Additive models

extension of chapter 1 [multiple linear regression]

Instructor: Dr. Sharandeep Singh Pandher*

? Response $y \sim$ Normal

1 Introduction

Stone (1985) introduced additive models and mathematical expression can be written as

$$y = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

. where f_j are smooth arbitrary functions. The advantage of the additive model approach is that the best transformations are determined simultaneously and without parametric assumptions regarding their form. Also, Categorical variables can be easily accommodated within the model using the usual regression approach with the following mathematical equation:

2 categorical

γ regression parameters

$$y = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \boldsymbol{Z}\gamma + \epsilon$$

HW5

. where $\boldsymbol{Z}$ is the design matrix for the variables that will not be modeled additively, where some may be quantitative and others qualitative. The γ are the associated regression parameters. We can also have an interaction between a factor and a continuous predictor by fitting a different function for each level of that factor. There are at least **three** ways of **fitting additive models** in *R*.

1. The Generalized Additive Models (**gam**) package originates from the work of Hastie and Tibshirani (1990).

2. The Mixed GAM Computation Vehicle with Automatic Smoothness Estimation (**mgcv**) package is part of the recommended suite that comes with the default installation of R and is based on methods described in Wood (2000).

3. The General Smoothing Splines (**gss**) package of Gu (2002) takes a splinebased approach.

- **Note**: The **gam** package allows **more choice** in the smoothers used while the **mgcv** package has an **automatic choice** in the amount of smoothing as well as wider functionality.

*Address: Department of Mathematics and Statistics, University of Alberta, Edomoton, AB,T6G 2G1, Canada, e-mail: sharand1@ualberta.ca;

1.1 Fitting algorithms for gam and mgcv

- The **backfitting** algorithm is used in the gam package as:

  1. Initialize $\beta_0 = \bar{y}$, and $\hat{f}_j(x) = \hat{\beta}_j x$ where $\hat{\beta}$ is some initial estimate, such as the least squares, for $j = 1, 2, ..., p$.
  2. We cycle $j = 1, 2, ..., p$

     $$f_j = S(x_j, y) - \beta_0 + \sum_{i \neq j} f_i(x_i)$$

     where $S(x, y)$ means the smooth on the data (x, y).The choice of S for nonparametric smoother are splines or loess while for a parametric fit are linear or polynomial.
  3. The algorithm is iterated until convergence.

  – **Note**: Hastie and Tibshirani (1990) show that convergence is assured under some rather loose conditions. The term $y - \beta_0 + \sum_{i \neq j} f_i(x_i)$ is a partial residual-the result of fitting everything except x_j, making the connection to linear model diagnostics.

- The mgcv package employs a **penalized smoothing spline** approach. Suppose we represent $f_j(x) = \sum_i^n \beta_i \phi_i(x)$ for a family of spline basis functions ϕ_i. we impose a penalty $\int [f_j''(x)]^2 dx$ which can be written in the form $\beta_j^\top S_j \beta_j$ for a suitable matrix S_j that depends on the choice of basis. We then maximize:

  $$logL(\beta) - \lambda \sum_j^n \lambda_j \beta_j^\top S_j \beta_j$$

  where $L(\beta)$ is likelihood with respect to β and the λ_j control the amount of smoothing for each variable. **GCV** is used select the λ_j.

1.2 Additive Models Using the gam Package

We use data from 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, we will reduce the predictors to just **three**: temperature measured at E1 Monte, **temp**, inversion base height at LAX, **ibh**, and inversion top temperature at LAX, **ibt**. A number of cases with missing variables have been removed for simplicity. The data were first presented by Breiman and Friedman (1985).
(Note: This data set discussed in chapter-3 for Generalized additive model but here we will discuss in details for Additive models)

- **First we fit a simple linear model for**:

```
rm(list=ls())
library(faraway)
data(ozone)
######## Fit Simple linear regression model#########
olm <- lm(O3~temp+ibh+ibt,ozone)
```

O3 vs temp.
<2e-16 < 0.05
reject Ho: $\beta_1 = 0$
Ha: $\beta_1 \neq 0$

O3 vs ibt
do not reject
Ho: $\beta_3 = 0$
ibt has no relation with O3

O3 temp has +ve relation with O3

```
summary(olm)
..........
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
---
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
```

temp - β_1

> 0.05

Here, **ibt** is **not significan**t in this model that suggested to find right transforms on the predictors. Additive model can help here, let try it.

