Computer Intensive Methods using R

Part 7: topics in non parametric regression modeling

Prof. Dr. Ziv Shkedy

Master of Statistics
Hasselt University

General Information

Overview of the course

- Selected topics:
  - Bootstrap Standard error: example of curve fitting.
  - Inference with nonparametric regression using bootstrap methods.

Overview of the course (part 1)

The Bootstrap algorithm

Introduction:
- Sampling from a population.
- The empirical distribution.
- Plug in principle.

Estimation:
- Accuracy of statistics.

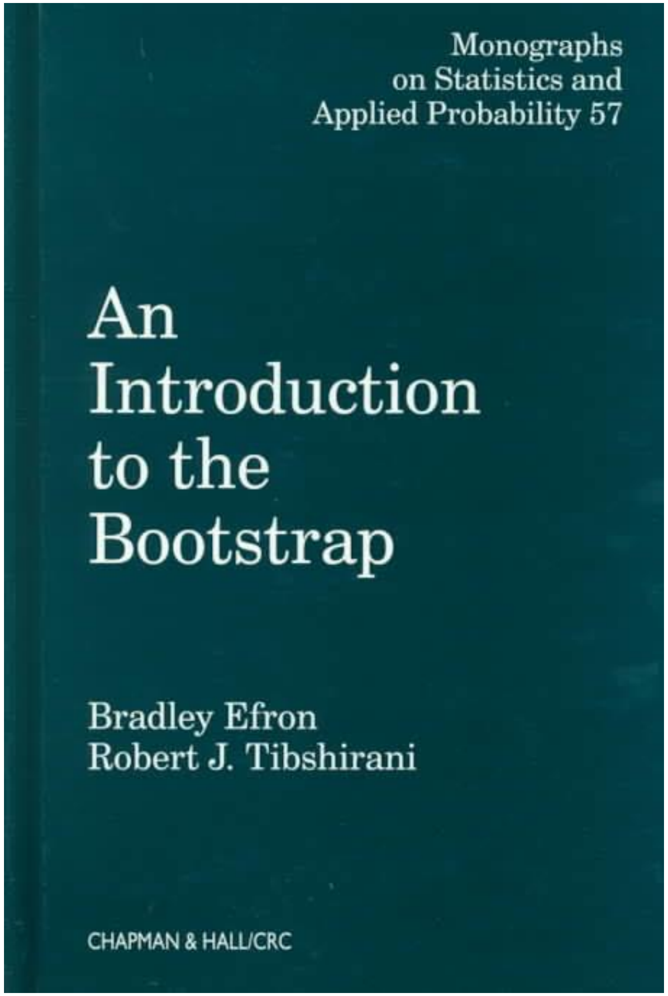
Inference:
- One sample tests.
- Two-samples tests.
- Bootstrap and permutation tests.

Modeling:
- Non parametric regression.

Reference

- Bradley Efron and Robert J. Tibshirani (1994): An introduction to bootstrap.

- Davison A.C. and Hinkley D.V: Bootstrap Methods and Their Application.



Course materials

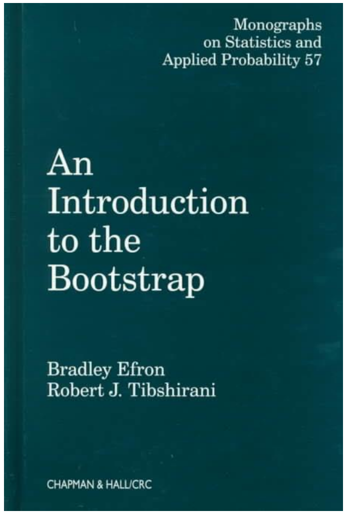
- Slides.
- R program.
- R datasets & External datasets.
- YouTube tutorials.
- Videos for the classes (highlights of each class in the course).

YouTube tutorials

- YouTube tutorials about bootstrap using R:
  1. One-sample bootstrap CI for the mean (host: LawrenceStats): https://www.youtube.com/watch?v=ZkCDYAC2iFg.
  2. Using the non-parametric bootstrap for regression models in R (host:Ian Dworkin):https://www.youtube.com/watch?v=ydtOTctg5So.
  3. Performing the Non-parametric Bootstrap for statistical inference using R (host: Ian Dworkin): https://www.youtube.com/watch?v=TP6r5CTd9yM
  4. Using the sample function in R for resampling of data - absolute basics (host: Ian Dworkin):https://www.youtube.com/watch?v=xE3KGVT6VLE
  5. Permutation tests in R - the basics (host:Ian Dworkin):https://www.youtube.com/watch?v=ZiQdzwB12Pk.
  6. Bootstrap Sample Technique in R software (host: Sarveshwar Inani):https://www.youtube.com/watch?v=tb6wb9ZdPH0
  7. Bootstrap confidence intervals for a single proportion (host: LawrenceStats):https://www.youtube.com/watch?v=ubX4QEPqx5o
  8. Bootstrapped prediction intervals (host:James Scott):https://www.youtube.com/watch?v=c3gD_PwsCGM.

- https://www.youtube.com/watch?v=gcPIyeqymOU

Bootstrap standard error: some examples



Topics

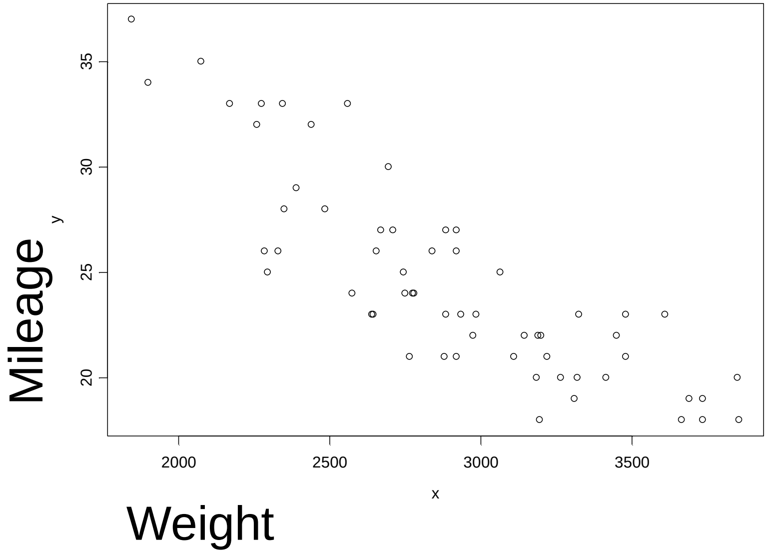
- Examples:
  - The score data:
    - Distribution of the covariance matrix.
    - Ratio between variables.
  - The fuel data:
    - Non parametric regression: a loess model for the fuel data.

Example 2:
curve fitting (part 1)

The dataset is not the same as the data in the book

Example: fuel consumption

- The dataset gives information on cars taken from the April, 1990 issue of Consumer Reports.
- Two variables:
  – Mileage: a numeric vector of gas mileage in miles/gallon as tested by CU.
  – Weight: a numeric vector of the car’s weight.



External data:

```
fuel.frame<-
read.table('C:/projects/cim/UpdatesSlides_2017/Data/fuel.txt')
```

The aim of the analysis

- Model the relationship between the car's weight and mileage.
- We would like to predict the mileage of a car which weight 3000 Kg.
- We would like to estimate the standard error for a prediction for a specific weight.

Model formulation

We assume that the mileage (y) is a function of the weight.

$$y_i = r(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

The function *r(x)* represents the dependency of the mileage on the car's weight.

$$r(x) = E(y \mid x)$$

Estimation of *r(x)*

The main question is how to estimate *r(x)*

$$r(x) = E(y \mid x)$$

Parametric

Nonparametric

Parametric approach

Consider a linear regression model of the form

$$y_i = r(x_i) + \varepsilon_i$$

$r(x)$ is quadratic function of the car's weight

$$r_\beta(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^{\,2}$$

OLS estimators

$$y_i = r(x_i) + \varepsilon_i \qquad\qquad r_\beta(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

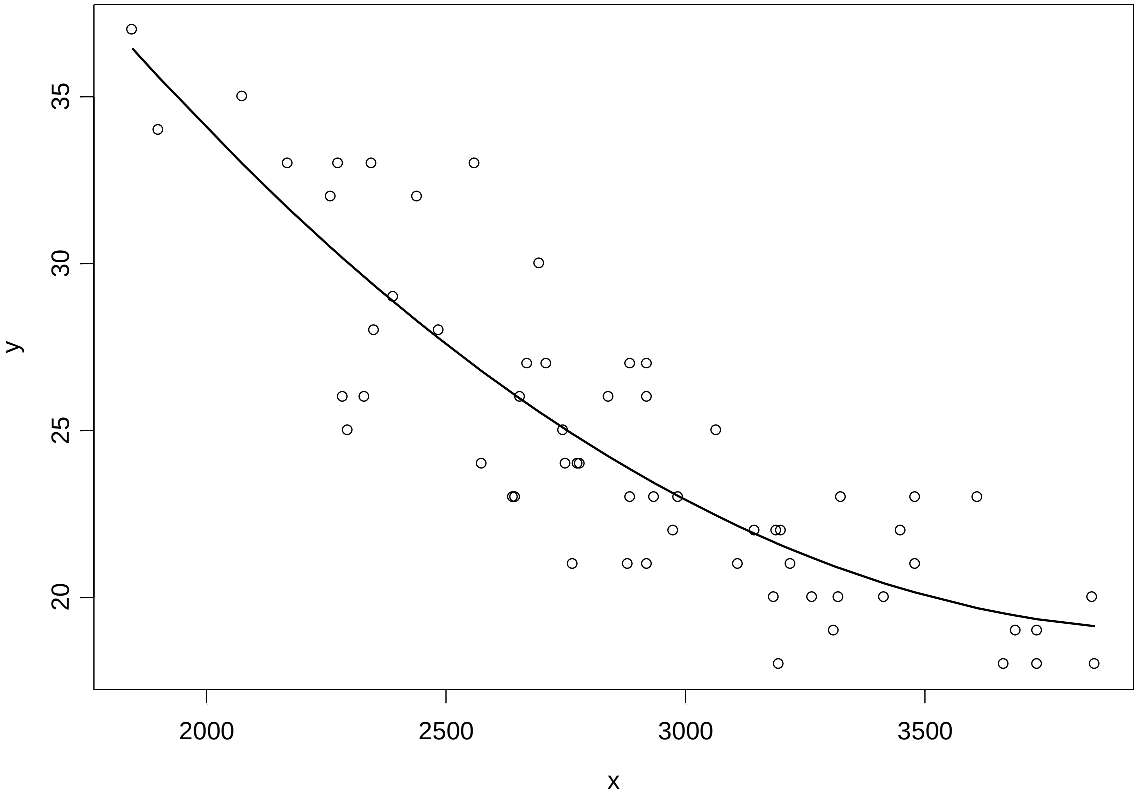
We minimize the residuals sum of squares

$$RSE(\beta) = \sum_{i=1}^{n} \left[y_i - r_\beta(x_i) \right]^2$$

We choose the vector of β which minimizes the residuals sum of squares

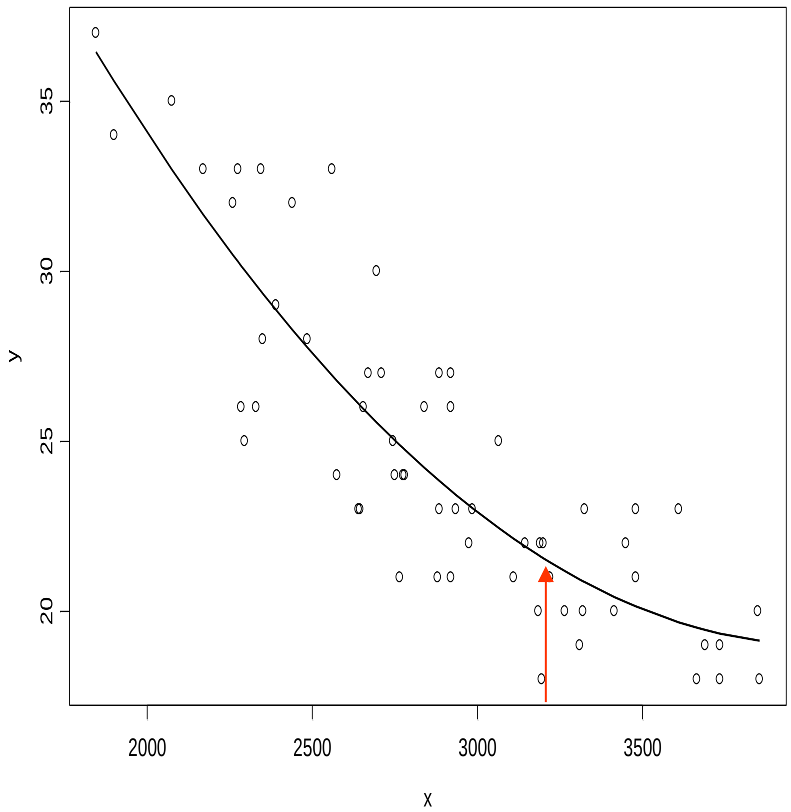
$$RSE(\hat{\beta}) = \min_\beta RSE(\beta)$$

