Lecture 16: Use of Randomization and Nonparametric Regression

Prof. Esther Duflo

14.310x

Uses of Randomization

- Clinical Trials
- Social Policy experiments
- Research
- A/B testing
- Marketing Experiments

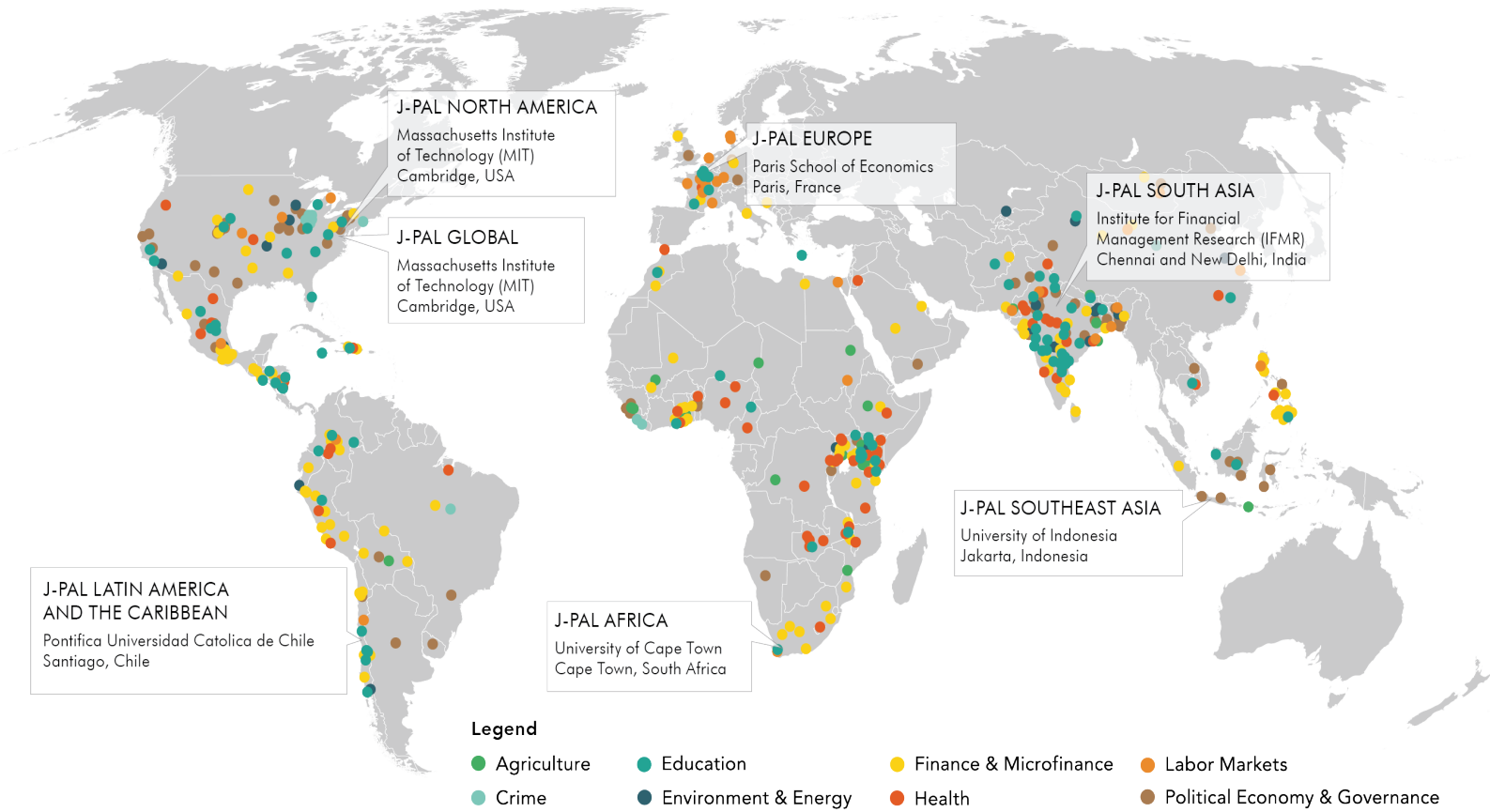
Clinical Trials

- It has been "gold Standard" for many years.
- FDA require randomized trials before approving drugs
- Some issues that the profession is dealing with
  - Selective reporting of results from experiments.
    - Why is that a problem?
    - Solution is registry of all experiments. Problem is not everybody reports all results in registry!
- Selective reporting of results by subgroups
  - Why is that a problem?
  - Solution is pre-analysis plan. Problem is not everybody follows the PAP!

