Regression analysis with right censored data : Varying coefficient model

Seong Jun YANG
Hankuk University of Foreign Studies

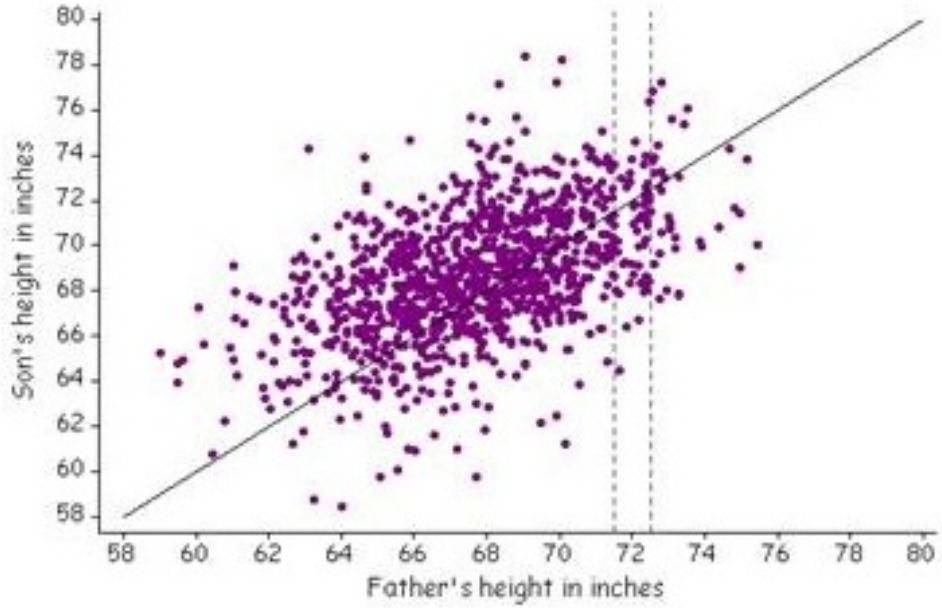
May 20, 2016, at

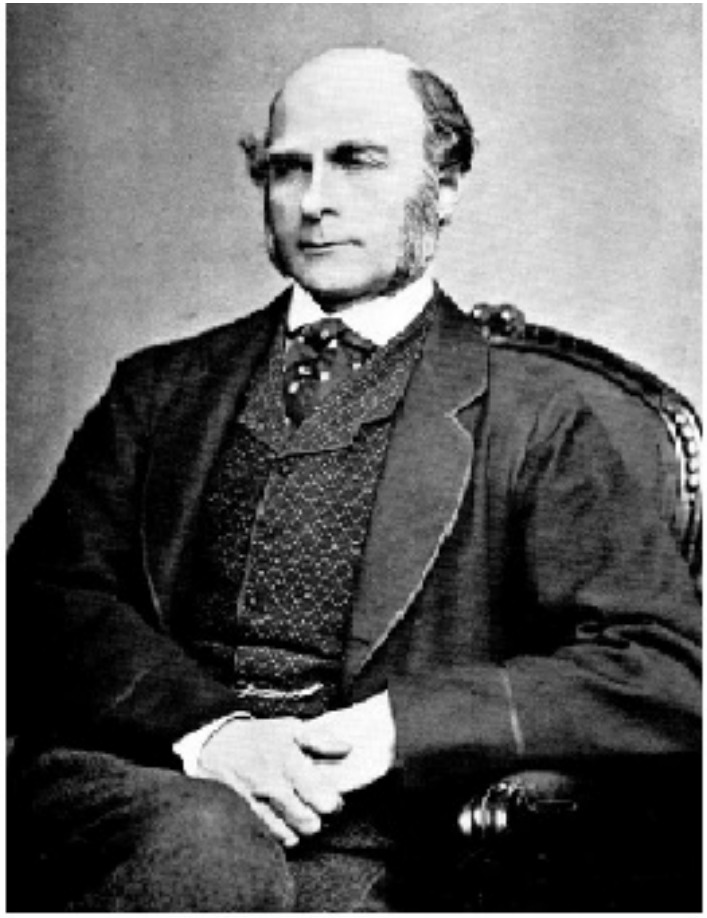
- Regression analysis
- Varying coefficient model
- Varying coefficient model with right censored data

- Regression analysis
- Varying coefficient model
- Varying coefficient model with right censored data

What is regression analysis?

- History : Francis Galton (1822-1911)





- Regression towards the mean : The son is predicted to be more like the average than the father.

What is regression analysis?

- Today : Relationships among variables

$$Y \sim f(X)$$

  X : input variable, Y : output variable
- We observe pairs of (X, Y).

$$(X_i, Y_i),\ i = 1, \ldots, n.$$

- Data generating process is contaminated by (unobservable) random error.

Parametric linear regression model

- Simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

β_0, β_1 : parameters, ϵ_i : error with mean zero
- One input, one output
- Linear function (without error)
- $E(Y_i|X_i = x) = \beta_0 + \beta_1 x \Leftarrow$ target function!

- How to estimate the model?

$$\text{Minimize} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

- Method of least squares

Parametric polynomial regression model

- Simple polynomial regression model

$$Y_i = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_i^p + \epsilon_i$$

$\beta_0, \beta_1, \ldots, \beta_p$: parameters, ϵ_i : error with mean zero
- One input, one output
- Polynomial function (without error)
- $E(Y_i|X_i = x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p \Leftarrow$ target function!

- How to estimate the model?

$$\text{Minimize} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i - \cdots - \beta_p X_i^p)^2$$

- Method of least squares

Parametric to Nonparametric

Possible choices of f

- $f(X) = \beta_0 + \beta_1 X$
- $f(X) = \beta_0 + \beta_1 X + \cdots + \beta_p X^p$
- $f(X) = \beta_3 \dfrac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)}$
- $\cdots$

$\Rightarrow$ too restrict and/or subjective

Parametric to Nonparametric

Candidates for f? $\Rightarrow f \in \mathcal{F}$, How large is it?

- $f(X) = \beta_0 + \beta_1 X \Rightarrow \dim(\mathcal{F})=2$
- $f(X) = \beta_0 + \beta_1 X + \cdots + \beta_p X^p \Rightarrow \dim(\mathcal{F})=p+1$
- $f(X) = \beta_3 \dfrac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)} \Rightarrow \dim(\mathcal{F})=3$

Parametric regression model $\Rightarrow \dim(\mathcal{F}) < \infty$

Parametric vs Nonparametric

Nonparametric regression model

- No pre-determined form of f
- $\dim(\mathcal{F})=\infty$
- f is estimated by information derived from data.

"Let the data speak for themselves"

Nonparametric regression model

- Model

$$Y_i = f(X_i) + \epsilon_i$$

$f(\cdot)$: smooth function
- One input, one output
- No specific functional form
- $E(Y_i|X_i = x) = f(x) \Leftarrow$ target function!

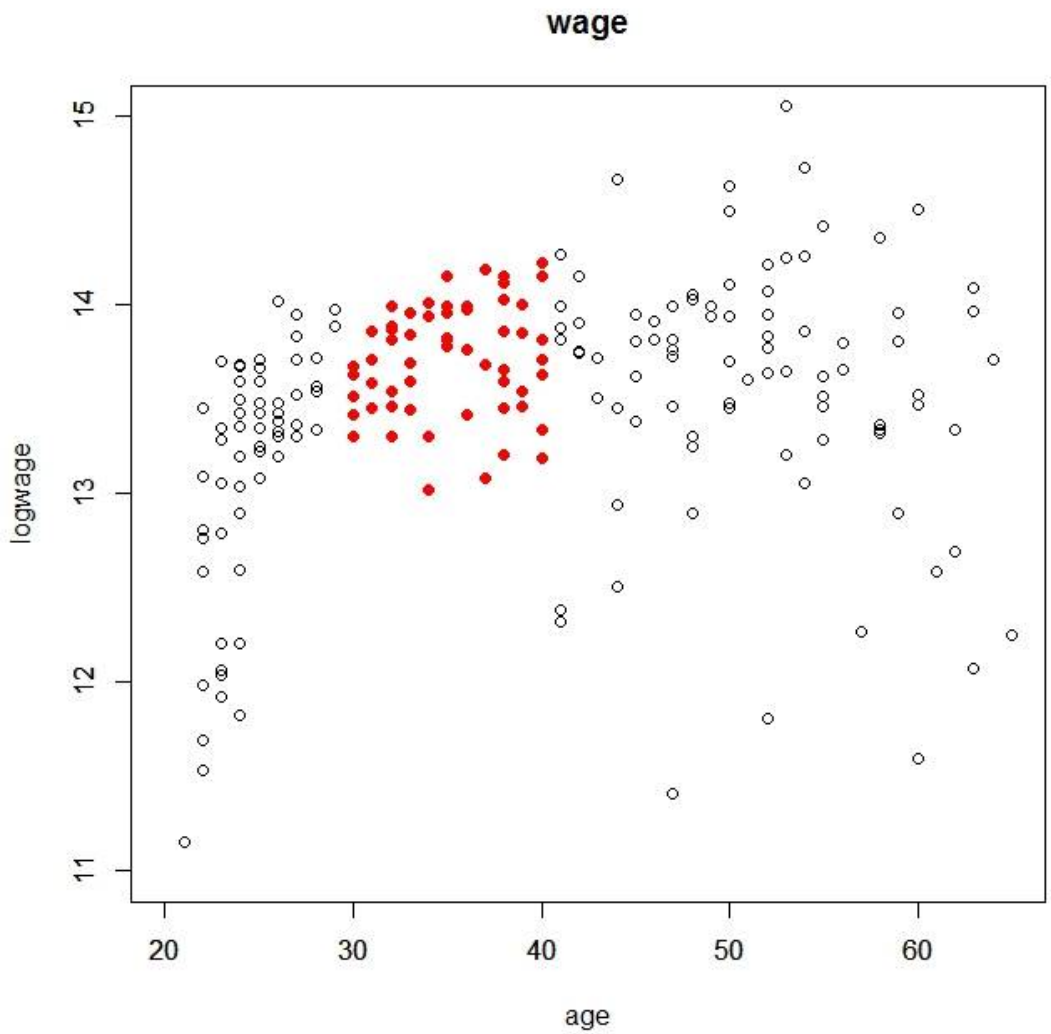
- How to estimate the model? : local modeling near x

