Glimpse of Advanced Regression model
- Non parametric approach

By Suman Kapoor , Bio-Statistician 2, QuintilesIMS

Today we are going to talk about..

- Why we need non parametric approach
- Smoothing
- Spline
- GAM Using SAS
- Case study

Linear Model

Linearity Assumption: Normality assumption:

Linear regression equation

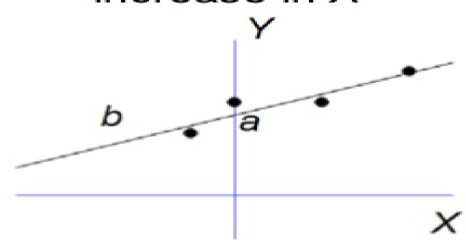
(without error)

$$\hat{Y} = bX + a$$

predicted values of *Y*

slope = rate of increase/decrease of Y hat for each unit increase in *X*

Y-intercept = level of *Y* when X is *0*.



Generalized Linear Model

- **Response Variable does not need to be normally distributed**
- **GLM does not assume a linear relationship, however it’s assume linear relationship between transformed response and the exploratory variable**

$$g(\mu) = \alpha + \beta_1 X_1 + \cdots + \beta_k X_k$$

$$g(\mu) = \mu$$

Identity link (form used in *normal* and *gamma* regression models):

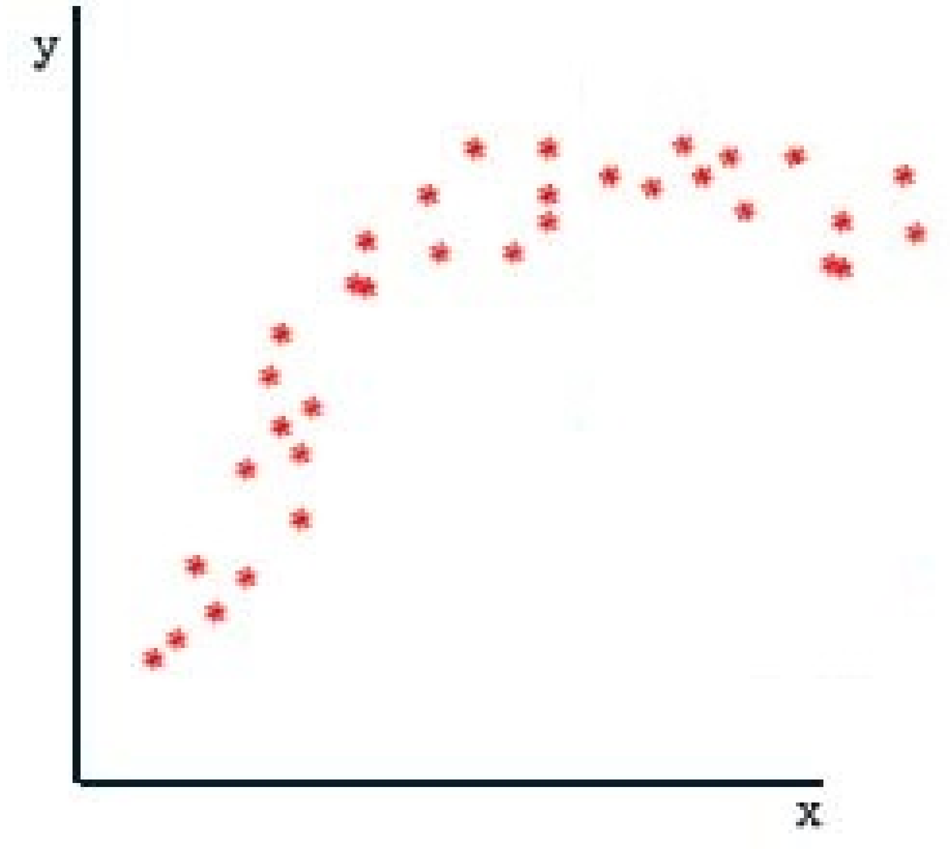
$$g(\mu) = \log(\mu)$$

Log link (used when μ cannot be negative as when data are *Poisson* counts):

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

Logit link (used when μ is bounded between 0 and 1 as when data are binary):

Solution 1: Non linear regression (Non linear least squares NLS)



Non linear relation ship

Jaw bone strength over the age of deer

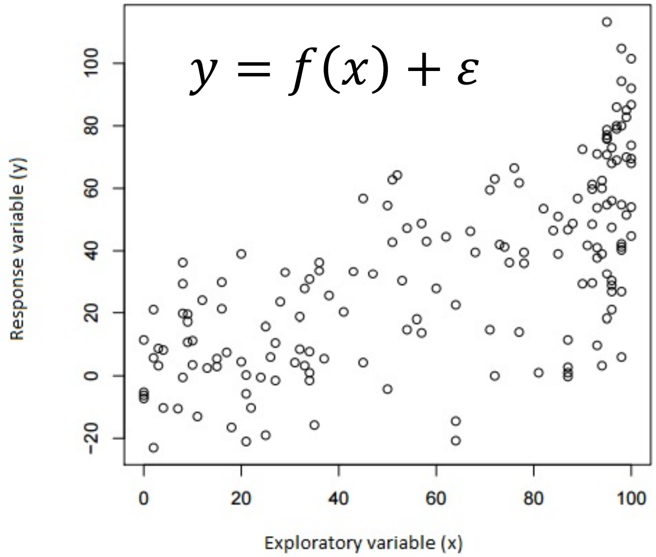
Advantage:
Estimate the parameters and standard errors of the parameters of specific non – linear equation from data

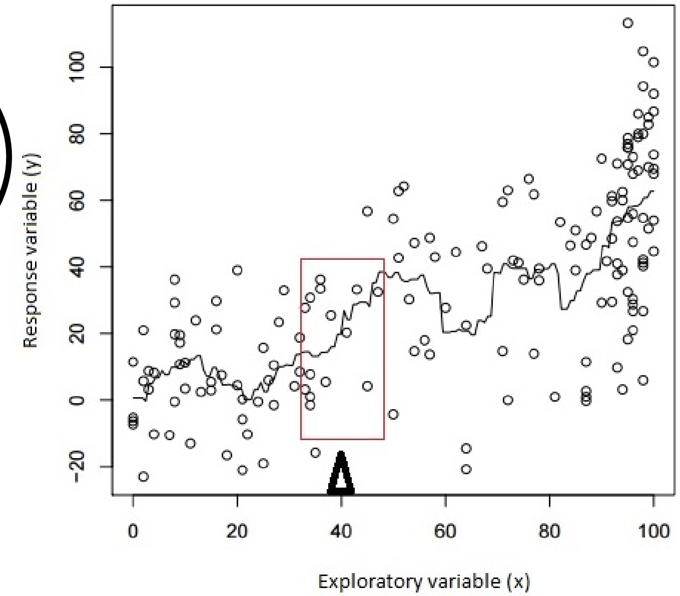
Disadvantage:
Function need to be specified prior