Social experiments

- RCT have a long history in social sciences as well.
- Rand Health insurance experiment; negative income tax experiment
- MDRC (Bob Solow was chairman of the board!) has conducted welfare experiments for decades, and done a lot to increase the acceptability of the method,
- They have continued all this while and have now greatly expanded in scope (geographical and in terms of theme), size and ambition

J-PAL 729 (and counting) evaluations in 67 countries



A/B testing

- Strictly speaking it is the comparison of two version of a web page (A vs B)
- Users are randomly directed to either A or B
- (they have invented another name, multivariate testing, when you will compare A, B, C or D)
- Outcome variable is defined by your need, but can be clicks, purchases, etc.
- A/B testing has become standard practice in web-based business: almost free, natural metric, lots of uncertainty.
- And is now used a bit more broadly to refer to any RCT in firm setting...

Marketing

- Some companies have always used RCT: Capital one was always known for experimenting with layout of letters.
- But use was surprisingly limited until fairly recently when A/B testing has implanted the idea with managers that you could more generally experiment.
- The last 10 years or so have seen an explosion of marketing experiment (Duncan Simester reviews 61 field experiments in marketing since 1995, including 37 since 2010!)
- Type of questions: Pricing; and Advertising.

Randomization as a research tool

- Social science researchers also use randomization to answer questions that they are interested in !
- Let's go back to the race question that we started the lecture with.
- Suppose you are interested in the question we first asked: "she cannot get a job interview because she is Black". As we explained we have to specify the question a bit better.
- If the question is about discrimination, then can we keep everything constant except the perception of race?
- Bertrand and Mullainathan: Resume study. Send resume with black (lakisha, Jamal) or white (emily, greg) sounding first names in response to job applications, and wait for call backs.

Resume Study

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the p-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

Social Science Research meets A/B testing and policy

- Blurring of boundaries between research, A/B testing, and policy
- White house launches SBST (Social and behavioral science Team): "A group of experts in applied behavioral science that translates findings and methods from the social and behavioral sciences into improvements in Federal policies and programs for the benefit of the American people".
- Example: "To help Federal student-loan borrowers stay on top of their payments, SBST and FSA sent a reminder email to over 100,000 borrowers who had missed their first payments. The reminder email led to a 29 percent increase in the fraction of borrowers making a payment in the first week after it was sent, from 2.7 to 3.5 percent",

Fixed Effects Meta-Analysis

- Awasthi et al (2013) used Fixed-Effects (FE) meta-analysis, which computes a weighted average of the treatment effects. Each estimate is weighted by the inverse of its sampling variance.
- So if we have K studies, with estimated treatment effect $\hat{\tau}_k$ and standard error se_k for each study, then

$$\hat{\tau}_{FE} = \sum_{k=1}^{K} \hat{\tau}_k \frac{(se_k^2)^{-1}}{\sum_{k=1}^{K}(se_k^2)^{-1}} \tag{1}$$

The uncertainty around this estimate is calculated as follows:

$$\hat{se}_{FE} = \sqrt{\frac{1}{\sum_{k=1}^{K}(se_k^2)^{-1}}} \tag{2}$$

- FE is only optimal if the underlying treatment effects are homogenous
- It performs very poorly in the presence of "precision outliers" because it cannot distinguish between precision and generalizability.

Random Effects/Hierarchical Models for Meta-Analysis

- Hierarchical models aggregate the evidence while allowing heterogeneous effects across studies
- The key idea: specify an effect for each site k, denoted τ_k, but have each effect drawn from a common "parent" distribution governed by mean $\tau = E[\tau_k]\ \forall k$ and variance σ_τ^2.
- The Rubin (1981) model aggregates evidence from K Normally-distributed estimates from a Normal parent distribution:

$$\begin{aligned} \hat{\tau_k} &\sim N(\tau_k, \hat{se}_k^2)\ \forall k \\ \tau_k &\sim N(\tau, \sigma_\tau^2)\ \forall k \end{aligned} \tag{3}$$

- The τ is the generalizable information, common across studies.
- The σ_τ^2 measures the genuine heterogeneity in effects
- Priors on the (τ, σ_τ^2) then allows us to simulate the full joint posterior distribution for inference - for complex likelihoods this will be much more tractable than multivariate optimization.

