Handout 9 Nonparemetric Model

Chunyu Qu

12/13/2021

Wage problem

consider a model that relates the wage (wage) to education (educ), experience (exper), and age (age):

$$wage = g(educ, exper, age) + u$$

a. First estimate a linear model with no interactions using the lm command in R. Are all variables statistically significant in explaining wage?

```
## [1] "D:/DPrograms/Google Drive/Fordham/2019 Spring/AE/TA/PS8"
```

```
##
## Call:
## lm(formula = wage ~ educ + exper + age, data = wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -879.71 -250.74  -42.82  193.49 2211.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -559.621    152.095  -3.679 0.000247 ***
## educ          71.434      6.530  10.940  < 2e-16 ***
## exper         12.127      3.774   3.213 0.001358 **
## age           12.552      4.730   2.654 0.008091 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 375.1 on 931 degrees of freedom
## Multiple R-squared:  0.1423, Adjusted R-squared:  0.1396
## F-statistic:  51.5 on 3 and 931 DF,  p-value: < 2.2e-16
```

All the p-values are less than 0.05, which means they are all significant in explaining wage.

b. Now, using the npregbw command from the NP package compute the optimal bandwidths using least-squares-cross-validation. What are the optimal bandwidths and how do they compare to those that you would compute using the ROT?

```
# Invoke the np package
library(np)
```

```
## Warning: package 'np' was built under R version 3.5.3
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-9)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
```

```
# Compute the least-squares cross-validated bandwidths for the local, with constant estimator (default)

bw = npregbw(formula=wage ~ educ + exper + age, data = wage)
```

```
##
Multistart 1 of 3 |
Multistart 1 of 3 |
Multistart 1 of 3 |
Multistart 1 of 3 /
Multistart 1 of 3 -
Multistart 1 of 3 \
Multistart 1 of 3 |
Multistart 1 of 3 |
Multistart 2 of 3 |
Multistart 2 of 3 |
Multistart 2 of 3 /
Multistart 2 of 3 -
Multistart 2 of 3 \
Multistart 2 of 3 |
Multistart 2 of 3 |
Multistart 3 of 3 |
Multistart 3 of 3 |
Multistart 3 of 3 /
Multistart 3 of 3 -
Multistart 3 of 3 \
Multistart 3 of 3 |
Multistart 3 of 3 |
Multistart 3 of 3 |
```

```
summary(bw)
```

```
##
## Regression Data (935 observations, 3 variable(s)):
##
## Regression Type: Local-Constant
## Bandwidth Selection Method: Least Squares Cross-Validation
## Formula: wage ~ educ + exper + age
## Bandwidth Type: Fixed
## Objective Function Value: 138198.4 (achieved on multistart 1)
##
## Exp. Var. Name: educ  Bandwidth: 0.8432384 Scale Factor: 1.511221
## Exp. Var. Name: exper Bandwidth: 7.436399  Scale Factor: 4.516764
## Exp. Var. Name: age   Bandwidth: 0.6332775 Scale Factor: 0.5414299
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 3
## Estimation Time: 120.42 seconds
```

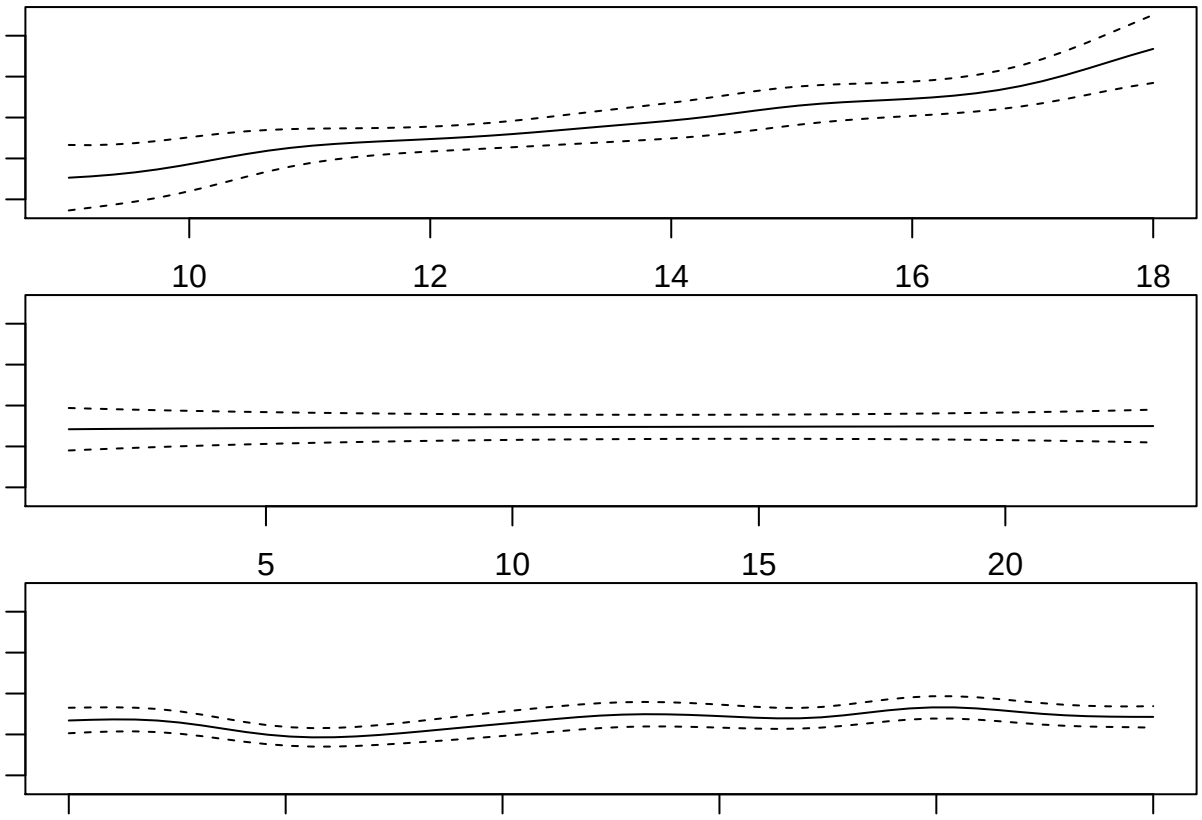
```
# Compute the ROT bandwidths for the local
rotbw=npcdensbw(formula=wage ~ educ + exper + age, data = wage, bwmethod="normal-reference")
summary(rotbw)
```

```
##
## Conditional density data (935 observations, 4 variable(s))
## (1 dependent variable(s), and 3 explanatory variable(s))
##
## Bandwidth Selection Method: Normal Reference
## Formula: wage ~ educ + exper + age
```

```
## Bandwidth Type: Fixed
##
## Exp. Var. Name: educ  Bandwidth: 0.6678214 Scale Factor: 1.059224
## Exp. Var. Name: exper Bandwidth: 1.970486  Scale Factor: 1.059224
## Exp. Var. Name: age   Bandwidth: 1.399877  Scale Factor: 1.059224
##
## Dep. Var. Name: wage  Bandwidth: 163.9504 Scale Factor: 1.059224
##
## Continuous Kernel Type: Second-Order Gaussian
## No. Continuous Explanatory Vars.: 3
## No. Continuous Dependent Vars.: 1
```