Data and predicted values



$$\hat{r}_{\beta}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2$$

Prediction for x=3200



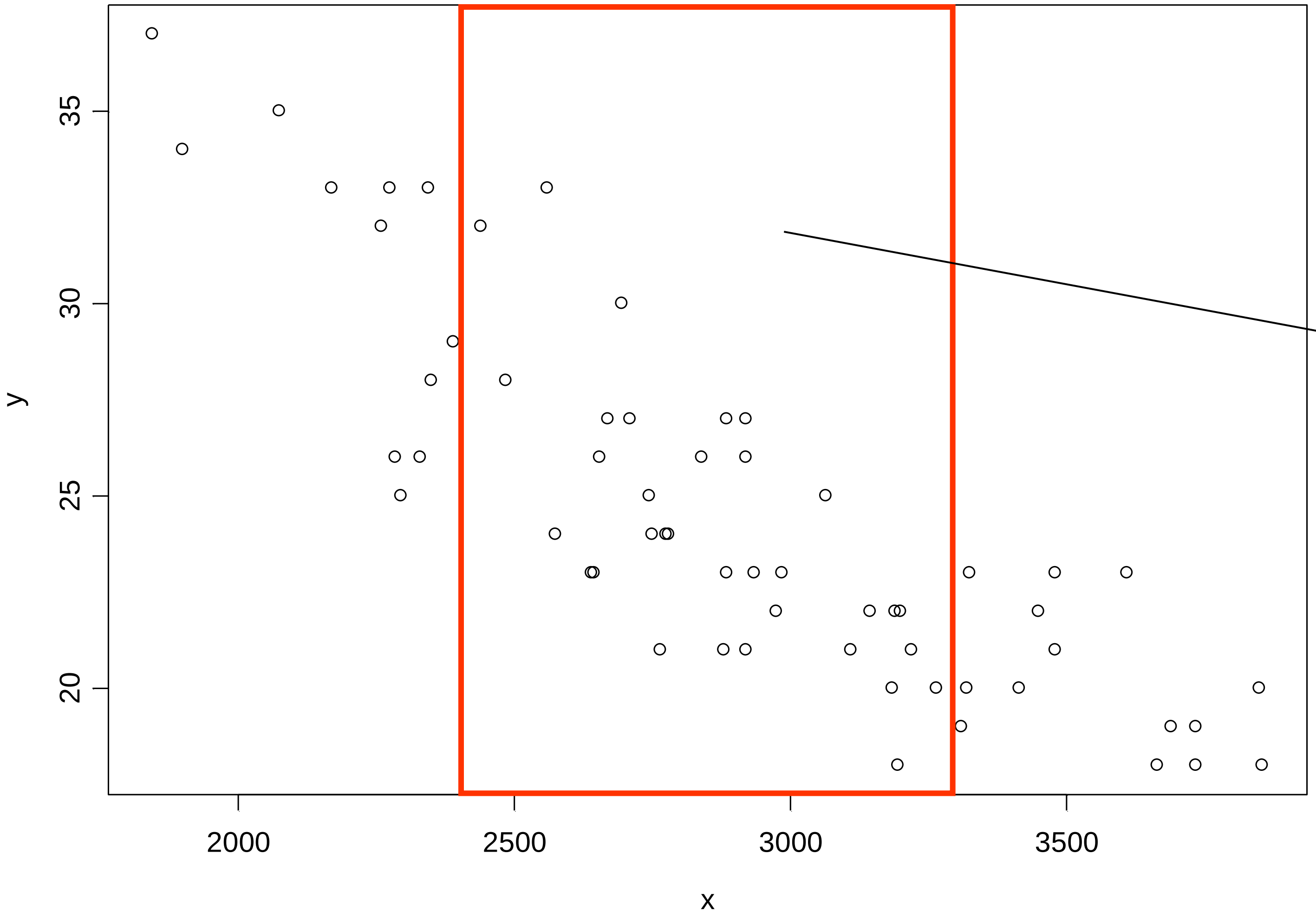

```
> x <- 3200
> newdat <- data.frame(x)
> predict(fit.lm, newdat)
       1
 21.56516
```

Non parametric approach: the Loess

A loess model is a local regression model

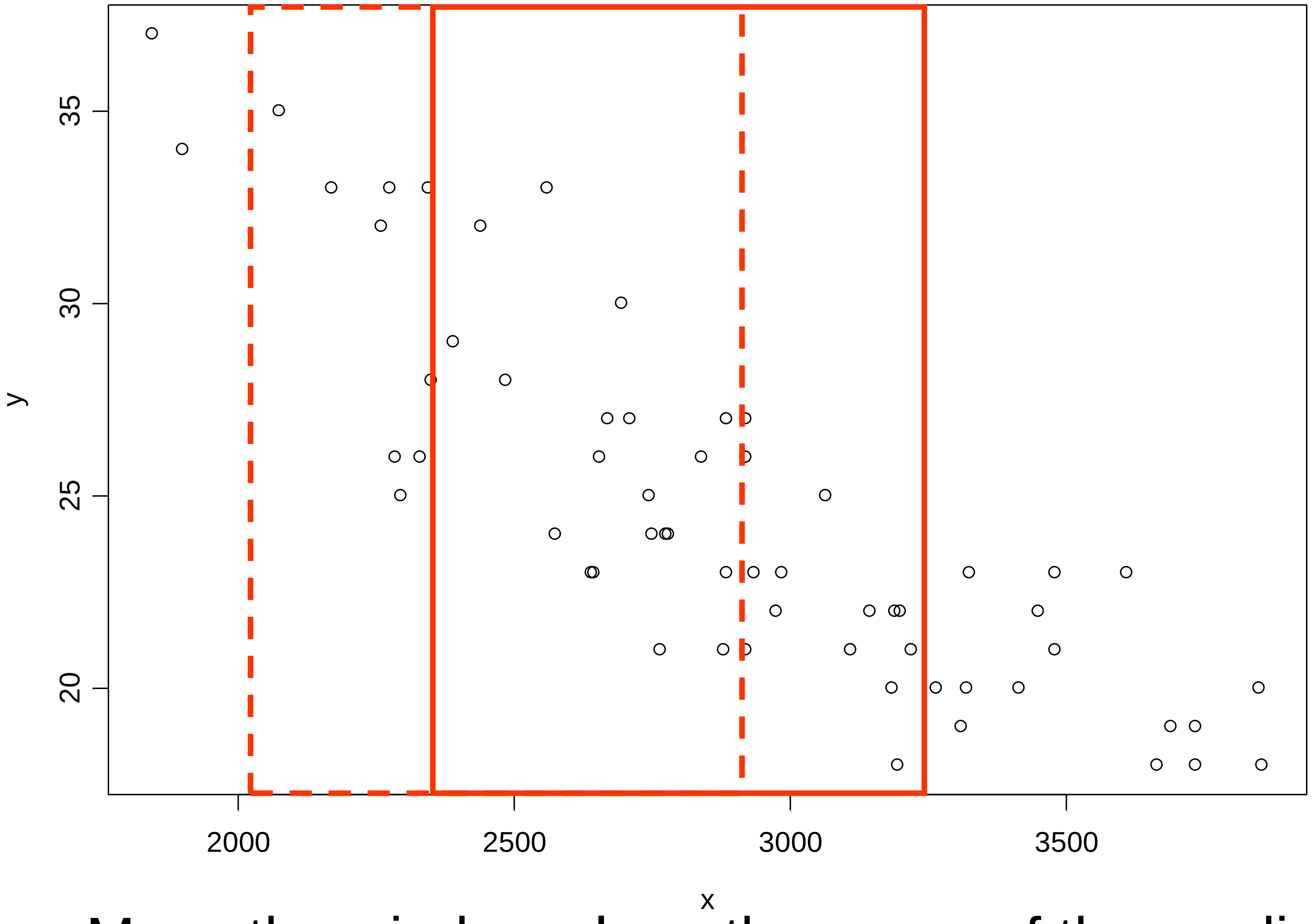
It is a non parametric regression model which fit a weighted regression model locally.

How the loess works ?



Fit in the window a weighted regression model

How the loess works ?



Move the window along the range of the predictor

The size of the window

The size of the window (λ) determines which proportion of the data will be used to estimate r(x) at any given point of x

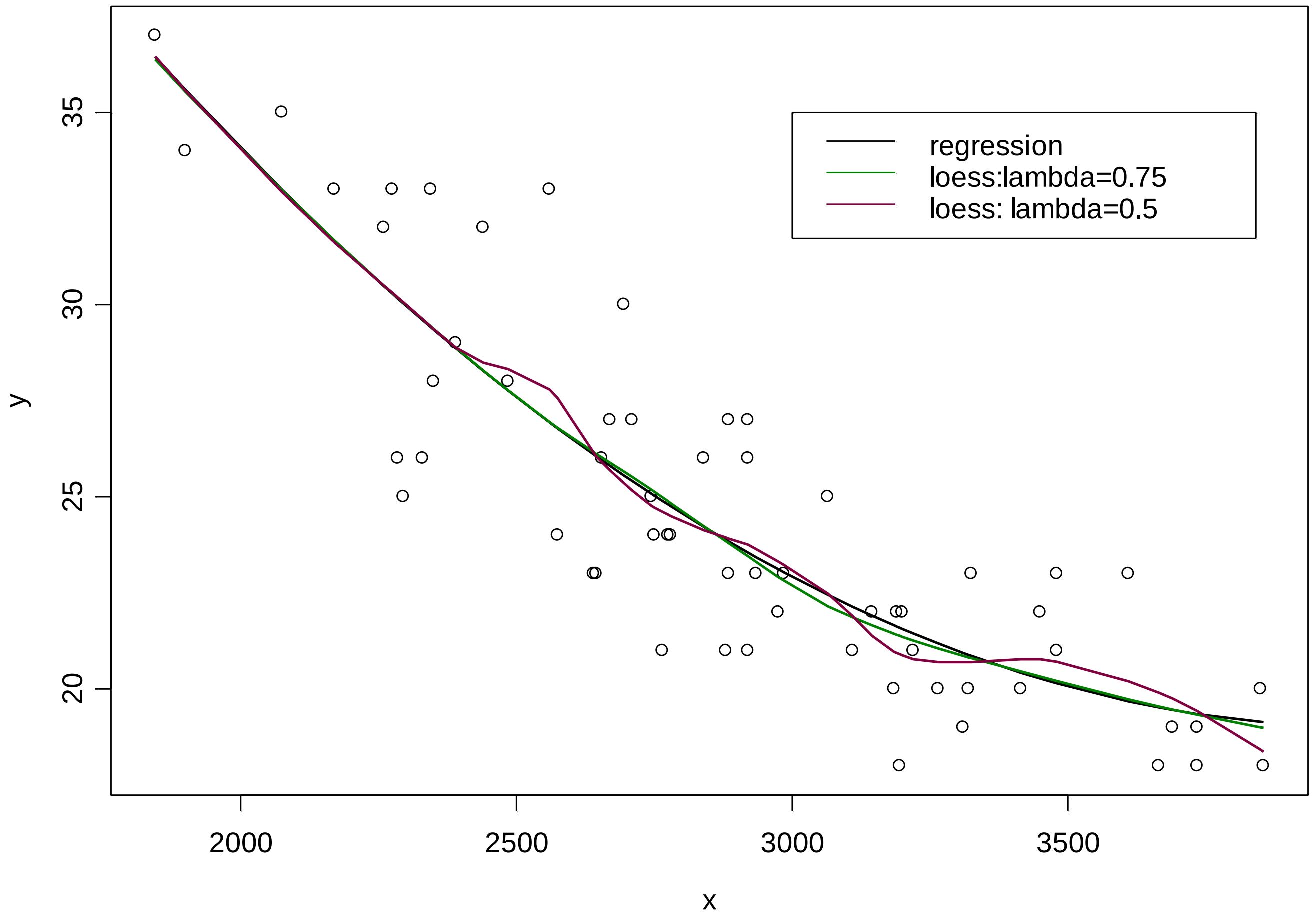
For a global model: we use all the data

The loess() function

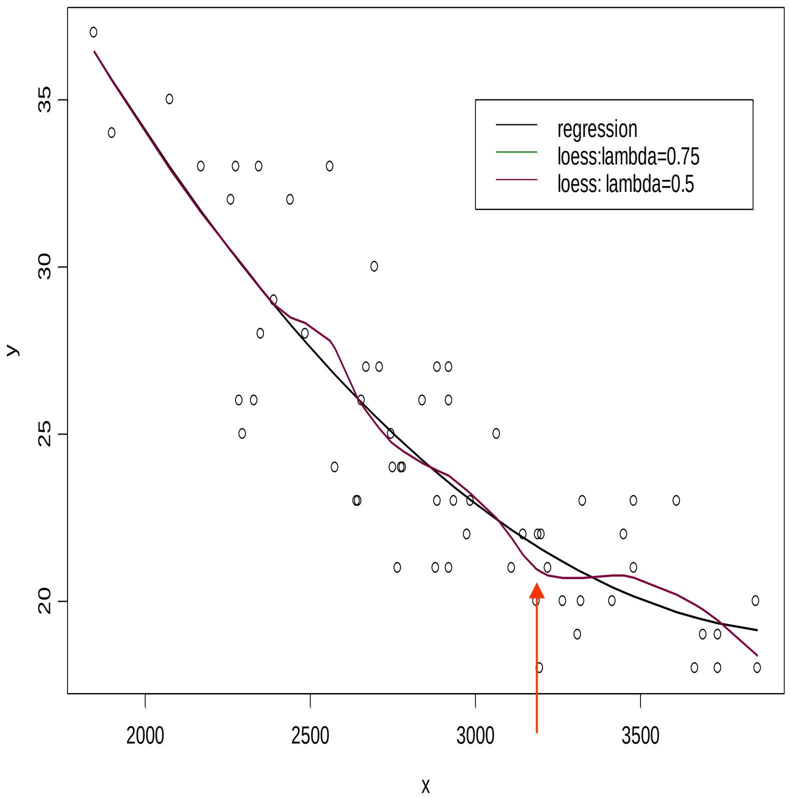
Linear regression and loess models with two smoothing parameters:

```
fit.lm<-lm(y~x+x^2)
fit.lo1<-loess(y~x)
fit.lo2<-loess(y~x,span=0.5)
```

Data and predicted values



Prediction for x=3200




```
> x <- 3200
> newdat <- data.frame(x)
> predict(fit.lm, newdat)
        1
 21.56516
> predict(fit.lo2, newdat)
[1] 20.87979
```

Bootstrap estimated for *r(x)*

We wish to apply the bootstrap method in order to estimate the standard error for the prediction of a spesific weight.