- **We fit an additive model with loess smoother on all three predictors using a Gaussian response.**

```
####fit an additive model with  loess smoother on all three predictors using
                           the a Gaussian response#####
library(gam)
amgam <- gam(O3 ~ lo(temp) + lo(ibh) + lo(ibt), data=ozone)
summary(amgam)

(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: NA

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 *** 0.001 ** 0.01 * 0.05 . 0.1   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 ***
```

$R^2 = 1 - \frac{5935.096}{21115.41}$

> 0.05

0.652×10 (. = A

0.652+A

```
#################  R^2 ##############
> 1-5935.096/21115.41
[1] 0.7189211
```

→ 0.652 → 0.7189211

Compared to the linear model, the R^2 of GAM is improved approximately 10 percentage . 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 $\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ for linear regression). 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**.

```
### Fit the model without the predictor of interest and then construct the F-test #####
> amgamr <- gam(O3 ~ lo(temp) + lo(ibh) , data=ozone)
>  anova(amgamr,amgam,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
```

H_0: model 1 = model2

H_a: model1 ≠ model2

>0.05

Although the p-value from F-test is still an approximation, we can see some evidence that ibt is not significant. **Lets 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=".")
```

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 degree** , 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.

1.3 Additive Models Using mgcv

Another method of fitting additive models is provided by the mgcv package of Wood (2000. We demonstrate its use on the same data. 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.

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

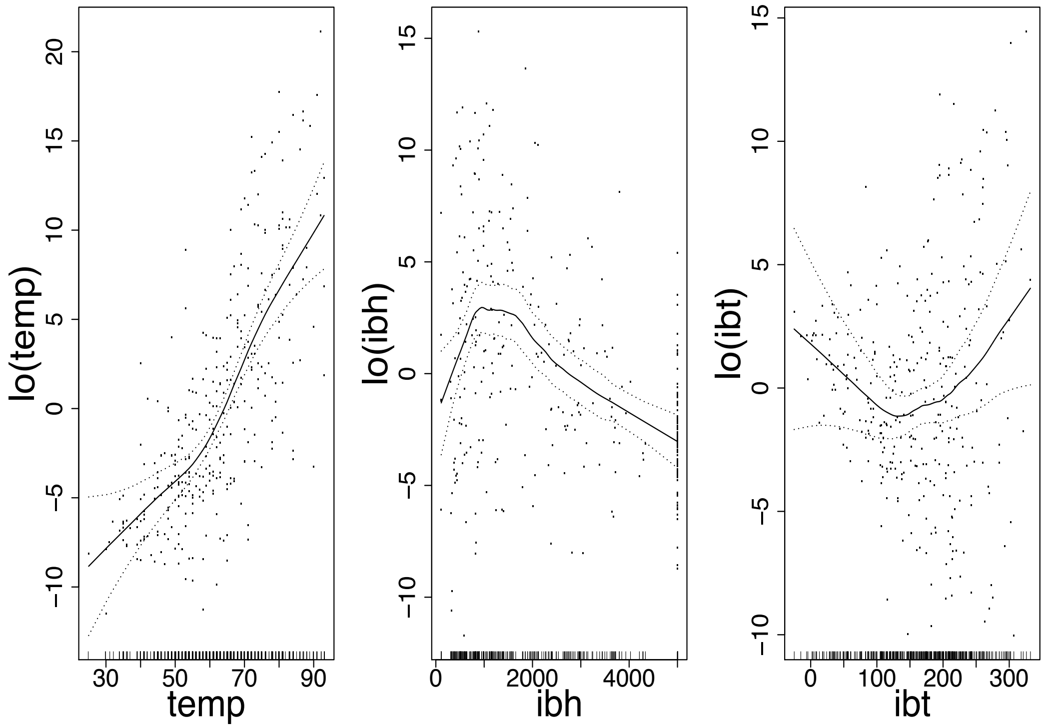

Figure 1: Transformations on the predictors chosen by the gam fit on the ozone data. Partial residuals and approximate 95 percentage pointwise confidence bands are shown.

```
Call: gam(formula = O3 ~ s(temp) + s(ibh) + s(ibt), data = ozone)
(Dispersion Parameter for gaussian family taken to be 18.7364)

    Null Deviance: 21115.41 on 329 degrees of freedom
Residual Deviance: 5939.441 on 316.9997 degrees of freedom
AIC: 1918.292

Number of Local Scoring Iterations: NA

Anova for Parametric Effects
           Df  Sum Sq Mean Sq  F value    Pr(>F)
s(temp)     1 12162.0 12162.0 649.1077 < 2.2e-16 ***
s(ibh)      1  1065.6  1065.6  56.8731 4.941e-13 ***
s(ibt)      1     2.0     2.0   0.1063    0.7446
Residuals 317  5939.4    18.7
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