Have to Identify non - linear function early

Solution 2: Smoothing

Aim is to estimate “f”
Trend /Smooth

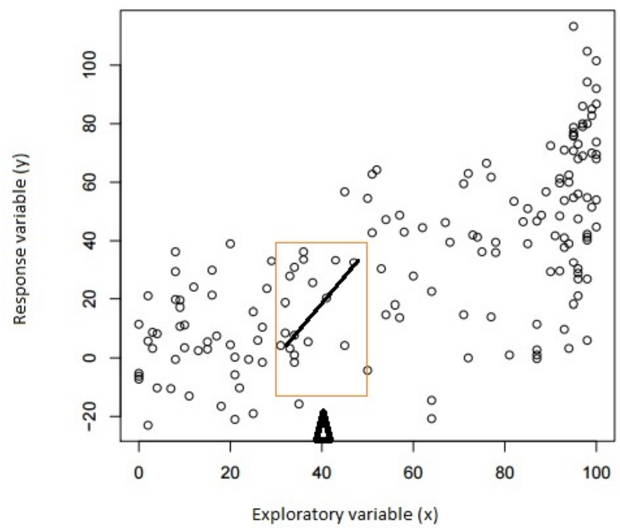




$$S(X_i) = \sum_{i \in N(X_i)} (y_i)/n_i$$

$$S(X_i) = \alpha_i + \beta_i x_i$$

Running line smoother:
Parameters are estimated using OLS based on N(xi) neighborhood of Xi.



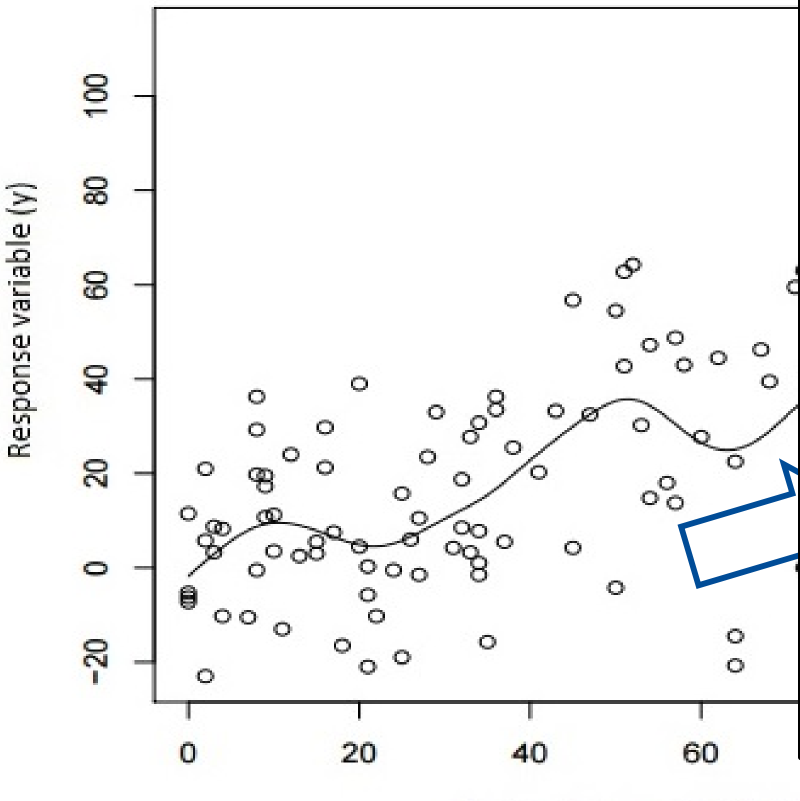
Running mean smoother :
N(xi) neighborhood with ni observation.
Symmetric neighborhood nearest 2k+1 points
Also called Moving average

Limitation: Use only for equally spaced points.
Weight can be given.

QuintilesIMS™

Smoothing..

LOWESS – Locally Weighte



- For each value X_i of X (Exploratory), a neighborhood (Window/Span) is defined
- First order or second order polynomials are fitted using weighted regression
- Values close to X_i get more weights than far from X_i.

$D_i = x_i - X / h$; X is target value , h is width , create weights using **tricube** function

$W_i = (1 - D_i^3)^3$

D is the distance to the target point expressed as a fraction

- Fitting process repeated several times

- Other than polynomial , smoothness is controlled by band width. If it's .25 , that means 25% of the data are used in each local regression.

Small SPAN : insufficient data near to X for accurate fit (Large variance)

Large SPAN: Over smoothed Resulting loss of information (Large BIAS)

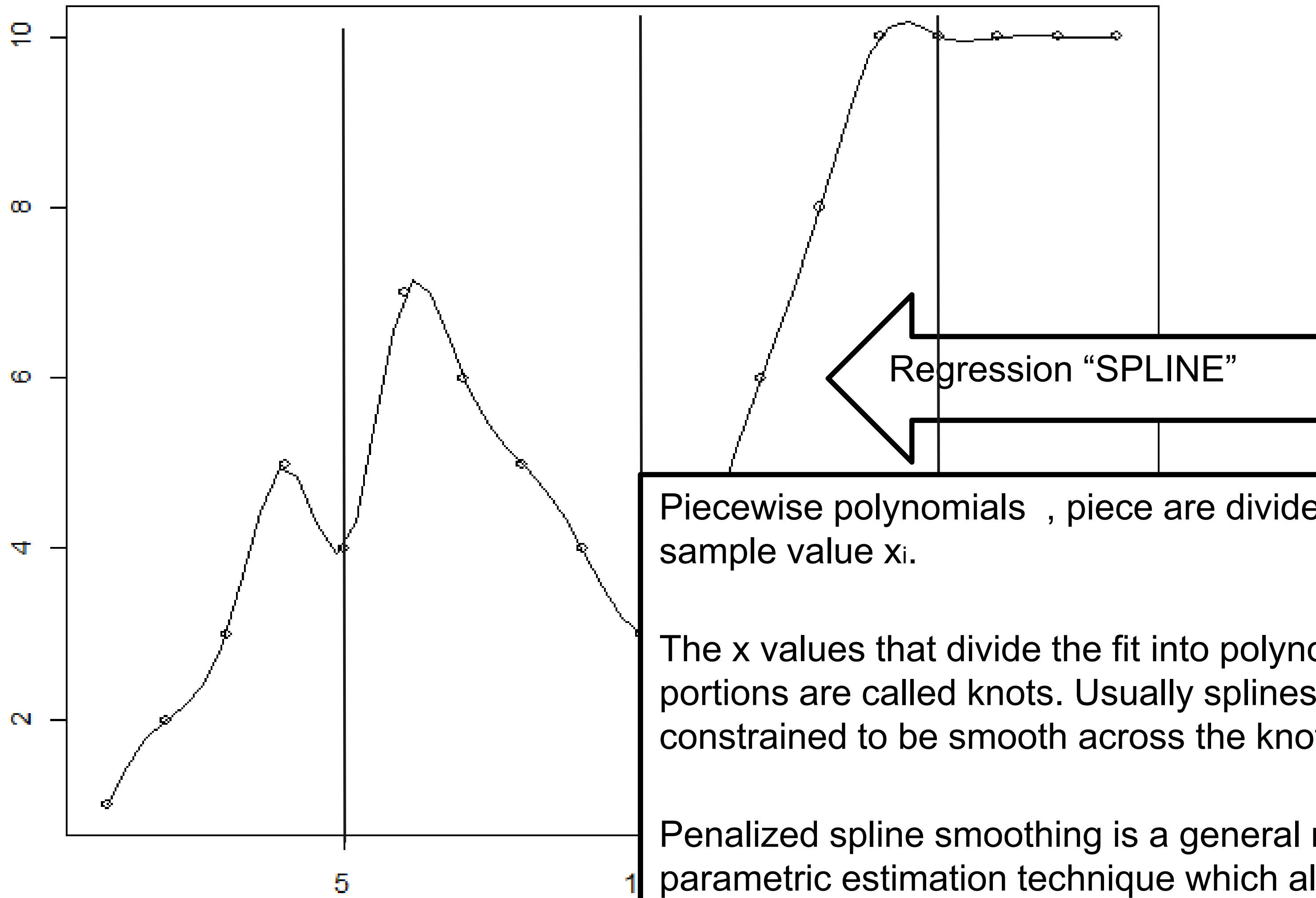
The shaded region indicates the window of values around the target value (arrow). A weighted linear regression (broken line) is computed, using weights given by the "tricube" function (dotted line). Repeating the process for all target values gives the solid curve.