Let $r(x)=E(y|x)$. We do not know $E(y|x)$ so…

$$\hat{r}(x) = \hat{E}(y \mid x)$$

We estimate the mean using linear regression or loess

Bootstrap estimate for $r(x)$

The estimate for the mean for a specific value of weight:

$$\hat{r}(x_i) = \hat{E}(y \mid x_i)$$

Our aim is to estimate the standard error of the predicted value $\hat{r}(x_i)$

Bootstrap estimate for *r(x)*

The data (x_i,y_i)

Bootstrap sample $(x_i,y_i)^*$

Bootstrap sample $(x_i,y_i)^*$

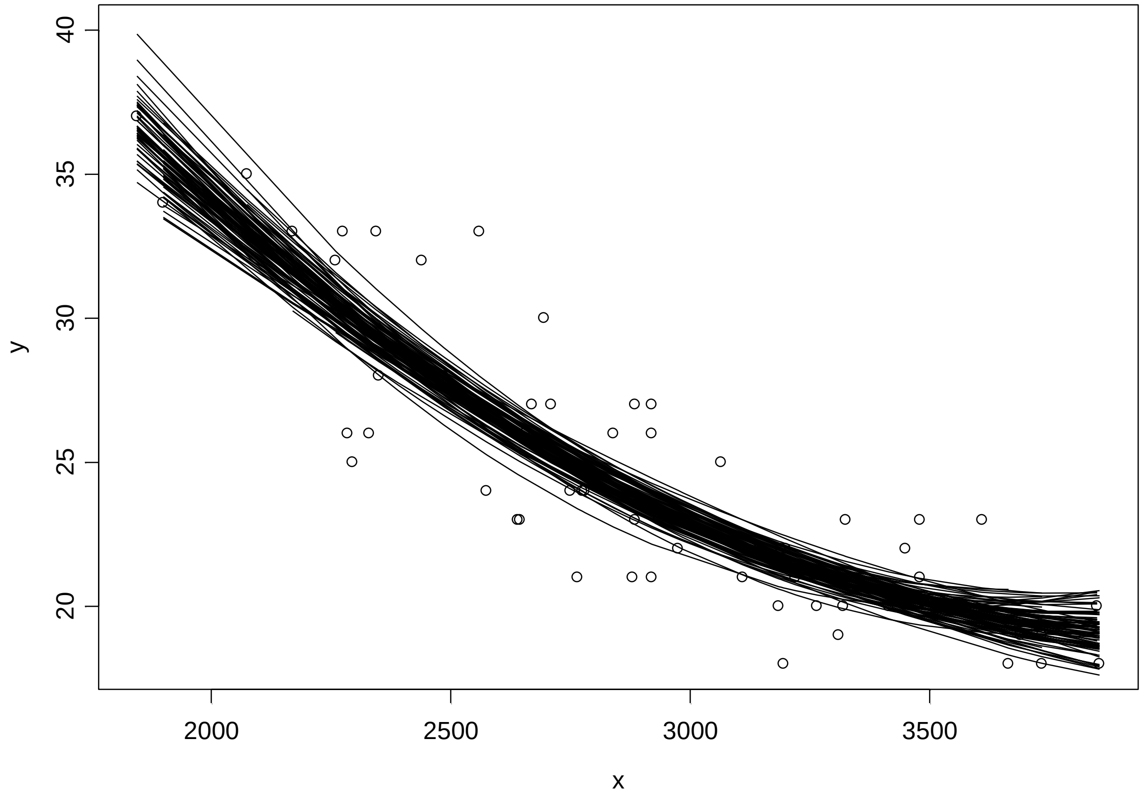
Bootstrap sample $(x_i,y_i)^*$

B bootstrap samples

$$\hat{r}(x_1^*) \qquad \hat{r}(x_2^*) \qquad \hat{r}(x_3^*) \qquad \hat{r}(x_B^*)$$

B bootstrap estimates for *r(x)*

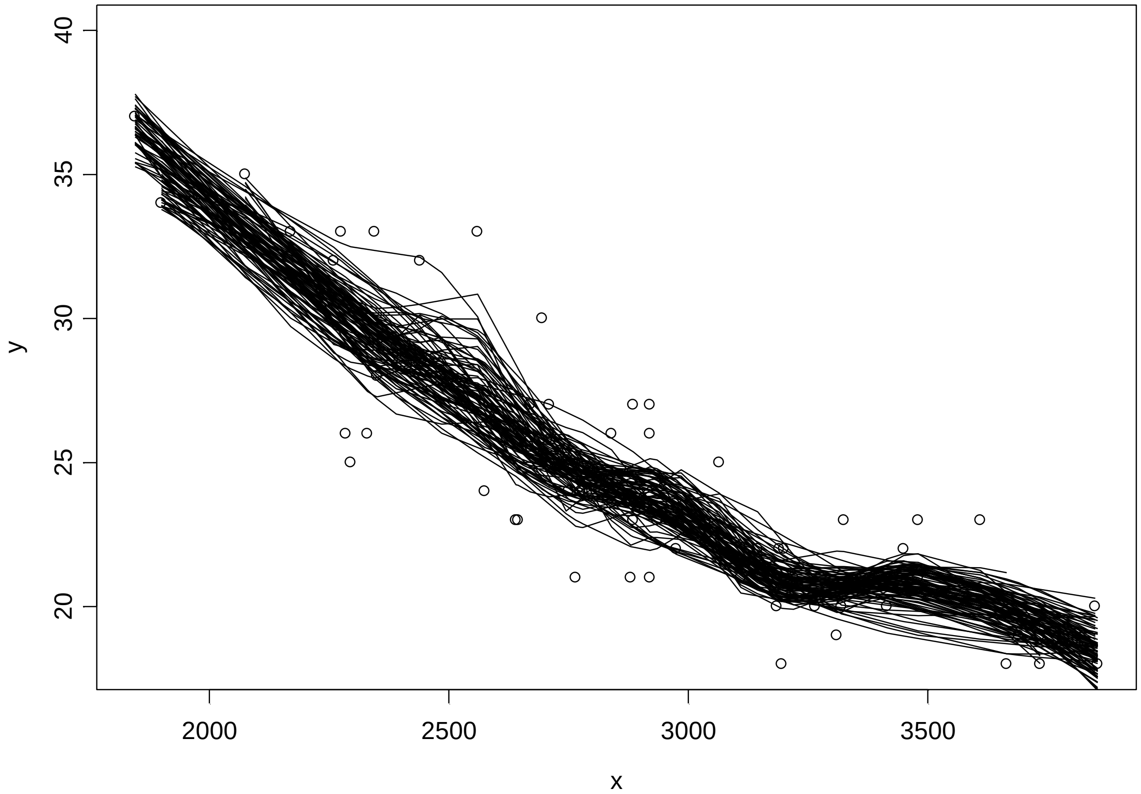
Bootstrap estimates for *r(x)*: linear regression



Each line: a bootstrap replicate for

$$\hat{r}^*_{\beta}(x_i)_b = \hat{\beta}^*_0 + \hat{\beta}^*_1 x_i + \hat{\beta}^*_2 x_i^2$$

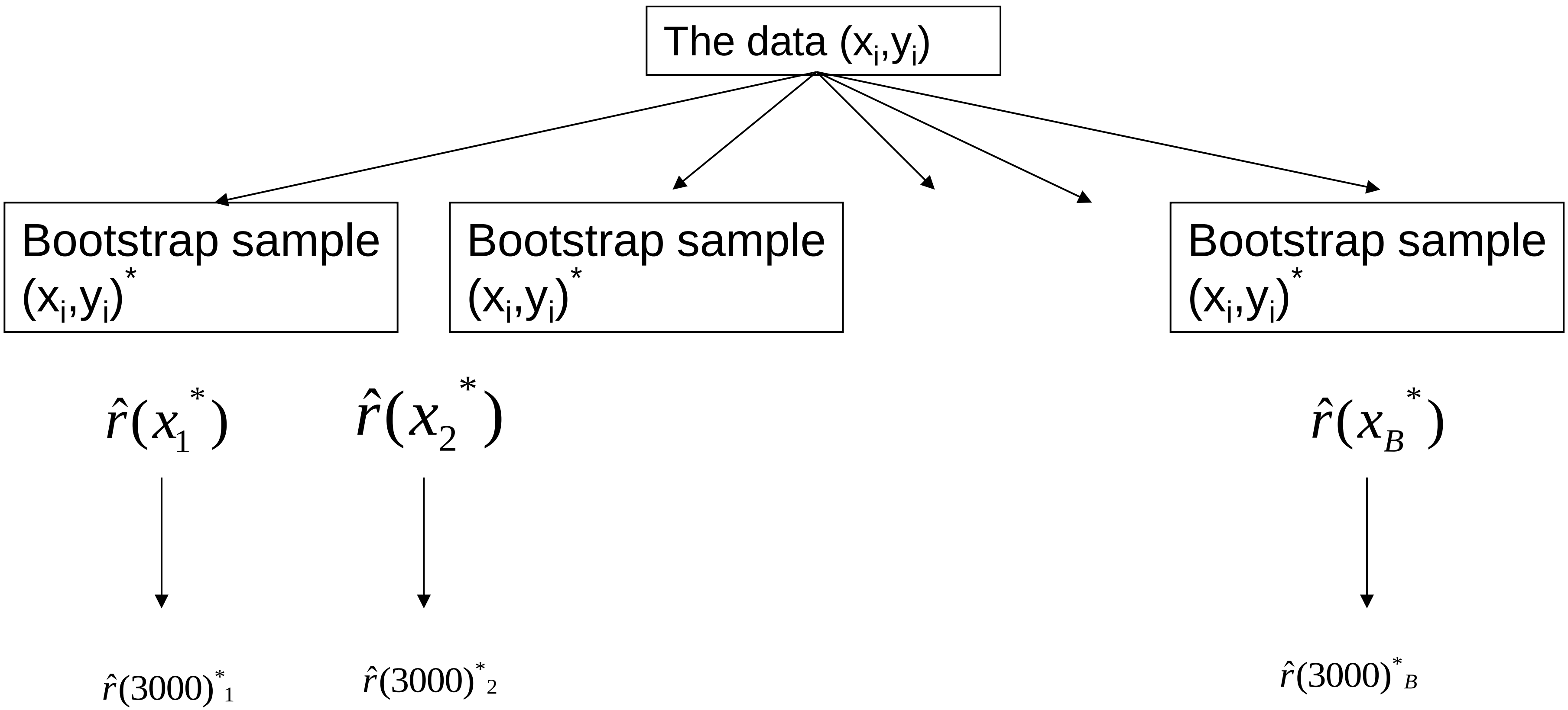
Bootstrap estimates for $r(x)$: loess (λ=0.5)



Each line: a bootstrap replicate for

$$\hat{r}^*(x_b^*)_b$$

Bootstrap estimates for *r(x)*



Bootstrap estimates for *r(x)*

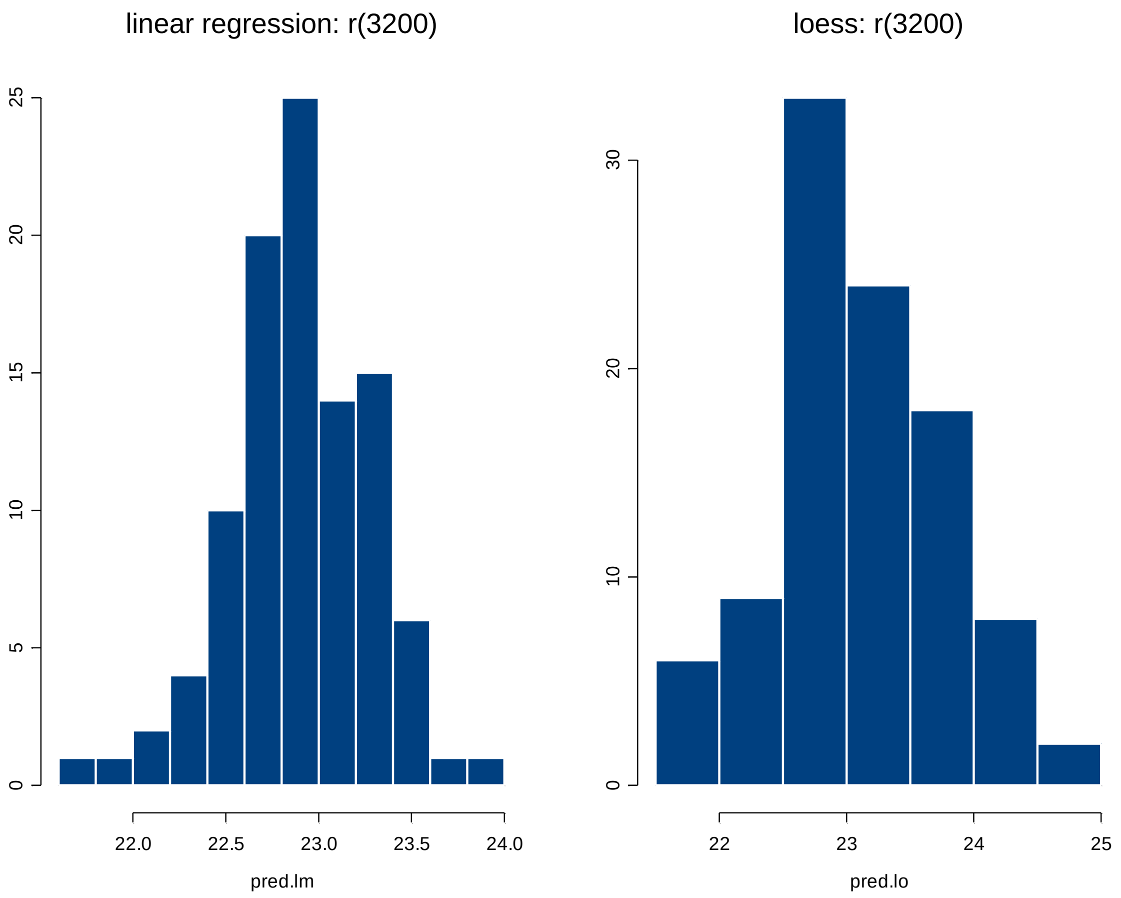
$\hat{r}(x_1^{\ *})$ → $\hat{r}(3000)^*{}_1$