$$\text{Minimize} \sum_{i=1}^{n} (Y_i - f(x))^2 w_i(x)$$

- $\hat{f}(x) = \dfrac{\sum_{i=1}^{n} w_i(x) Y_i}{\sum_{i=1}^{n} w_i(x)}$: weighted average of Y_i

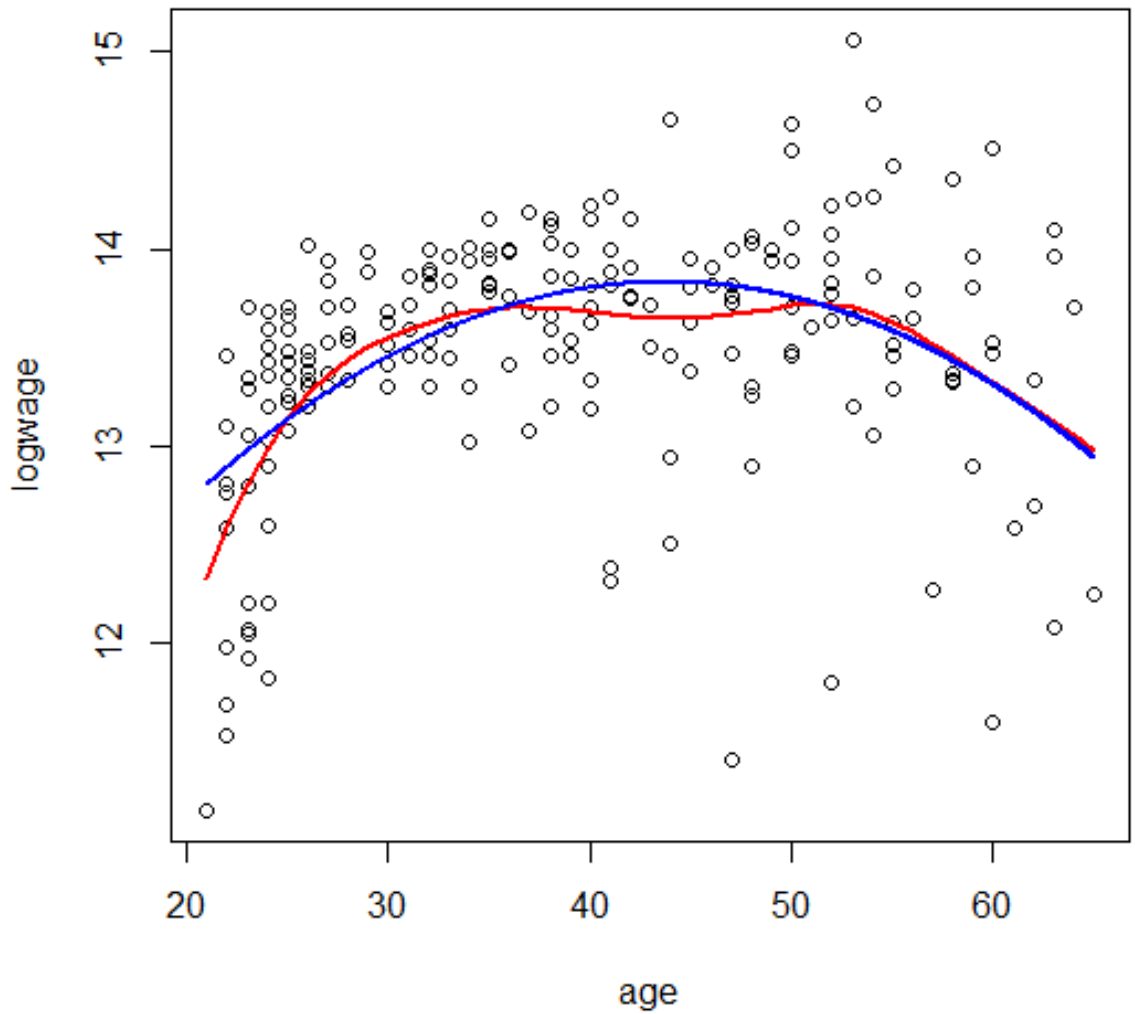
Nonparametric regression model

- wage $\sim f(\text{age})$
- $f(35)$? : local average of "wage" near age=35



Quadratic model vs Nonparametric model

- Quadratic (blue) vs Nonparametric (red)



- Regression analysis
- Varying coefficient model
- Varying coefficient model with right censored data

Extension of linear regression model

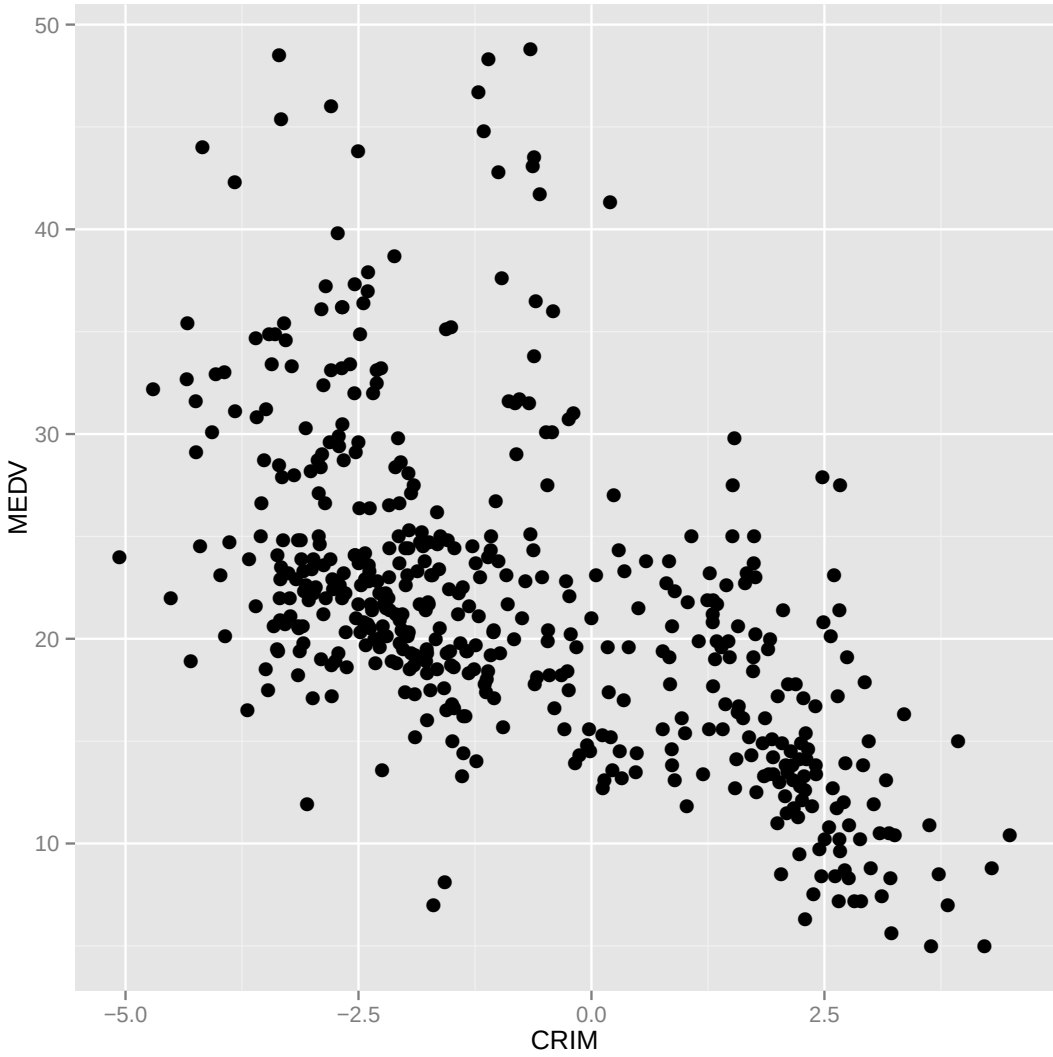
- Constant coefficients assumption : too restrict?
- Example : Boston housing data