> 0.05

Insignificant

Insignificant

```
Anova for Nonparametric Effects
```

```
              Npar Df Npar F     Pr(F)
(Intercept)
s(temp)             3 7.2980 9.543e-05 ***
s(ibh)              3 6.4328  0.000306 ***
s(ibt)              3 5.5788  0.000968 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

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**. Also, **see the similar pattern in fig-2**.

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

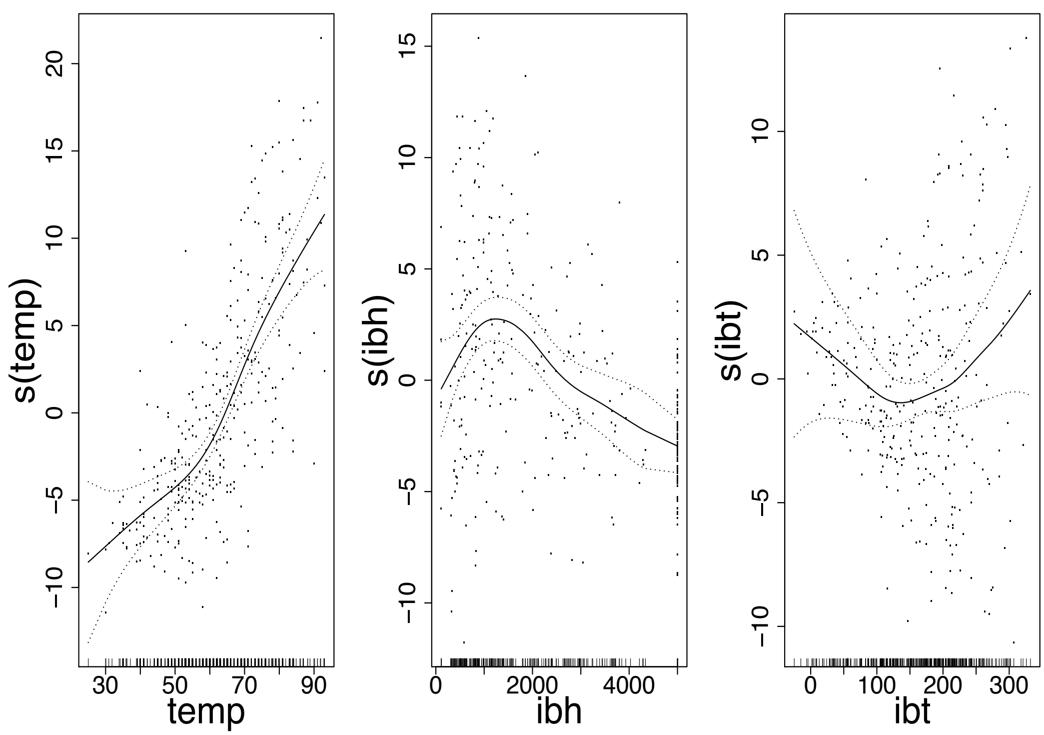

Figure 2: Transformation functions for the model fit by mgcv. Note how the same scale has been deliberately used on all three plots. This allows us to easily compare the relative contribution of each variable.

(y ~ x1 + x2)
{y ~ s(x1) + s(x2)}

- **Now, we 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 below test result confirms that there is really a change in the trend of temperature.**

```
> fitting a model with a linear term of temp and then make the F-test####
> 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    322.74       6950
2    319.11       6054 3.6237   895.98 13.109 3.146e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

reject ← Ho: model 1 = model 2

Accept Ha: model 1 ≠ model 2

< 0.05

reject Ho

The p-value is only approximate, but it certainly seems there really is a change in the trend.

Interaction b/w x_1 & x_2

- Check **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 using F-test. 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

```
> ###### bivariate transformations####
> 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 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
               edf Ref.df      F  p-value
s(temp,ibh) 6.346  8.040 14.860  < 2e-16 ***
s(ibt)      2.917  3.679  9.805 1.32e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

→ Significant

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

.

```
> ######## Use F-test ###########
>
> 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    316.93     5977.9
2    317.28     6123.6 -0.34665  -145.66 22.445 0.00113 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Insignificant

Significant

- One use for **additive models** is as an **exploratory tool** for **standard parametric regression modeling**. We can use the fitted functions to help usfind suitable simple transformations of the predictors. One idea here is to model the **temp** and **ibh** effects using **piecewise linear regression**(also known as **broken stick** or segmented regression). We define the **right** and **left** hockey-stick functions as below 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 previous analysis of figures above.

