Applied Statistical Regression
HS 2010 – Week 13

Marcel Dettling

Institute for Data Analysis and Process Design

Zurich University of Applied Sciences

marcel.dettling@zhaw.ch

http://stat.ethz.ch/~dettling

ETH Zürich, December 20, 2010

Non-Parametric Regression

Given fixed predictor values $x_1, ..., x_n$, we observe responses $y_1, ..., y_n$ with the relation:

$$y_i = f(x_i) + \varepsilon_i \text{, for all } i = 1, ..., n$$

What is unknown?

- errors ε_i : we require iid property, zero mean, constant variance
- functional relation $f(\cdot)$

→ $f(\cdot)$ was parametric so far. This was a very versatile tool, **see the blackboard for some examples…**

Parametric or Non-Parametric?

Advantages of parametric models:

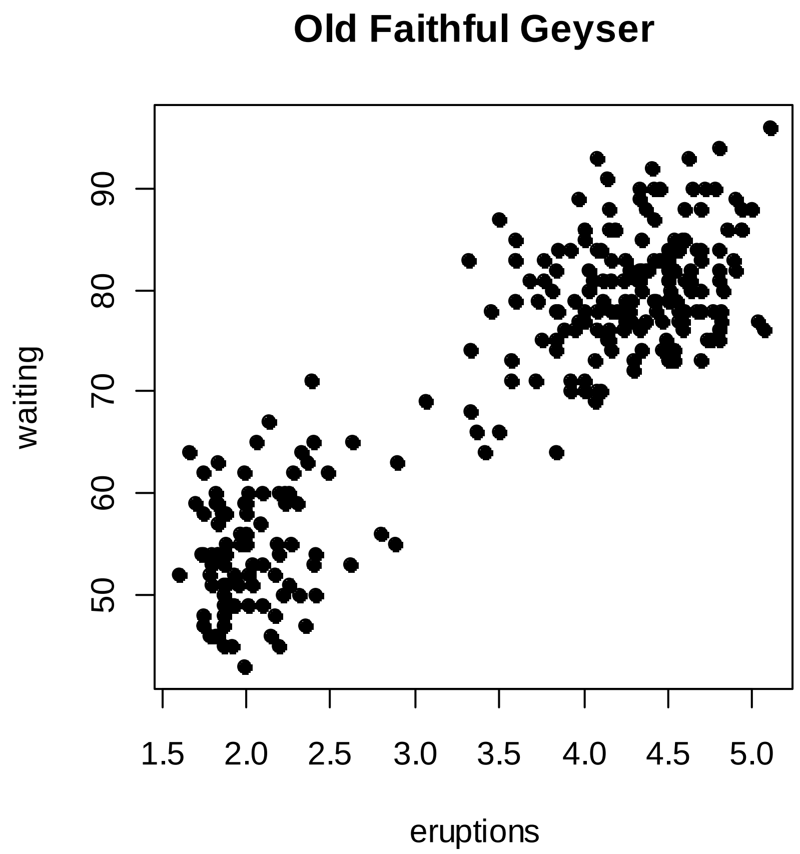
- Parametric models are more efficient
- Clear formulae make for clear interpretation
- Formal inference is possible
- Prediction/interpolation is possible

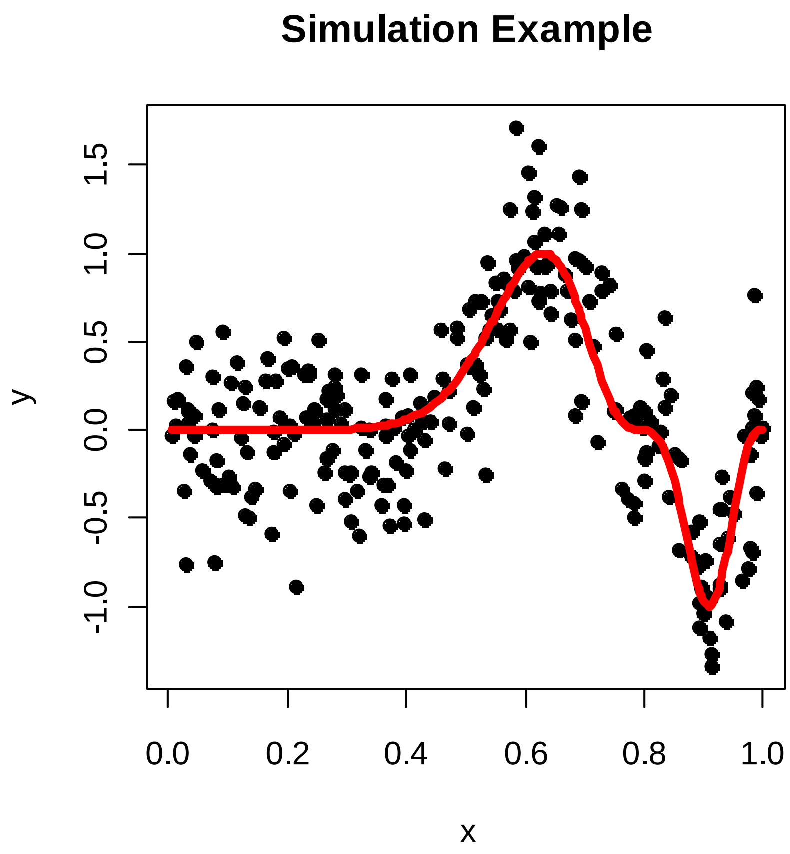
Advantages of non-parametric models:

- Flexibility, no prior knowledge required
- Less assumptions, less prone to bad mistakes