Advantage and Disadvantage of LOWESS

- Does not require specific function to fit a model
- Need to pass only smoothing parameter and degree of local polynomial

- Fairly large data require and densely sampled data sets in order to produce good fit as it's relies on local data structure

Motivation…



Piecewise polynomials , piece are divided into sample value x_i.

The x values that divide the fit into polynomial portions are called knots. Usually splines are constrained to be smooth across the knots.

Penalized spline smoothing is a general non-parametric estimation technique which allows to fit smooth but else unspecified functions to empirical data

Smoothing Spline

$$\sum_{i=1}^{n}(y_i - s(x_i))^2 + \lambda \int (s''(x_i))^2 dx$$

Penalized sum of square [Sum of square of X + Penalty on X]

Spline function s(x)

Aim is to estimate unknown function s(.)

Penalty Value "Lamda" Programmatically need to estimate best "Lamda" / or provide manually. Control tradeoff.

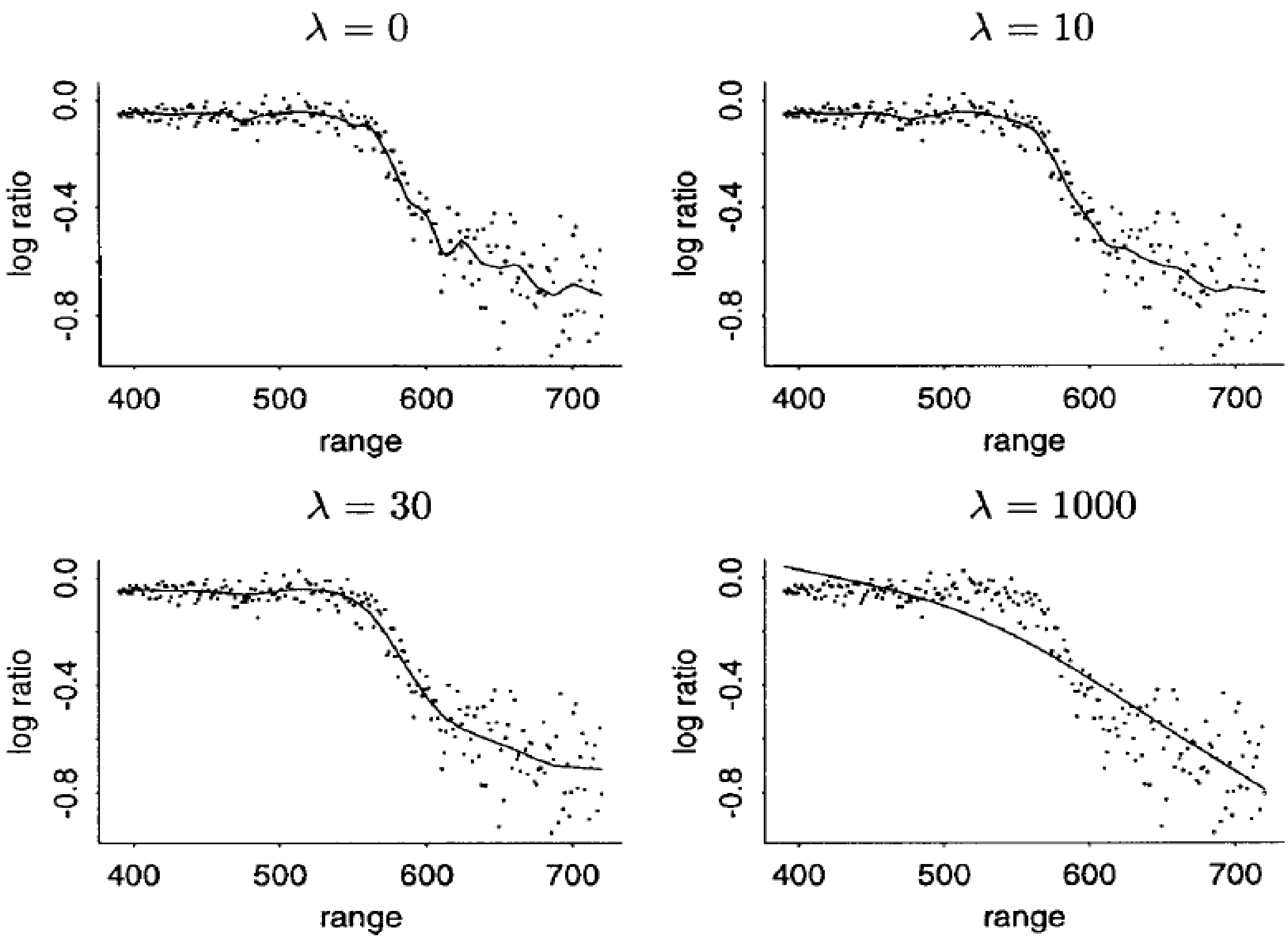
Second order derivative of the Spline function [measure of roughness of the fitted curve]

If $\lambda = 0$ then it's a standard least square , leads to join all the point
If $\lambda = \infty$ then then penalty term dominates , so the graph of optimal function is straight line