$\{ temp_i - 60 \; : temp_i > 60$; 0 otherwise $\}$

$f(x) = \{ x-c \; ; x > c$; 0 otherwise $\}$

```
######### Define the right and left hockey-stick functions#####
rhs <- function(x,c) ifelse(x > c, x-c, 0)
lhs <- function(x,c) ifelse(x < c, c-x, 0)
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)
```

$c=60$ $c=1000$

```
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 *** 0.001 ** 0.01 * 0.05 . 0.1   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
```

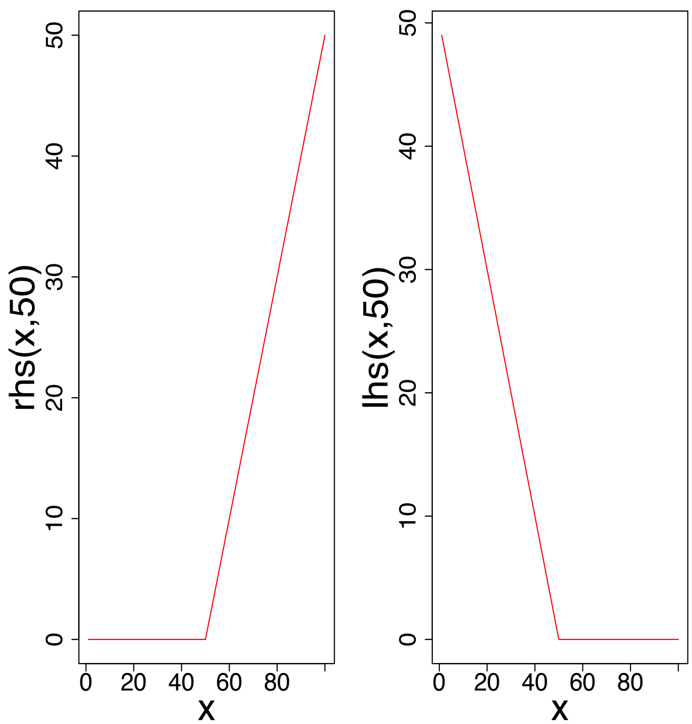

Figure 3:

Compared this model with the **first ordinary** linear regression model, the fit is **better** and about as good as the **additive Model** fit. It is unlikely for us to discover these transformation without the help of the intermediate additive models. Furthermore, the linear model has the advantage that we can write the prediction formula in a compact form. We can use additive models for building a linear model as above, but they can be used for inference in their own right. **For example, we can predict new values with standard error:**

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

Inside the original range of data

```
         1
11.01278

$se.fit
         1
0.9727755
```

Increase

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.

$f(x) = x^2, 0 \le x \le 3$

$3.5 \le x < 4$

```
 > ###########Outside of range data ######
> predict(ammgcv,data.frame(temp=150,ibh=2000,ibt=100),se=T)
$fit
         1
47.04228

$se.fit
         1
10.00962
```

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

```
############# examine the usual diagnostics #######
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="")
```

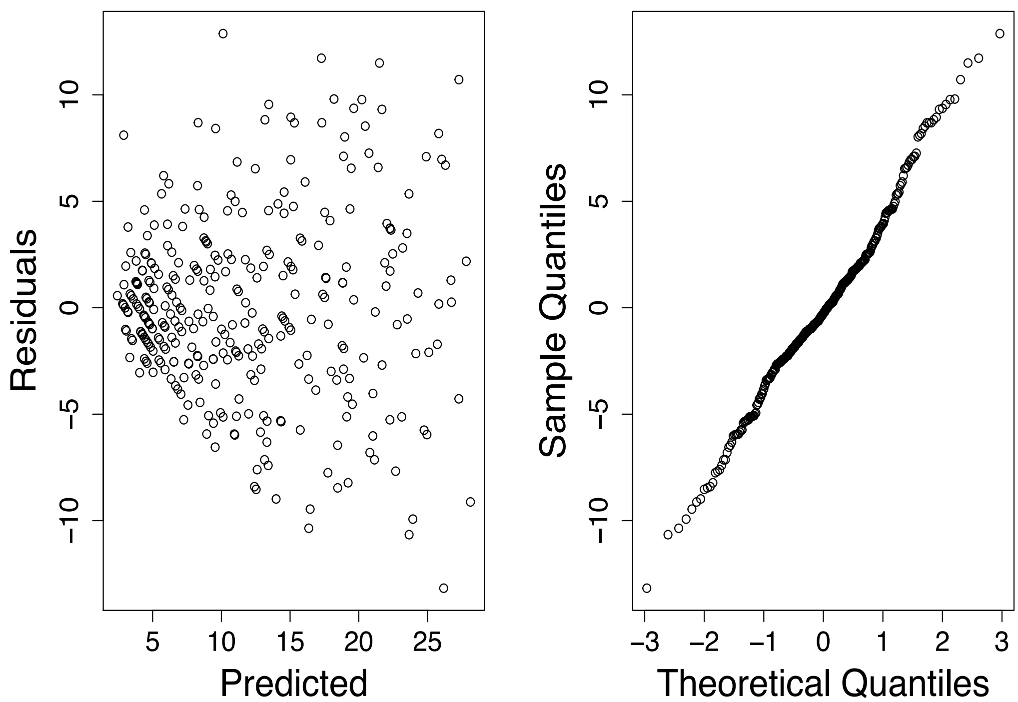

Figure 4: Residuals plots for the additive model.