Applied Statistical Regression
HS 2010 – Week 13

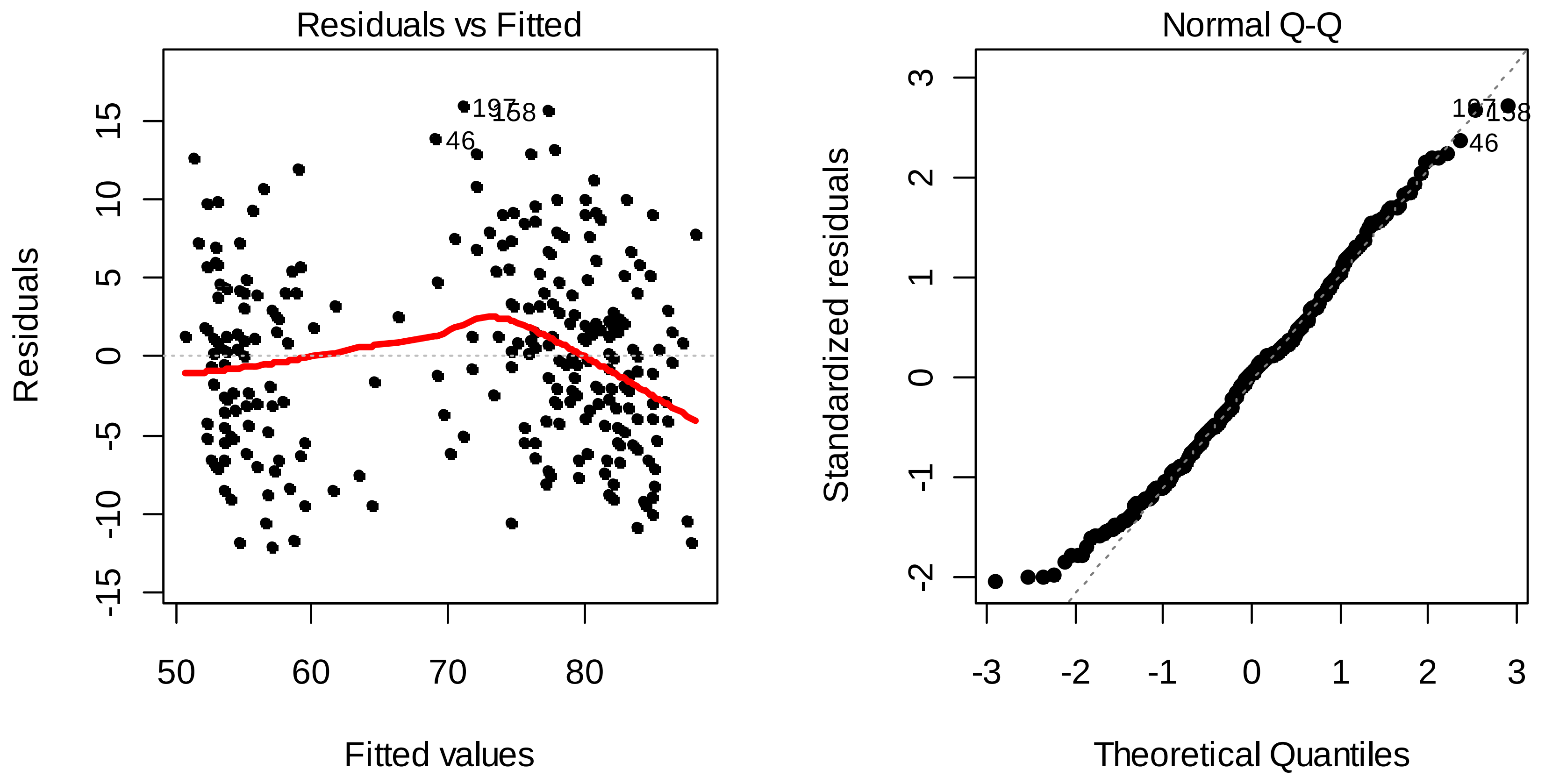
Examples





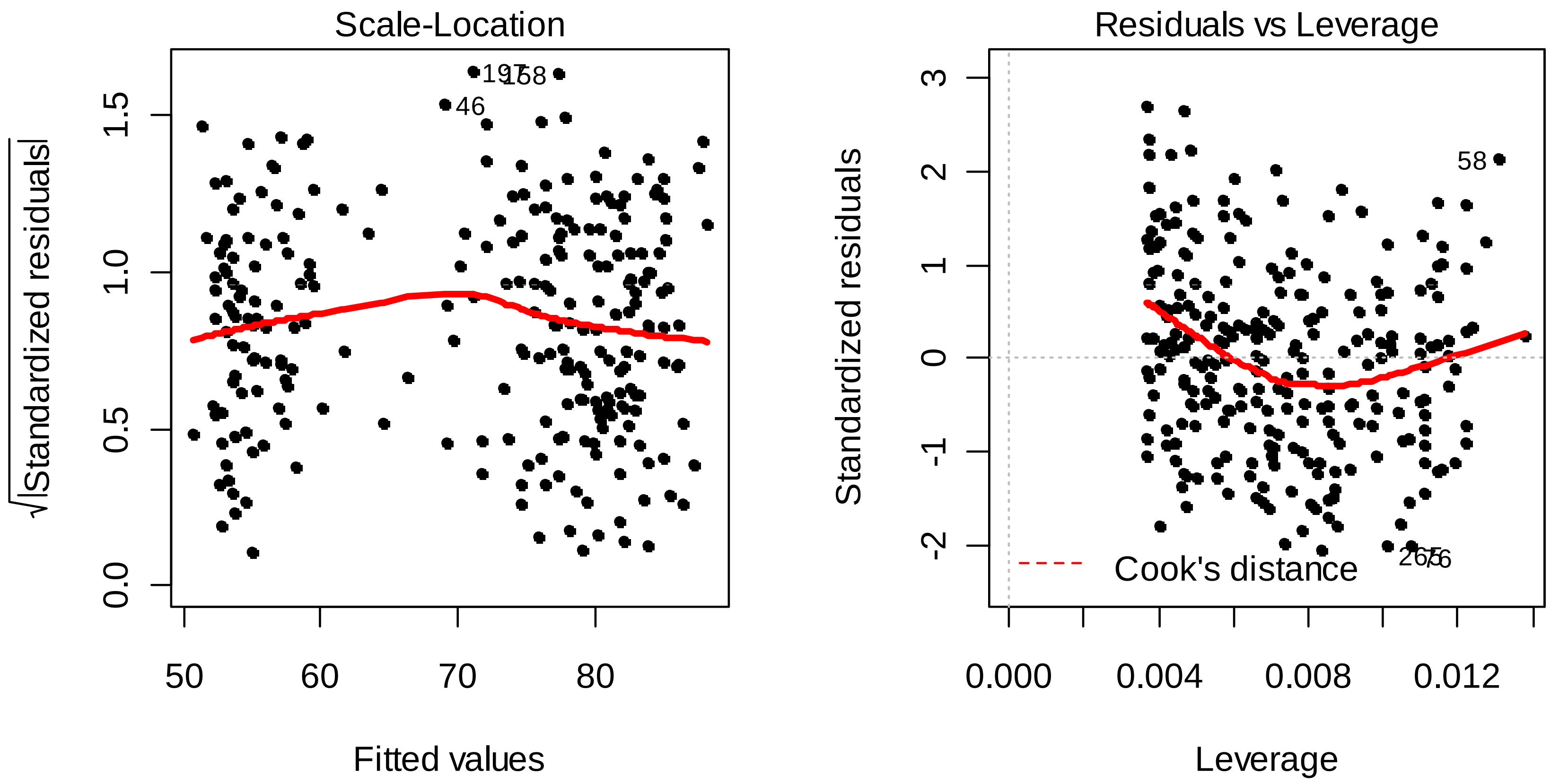
Applied Statistical Regression
HS 2010 – Week 13

Linear Model for Old Faithful?



Applied Statistical Regression
HS 2010 – Week 13

Linear Model for Old Faithful?



Kernel Smoothers

Kernel smoothing = weighted averaging of y-values over a fixed size window of x-values.

The estimate of $f(\cdot)$, denoted by $\hat{f}_\lambda(\cdot)$ is defined as:

$$\hat{f}_\lambda(x) = \frac{1}{n}\sum_{j=1}^{n} w_j Y_j \text{ with weights } w_j = \frac{1}{\lambda}\cdot K\left(\frac{x-x_j}{\lambda}\right)$$

- For the kernel, we require $\int K = 1$
- We can have rectangular kernels, Gaussian kernels, …
- λ, called the bandwith, is the smoothing parameter

Nadaraya-Watson Kernel Estimator

If the predictor values are spaced very unevenly, the general Kernel estimator can yield poor results. This problem can be mitigated somewhat by the ***Nadaraya-Watson estimator***:

$$\hat{f}_\lambda(x) = \frac{\sum_{j=1}^{n} w_j Y_j}{\sum_{j=1}^{n} w_j}$$

This estimator is a modified version of the kernel estimator. Its advantage is that the weights for the fitted value at each observation x_i will sum up to one.