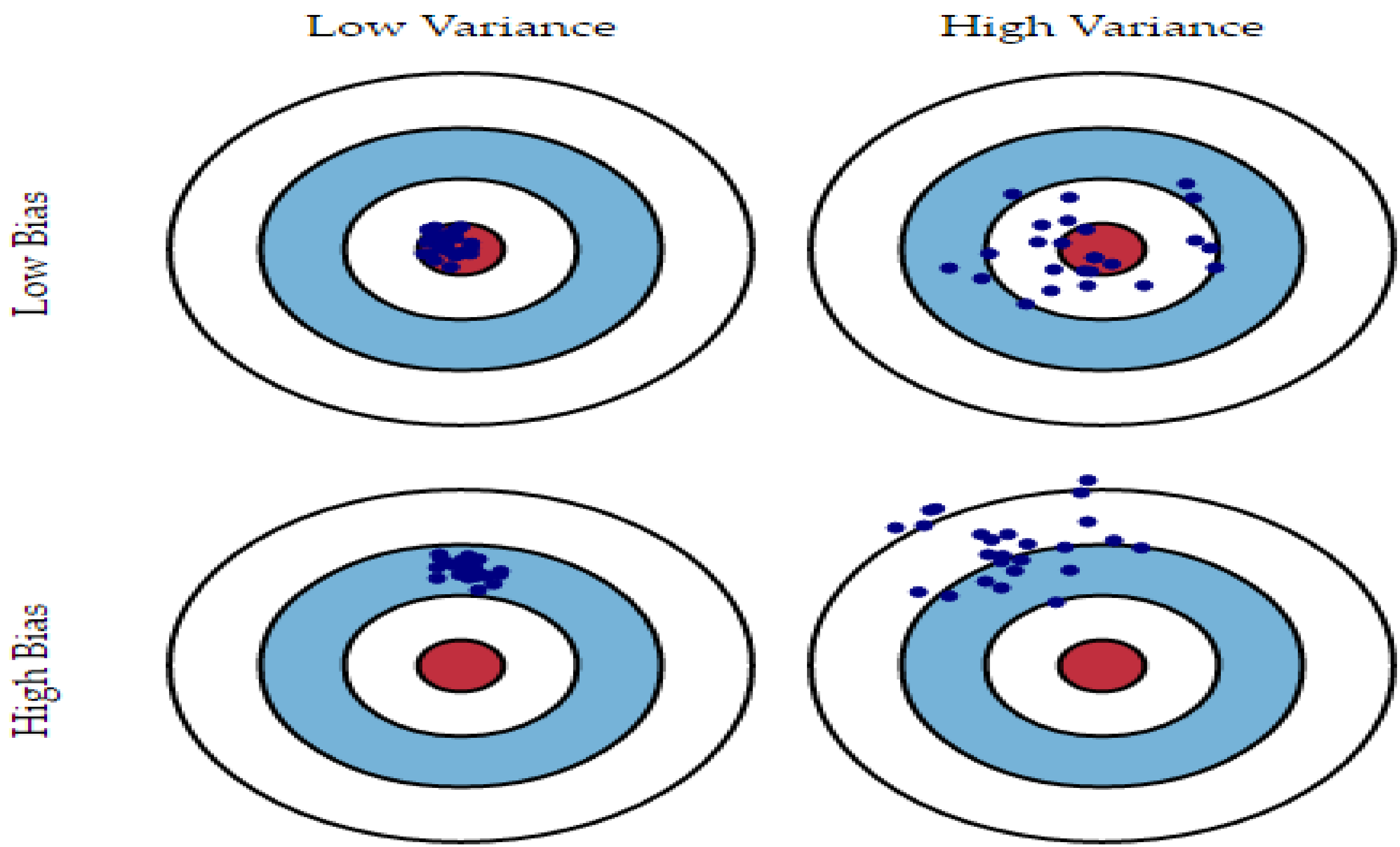
Usually determined indirectly through choice of **degrees of freedom**
Number of linear constraints

Most crucial steps of Generalized Additive Model is to select **DF**
Because: Smoothness and fit must be balanced

Characteristics of “Lamda”



Optimal “Lamda”



Optimal “Lamda” and how to find it..

- To find the Optimal amount of smoothing …

$$\mathrm{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Cross Validation:

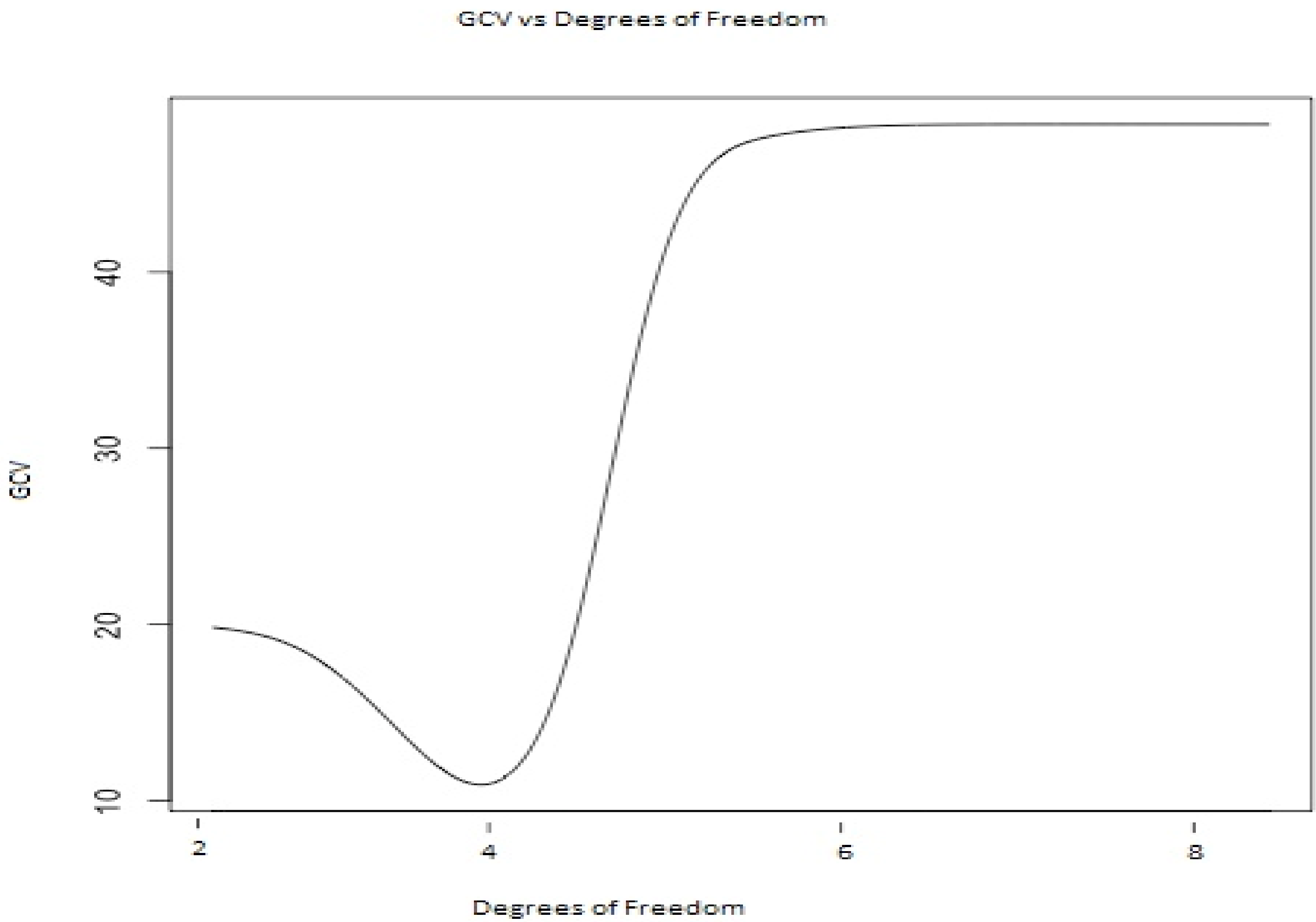
$$CV(\lambda) = \sum_{i=1}^{n} \{y_i - \hat{f}_{-1}(x_i;\lambda)\}^2$$