$\hat{r}(x_2^{\ *})$ → $\hat{r}(3000)^*{}_2$

$\hat{r}(x_B^{\ *})$ → $\hat{r}(3000)^*{}_B$

$$S.E._B = \left\{ \frac{1}{B-1} \sum_{b=1}^{B} [\hat{r}_b(3000)^* - \hat{r}\ (3000)^*]^2 \right\}^{0.5}$$

Predicted value for *r(3200)*: loess and linear regression



Standard error for *r(3000)*

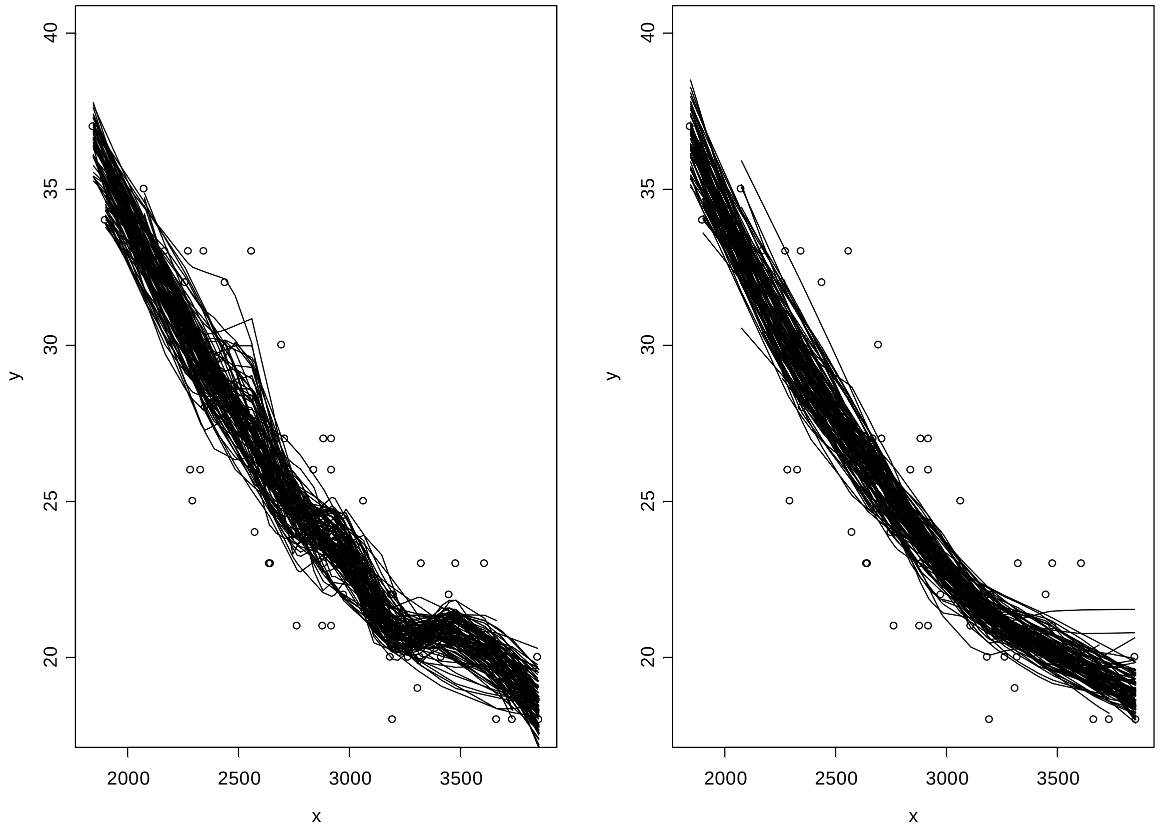
	$\hat{r}_{reg}(3200)$	$\hat{r}_{lo}(3200)$
value	21.56	20.87
$S.E[\hat{r}\,(3200)]$	0.142	0.454

```
> var(pred.lm)
[1] 0.1422859
> var(pred.lo)
[1] 0.454642
```

The predicted values obtained from loess less accurate than the predicted value obtained from the regression model

How come ??

Bootstrap with two loess models (with λ=0.5 and λ=0.75)



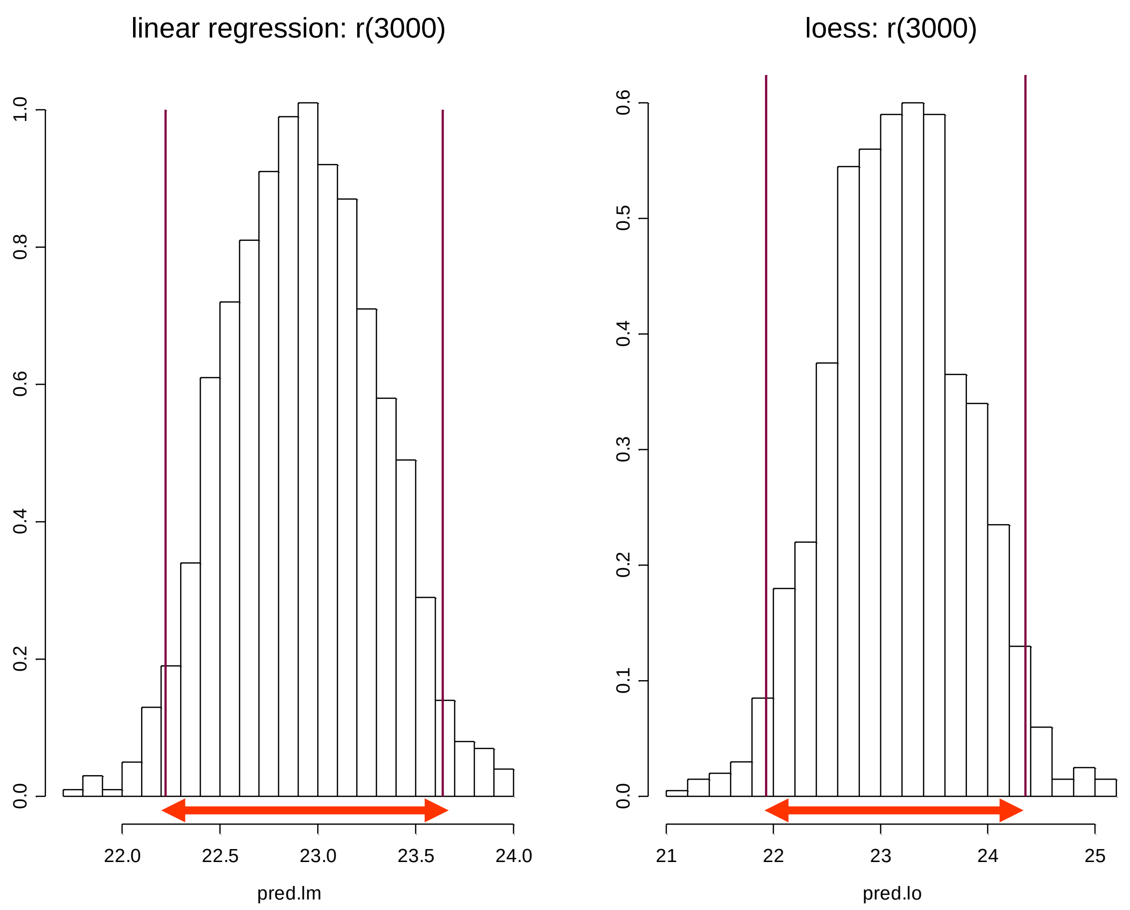
Standard error for $r(3200)$

	$\hat{r}_{reg}(3200)$	$\lambda=0.75$ $\hat{r}_{lo}(3200)$	$\lambda=0.5$ $\hat{r}_{lo}(3200)$
value	21.56	21.361	20.87
$S.E[\hat{r}\,(3200)]$	0.142	0.2516	0.454

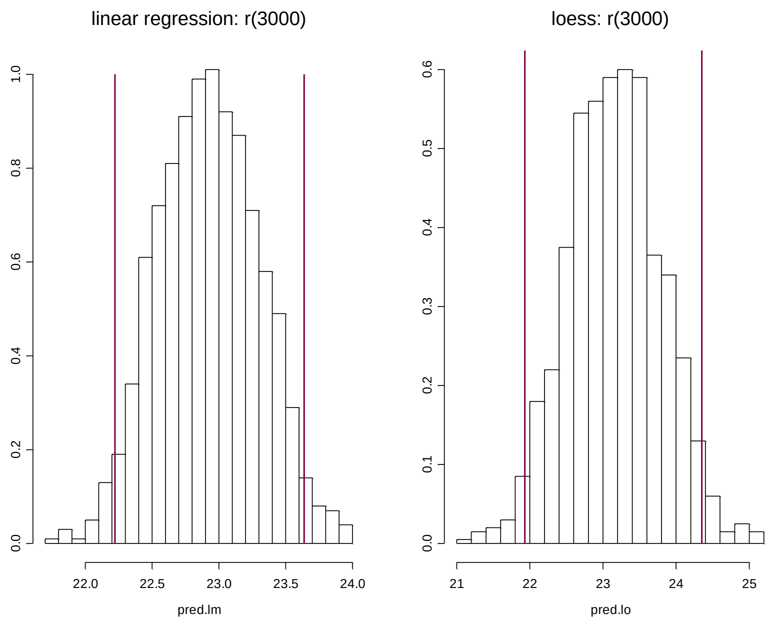
> predict(fit.lo1, newdat)
[1] 21.36145

> var(pred.lo)
[1] 0.2516016

Bootstrap C.I. for $\hat{r}(3000)$



95% bootstrap C.I for $\hat{r}(3000)$




```
> cir
     2.5%    97.5%
 22.22285 23.63814
> cilo
     2.5%    97.5%
 21.93076 24.35096
```

R code for the fuel data

```
x<-fuel.frame$Weight
y<-fuel.frame$Mileage
y<-y[order(x)]
x<-sort(x)
plot(x,y)
fit.lm<-lm(y~x+x^2)
par(mfrow=c(1,1))
plot(x,y)
lines(x,fit.lm$fit,lwd=2)
fit.lo1<-loess(y~x)
fit.lo2<-loess(y~x,span=0.5)
lines(x,fit.lo1$fit,lwd=2,col=4)
lines(x,fit.lo2$fit,lwd=2,col=3)
legend(3000,35,c("regression","loess:lambda=0.75","loess:
    lambda=0.5"),col=c(1,4,3),lty=c(1,1,1))
```

```
x<-3200
newdat<-data.frame(x)
predict(fit.lm,newdat)
predict(fit.lo2,newdat)
predict(fit.lo1,newdat)
```

R code for the bootstrap

```
x<-fuel.frame$Weight
y<-fuel.frame$Mileage
y<-y[order(x)]
x<-sort(x)
plot(x,y)
n<-length(x)
index<-c(1:n)
B<-1000
x.boot<-y.boot<-fit.b.lm<-fit.b.lo<-matrix(0,n,B)
x.b<-3000
newdat<-data.frame(x.b)
pred.lm<-pred.lo<-c(1:B)
for(i in 1:B){
cat(i)
boot.i<-sample(index,n,replace=T)
x.b<-x[boot.i]
y.b<-y[boot.i]
y.b<-y.b[order(x.b)]
x.b<-sort(x.b)
fit.lm.i<-lm(y.b~x.b+x.b^2)
fit.lo.i<-loess(y.b~x.b,span=0.5)
x.boot[,i]<-x.b
y.boot[,i]<-y.b
fit.b.lm[,i]<-fit.lm.i$fit
fit.b.lo[,i]<-fit.lo.i$fit
pred.lo[i]<-predict(fit.lo.i,newdat)
pred.lm[i]<-predict(fit.lm.i,newdat)
}
```