Applied Statistical Regression
HS 2010 – Week 13

Choosing the Kernel

We require that the kernel is:

- smooth
- compact
- easy to compute

A good and popular choice is the ***Epanechnikov kernel***:

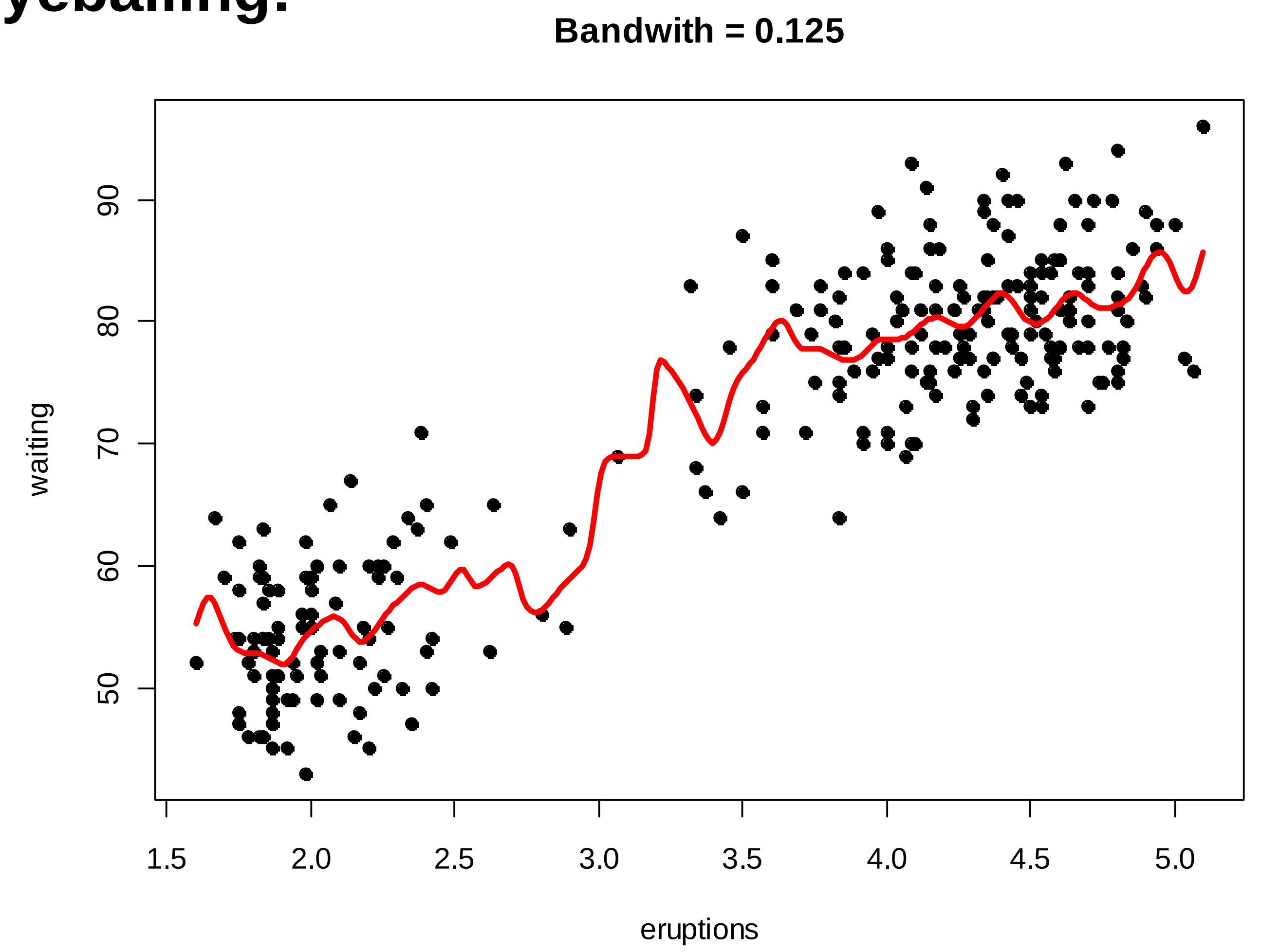
$$K(x) = \begin{cases} \frac{3}{4}(1-x^2), & if \ |x|<1 \\ 0 & else \end{cases}$$

But smoothing usually is not too dependent on the kernel…

Applied Statistical Regression
HS 2010 – Week 13

Choice of the Bandwith

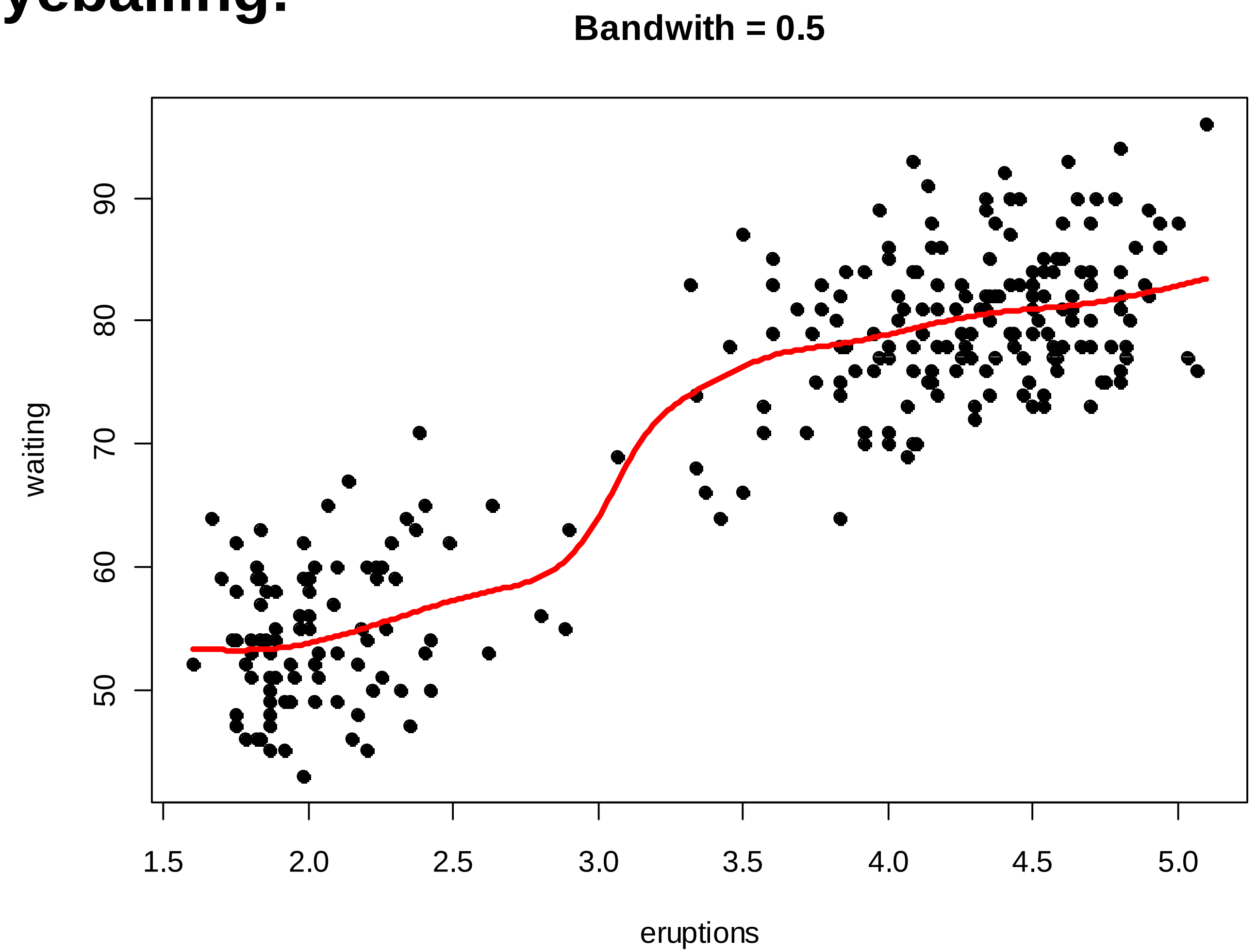
By eyeballing:



Applied Statistical Regression
HS 2010 – Week 13

Choice of the Bandwith

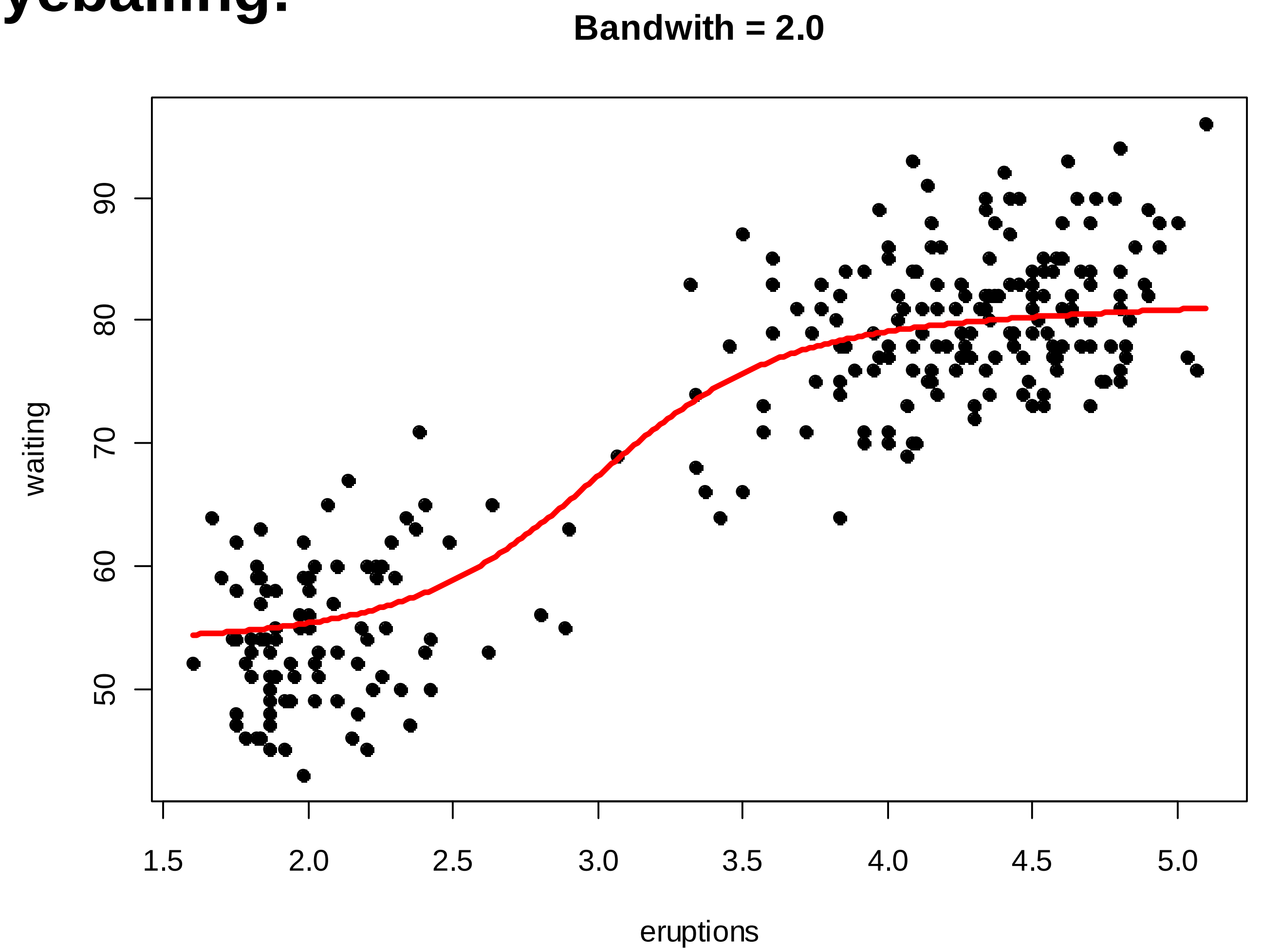
By eyeballing:



Applied Statistical Regression
HS 2010 – Week 13

Choice of the Bandwith

By eyeballing:



Choice of the Bandwith

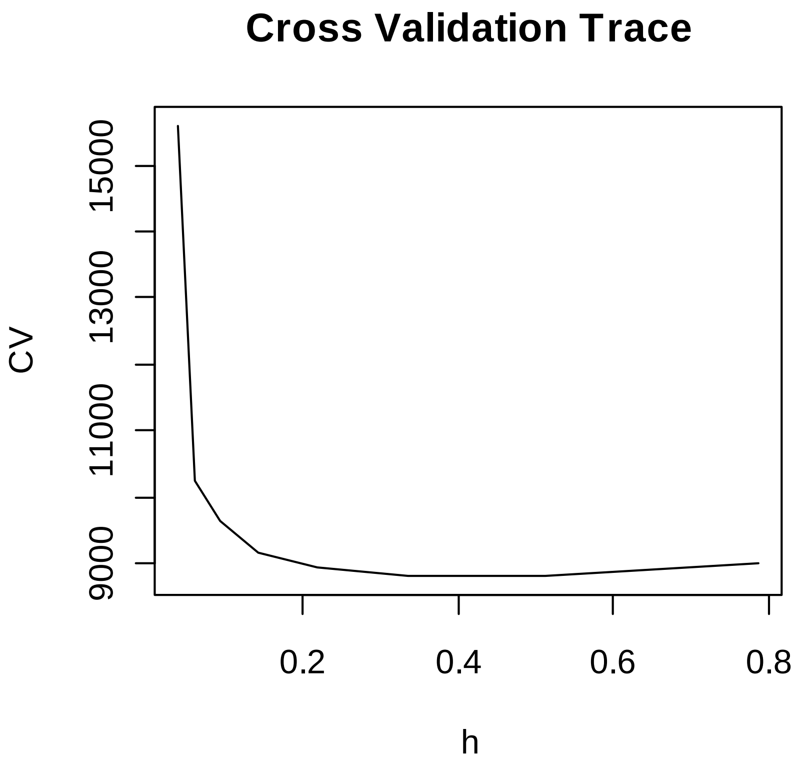
By cross validation:

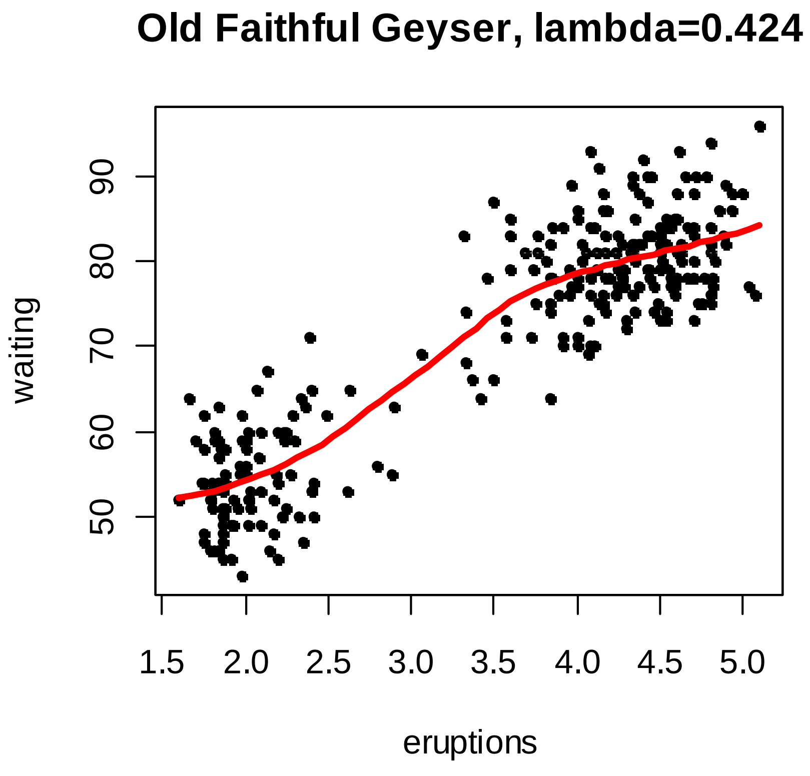
$$CV(\lambda) = \frac{1}{n}\sum_{j=1}^{n}(y_i - \hat{f}_{\lambda(j)}(x_j))^2$$