Spline fit without the (X_i, Y_i) point

1. Leave One observation out
2. Estimate smoothing curve using n-1 remaining observation
3. Predict the value at the omitted point Xi
4. Compare the predicted value at Xi with the real value Yi
5. The difference between the original and predicted value at Xi
6. Repeat the process
7. Adding up all squared residuals give cross validation error

Akaike’s Information Criterion (**AIC**) , Corrected AIC (**AICC**), Generalized cross Validation **(GCV)** (Matrix approach)

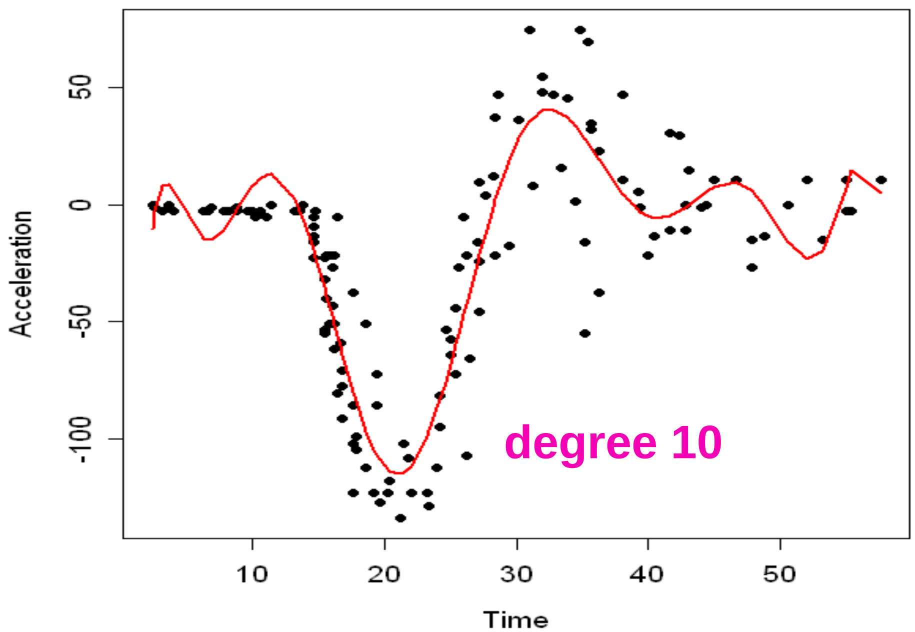
Plot



GCV vs Degrees of Freedom
GCV
10
20
30
40
2
4
6
8
Degrees of Freedom

Which Spline

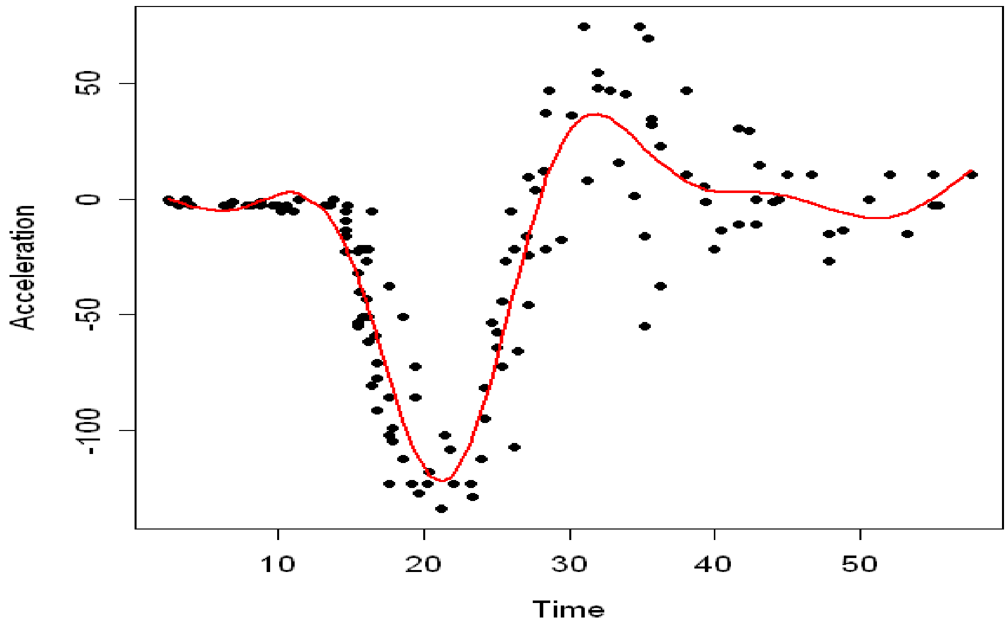
- Cubic Regression Spline (Special case of polynomial .. Third order)



- Third order polynomials
- Computationally very efficient

Which Spline..

- Thin Plate Regression Splines (default) **PROC GAMPL**



- Works well, lend to slightly more linear than cubic spline
- Computationally more costly to set up
- Recommended to use polynomial spline in case of huge data

B spline, natural spline, P spline are also present.

GAM

GLM:

$$g(y-\varepsilon_i)=\eta=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p$$

GAMs

Smoothing function of X_i

$$g(y-\varepsilon_i)=\eta=\beta_0+s(x_1)+s(x_2)+\ldots+s(x_p)$$

Case study: PROC GAMPL

Observational studies have suggested that low dietary intake or low plasma concentrations of **Beta-carotene**, or other carotenoids might be associated with increased risk of developing certain types of cancer.

A cross-sectional study , which want to investigate the relationship between personal characteristics and dietary factors, and plasma concentrations of **Beta-carotene**

Study subjects (N = 314) were patients who had an elective surgical procedure during a three-year period to biopsy or remove a lesion of the lung, colon, breast, skin, ovary or uterus that was found to be non-cancerous.