$$\text{MEDV} \sim \beta_0 + \beta_1 \text{CRIM}$$

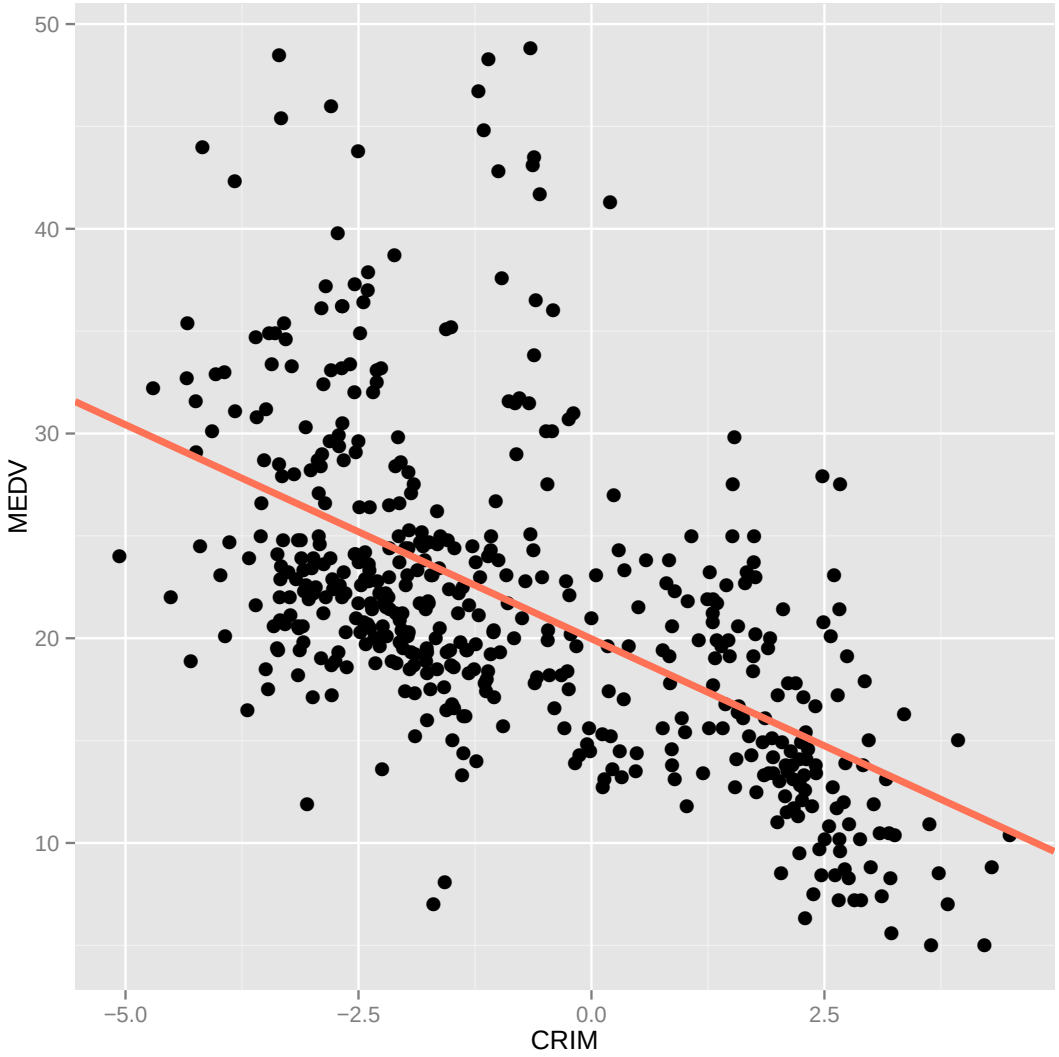
  - MEDV : median value of owner-occupied homes
  - CRIM : (log-) per capita crime rate by town

- We want to model the relationship between MEDV and CRIM.

Scatterplot : MEDV versus CRIM

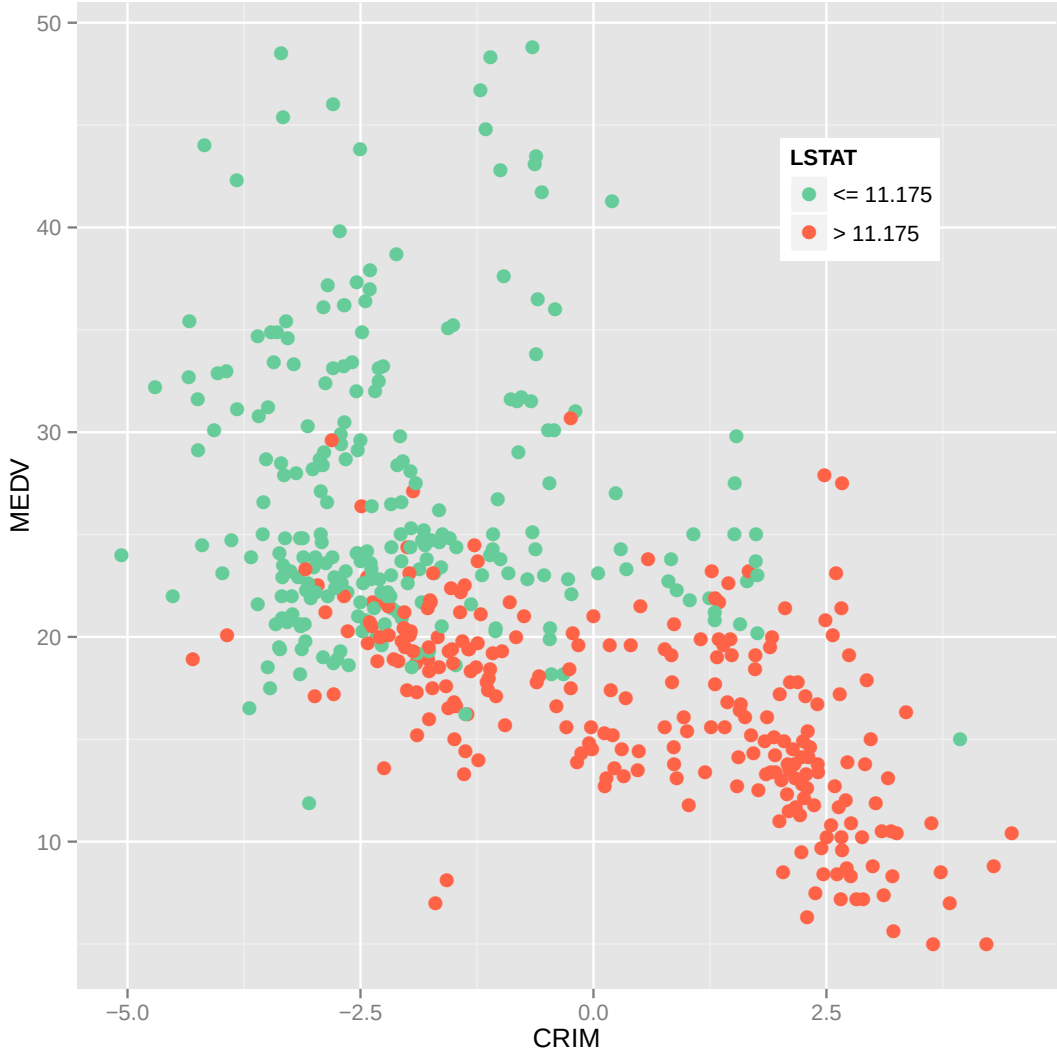


Fitting the simple linear regression model (M1)