where $\hat{f}_{\lambda(j)}(\cdot)$ is the fit that is obtained when the j^{th} data point was omitted from the fitting process. Thus, we fit *j* smoothers and for each *j*, we compute the discrepancy between the fit for x_i and the observed response y_j. Of course, this needs to be done for a set of candidate λ that may seem suitable according to some eyeballing.

Choice of the Bandwith

By cross validation:

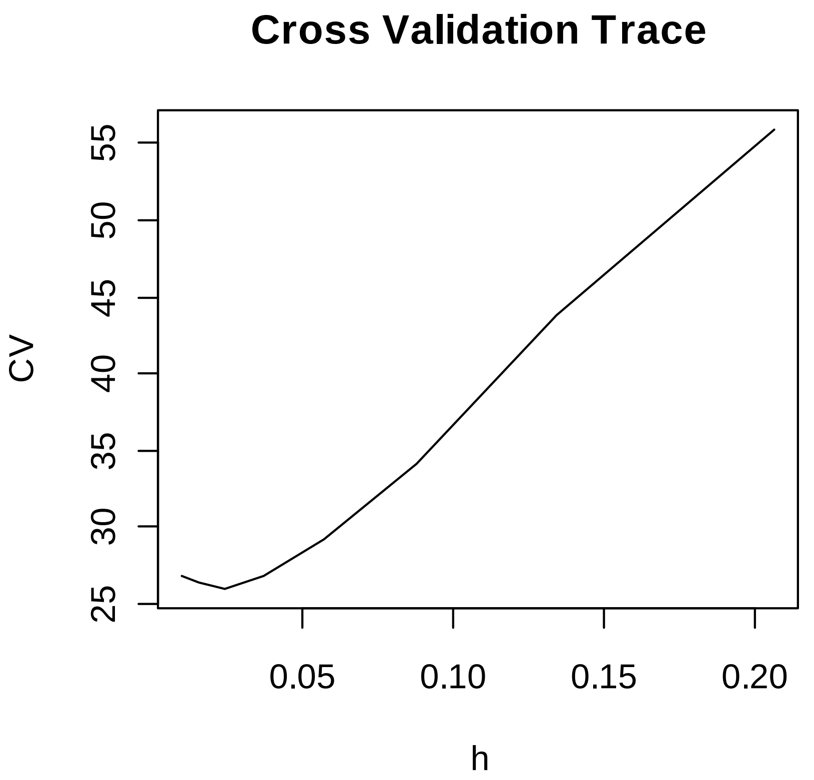


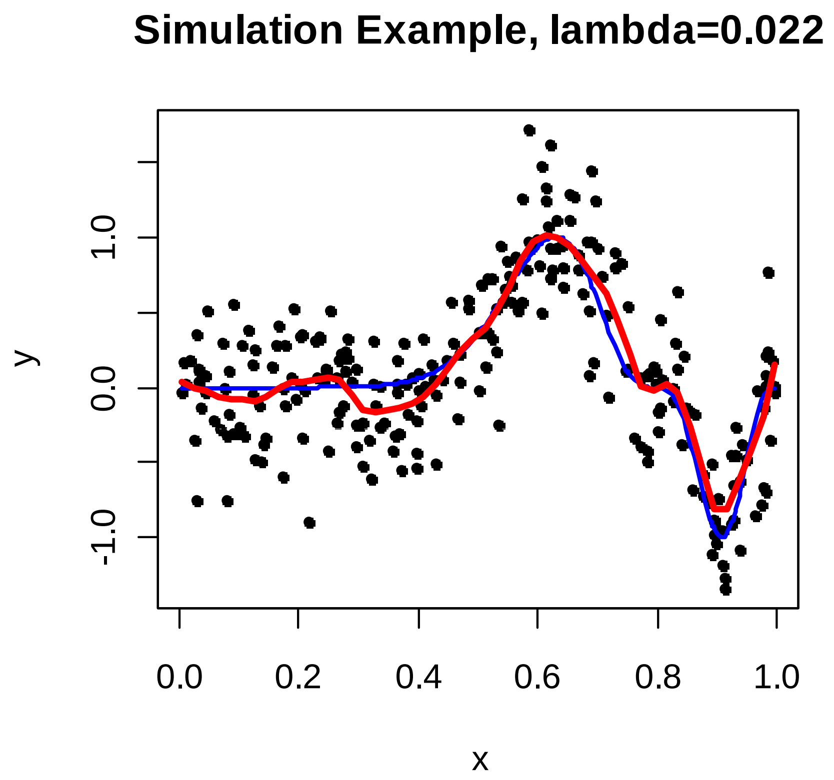


Applied Statistical Regression
HS 2010 – Week 13

Choice of the Bandwith

By cross validation:





Smoothing Splines

The basic notion behind the non-parametric regression is that there is that the relation between predictor and response is:

$$Y_i = f(x_i) + \varepsilon_i$$

The goal is now to minimize the sum of squared errors. This requires some additional penalty on the smoothness of $f(\cdot)$

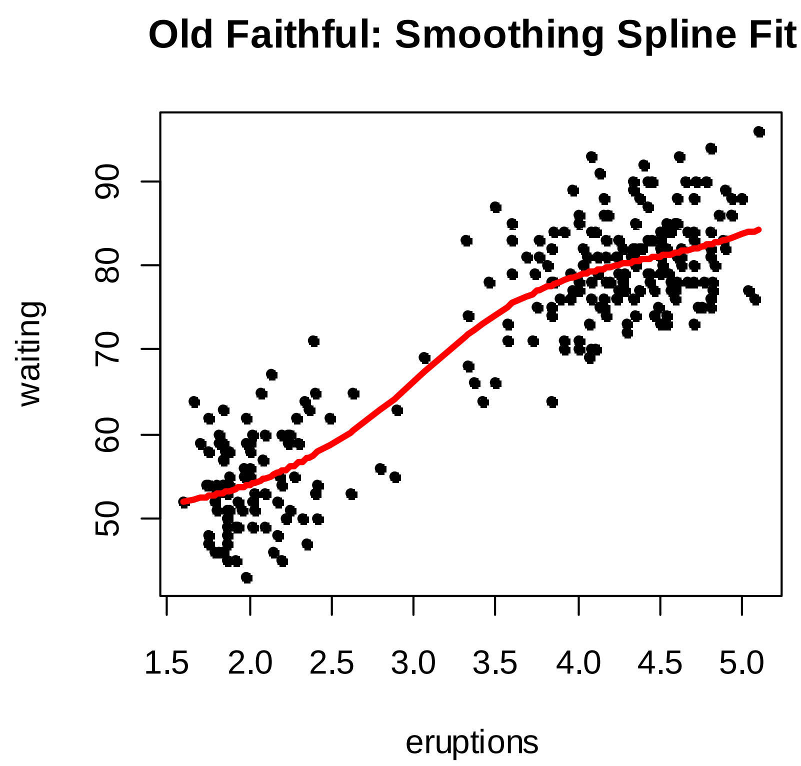
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \int \left(f''(x) \right)^2 dx$$