R code for the figures

```
plot(x,y,ylim=c(18,40))
for(i in 1:B)
{
lines(x.boot[,i],fit.b.lm[,i])
}
```

```
plot(x,y,ylim=c(18,40))
for(i in 1:B)
{
lines(x.boot[,i],fit.b.lo[,i])
}

par(mfrow=c(1,2))
hist(pred.lm,col=0)
title("linear regression: r(3200)")
hist(pred.lo,col=0)
title("loess: r(3200)")

var(pred.lm)
var(pred.lo)

quantile(pred.lm,probs=c(0.025,0.975))
quantile(pred.lo,probs=c(0.025,0.975))
```

R code for the histograms and C.I

```
cir<-quantile(pred.lm,probs=c(0.025,0.975))
cilo<-quantile(pred.lo,probs=c(0.025,0.975))
par(mfrow=c(1,2))
hist(pred.lm,col=0,nclass=25,probability=T)
lines(c(cir[1],cir[1]),c(0,1),lwd=2,col=3)
lines(c(cir[2],cir[2]),c(0,1),lwd=2,col=3)
title("linear regression: r(3000)")
hist(pred.lo,col=0,nclass=25,probability=T)
lines(c(cilo[1],cilo[1]),c(0,1),lwd=2,col=3)
lines(c(cilo[2],cilo[2]),c(0,1),lwd=2,col=3)
title("loess: r(3000)")
```

Example : Inference with nonparametric regression using bootstrap methods

This example is not in E&T book

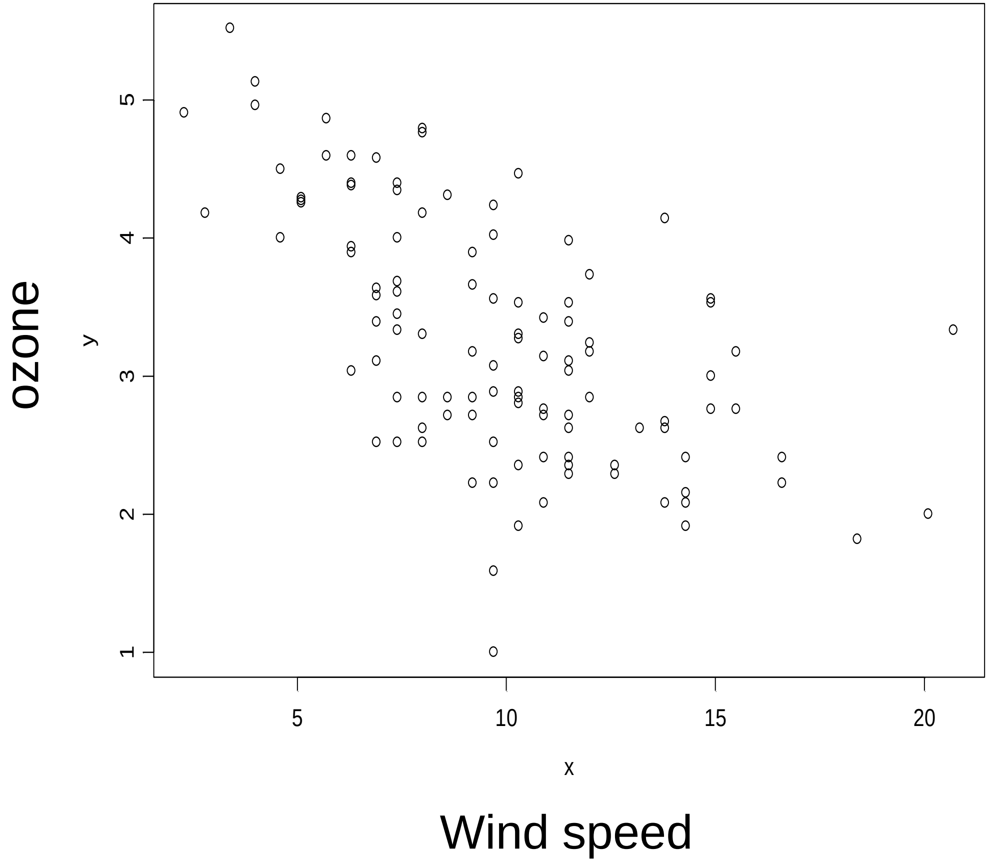
Inference with nonparametric regression

Within the setting of linear regression model we use non parametric regression model to check the assumption of the about the mean structure of the parametric models

Loess models as an example of a non parametric regression models

We use bootstrap simulation in order to approximate the (unknown) distribution of the test statistic under the null hypothesis

The data: the air dataset



111 observations taken from an environmental study that measured the four variables ozone, solar radiation, temperature, and wind speed for 111 consecutive days.

Simple linear regression: model formulation

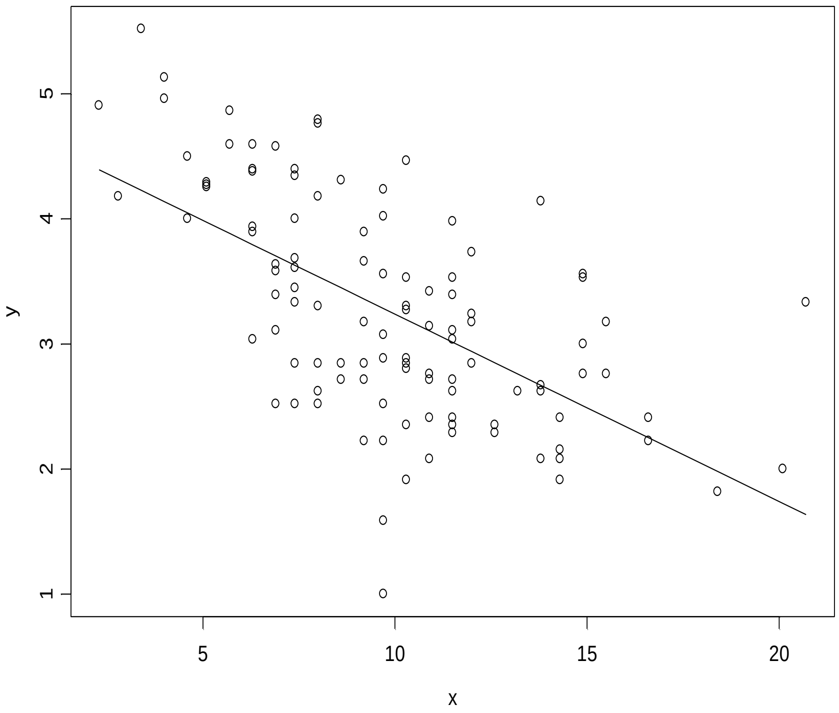
We assume that the ozone concentration (y) is a function of the wind speed (x)

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

The mean of y

$$E(y \mid x) = \alpha + \beta x$$

Data and predicted values




```
> fit.lm <- lm(y ~ x)
> summary(fit.lm)

Call: lm(formula = y ~ x)
Residuals:
    Min      1Q   Median     3Q   Max
 -2.284 -0.5144 -0.01934 0.5041 1.697

Coefficients:
               Value Std. Error  t value Pr(>|t|)
(Intercept)   4.7369   0.2025    23.3961   0.000
          x  -0.1498   0.0192    -7.8084   0.0000

Residual standard error: 0.7163 on 109 degrees of
     freedom
Multiple R-Squared: 0.3587
F-statistic: 60.97 on 1 and 109 degrees of
     freedom, the p-value is 3.823e-012

Correlation of Coefficients:
  (Intercept)
x -0.9419
```

$$\hat{\beta} = -0.1498, \qquad t = -7.8084$$

Test of hypotheses

We assume that the relationship between the ozone and the wind is linear

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

We wish to test the null hypothesis that the ozone level does not depend on the wind speed

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Testing of hypotheses: simple linear regression model

In terms of the mean of the ozone, the hypotheses can be formulated as

$$H_0 : E(y_i) = \beta_0$$

$$H_1 : E(y_i) = \beta_0 + \beta_1 x_i$$

F-test

$$F = \frac{(RSS_0 - RSS_1)/(df_1 - df_0)}{RSS_0 / df_0}$$

```
> anova(fit.lm)
Analysis of Variance Table

Response: y

Terms added sequentially (first to last)
          Df Sum of Sq  Mean Sq F Value   Pr(F)
    x      1  31.28305 31.28305 60.9711 3.823164e-012




Residuals 109  55.92571  0.51308
```

Nonparametric inference: model formulation

We assume that the ozone (y) is a function of the wind speed

$$y_i = r(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

The function $r(x)$ represents the dependency of the ozone on the wind

We do not specify a parametric structure to r(x)

$$r(x) = E(y \mid x)$$

Nonparametric inference: testing for no effect

For a simple linear regression model, the alternative hypothesis assume that the association is linear

We can relax this assumption and assume that there is an association between the ozone and the wind with out making any assumption about the mean structure

$$\boxed{\begin{aligned} &H_0 : E(y_i) = \beta_0 \\ &H_1 : E(y_i) = \beta_0 + \beta_1 x_i \end{aligned}} \longrightarrow \boxed{\begin{aligned} &H_0 : E(y_i) = \beta_0 \\ &H_1 : E(y_i) = r(x_i) \end{aligned}}$$

The no effect hypothesis

If there is no association between the ozone and the wind, i.e., the ozone level does not depend on the wind.

This means that r(x) does not depend on wind, or that r(x) is constant

$$H_0 : E(y_i) = r(x_i) = \beta_0$$
$$H_1 : E(y_i) = r(x_i)$$

Smooth function of the wind speed

Two alternative bootstrap approaches

The regression model $y_i = r(x_i) + \varepsilon_i$ can arise in two different ways

The first possibility is that the pair (x_i,y_i) were randomly sampled from a bivaraite distribution F for (X,Y).

In this case, the linear regression model refers to the conditional mean of y given x.

$$E(y \mid x) = r(x)$$

Two alternative bootstrap approaches

The second possibility is that the response y can be sample from a distribution

$$F_x(y)$$

Mean

$$r(x) = \alpha + \beta x$$

Variance

$$\sigma^2(x)$$

In this case x is not a random variable

Bootstrap algorithm in regression

Resampling cases

Model based re sampling

We will elaborate on this issue later in the course when we will discuss the topic bootstrap for linear models.

Bootstrap estimate for $r(x)$: resampling cases (1)

The data (x_i, y_i)

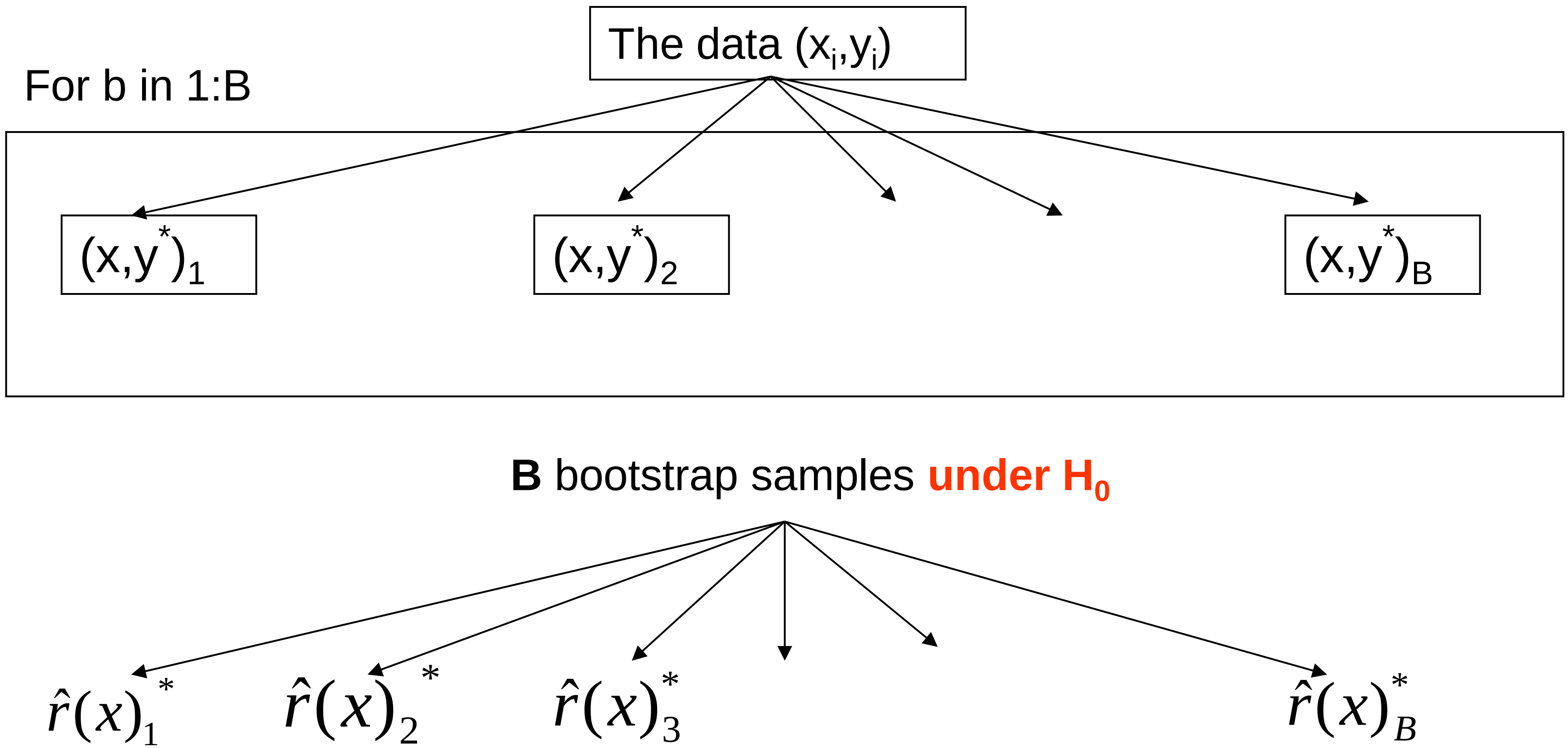
x_1, y_1^* x_2, y_2^* x_n, y_n^*

Note that we do not resample the pair but we fix x and resample from y

We obtain a bootstrap sample **under H_0**

$\hat{r}(x)^*$ is an estimate for $r(x)$ **under H0**

Bootstrap estimate for *r(x)*: resampling cases (1)