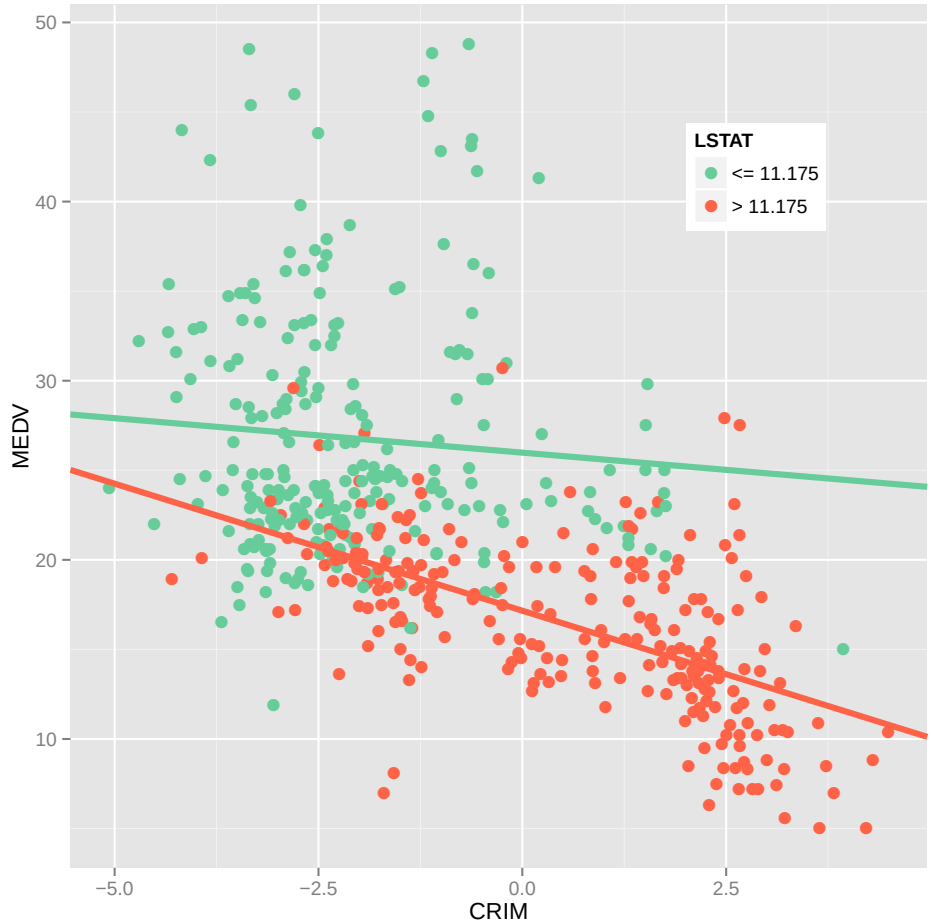
- Interpretation : CRIM has a negative effect on MEDV.

Taking a closer look : splitting by LSTAT



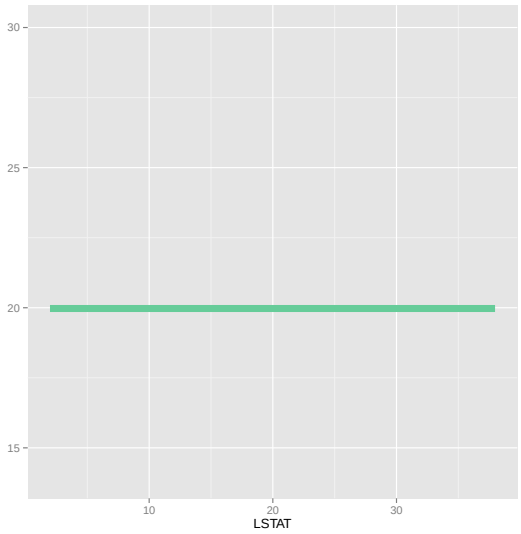
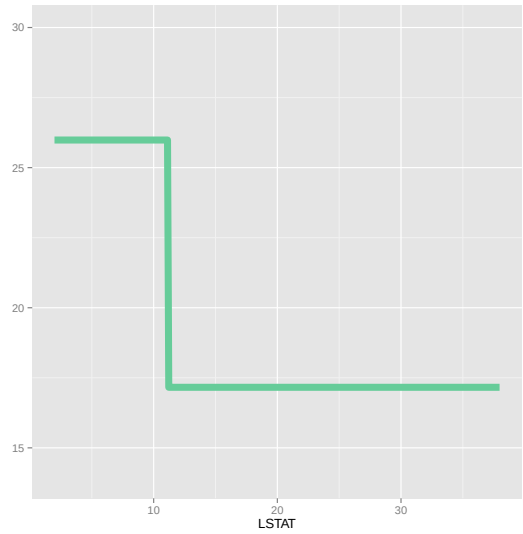
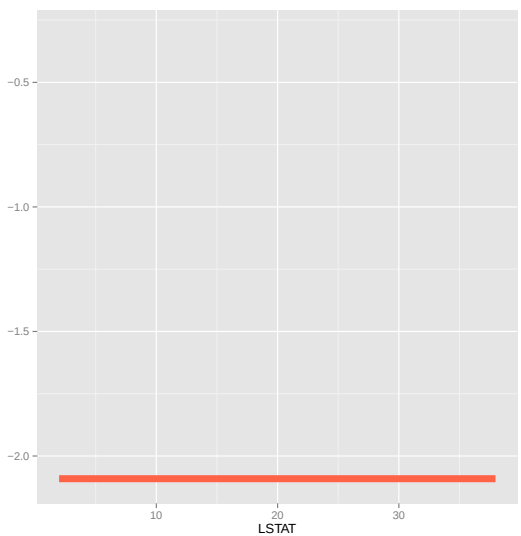
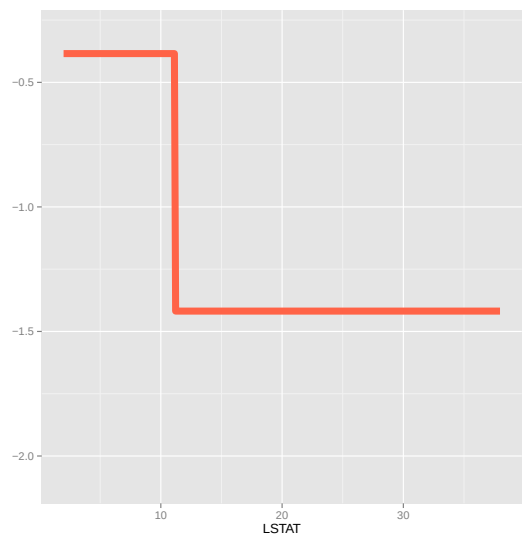
- LSTAT : lower status of the population (percent)

Fitting two different linear models (M2)



- CRIM has a more negative effect on MEDV when LSTAT is high.
  - LSTAT ≤ 11.175 : MEDV $\sim 25.99 - 0.38$CRIM
  - LSTAT > 11.175 : MEDV $\sim 17.16 - 1.42$CRIM

Comparison : M1 and M2

	M1	M2
Intercept		
Slope		

- In M2, the estimated coefficients are changing over LSTAT.
- An additional information can be extracted from the data by employing M2.
- Questions
  - We considered only one changing point in M2.
  - Why not 2 or more?
  - Why not infinitely many (or smooth) changes?

Varying coefficient model (M3)

- Model :

$$\text{MEDV} \sim \beta_0(\text{LSTAT}) + \beta_1(\text{LSTAT})\text{CRIM}$$

  $\beta_0(\cdot), \beta_1(\cdot)$: (smooth) coefficient functions.
- The coefficients are varying over LSTAT.
- If LSTAT is fixed, this reduces to the classical linear model.
- This is a nonparametric model since the parameters of interest are not constants but smooth functions.

Estimation of varying coefficient model (M3)

- Model : n copies of (Y, X, U)

$$Y_i = \beta_0(U_i) + \beta_1(U_i)X_i + \epsilon_i$$

X : input var. Y : output var. U : smoothing var.