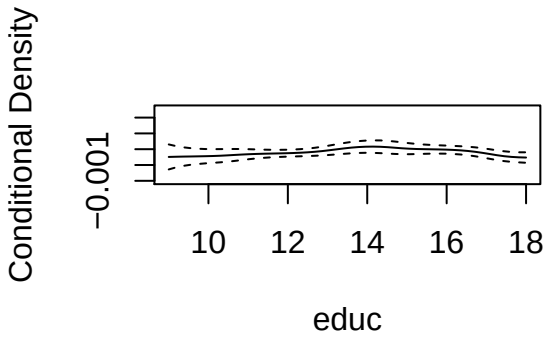
c. Now plot the relationship between wage, education, experience, and age using the command plot(bw,plot.errors.method="asymptotic"). Which variables seem to have a statistically significant relationship with wage? How do you know?

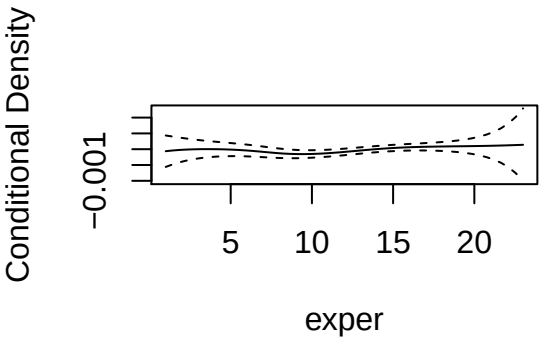
```
par(mar = rep(1, 4))
plot(bw,plot.errors.method="asymptotic")
```

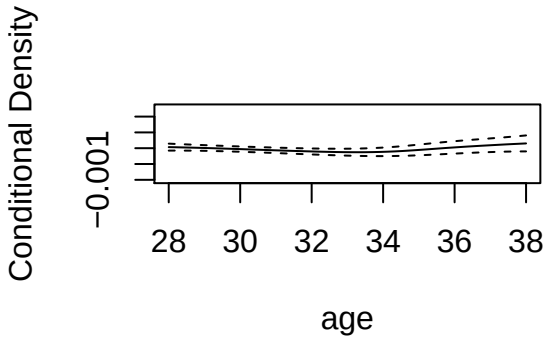


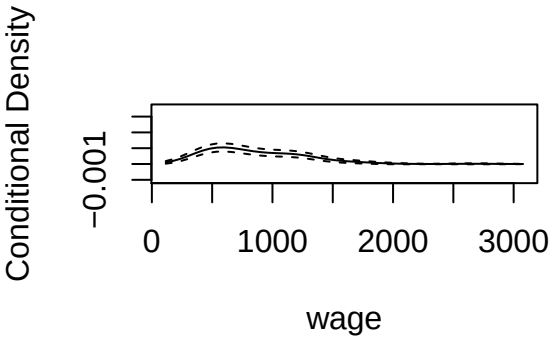
d. Restimate the nonparametric model imposing the ROT bandwidths and replot the nonparametric regression. Do your conclusions from part c. change?

```
plot(rotbw,plot.errors.method="asymptotic")
```


Conditional Density
−0.001
10
12
14
16
18
educ


Conditional Density
−0.001
5
10
15
20
exper


Conditional Density
−0.001
28
30
32
34
36
38
age


Conditional Density
−0.001
0
1000
2000
3000
wage