Poisson, it seems better to allow for over-dispersion. The default of not specifying `scale` would fix the dispersion at one. We see that the estimated dispersion is indeed somewhat bigger than one. We see that IBT is not significant and the selected transformations are quite similar to those fitted previously.

```
> par(mfrow=c(1,3),mar=c(5,5,2,2),cex.lab=3,cex.axis=2)
> plot(gammgcv,residuals=TRUE)
```




```
> detach(package:mgcv, unload = TRUE) #close the mgcv library
```

Reference

Julian J. Faraway (2006) *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models* (1st edition) by Chapman and Hall/CRC.