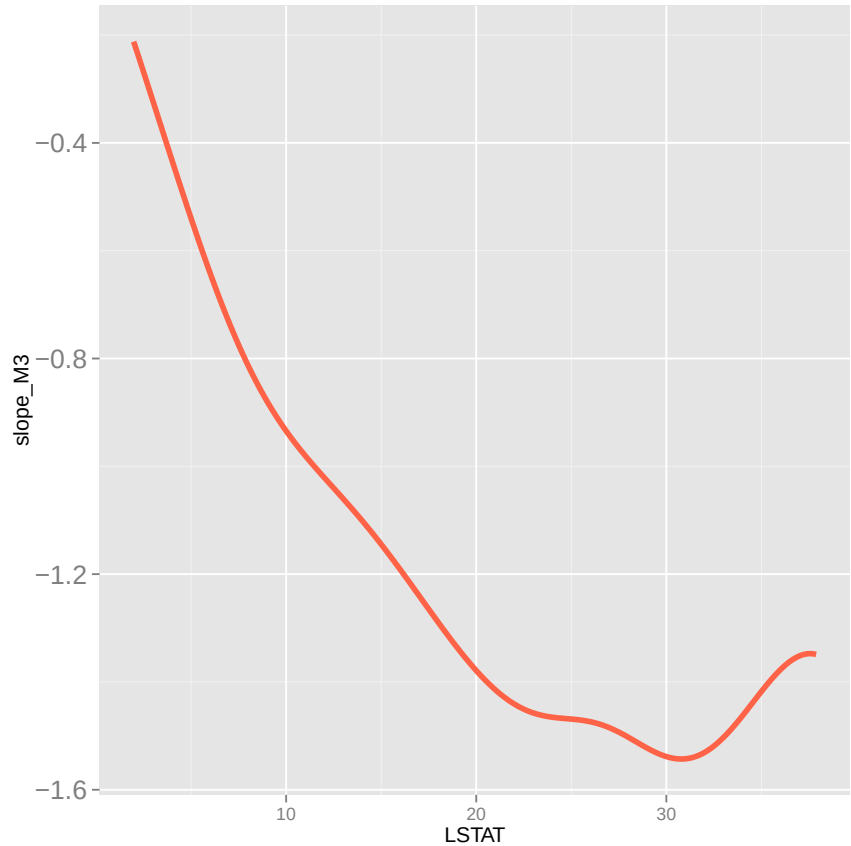
- Weighted least squares

$$\text{Minimize} \sum_{i=1}^{n}(Y_i - \beta_0(u) - \beta_1(u)X_i)^2 w_i(u)$$

- In the parametric linear model we estimate the model as follows:

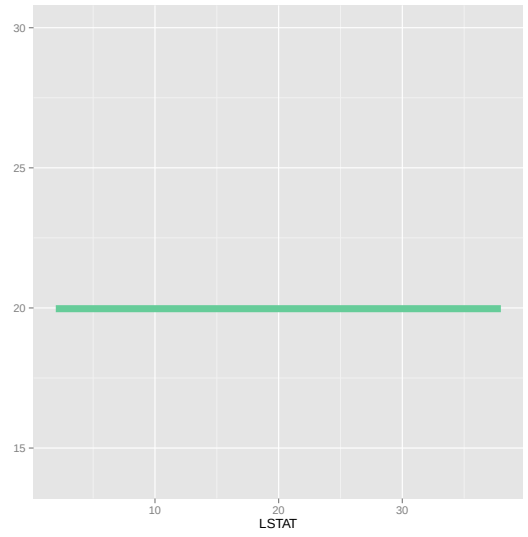
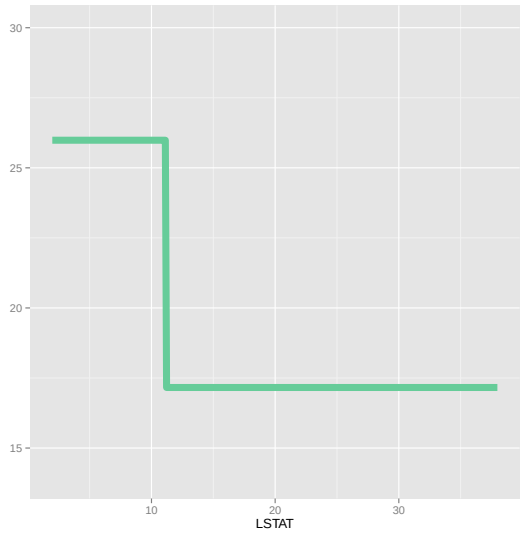
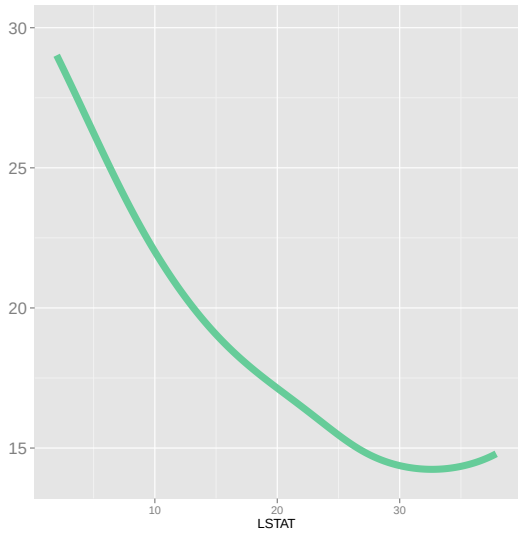
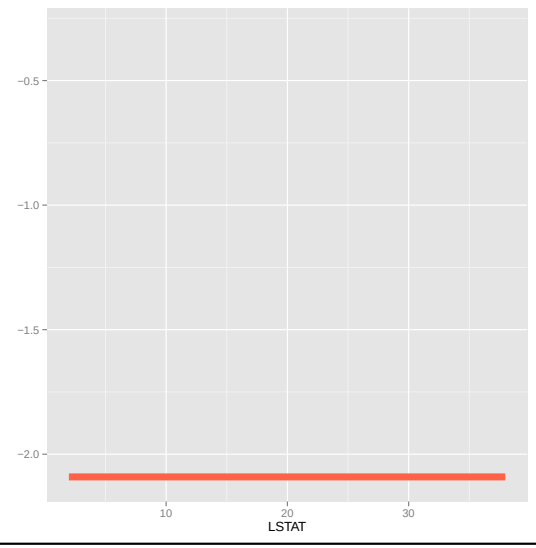
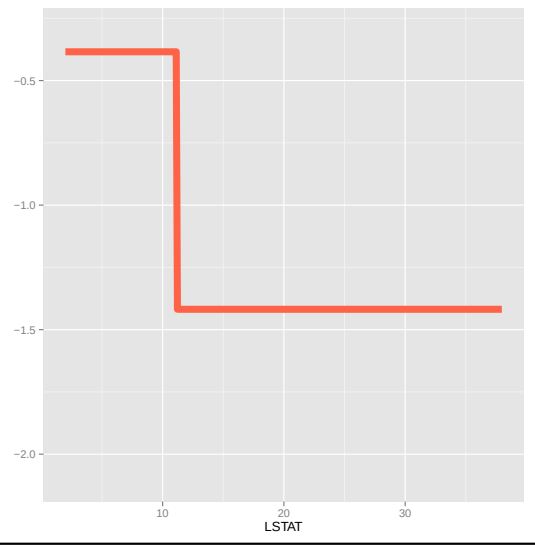
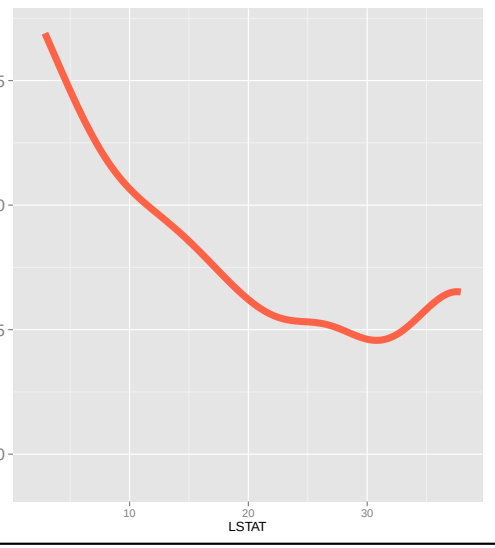
$$\text{Minimize} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

Estimated slope function in Boston housing data



- Interpretation
  - CRIM always has a negative effect on MEDV regardless of LSTAT
  - The decrease in MEDV is sharper at a higher level of LSTAT than for a lower level.

Comparison : M1, M2 and M3

	M1	M2	M3
I			
S			

Varying coefficient model (with d components)

- There may be multiple input variables.
- Model :

$$Y = \beta_1(U_1)X_1 + \cdots + \beta_d(U_d)X_d + \epsilon$$

  - U_j's can be equal or different.

- If $U_1 = \cdots = U_d = T$ where T is time.

$$Y = \beta_1(T)X_1 + \cdots + \beta_d(T)X_d + \epsilon$$

  This is called "time varying coefficient model".