Non Parametric (bi-variate) Regression

You have two random variable, X and Y and and express the conditional expectation of X given Y as : $E[Y|X] = g(X)$
Therefore, for any x, and y,

$$y = g(x) + \epsilon$$

where ϵ is the prediction error.
You may think that this relationship is causal or not. Problem is to estimate $g(x)$ without imposing a functional form.

The Kernel regression: A common non-parametric regression

g(x) is the conditional expectation of y given x.

$$E(Y|X = x) = \int yf(y|x)dy$$

By Bayes's rule:

$$\int yf(y/x)dy = \int \frac{yf(x,y)dy}{f(x)} = \frac{\int yf(x,y)dy}{f(x)}$$

Kernel Estimator

Kernel estimator replace $f(x, y)$ and $f(x)$ by their empirical estimates.

$$\hat{g}(x) = \frac{\int y \hat{f}(x, y) dy}{\hat{f}(x)}$$

- Denominator: estimating the density of x (we have seen this!)

$$\hat{f(x)} = \frac{1}{N * h} \sum_{i=1}^{n} K(\frac{x - x_i}{h}),$$

where h is a positive number (the bandwith) is the kernel estimate of the density of x. $K(.)$ is a density, i.e. a positive function that integrates to 1
It is a weighted proportion of observations that are within a distance h of the point x.

Kernel Estimator

Kernel estimator replace $f(x, y)$ and $f(x)$ by their empirical estimates.

$$\hat{g}(x) = \frac{\int y \hat{f}(x, y) dy}{\hat{f}(x)}$$

- Numerator

$$\frac{1}{N * h} \sum_{i=1}^{n} y_i K(\frac{x - x_i}{h})$$

Combine the two

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} y_i K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})} \tag{4}$$

$\hat{g}(x)$ is a weighted average of Y over a range of points close to x.
The weights are declining for points further away from x.
In practice, you choose a grid of points (ex. 50 points) and you calculate the formula given in equation 1 for each of these points.

Large sample properties

- as h goes to zero, bias goes to zero
- as nh goes to infinity, variance goes to zero.
- So as you increase the number of observation, you "promise" to decrease the bandwidth

Choices to make

- Choice of Kernel
  1. Histogram: $K(u) = 1/2$ if $\mid u \mid \leq 1$, $K(u) = 0$ otherwise.
  2. Epanechnikov $K(u) = \frac{3}{4}(1 - u^2)$ if $\mid u \mid \leq 1$ $K(u) = 0$ otherwise
  3. Quartic
     $K(u) = (\frac{3}{4}(1 - u^2))^2$ if $(u \leq 1)$, $K(u) = 0$ otherwise
- Choice of bandwidth : Trade off Bias, and Variance
  - A large bandwidth will lead to more bias (as we are missing important features of the conditional expectation).
  - A small bandwidth will lead to more variance (as we start to pick up lots of irrelevant ups and downs)

Cross Validation

One way to formalize this choice is cross validation.
First, define for each observation i define the prediction error as:

$$e_i = y_i - g\hat{(x_i)}$$

and the leave out prediction error as:

$$e_{i,-i} = y_i - g_{-i}\hat{(x_i)}$$

where $g_{-i}\hat{(x_i)}$ is the prediction of y based on kernel regression using all the observations except i.
An optimal bandwidth will minimize