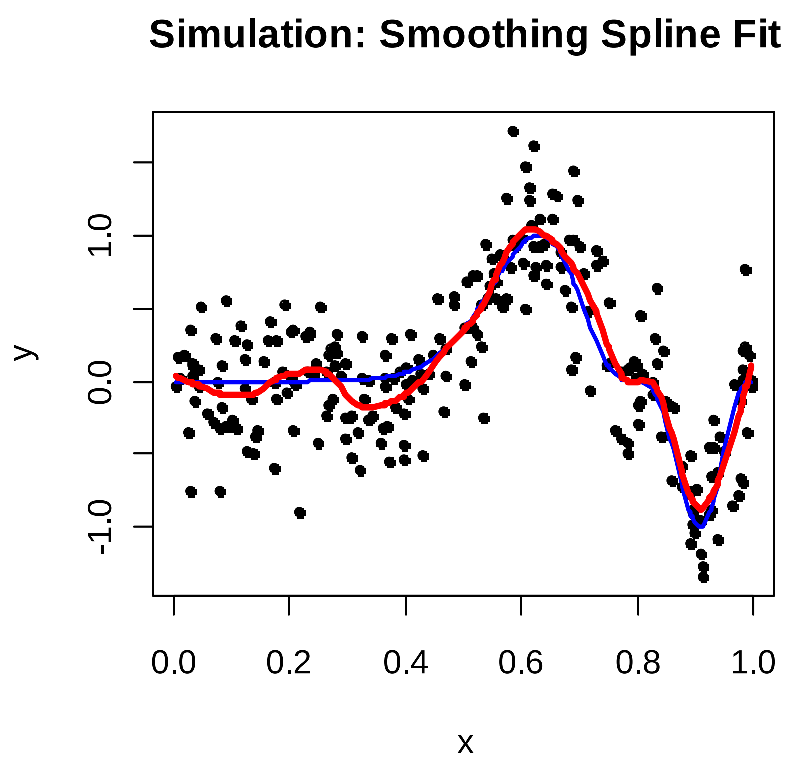
The solution are piecewise cubic polynomials in every interval (x_i, x_{i+1}). This yields: continuous function & parametric problem.

Applied Statistical Regression
HS 2010 – Week 13

Results with smooth.spline()

→ the function offers a GCV approach for the choice of λ





Loess Smoother

The loess smoother is more robust than kernel estimators and smoothing splines. This makes it an attractive alternative!

It works as follows:

1) Select a window of pre-defined size

2) Fit a polynomial (of degree 2 or 1) within this window, using a robust estimation method

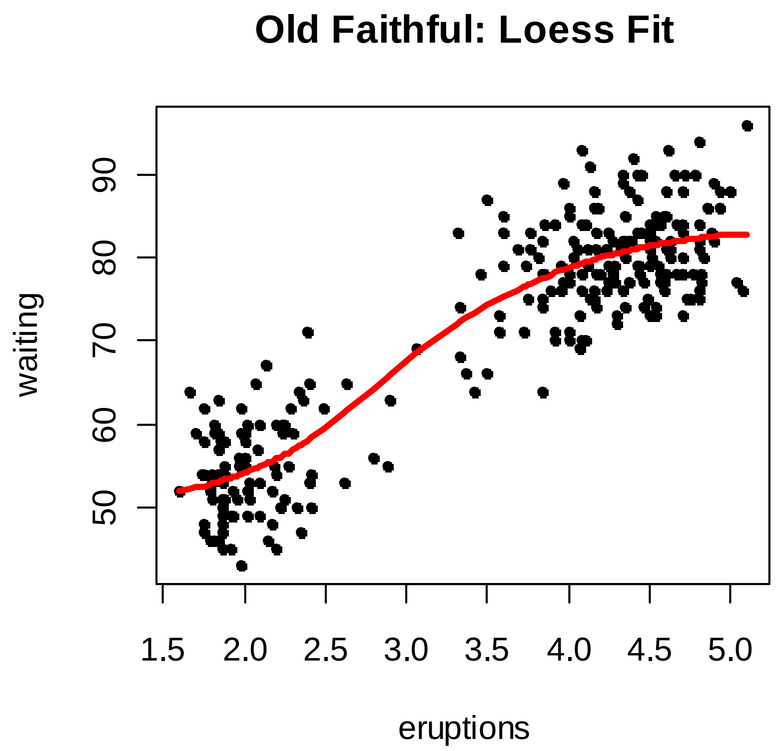
3) Predicted response at the window center := fitted value

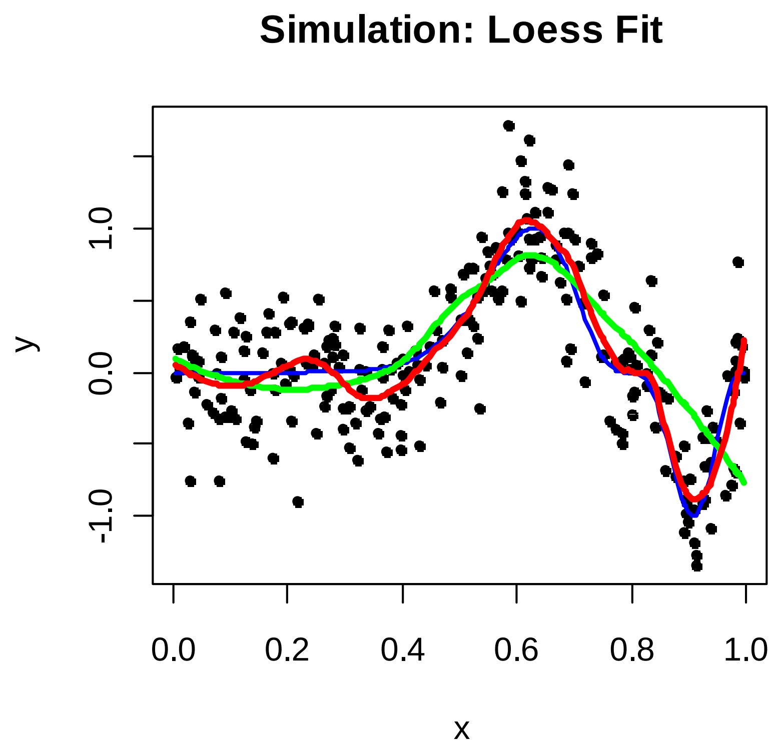
4) Slide the window over the entire x-range

Applied Statistical Regression
HS 2010 – Week 13

Results with loess()

→ λ , i.e. the window size needs to be chosen by the user





Applied Statistical Regression
HS 2010 – Week 13

Additive Models

Due to the curse of dimensionality, non-parametric smoothing methods are not straightforwardly generalizable to problems with multiple predictors.

Instead, we can use the **additive model**:

$$Y_i = \beta_0 + f_1(x_1) + \ldots + f_p(x_p) + \varepsilon_i$$

- f_j are smooth, potentially non-parametric functions
- the errors are i.i.d. with zero mean and constant variance
- flexibility, interpretability and efficient fitting are given
- this is a versatile model: parametric terms, interactions, …

Applied Statistical Regression
HS 2010 – Week 13

Software for Fitting Additive Models

There are several packages in R for fitting (G)AMs:

library(gam):

- free choice of the smoother which is used
- based on backfitting, which is an iterative procedure
- different smoothers and amounts of smoothing possible

library(mgcv):

- penalized smoothing spline approach
- automatic choice of smoothing parameters is possible

Example

The data are from a study of the relation between atmospheric ozone concentration and some meteorological predictors and originate from the LA basin. They were recorded in 1976.