Examples

Boston housing data :

- Output : median value of owner occupied home (Y)
- Input : crime rate (X_2), average number of rooms (X_3), tax (X_4), nitric oxide concentration (X_5), pupil-teacher ratio (X_6), percentage of lower income status (X_7)

- Fan and Huang (2005) considered the model

$$\begin{aligned} Y &= \beta_1(X_7) + \beta_2(X_7)X_2 + \beta_3(X_7)X_3 \\ &\quad + \beta_4(X_7)X_4 + \beta_5(X_7)X_5 + \beta_6 X_6 + \epsilon \end{aligned}$$

- Lee, Mammen and Park (2012) considered the model

$$Y = \beta_1(X_7) + \beta_2(X_3)X_2 + \beta_3(X_6)X_4 + \epsilon$$

Examples

Canadian lynx data : Fur return (x_t) at auction in 1982–1934 (log-scale)

- Threshold autoregressive model :

$$\begin{aligned} x_t &= 0.62 + 1.25x_{t-1} - 0.43x_{t-2} + \epsilon_t^1, \quad x_{t-2} \leq 3.25 \\ &= 2.25 + 1.52x_{t-1} - 1.24x_{t-2} + \epsilon_t^2, \quad x_{t-2} > 3.25 \end{aligned}$$

- varying coefficient model :

$$x_t = \beta_0(x_{t-2}) + \beta_1(x_{t-2})x_{t-1} + \beta_2(x_{t-2})x_{t-2} + \epsilon$$

Examples

Environmental data :

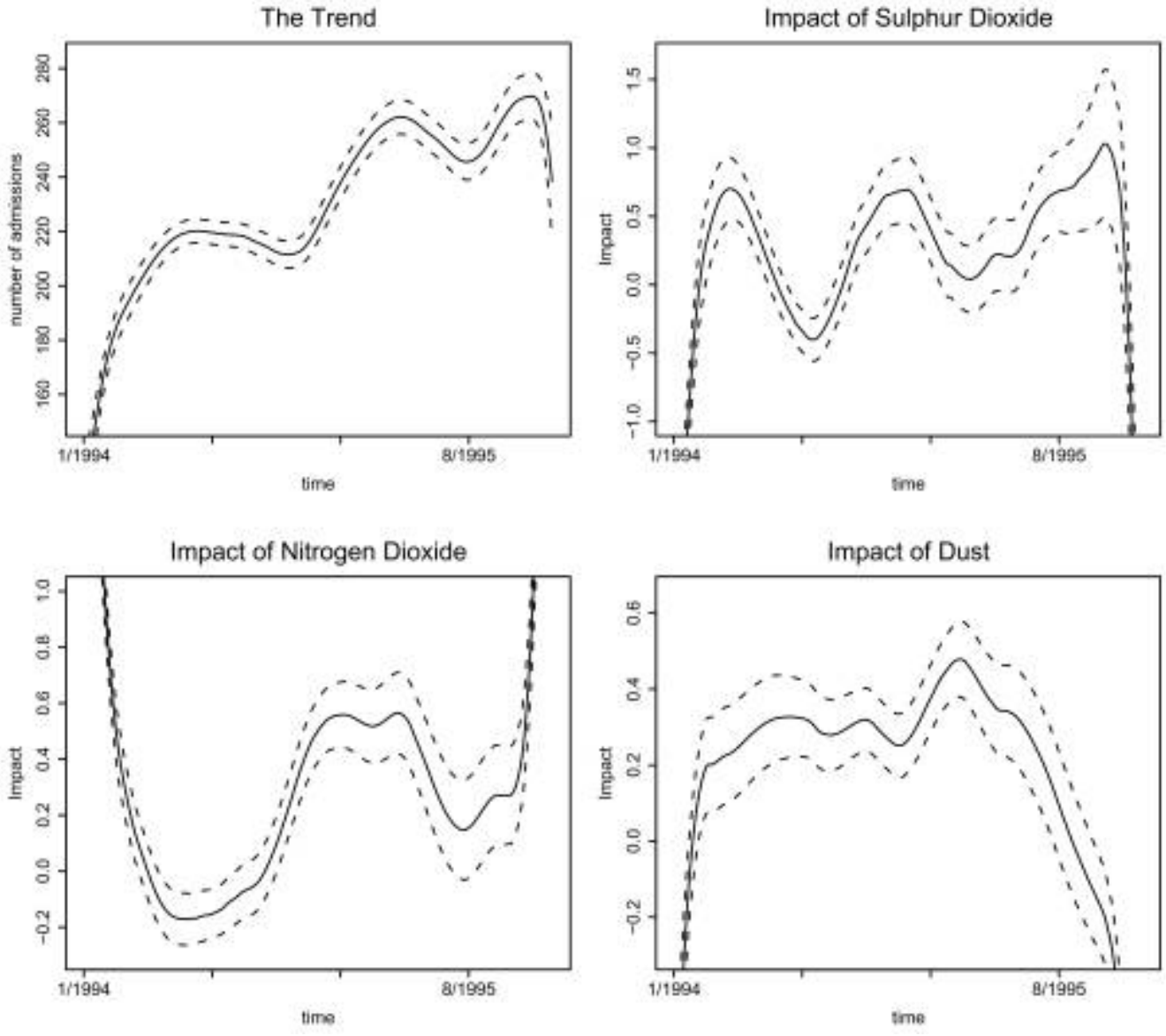
- Output : number of daily total hospital admissions for circulatory and respiratory problems (Y)
- Input : level of Sulphur Dioxide (X_2), level of Nitrogen Dioxide (X_3), level of dust (X_4)

- Fan and Zhang (2008) considered the model

$$Y = \alpha_1(T) + \alpha_2(T)X_2 + \alpha_3(T)X_3 + \alpha_4(T)X_4 + \epsilon$$

where T is time.

Examples


The Trend
number of admissions
160
180
200
220
240
260
280
1/1994
8/1995
time
Impact of Sulphur Dioxide
Impact
−1.0
−0.5
0.0
0.5
1.0
1.5
1/1994
8/1995
time
Impact of Nitrogen Dioxide
Impact
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1/1994
8/1995
time
Impact of Dust
Impact
−0.2
0.0
0.2
0.4
0.6
1/1994
8/1995
time

- Regression analysis
- Varying coefficient model
- Varying coefficient model with right censored data

Varying coefficient models in survival data

We are interested in studying the following varying coefficient model

$$Y = \beta_1(U_1)X_1 + \cdots + \beta_d(U_d)X_d + \epsilon$$

when Y is survival time.

Examples of survival data

- Medical study : time to death in patients
- Economics : time to finding a job in unemployed people
- Reliability theory : time to breakdown in machines
- and many more

$\Longrightarrow$ It records the length of time until some event occurs.

Right censoring

In survival data, right censoring is often encountered. It refers to the situation that :

- Y is only partially observed.
- Instead of observing Y, we observe

$$T = \min(Y, C) \quad \text{and} \quad \Delta = I(Y \leq C),$$

where C is the censoring time.

Right censoring occurs when a subject leaves the study by various reasons or the study is terminated before the event occurs.

Estimation without censoring

Without censoring, it is known that the coefficient functions can be estimated by minimizing the following objective function:

$$\int \sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{d} \beta_j(u_j) Z_{ij} \right]^2 w_i(\mathbf{u}) d\mathbf{u},$$

Problem : We do not observe Y.

Synthetic response

The main idea to solve the problem is to replace unobservable Y with Y^G where

$$Y^G = \frac{\Delta T}{1 - G(T-)}.$$

where $G(y) = P(C \leq y)$, which is observable if G would be known.