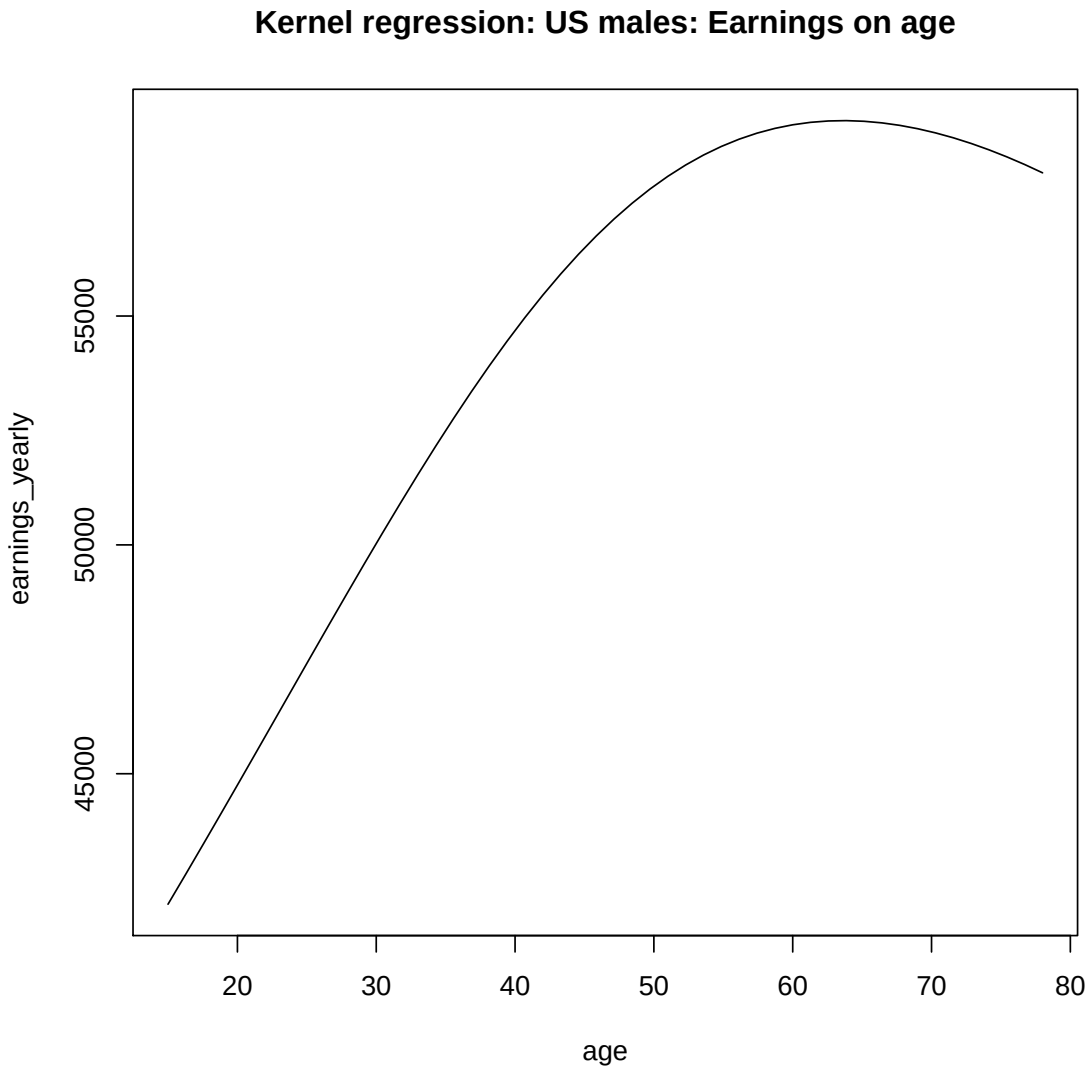
$$CV = \frac{1}{N}\sum_{i=1}^{N} e_{i,-i}^2$$

(or often in practice $CV = \frac{1}{N}\sum_{i=1}^{N} e_{i,-i}^2 M(X)$) where $M(X)$ is a trimming function to avoid influence of boundary points)