We consider three predictors:

- **temp**, the temperature measured at El Monte
- **ibh**, the inversion base height at the LAX airport
- **ibt**, the inversion top temperature, again at LAX.

→ We will fit both a multiple linear regression model for reference and an additive model to show improvements.

Summary Output

```
> summary(fit)
Coefficients:
       (Intrcpt)     age   income   educ.L   educ.Q   educ.C
Indpt     -5.136   0.005    0.016    5.244   -6.341    4.693
Republ    -1.409   0.010    0.013    0.564   -0.720    0.017

          educ^4  educ^5   educ^6
Indpt     -2.552   1.291   -0.539
Republ     0.000  -0.103   -0.129

Std. Errors: ...

Residual Deviance: 1511.612
AIC: 1547.612
```

Inference

No individual hypothesis tests, although standard errors are provided in the summary output!

Reason: all parameters $\beta_{k2},...,\beta_{kJ}$ simultaneously need to be equal to zero, which cannot be tested with an individual hypothesis test.

Way out: resort to a comparison of nested models, which will as before be based on log-likelihood ratios, resp. deviance differences.

Inference: Example

```
> fit.age.inc <- multinom(party ~ age + income, data=nes)
> deviance(fit.age.inc) - deviance(fit)
[1] 13.70470
> pchisq(13.70470, fit$edf - fit.age.inc$edf, lower=FALSE)
[1] 0.3199618
```

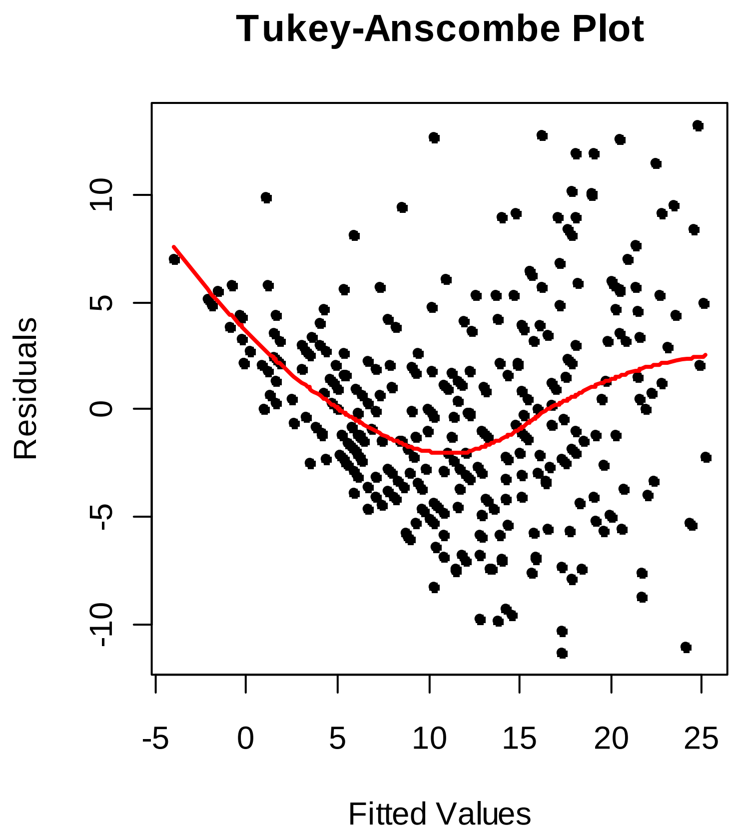
- p-value is 0.32, thus, **`education`** is not significant
- Is this a surprise, given the mosaic plot from above?
- no, the biggest differences in party affiliation are among the young people below 25 years of age, which represent only a very small fraction of the observations

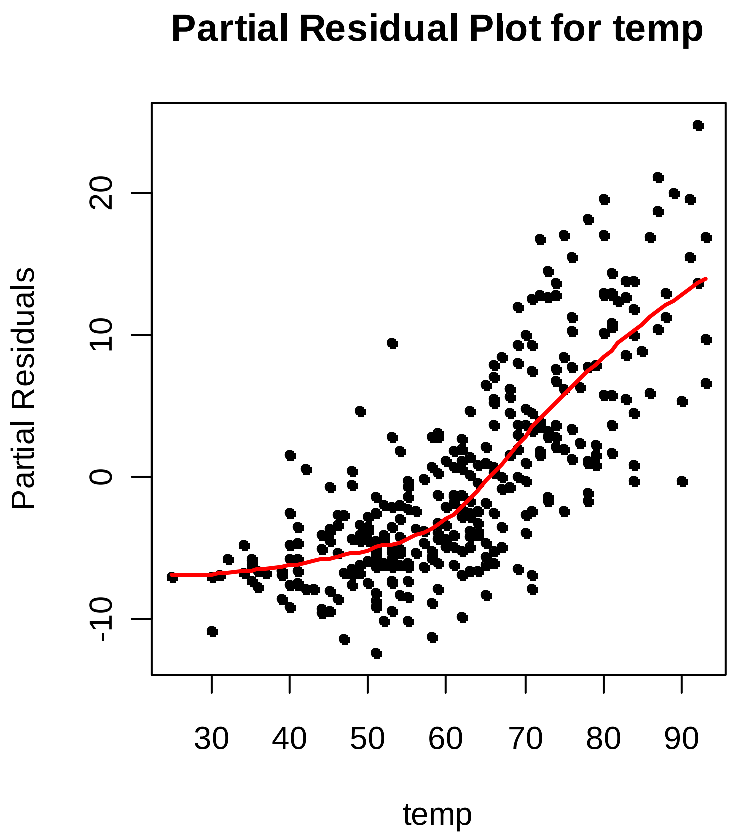
Multiple Linear Regression

```
>  summary(lm(O3 ~ temp + ibh + ibt, data = ozone))
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
```