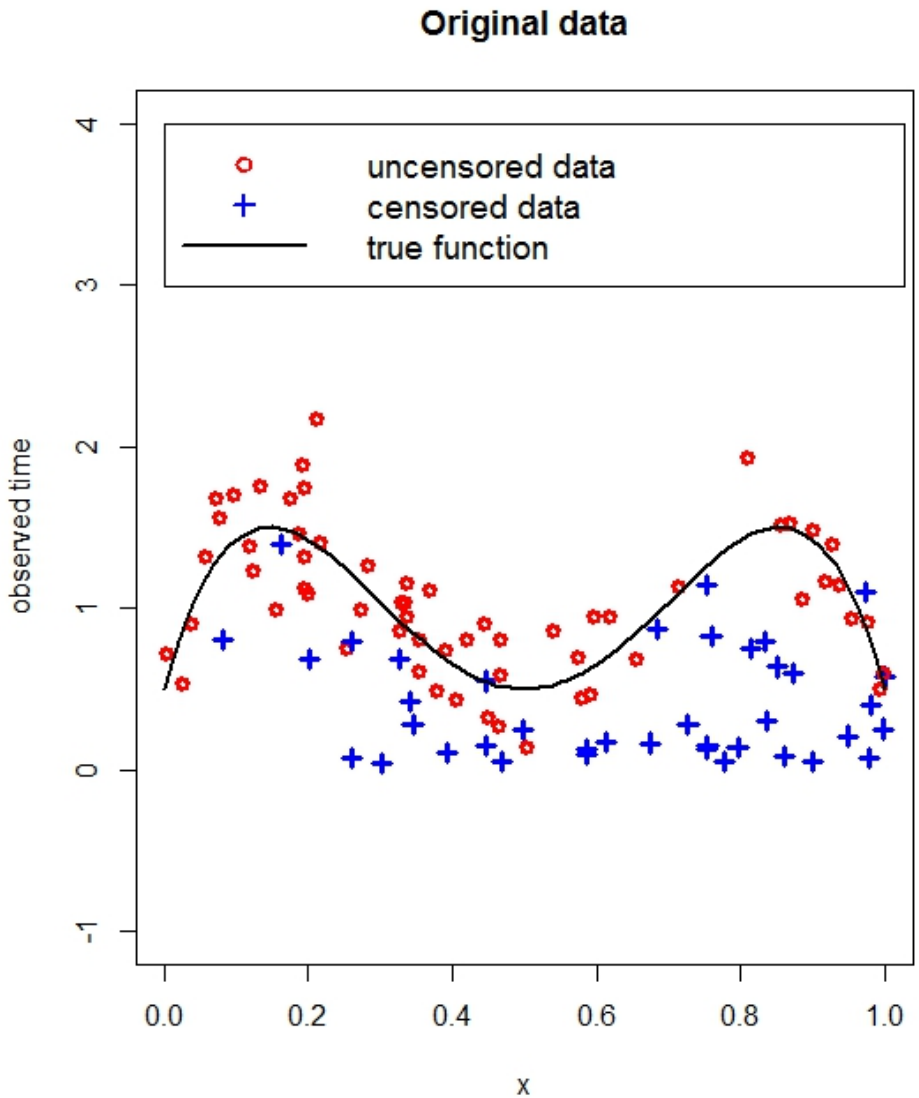
Under suitable conditions,

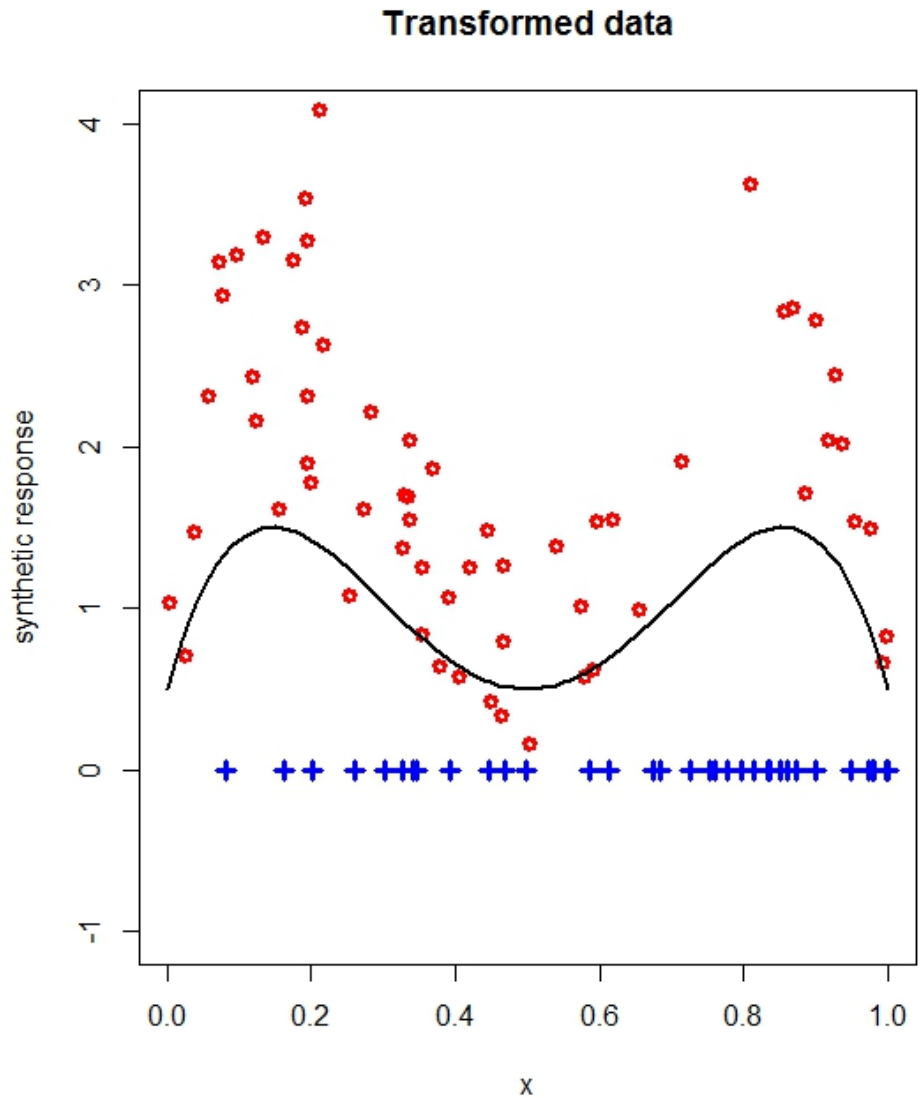
$$E(Y|\mathbf{X}, \mathbf{U}) = E(Y^G|\mathbf{X}, \mathbf{U})$$

Note that $E(Y|\mathbf{X}, \mathbf{U})$ is the regression function, which is out target.

$$E(Y|\mathbf{X}, \mathbf{U}) = \beta_1(U_1)X_1 + \cdots + \beta_d(U_d)X_d = E(Y^G|\mathbf{X}, \mathbf{U})$$

Illustration





- With the synthetic response Y^G, we can apply existing regression techniques as if there were no censoring.
- Note that

$$E(Y^G|\mathbf{X},\mathbf{U}) = E(Y|\mathbf{X},\mathbf{U})$$

but

$$\text{Var}(Y^G|\mathbf{X},\mathbf{U}) > \text{Var}(Y|\mathbf{X},\mathbf{U})$$

$\Rightarrow$ Price to pay!

Estimation

We apply existing techniques on the estimated 'synthetic response' $Y^{\hat{G}}$.

We minimize the following objective function

$$\int \frac{1}{n} \sum_{i=1}^{n} \left[Y_i^{\hat{G}} - \sum_{j=1}^{d} \beta_j(u_j) Z_{ij} \right]^2 w_i(\mathbf{u}) d\mathbf{u},$$

where $\hat{G}$ is a "good" estimator of G.

Theorem 1.

Under certain regularity conditions and G is continuous,

- $\widehat{\beta}_1^{\widehat{G}}(u_1), \ldots, \widehat{\beta}_d^{\widehat{G}}(u_d)$ *are asymptotically independent*
- *For* $j = 1, \ldots, d$,

$$\widehat{\beta}_j^{\widehat{G}}(x_j) - \beta_j(x_j) = n^{-2/5} W_j + o_p(n^{-2/5})$$

where

$$W_i \sim N(b_j(u_j), \mathbf{V}_j(u_j))$$

$b_j(\cdot)$ *&* $\mathbf{V}_j(\cdot)$ *are suitably defined bias and variance functions.*

Real data example

data on drug abuse :

- $Y = \log(\text{time to return to drug use in days} / 365.25)$
- $AGE =$ age in years
- $BECK =$ Beck's depression score : score between 0 (not depressed) and 63 (severely depressed), which is a 21-question multiple-choice self-report inventory
- $IVHX =$ drug use history ($0 =$ never, $1 =$ present)
- $NDT = \log(\text{number of prior drug treatments})$
- $LOT =$ length of treatment in days

Sample size : $n = 398$

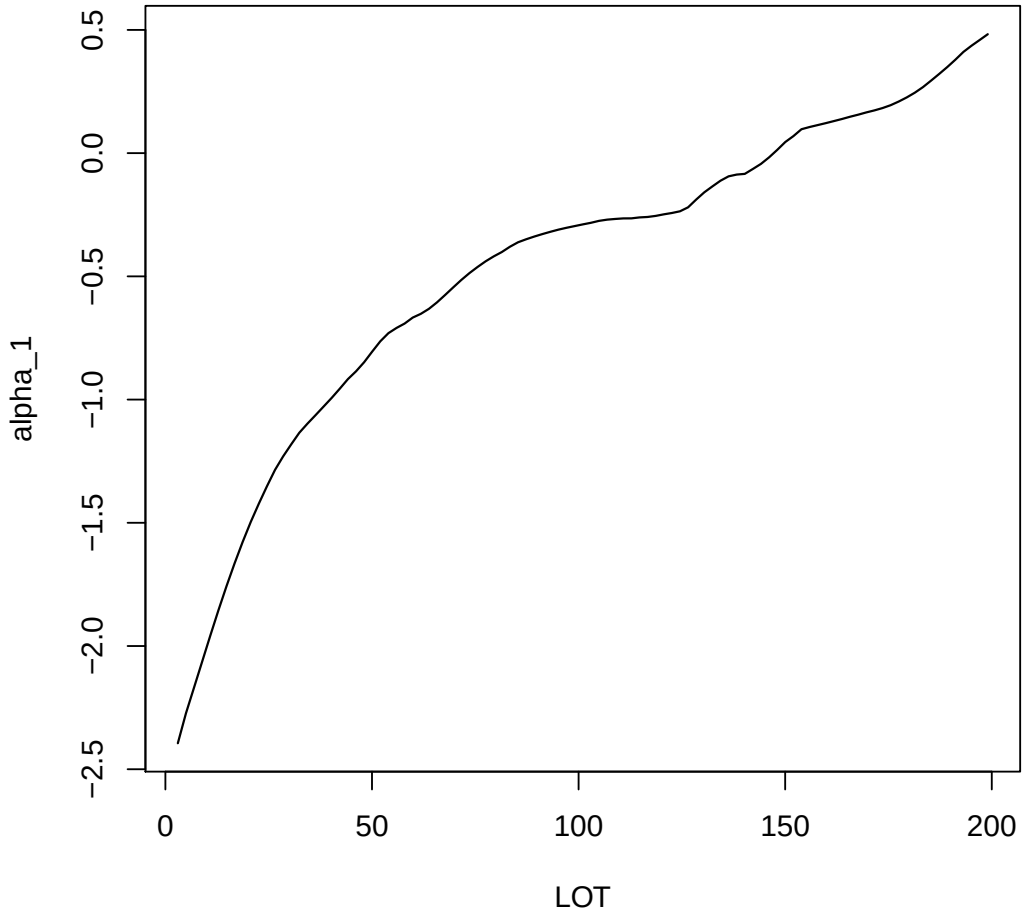
The response Y is subject to random right censoring, and $P(\Delta = 0) \approx 0.20$