B bootstrap estimates for *r(x)* **under H_0**

A reference band for the no effect model

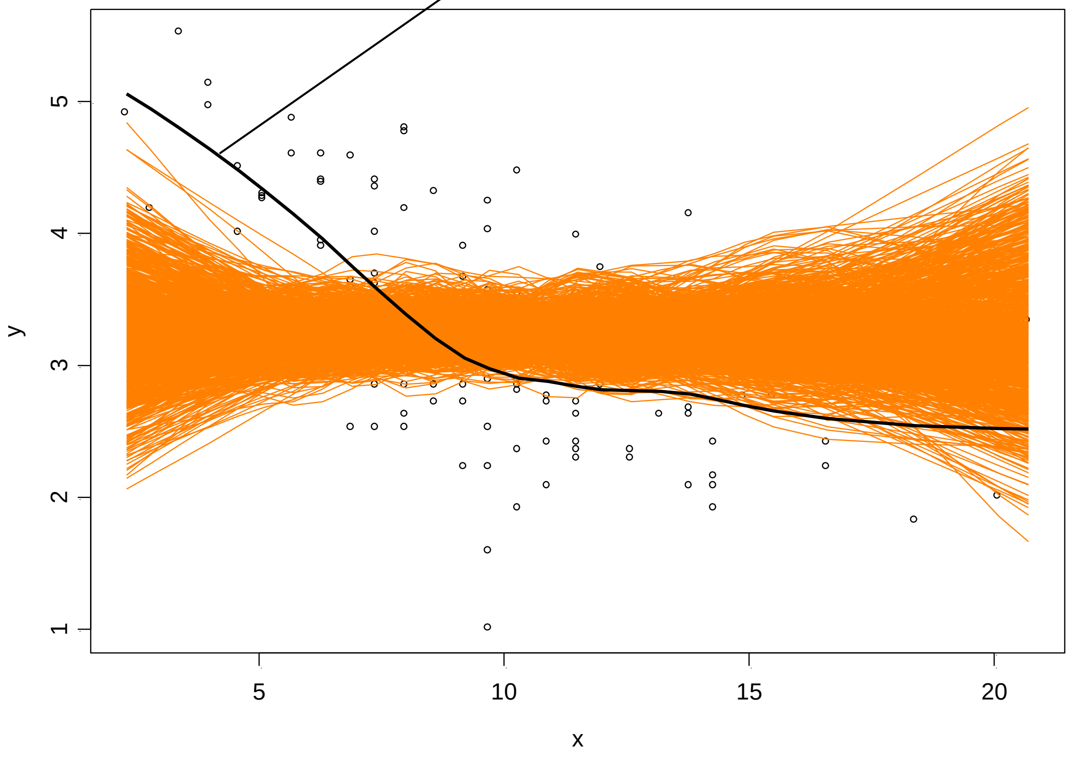
We wish to construct a confidence band under the null model

How can we re sample under the null ?

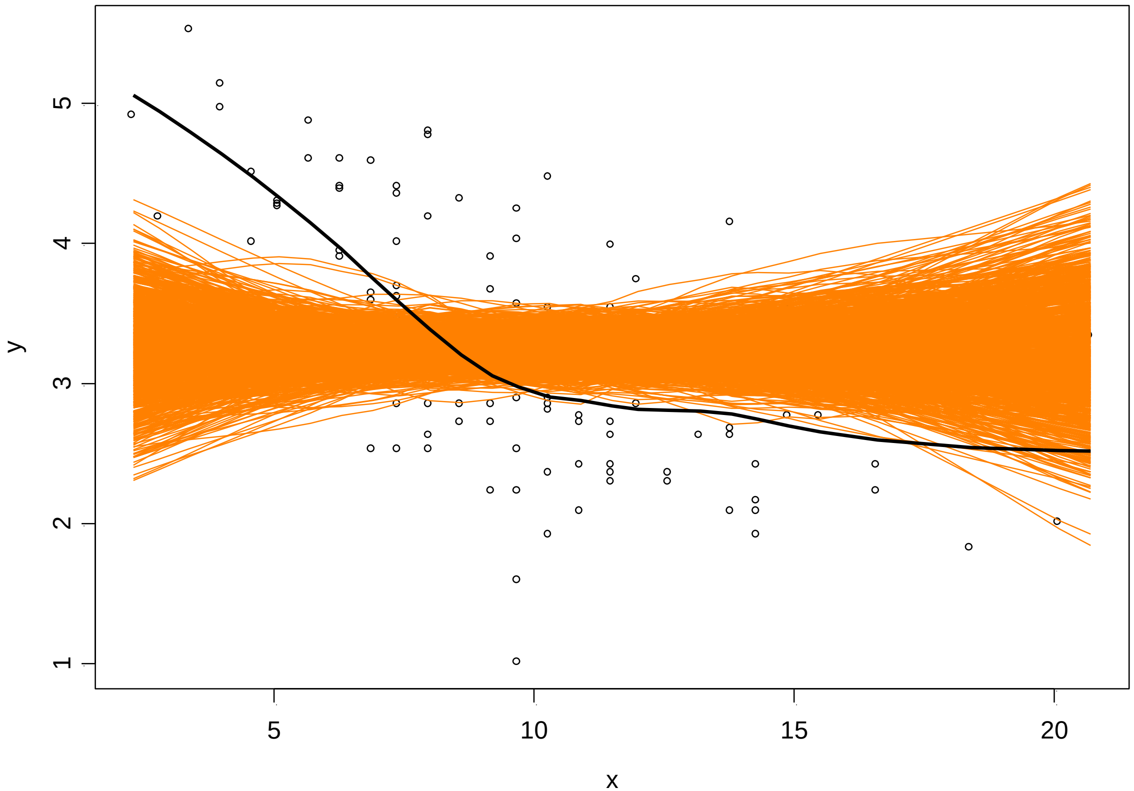
Assuming that we can re sample under the null and we smooth the data with loess, does the estimated model depend on the smoothing parameter ?

A reference band for the no effect model (lambda=0.5)

Loess model fitted to the data



A reference band for the no effect model (lambda=0.75)



R code

```
nx<-length(x)
B<-1000
boot.x<-boot.fit<-matrix(0,length(x),B)
for(i in 1:B)
{
#x.boot<-sample(x, size=nx, replace=T)
y.boot<-sample(y, size=nx, replace=T)
boot.lo<-loess(y.boot ~ x, degree = 1, span = 0.75)
boot.fit[,i]<-boot.lo$fit
#boot.x[,i]<-x.boot
cat(i)
}

plot(x,y)
fit.lo<-loess(y~x)
lines(x,fit.lo$fit)
for(i in 1:B)
{
#lines(sort(boot.x[,i]),boot.fit[,i][order(boot.x[,i])],col=5)
lines(sort(x),boot.fit[,i][order(x)],col=5)
}
lines(x,fit.lo$fit,lwd=3)
```

Testing the no effect model

We consider two possible models

1. A model in which the response does not depend on the predictor (the no effect model)
2. A model in which the response is a smooth function of the predictor

$$H_0 : E(y_i) = \beta_0$$
$$H_1 : E(y_i) = r(x_i)$$

The residuals sum of squares

We calculate the residual sum of squares under the reduced (null) and full (alternative) models

$$H_0 : E(y_i) = r(x_i) = \beta_0 \quad \longrightarrow \quad RSS_0 = \sum_{i=1}^{n} \left(y_i - \beta_0\right)^2$$

$$H_1 : E(y_i) = r(x_i) \quad \longrightarrow \quad RSS_1 = \sum_{i=1}^{n} \left(y_i - r(x_i)\right)^2$$

Note that under the null hypothesis RSS_0 is simply the square deviance from the sample mean.

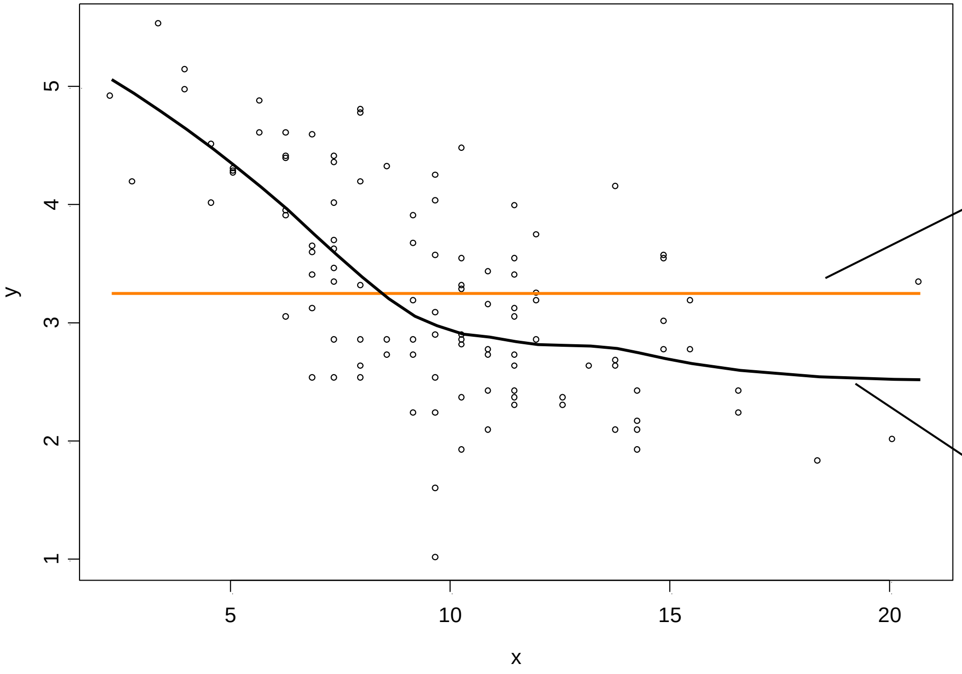
The pseudo likelihood ratio test

Similar to the linear regression case we can define a test statistics which quantify the difference between the residuals sum of squares under each model

$$H_0 : E(y_i) = r(x_i) = \beta_0$$
$$H_1 : E(y_i) = r(x_i)$$

$$F = \frac{RSS_0 - RSS_1}{RSS_1}$$

Graphical interpretation



$$RSS_0 = \sum_{i=1}^{n} \left(y_i - \beta_0 \right)^2$$

$$RSS_1 = \sum_{i=1}^{n} \left(y_i - r(x_i) \right)^2$$

The observed statistic

$$F = \frac{RSS_0 - RSS_1}{RSS_1} = 0.873$$

The observed statistics is equal to 0.873

Should we reject or not reject the null hypothesis ?

```
> fit.null <- lm(y ~ 1)
> ei <- fit.null$resid
> RSS0.null <- sum((ei^2))
> RSS0.null
[1] 87.20876
> fit.smooth <- loess(y ~ x, span =
   0.5)
> RSS1.null <-
   sum(fit.smooth$resid^2)
> RSS1.null
[1] 46.56092
> tn <- (RSS0.null -
   RSS1.null)/RSS1.null
> tn
[1] 0.8730034
```

Intuitively we should reject the null hypothesis if F is “large”.

F=0.872 ???

We need to approximate the distribution of F under the null

Bootstrap algorithm for testing no effect

For b=1 to B

1. Sample with replacement from X and Y

2. Calculate RSS_0 and RSS_1 for the bootstrap sample

$$(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots\ldots\ldots, (x_n^*, y_n^*),$$

3. Calculate the pseudo likelihood ratio test

$$F_b^* = \frac{RSS_{0b}^* - RSS_{1b}^*}{RSS_{1b}^*}$$

Bootstrap algorithm for testing no effect

How to calculate the the pseudo likelihood ratio test ?