Aim is to identify factors which influence plasma concentrations of **Beta-carotene**

Variable Description:

plasma concentrations of Beta-carotene (Response)

Independent variables: (Continuous)

Age (Years)

QUETELET (weight/Height^2)

FIBER (Grams consumed/Day)

ALCOHOL (Number of alcoholic drinks consumed per week)

CHOLESTEROL consumed (mg per day)

BETADIET: Dietary beta-carotene consumed (mcg per day)

Independent variables: (Categorical)

Sex (1=Male, 2=Female).

Smoking status (1=Never, 2=Former, 3=Current Smoker)

Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No)

Analysis..

6.1.2.3 Model

General Linear Model: log beta versus Sex, Smoking status, Vitamin

Factor	Type	Levels	Values
Sex	fixed	2	1, 2
Smoking status	fixed	3	1, 2, 3
Vitamin	fixed	3	1, 2, 3

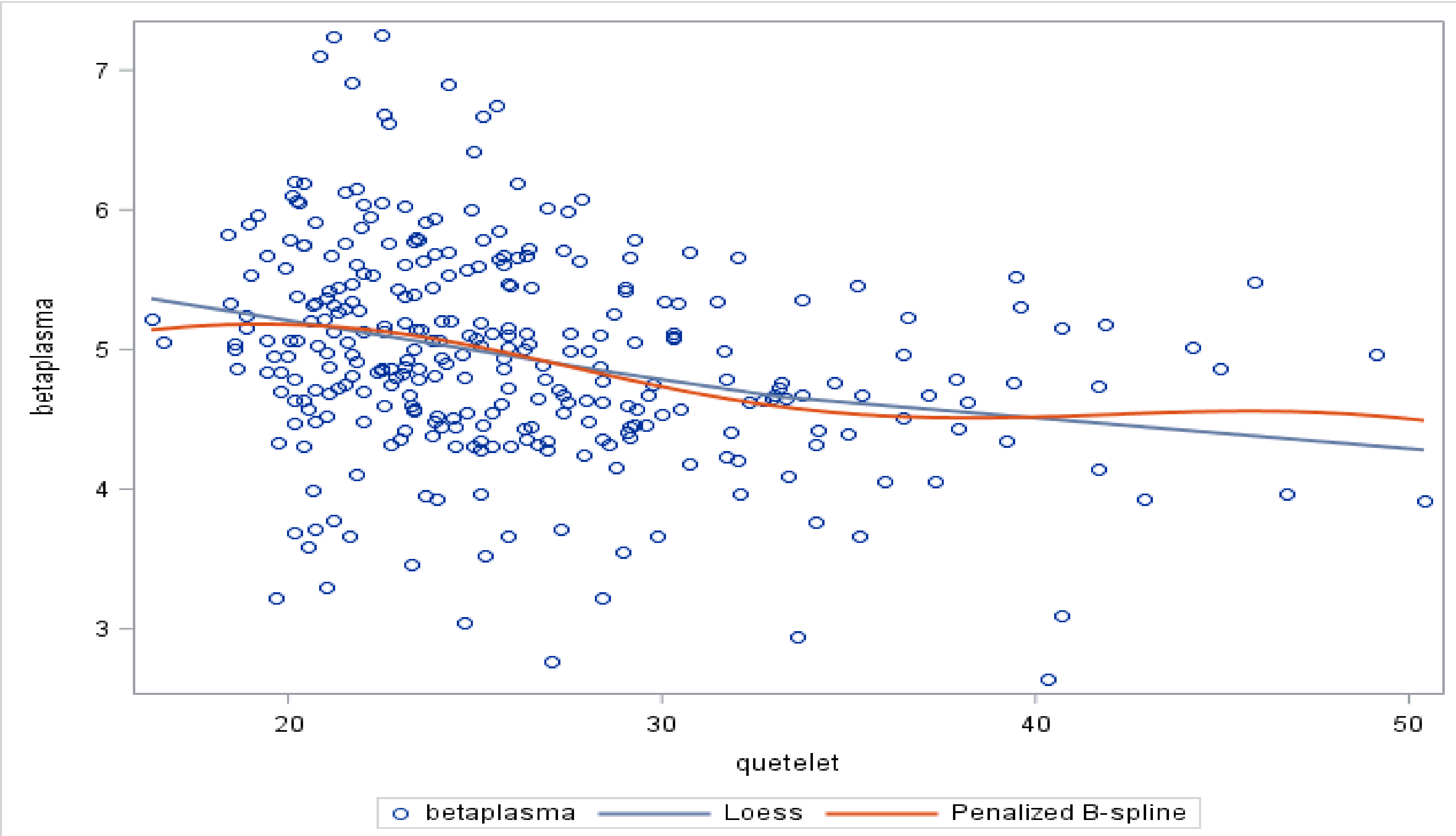
Analysis of Variance for log beta, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Age	1	0.40150	0.23897	0.23897	2.90	0.090
Sex	1	0.79890	0.17571	0.17571	2.13	0.145
Smoking status	2	0.66772	0.48616	0.24308	2.95	0.054
weight/(height^2)	1	2.63996	1.99036	1.99036	24.18	0.000
Vitamin	2	1.01660	0.06828	0.03414	0.41	0.661
CALORIES	1	0.00009	0.06235	0.06235	0.76	0.385
FAT	1	0.15930	0.00784	0.00784	0.10	0.758
FIBER	1	0.63890	0.18163	0.18163	2.21	0.139
ALCOHOL	1	0.01623	0.00839	0.00839	0.10	0.750
CHOLESTEROL	1	0.00013	0.02228	0.02228	0.27	0.603
BETADIET	1	0.26242	0.21037	0.21037	2.56	0.111
Error	287	23.62834	23.62834	0.08233		
Total	303	30.83984				

S = 0.286930 R-Sq = 23.38% R-Sq(adj) = 19.11%

Apply GAM – Diagnostics

Check relation of Response and each dependent variables



Apply GAM

```
ods graphics on;
proc gampl data=Plasma plots seed=12345;
  class smokstat vituse ;
  model betaplasma = spline(fiber) spline(betadiet)
spline(quetelet) spline(age) param(smokstat) spline(cholesterol)
param(vituse) spline(alcohol)  / dist= GAUSSIAN;
  id betaplasma;
output out=Predict;
run;
```

Response distribution

Result and Output

Model Information	
Data Source	WORK.PLASMA
Response Variable	betaplasma
Class Parameterization	GLM
Distribution	Normal
Link Function	Identity
Fitting Method	Performance Iteration
Fitting Criterion	GCV
Optimization Technique for Smoothing	Newton-Raphson
Random Number Seed	12345

Number of Observations Read	314
Number of Observations Used	314

Class Level Information		
Class	Levels	Values
smokstat	3	Curren Former Never
vituse	3	1 2 3

The performance iteration converged after 11 steps.