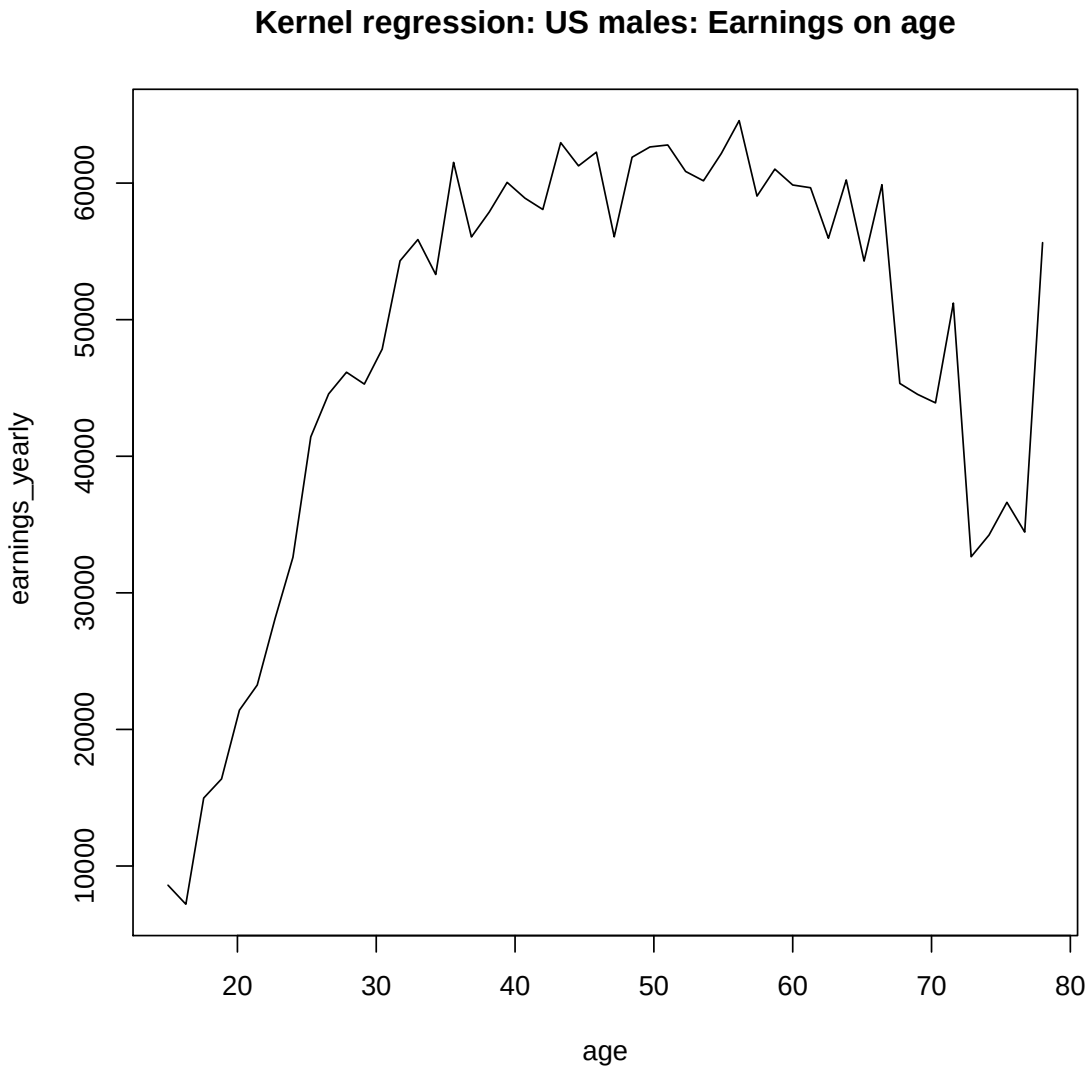
Kernel regression using npreg in R

```
############
## Example 2: CPS Data
############

## Load data
cps <- read.csv("cps_wage_data.csv")
cps_males <- subset(cps,sex=="MALE" & age<79)
attach(cps_males)

## Oversmoothing
cps_bw_over <- npregbw(xdat=age,ydat=earnings_yearly,bws=15,bandwidth.compute=FALSE)
cps_bw_over_model <- npreg(cps_bw_over)
pdf("US Earnings Oversmoothed.pdf")
plot(cps_bw_over,
    main="Kernel regression: US males: Earnings on age")
hide<-dev.off()

## Undersmoothing
pdf("US Earnings Undersmoothed.pdf")
cps_bw_under <- npregbw(xdat=age,ydat=earnings_yearly,bws=0.1,bandwidth.compute=FALSE)
cps_bw_under_model <- npreg(cps_bw_under)
plot(cps_bw_under_model,
    main="Kernel regression: US males: Earnings on age")
hide<-dev.off()

## Kernel regression with correct bandwidth
## (Calculated by npregbw)
pdf("US Earnings.pdf")
cps_bw <- npregbw(xdat=age,ydat=earnings_yearly)
cps_model <- npreg(cps_bw)
plot(cps_model,
    main="Kernel regression: US males: Earnings on age")
hide<-dev.off()

## Adding confidence bands
pdf("US Earnings Confidence Bands.pdf")
plot(cps_model,
    main="Kernel regression: US males: Earnings on age",
    plot.errors.method="asymptotic",
    plot.errors.style="band",
    plot.errors.quantiles = c(0.05,0.95))
hide<-dev.off()
```