For b=1 to B

For each bootstrap sample we fitted the two models

$$(x_1^*, y_1^*), (x_2^*, y_2^*), \ldots\ldots\ldots, (x_n^*, y_n^*)$$

$$\downarrow$$

$$1)\, y_i = \beta_0 + \varepsilon_i$$

$$2)\, y_i = r(x_i) + \varepsilon_i$$

Bootstrap P value

We can approximate the distribution of F under the null using the B bootstrap replicates for F

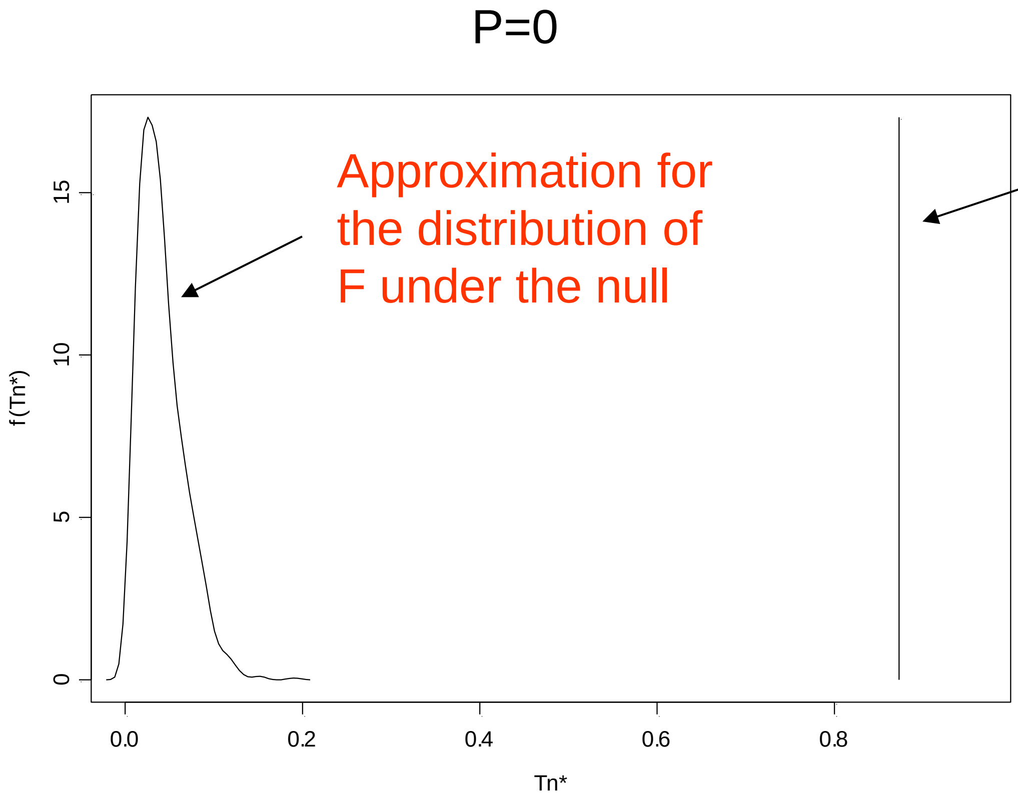
$$F_1^*, F_2^*, \ldots, F_b^*, \ldots\ldots F_B^*$$

Bootstrap P value

$$P = \frac{\#\left\{\hat{F}^* > \hat{F}\right\}}{B+1}$$

observed test statistic

Results for B=100



observed

F=0.873

Testing linear relationship

Suppose that our null model does not assume that there is no effect but that there is a liner association between the ozone and the wind

$$E_{H_0}(y_i) = \alpha + \beta x_i$$

How can we test this model ???

Null model: model formulation

Under the null hypothesis we assume that the ozone level (y) a linear function of the wind (x)

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$$

What is the alternative model ?

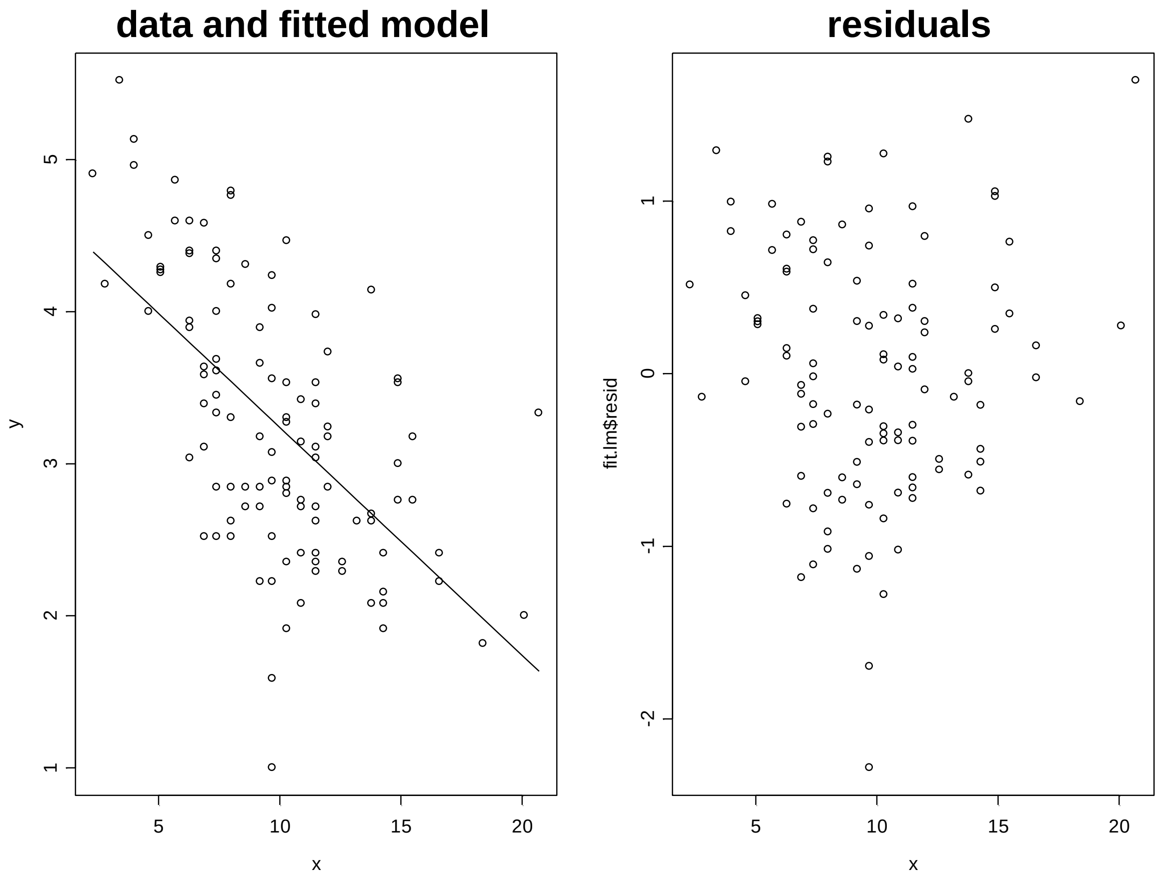
Residuals under the null model

We assume that the ozone (y) is a linear function of the wind speed

The i'th residual can be estimated by

$$e_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

Data, estimated model and residuals



The residuals

The residual:

$$e_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

The hypotheses:

$$H_0 : E(y_i) = \alpha + \beta x_i$$
$$H_1 : E(y_i) = r(x_i)$$

If the null hypothesis is correct we expect that the linear model will capture all the structure from the data

What happen if the null model is not correct ????

Reformulation of the hypotheses

We reformulate the hypotheses in terms of the residuals.

Under the null hypothesis $E(e)=0$.

Under the null hypotheses we do not expect to see any structure among the residuals

$$H_0 : E(e_i) = 0$$
$$H_1 : E(y_i) = g(x_i)$$

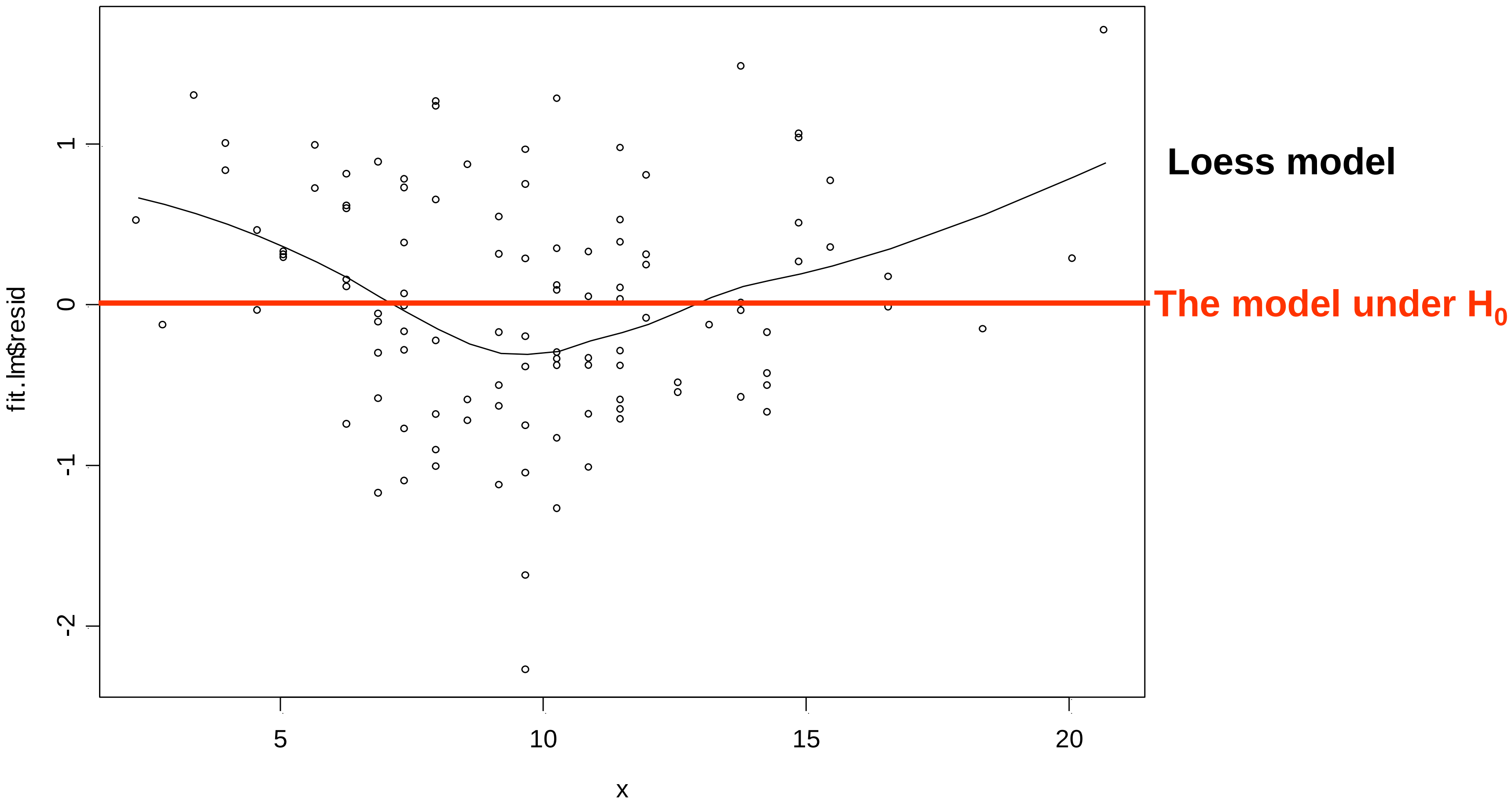
Reformulation of the hypotheses

Under H_1, we do not expect that the null model will be able to capture the structure in the data

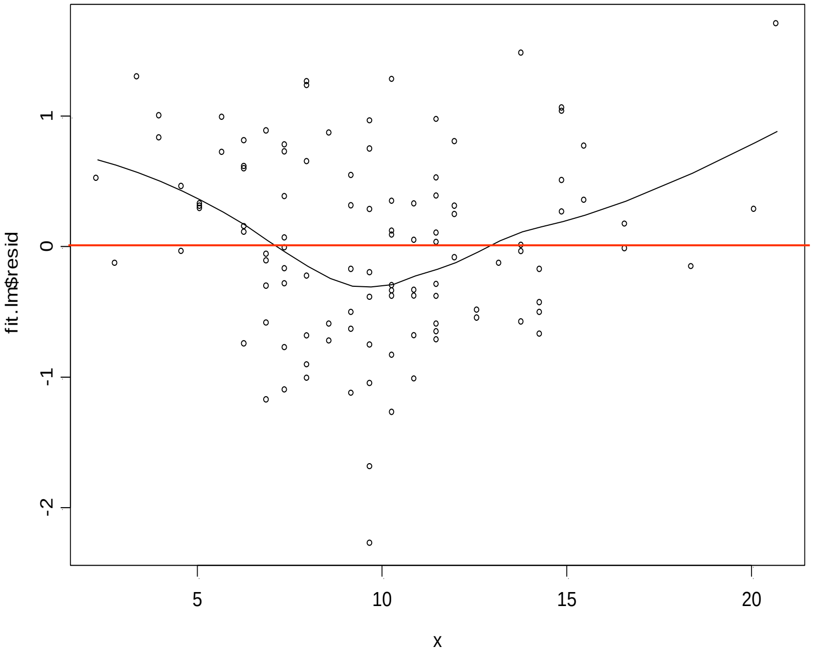
Under H_1 , $E(e)$ is a smooth function of x

$$H_0 : E(e_i) = 0$$
$$H_1 : E(e_i) = g(x_i)$$

Graphical interpretation



The observed test statistics




```
> x <- air$wind
> y <- air$ozone
> y <- y[order(x)]
> x <- sort(x)
> n <- length(y)
> tn.boot <- c(1:b) * 0
> fit.null <- lm(y ~ x)
> ei <- fit.null$resid
> RSS0.null <- sum((ei^2))
> sigma.null <- sum((ei^2))/(n - 2)
> fit.smooth <- loess(ei ~ x, degree = 1,
   span = 0.5)
> RSS1.null <- sum(fit.smooth$resid^2)
> tn <- (RSS0.null - RSS1.null)/RSS1.null
> tn
[1] 0.1801389
```

Residuals from the model under the null hypothesis

$\boxed{\text{ei <- fit.null\$resid}}$

Testing the no effect hypothesis for e

$$H_0 : E(e_i) = 0$$

$$H_1 : E(e_i) = g(x_i)$$

We wish to approximate the distribution of the test statistic under H_0

Under the null hypothesis the residuals from the linear model do not have any structure.