Applied Statistical Regression
HS 2010 – Week 13

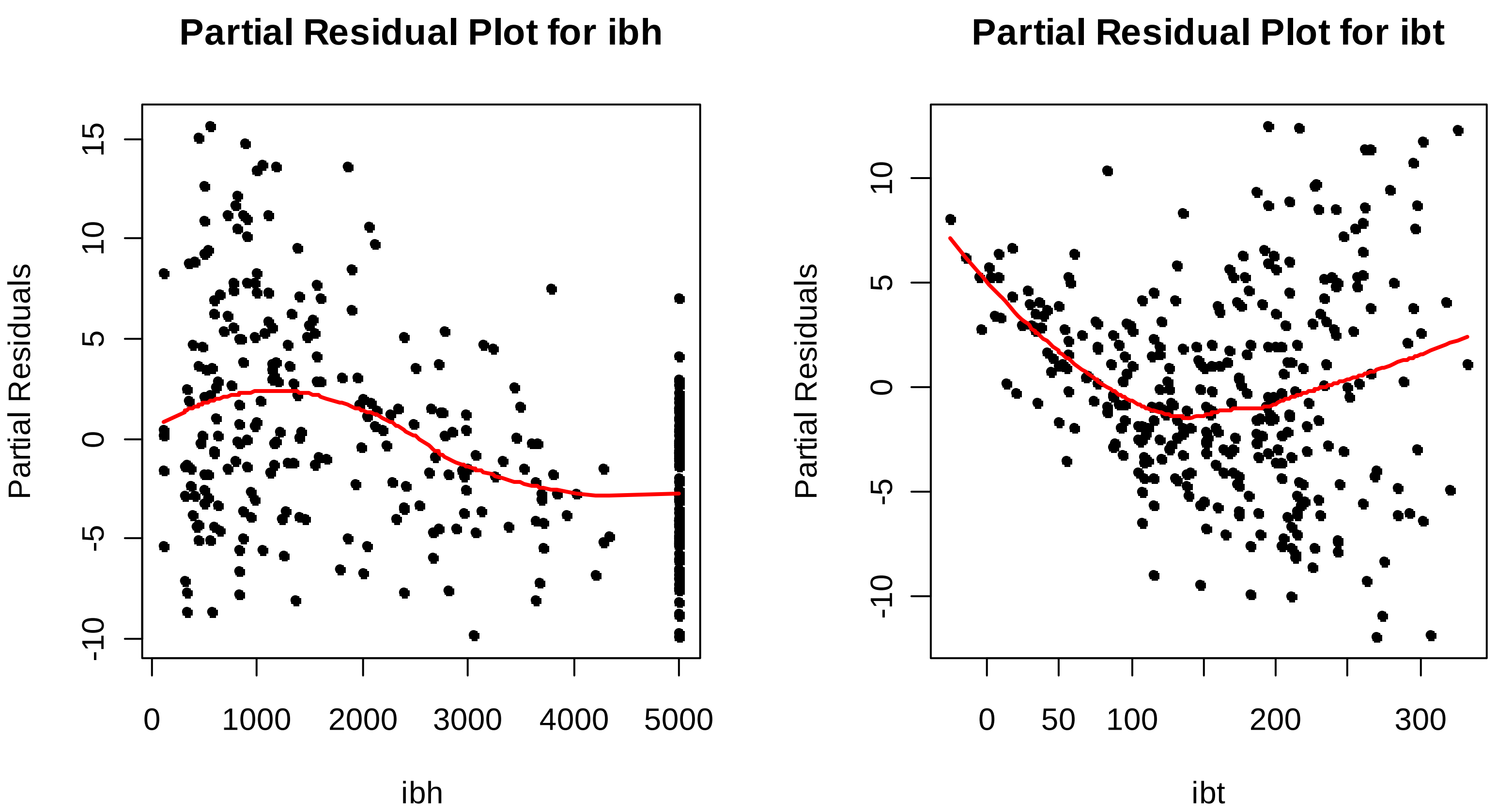
Diagnostic Plots





Applied Statistical Regression
HS 2010 – Week 13

Diagnostic Plots



Fitting with gam()

```
> summary(fit.gam)
Call: gam(O3 ~ lo(temp) + lo(ibh) + lo(ibt), data=ozone)

Null Deviance: 21115.41 on 329 degrees of freedom
Residual Deviance: 5935.096 on 318.0005 degrees of freedom
AIC: 1916.049
            Df Npar Df Npar F     Pr(F)
(Intercept)  1
lo(temp)     1      2.5 7.4550 0.0002456 ***
lo(ibh)      1      2.9 7.6205 8.243e-05 ***
lo(ibt)      1      2.7 7.8434 9.917e-05 ***
```

Inference for gam()

Deviance explained:

```
> 1-5935.096/21115.41
[1] 0.7189211
```

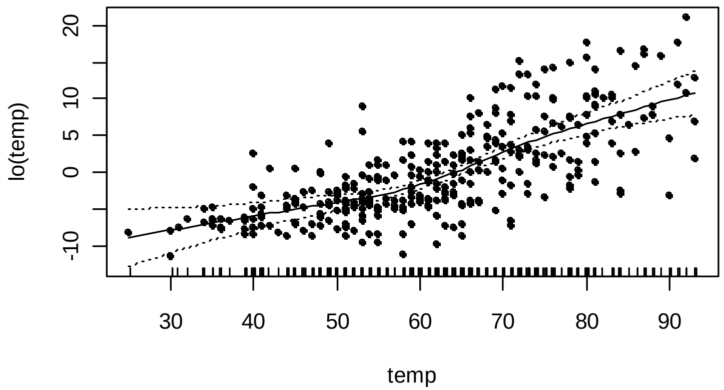
For multiple regression, the result was 0.652. However, we now spend more degrees of freedom.

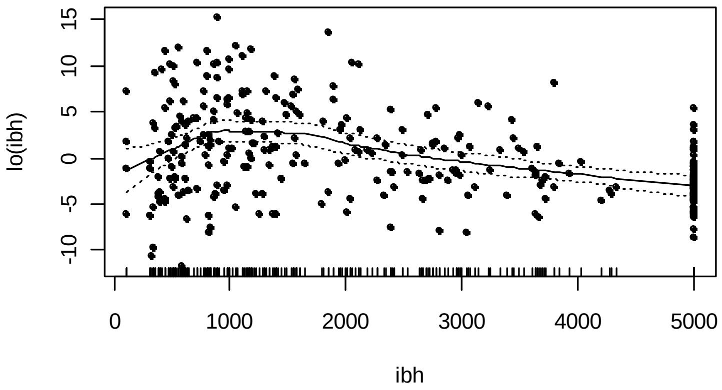
Individual hypothesis test are best done using nested model comparisons with an F-test, rather than with the score test:

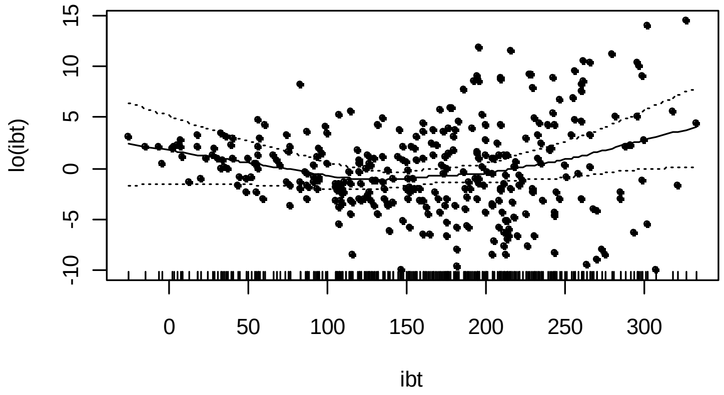
```
> fit.gam.small <- gam(O3 ~ lo(temp)+lo(ibh), data=ozone)
> anova(fit.gam.small, fit.gam, test="F")
>     318.00      5935.1 3.6648   109.47 1.6005  0.179
```

Applied Statistical Regression
HS 2010 – Week 13

Graphical Output of gam()







Informations on the Exam

- The exam will be on January 26, 2011 (provisional) and lasts for 120 minutes. But please see the official announcement.

- It will be open book, i.e. you are allowed to bring any written materials you wish. You can also bring a pocket calculator, but computers/notebooks and communcation aids are forbidden.

- Topics include everything that was presented in the lectures, from the first to the last, and everything that was contained in the exercises and master solutions.

- You will not have to write R-code, but you should be familiar with the output and be able to read it.

Applied Statistical Regression
HS 2010 – Week 13

Informations on the Exam

- With the exam, we will try our best to check whether you are proficient in applied regression. This means choosing the right models, interpreting output and suggesting analysis strategies.

- Old exams will not be available for preparation. I recommend that you make sure that you understand the lecture examples well and especially focus on the exercises.

- There are 2 question hours in January. See the course webpage or exercise sheet 7 for time and location.

- There are some additional points for doctoral students, which will also be communicated via e-mail.

Applied Statistical Regression
HS 2010 – Week 13

End of the Course

→ **Happy holidays and all the best for the exams!**