Model :

After processing with some simple selection procedures, we select the final model as

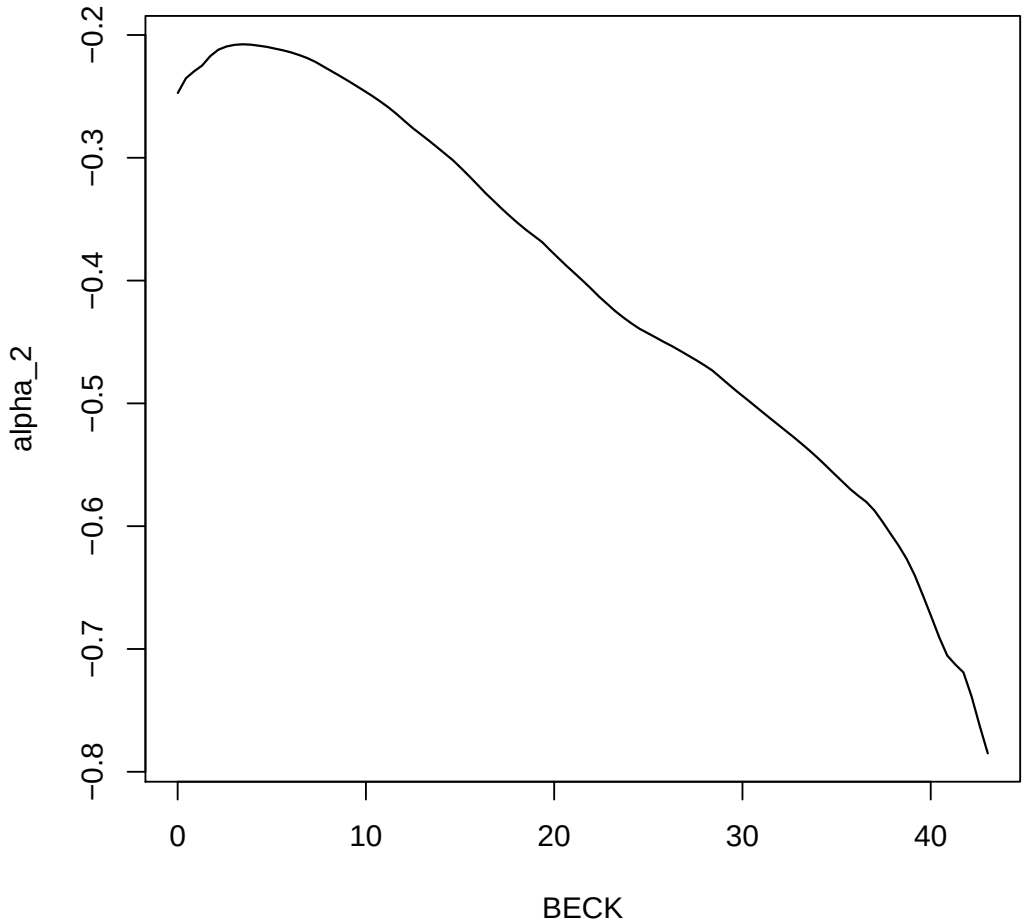
$$Y = \alpha_1(LOT) + \alpha_2(BECK)IVHX + \alpha_3(AGE)NDT + \varepsilon.$$

Estimation of α_1 : $Y \sim \alpha_1(LOT)$



$\Rightarrow$ Time to return to drug use increases as LOT increases

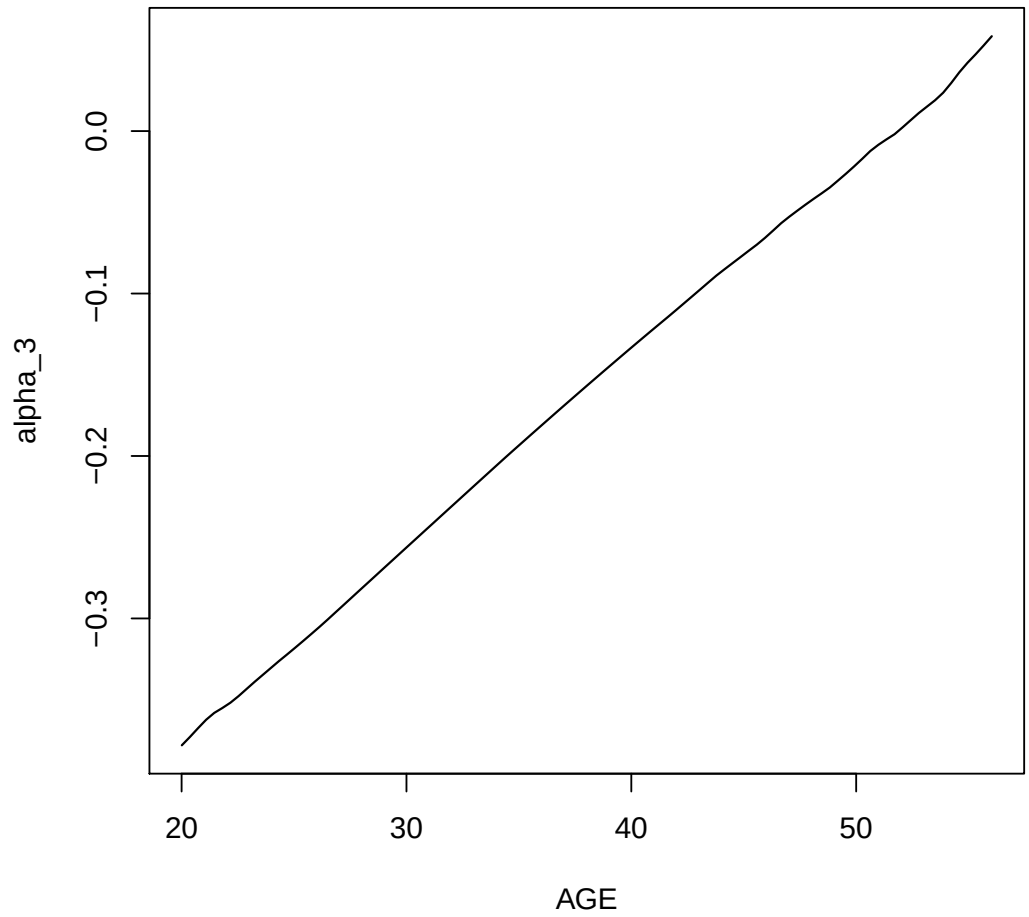
Estimation of α_2 : $Y \sim IVHX\,\alpha_2(BECK)$



Coefficient of IVHX ≤ 0 for all values of BECK

$\Rightarrow$ Patients who have drug history return to drug use earlier, and time to return to drug use decreases with BECK depression score

Estimation of α_3 : $Y \sim NDT\,\alpha_3(AGE)$



Trend is nearly linear

$\Rightarrow$ AGE and NDT have linear interaction effect

Note that the function passes through 0 around AGE=46. Hence,

- For young patients NDT has a negative effect on time to return to drug use (i.e. they tend to return to drug use earlier if they experienced many drug treatments)
- An opposite trend is observed for older patients

This seems reasonable, since large values of NDT (Number of prior drug treatments) for young people means that they are strongly addicted to drugs.

Thank you!