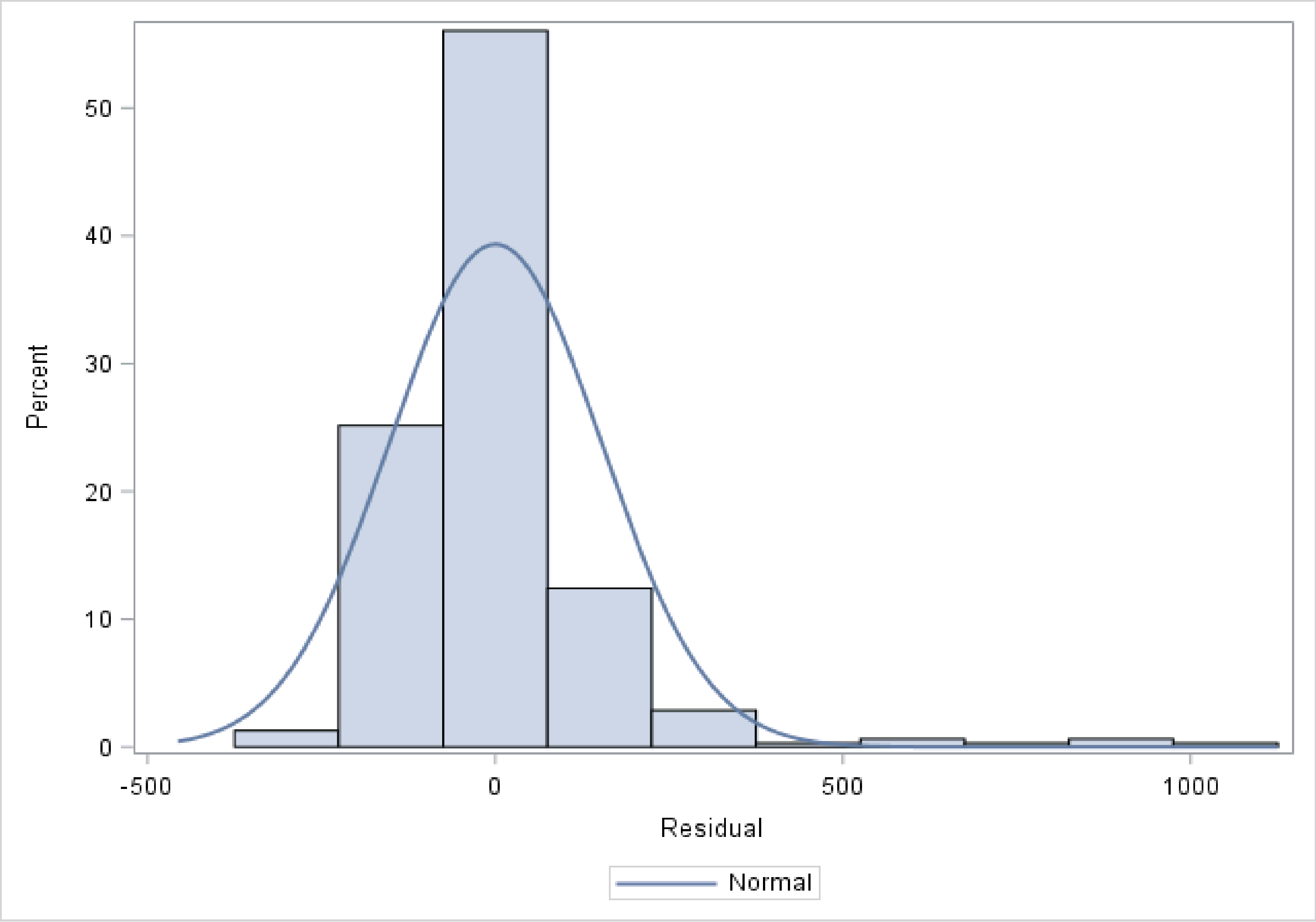
Fit Statistics	
Penalized Log Likelihood	-54245
Roughness Penalty	104444
Effective Degrees of Freedom	34.22618
Effective Degrees of Freedom for Error	278.14907
AIC (smaller is better)	4114.10717
AICC (smaller is better)	4122.75689
BIC (smaller is better)	4242.43458
GCV (smaller is better)	28856

Result and Output

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	156.781992	18.207986	74.1427	<.0001
smokstat Curren	1	-56.189485	28.409770	3.9118	0.0479
smokstat Former	1	-4.971768	19.890133	0.0625	0.8026
smokstat Never	0	0	.	.	.
vituse 1	1	81.248067	21.227798	14.6493	0.0001
vituse 2	1	45.650483	23.483014	3.7791	0.0519
vituse 3	0	0	.	.	.
Dispersion	1	23064	32605		

Tests for Smoothing Components				
Component	Effective DF	Effective DF for Test	F Value	Pr > F
Spline(fiber)	6.76129	8	22.43	<.0001
Spline(betadiet)	8.00000	8	19.22	<.0001
Spline(quetelet)	1.00000	1	16.98	<.0001
Spline(age)	3.46489	5	9.86	<.0001
Spline(cholesterol)	8.00000	8	11.79	<.0001
Spline(alcohol)	1	1	0.23	0.6352