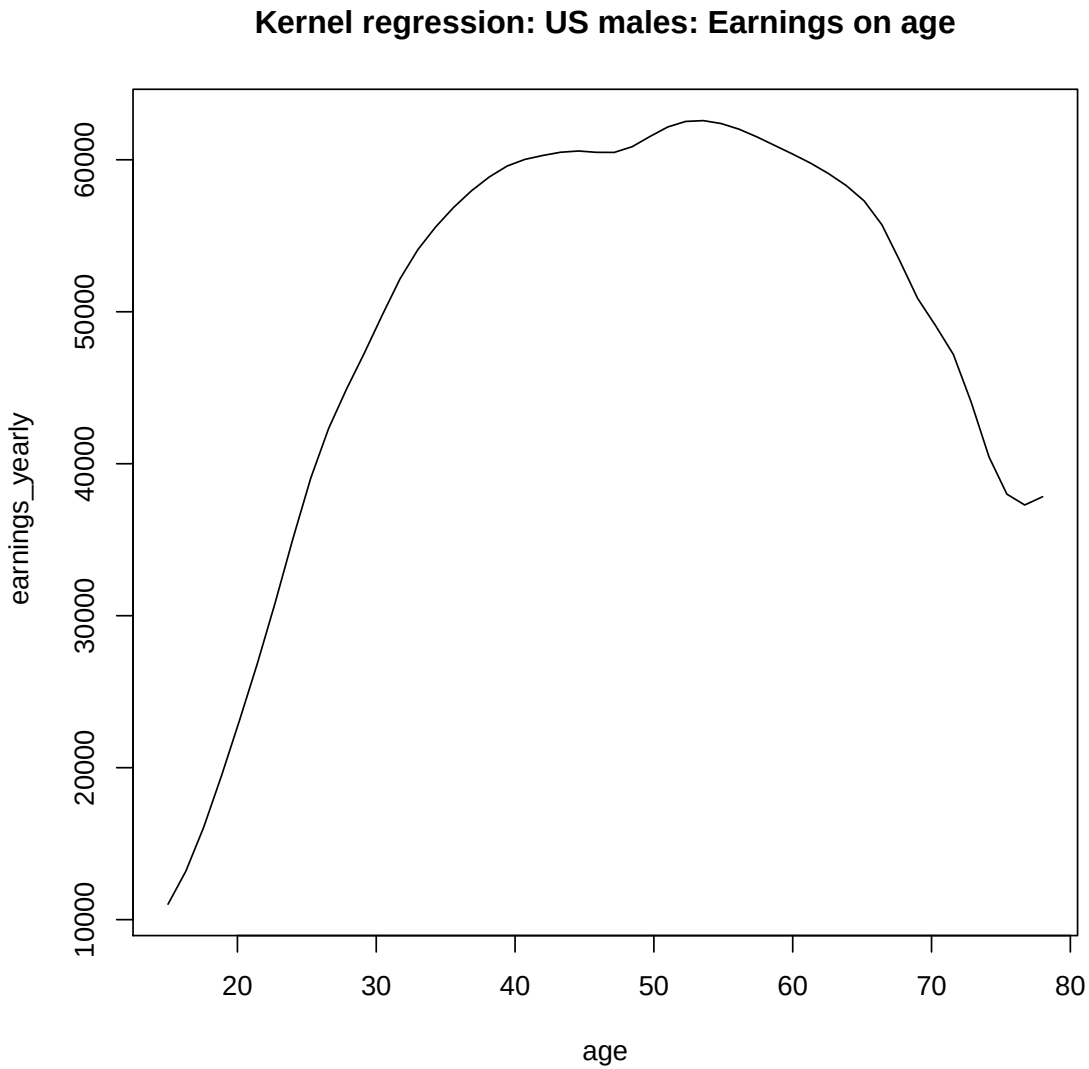
Kernel regression, choice of bandwidth too large