Under the null hypothesis the linear model capture all the structure from the data and the residuals are just random noise.

Under the alternative the residuals are a smooth function of x

First step: fit the null model

Step 1:

$$\hat{e}_i = y_i - (\hat{\alpha} + \hat{\beta} x_i)$$

$$\hat{e}_1, \hat{e}_2, \ldots\ldots\ldots, \hat{e}_n$$

Step 2: bootstrap………

The bootstrap algorithm

Residuals from the null model (the “observed data”)

$$\hat{e}_1, \hat{e}_2, \ldots\ldots\ldots, \hat{e}_n$$

B bootstrap samples

$\hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^*$ $\hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^*$ $\hat{e}_1^*, \hat{e}_2^*, \ldots\ldots\ldots, \hat{e}_n^*$

The bootstrap algorithm

For b=1,2,…,B

$$\hat{e}_1^*, \hat{e}_2^*, \ldots\ldots, \hat{e}_n^*$$

$$e_i^* = \beta_0 + \eta_i$$

$$e_i^* = g(x_i) + \eta_i$$

$$RSS_{0b}^*$$

$$RSS_{1b}^*$$

$$F_b^* = \frac{RSS_{0b}^* - RSS_{1b}^*}{RSS_{1b}^*}$$

The bootstrap algorithm

We can approximate the distribution of the test statistic under the null using the bootstrap replicates for F

$$\hat{F}_1^*, \hat{F}_2^*, \ldots\ldots\ldots, \hat{F}_B^*$$

Bootstrap P value

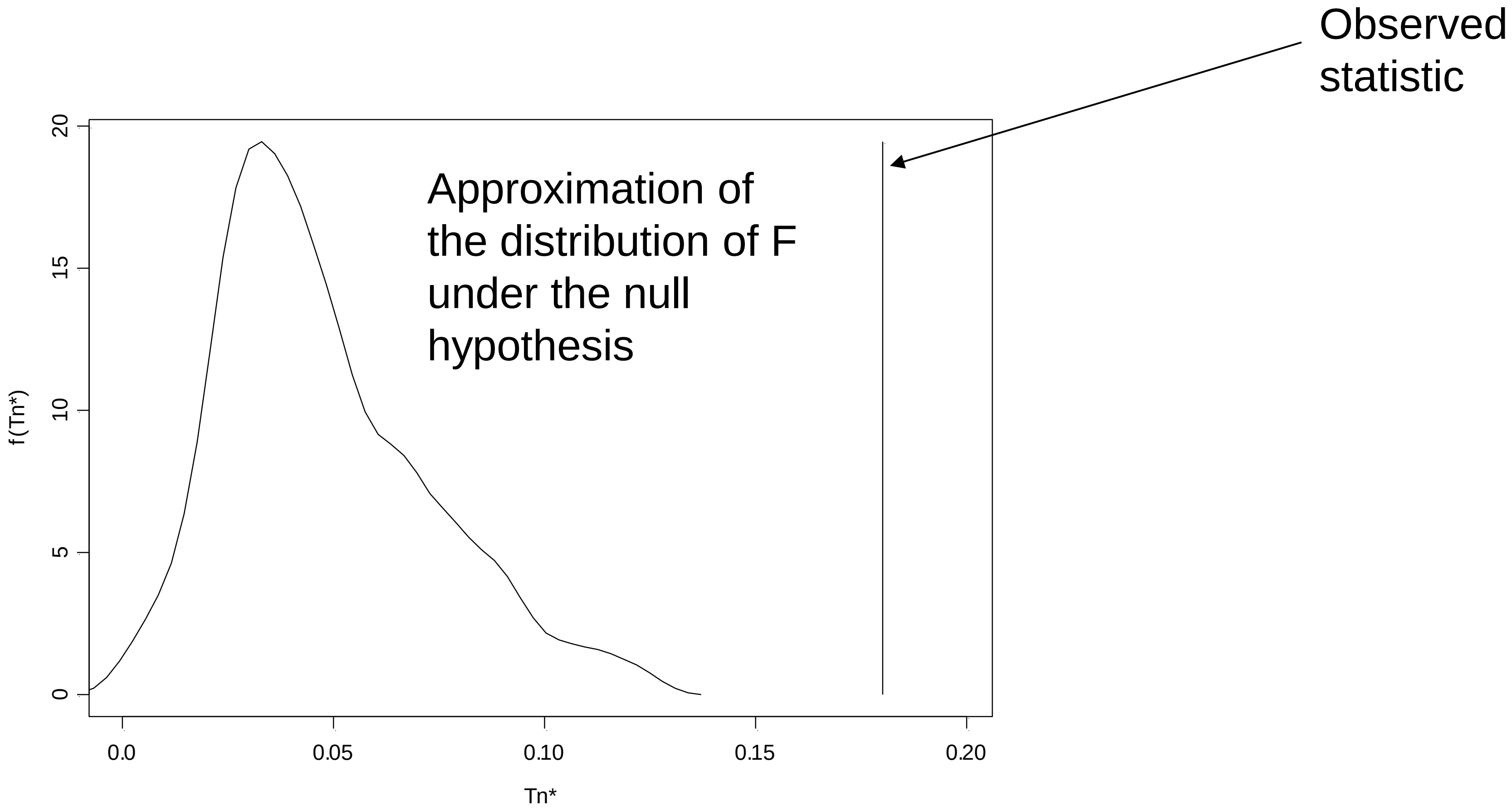
$$P = \frac{\#\left\{\hat{F}^* > \hat{F}\right\}}{B+1}$$

R code

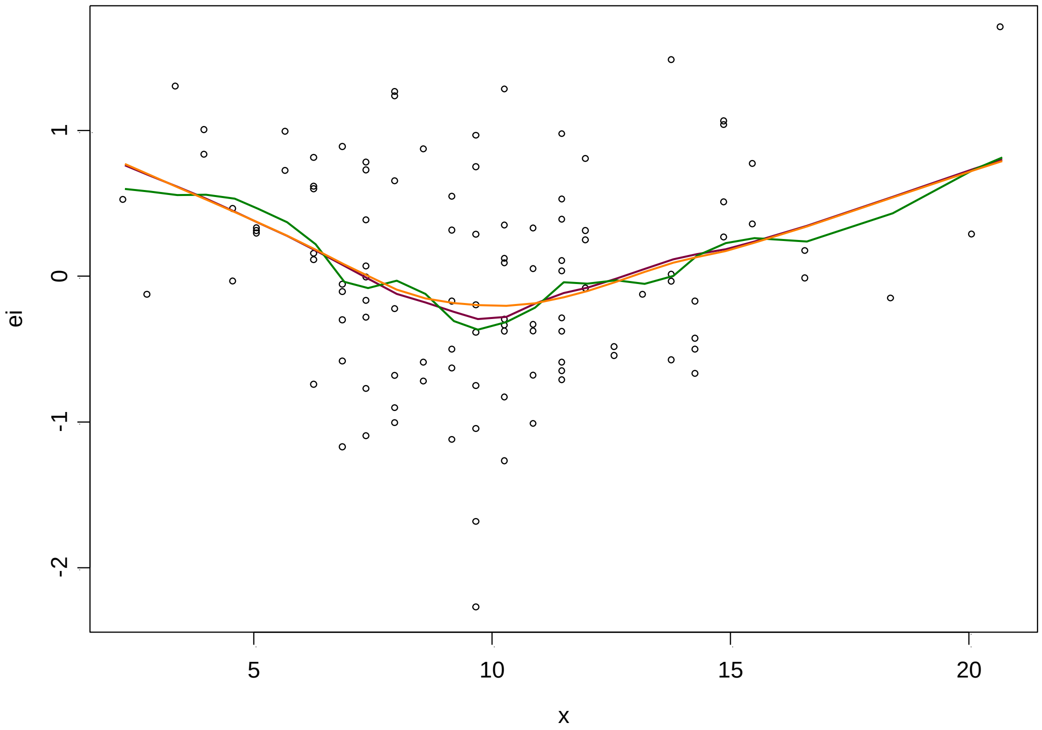
```
b<-100
mat.res<-matrix(0,n,b)
for(i in 1:b) {
          ei.boot <- sample(ei, size = n, replace = TRUE)
          fit.boot.null <- lm(ei.boot ~ 1)
          ei.boot.1 <- fit.boot.null$resid
          RSS0 <- sum((ei.boot.1^2))
          fit.boot.smooth <- loess(ei.boot ~ x, degree = 1, span = 0.5)
          RSS1 <- sum(fit.boot.smooth$resid^2)
          tn.boot[i] <- (RSS0 - RSS1)/RSS1
          mat.res[,i]<-fit.boot.smooth$fit
          cat(i)
    }
    tn.boot <- sort(tn.boot)
    p.val <- c(1:b) * 0
    for(i in 1:b) {

          if(tn <= tn.boot[i])
                    p.val[i] <- 1
    }
    p.value <- sum(p.val)/b
    p.value
```

The bootstrap P value



The effect of the smoothing parameter



The larger the smoothing parameter, the smoother the estimated model

The significance trace plot

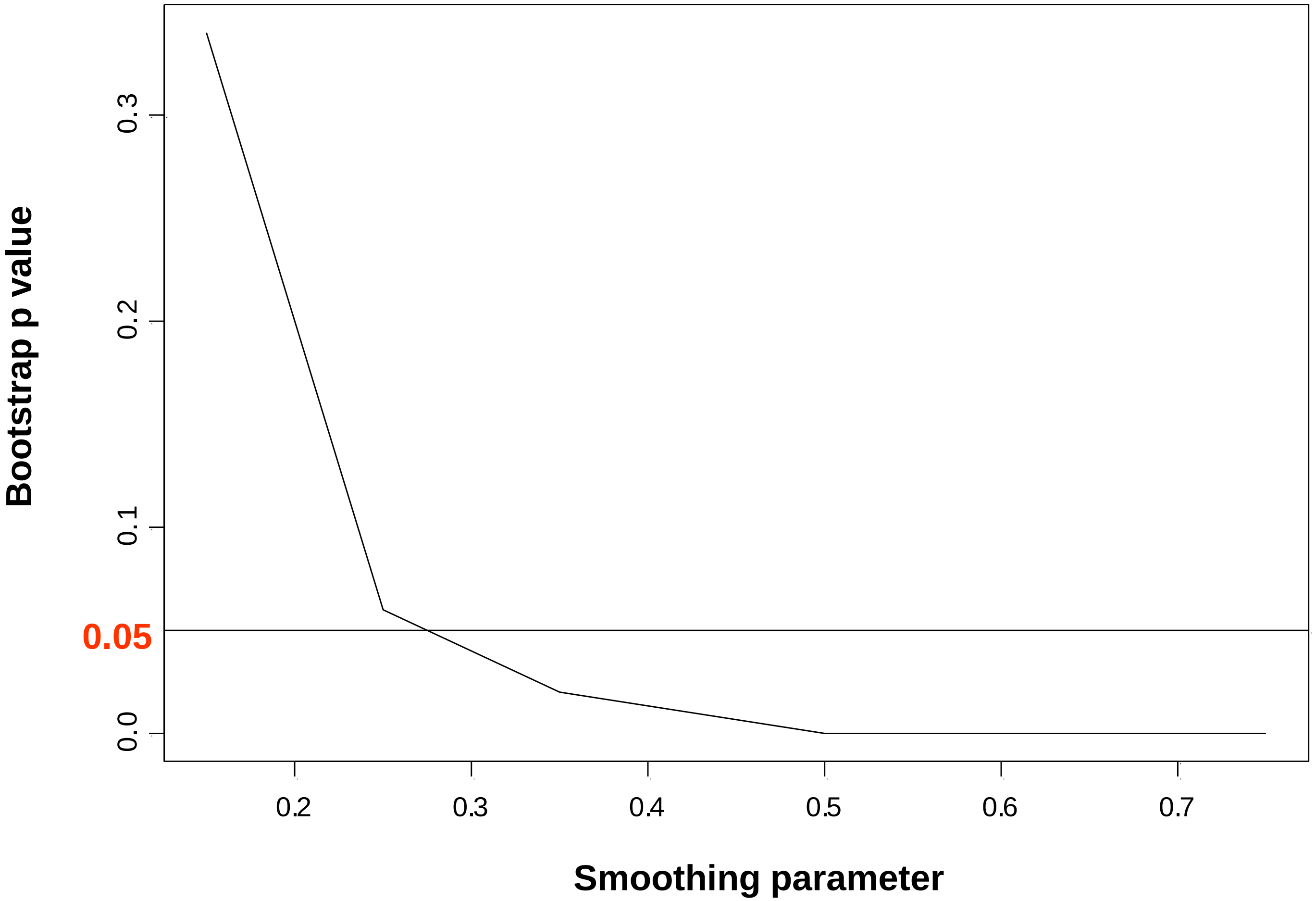
The choice of the smoothing parameter will influence the “smoothness” of the estimated model

The analysis discussed before can be repeated with difference values of smoothing parameter, for each smoothing parameter we calculate the bootstrap p value

We plot the bootstrap p values versus the smoothing parameters.

We call this plot: the significance trace plot

The significant trace plot



What do we see here ???

correction

The pseudo likelihood ratio test

Similar to the linear regression case we can define a test statistics which quantify the difference between the residuals sum of squares under each model

$$H_0 : E(y_i) = r(x_i) = \beta_0$$
$$H_1 : E(y_i) = r(x_i)$$

$$\boxed{F = \frac{RSS_0 - RSS_1}{RSS_1}}$$