Residual Analysis

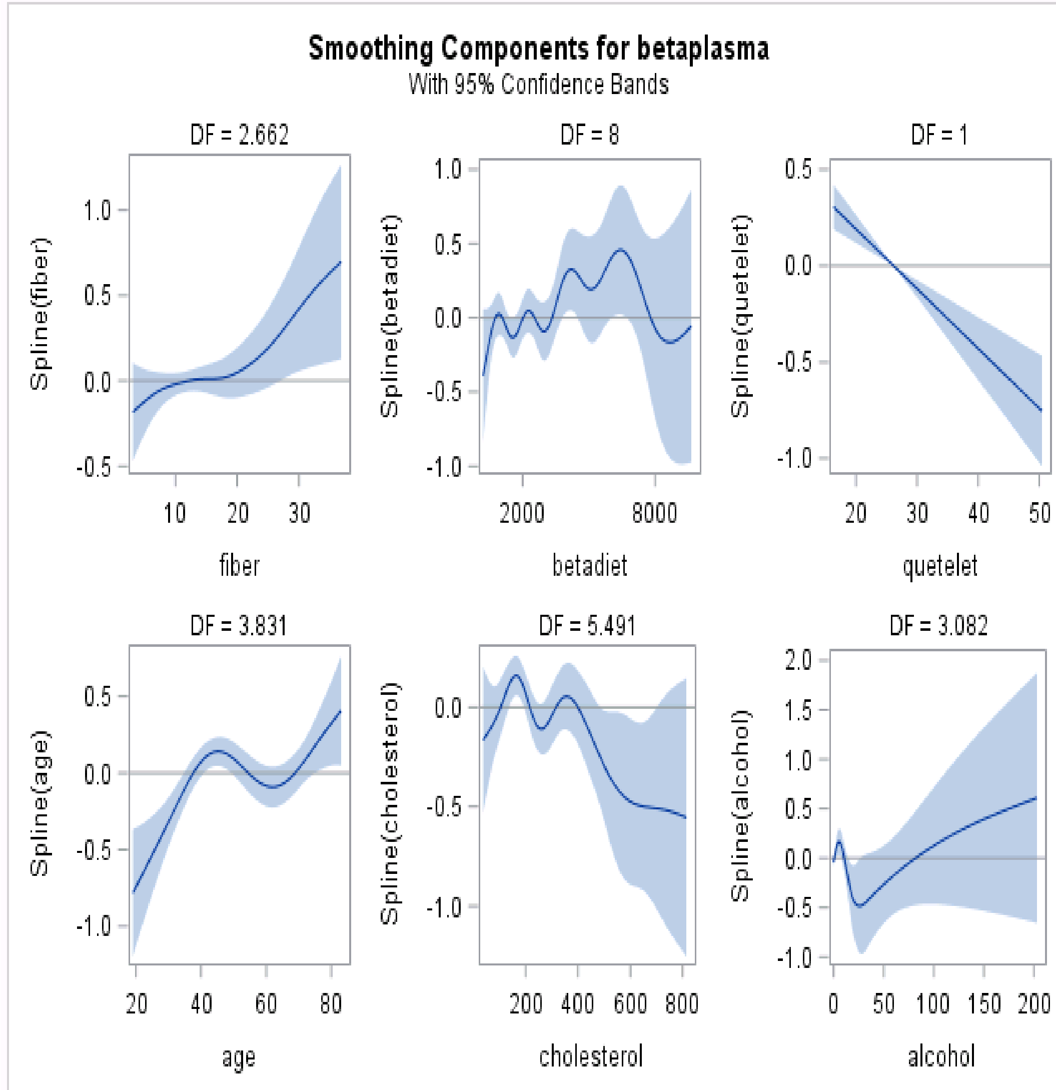


Transformation - LOG

Fit Statistics	
Penalized Log Likelihood	-285.67279
Roughness Penalty	3.58378
Effective Degrees of Freedom	30.06563
Effective Degrees of Freedom for Error	280.74815
AIC (smaller is better)	627.89306
AICC (smaller is better)	634.49535
BIC (smaller is better)	740.62092
GCV (smaller is better)	0.43368

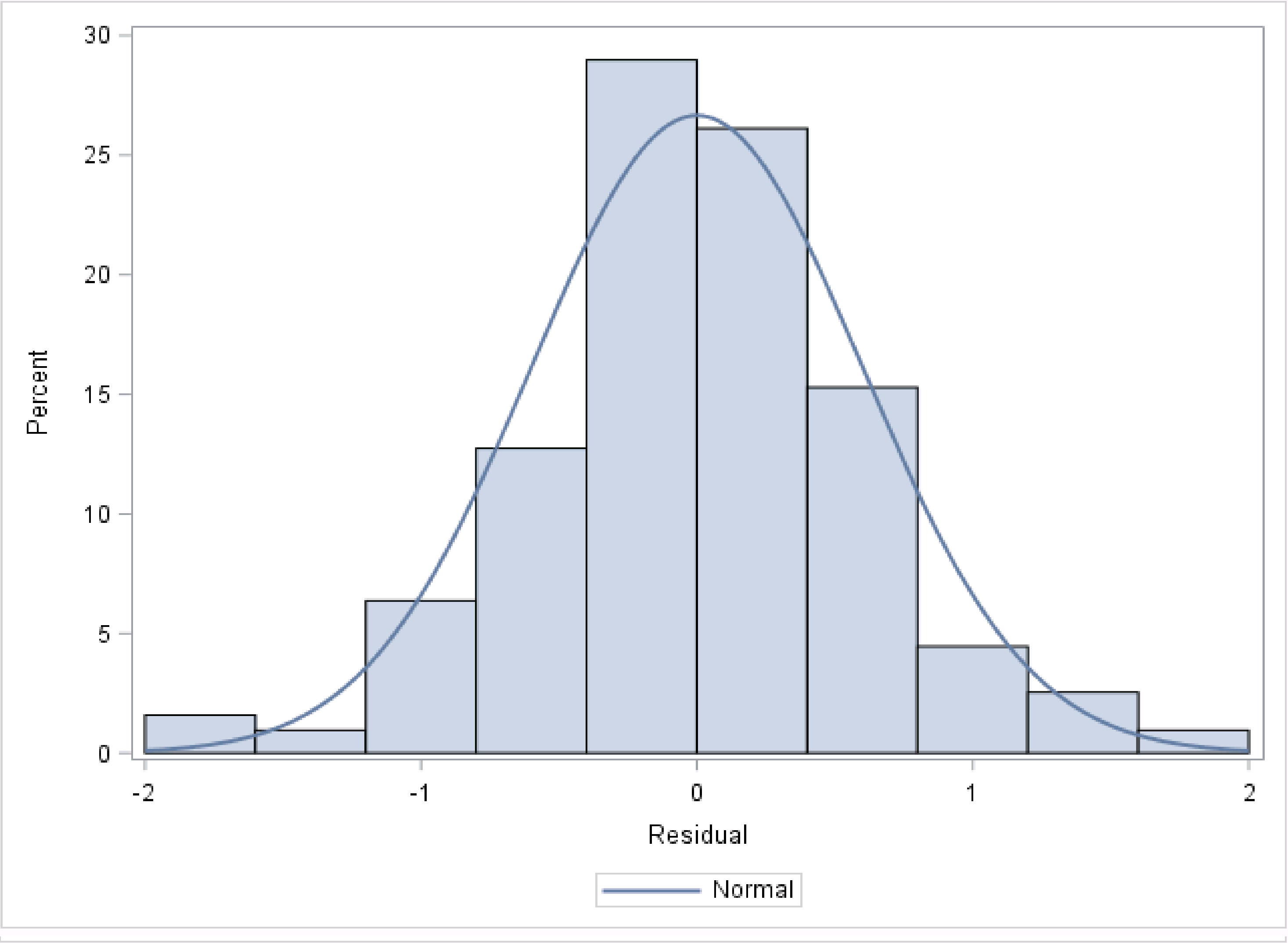
Tests for Smoothing Components

Component	Effective DF	Effective DF for Test	F Value	Pr > F
Spline(fiber)	2.66236	4	6.52	<.0001
Spline(betadiet)	8.00000	8	12.03	<.0001
Spline(quetelet)	1.00000	1	26.51	<.0001
Spline(age)	3.83059	5	26.96	<.0001
Spline(cholesterol)	5.49054	7	16.28	<.0001
Spline(alcohol)	3.08214	4	11.58	<.0001



QuintilesIMS™

Residual Analysis After Log transformation:



Summary

Can be use both liner and non linear exploratory variables in the model

When GLM fail to predict, GAM can be used

Discussion..

Thank you so much