Kernel regression, choice of bandwidth too small



Kernel regression with appropriate bandwidth



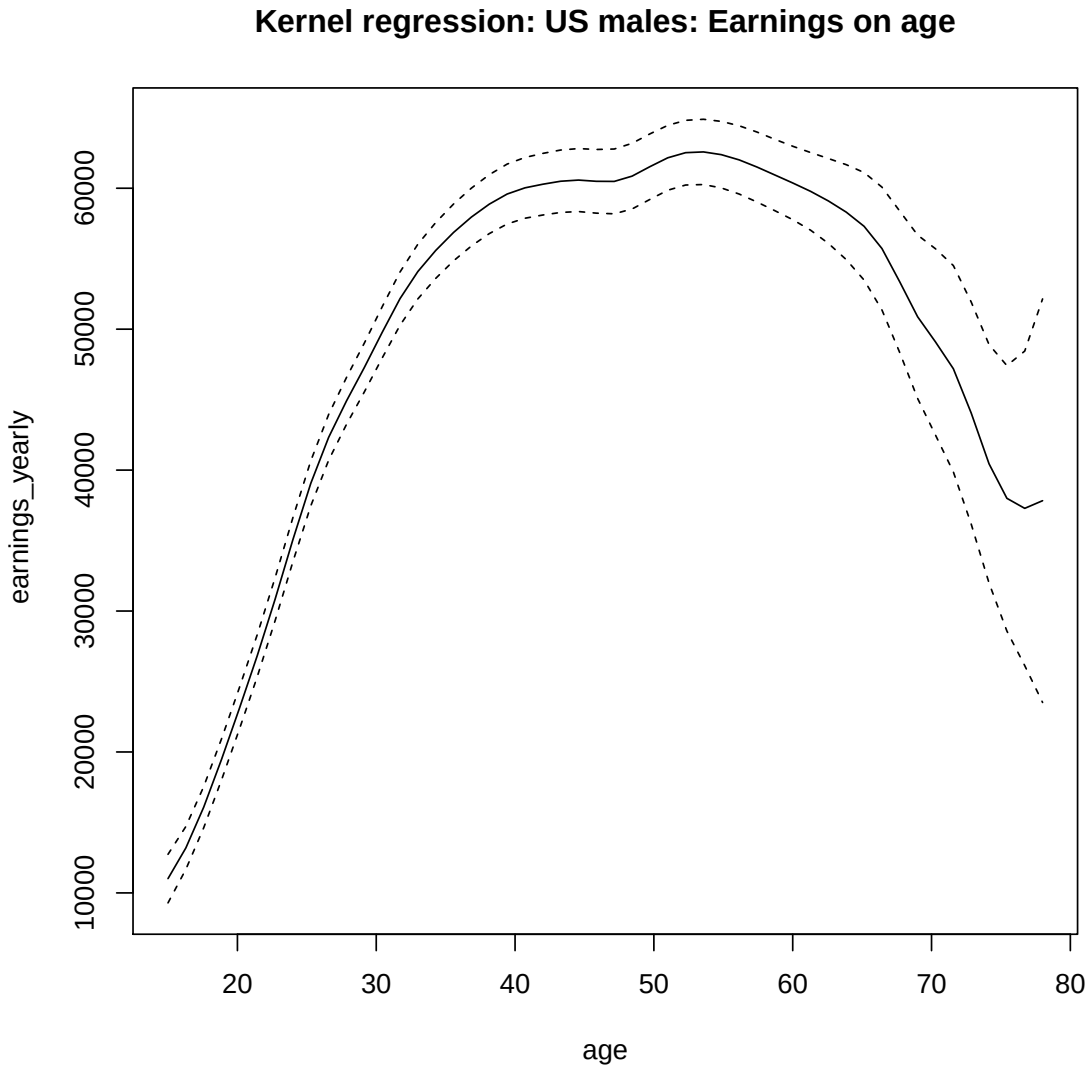
Confidence bands

$y_i = g(X_i) + e_i$ and $E[e_i|X_i] = 0$
$e_i^2 = \sigma_i^2(X_i) + \eta_i$ and $E[\eta_i|X_i] = 0$
So a Kernel estimate of $\sigma_i^2(X_i)$ is :

$$\hat{\sigma^2}(x) = \frac{\sum_{i=1}^{n} e_i^2 K(\frac{x-x_i}{h})}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})} \tag{5}$$

Point-wise confidence interval can be drawn using this estimate.

Kernel regression with confidence bands



Other non parametric methods

- Series estimation (approximate the curves by polynomes)
- Local linear regression(instead of taking the mean, in each interval, take predicted value from a regression,
- To see them we need to see regressions!