Using Cross Validation To Find An Appropriate Model–Take "Wage1" Dataset As An Example

Qihong Ruan

WISE

May 19th, 2018

I. "Wage1" Dataset

- Source: Wooldridge (2003, p. 226)
- Dependent variable is log(wage)(lwage).
- The explanatory variables:
  1. educ(years of education)
  2. exper (the number of years of potential experience)
  3. tenure (the number of years with their current employer)
  4. female('Female'/'Male')
  5. married('Married'/'Notmarried').
- n = 526 observations.

II. Motivation

- Hayfield and Racine(2008) analyze "wage1" dataset by using various nonparametric and semi-parametric methods in the "np" package.
- However, they use R^2 to compare the goodness-of-fit among the models, which doesn't make sense for choosing an appropriate model.
- I aim to choosing the model with best predictive ability.
  1. 5-fold cross validation, each time I get 1 MADE of the whole sample.
  2. Repeat 100 times.
  3. Boxplot MADE and choose the best model.
     Issue: This method is compute-intensive.(1 hour each time)
     Solution: I use HPC and parallel computation.
- **Finding:** Nonparametric kernel regression(Racine and Li, 2004; Li and Racine, 2004) has the best predictive ability for "wage1" dataset.

III. Models

1. OLS Model:

$$\begin{aligned}\ln wage_i &= \mu + \beta_1 female_i + \beta_2 married_i + \beta_3 educ_i \\ &\quad + \beta_4 exper_i + \beta_5 exper_i^2 + \beta_6 tenure_i + \beta_7 tenure_i^2 + \epsilon_i\end{aligned}$$

2. Kernel Regression (Racine and Li, 2004; Li and Racine, 2004):

$$\ln wage_i = g(female_i, married_i, educ_i, exper_i, tenure_i) + \epsilon_i$$

3. Partially Linear Model (Li and Racine, 2003):

$$\ln wage_i = \beta_1 female_i + \beta_2 married_i + \beta_3 educ_i + \beta_4 tenure_i + g(exper_i) + \epsilon_i$$

4. Semiparametric Single-index Model (Ichimura, 1993):

$$\ln wage_i = g(female_i + \beta_1 married_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 tenure_i) + \epsilon_i$$

5. Varing Coefficients Model (Li and Racine, 2010):

$$\begin{aligned}\ln wage_i &= \mu(female_i) + \beta_1(female_i) \cdot married_i + \beta_2(female_i) \cdot educ_i \\ &\quad + \beta_3(female_i) \cdot exper_i + \beta_4(female_i) \cdot tenure_i + \epsilon_i\end{aligned}$$

OLS Model

Table 1: Summary of OLS Model

```
Call:
lm(formula = lwage ~ female + married + educ + exper + expersq +
    tenure + tenursq, data = wage1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.81906 -0.24904 -0.02119  0.24525  1.12752

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.1805217  0.1065639   1.694   0.0909 .
femaleMale         0.2901837  0.0361121   8.036 6.33e-15 ***
marriedNotmarried -0.0529219  0.0407561  -1.299   0.1947
educ               0.0791547  0.0068003  11.640  < 2e-16 ***
exper              0.0269535  0.0053258   5.061 5.80e-07 ***
expersq           -0.0005399  0.0001122  -4.813 1.95e-06 ***
tenure             0.0312962  0.0068482   4.570 6.10e-06 ***
tenursq           -0.0005744  0.0002347  -2.448   0.0147 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3995 on 518 degrees of freedom
Multiple R-squared:  0.4426,    Adjusted R-squared:  0.4351
F-statistic: 58.76 on 7 and 518 DF,  p-value: < 2.2e-16
```

OLS Model

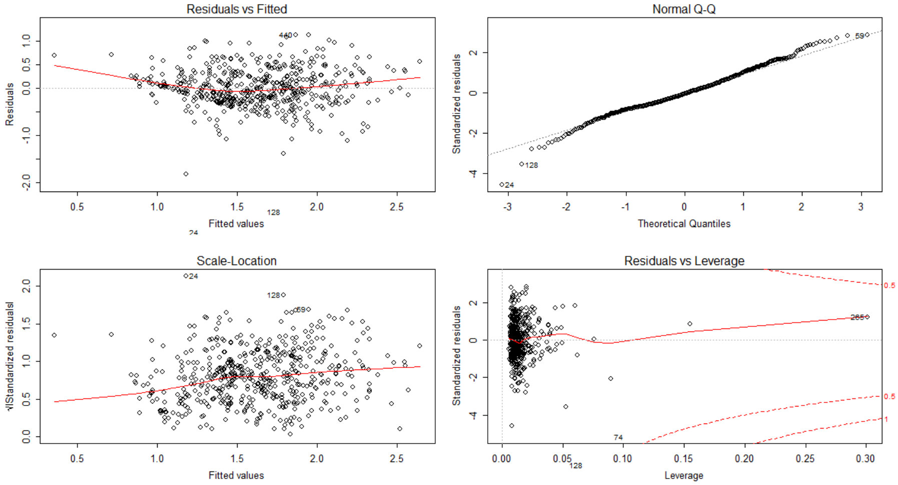

Figure1: Plots of Residuals

Kernel Regression

Kernel Regression Estimator: Local-Linear
Bandwidth Selection Method: CV.AIC
Continuous Kernel Type: Second-Order Gaussian
Unordered Categorical Kernel Type: Aitchison and Aitken

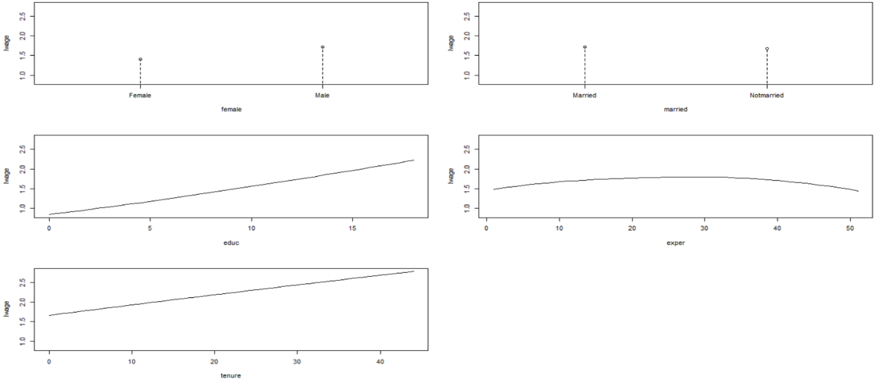

Figure2: Plots of Kernel Regression Model

Partially Linear Model

Note: More computationally burdensome than fully nonparametric models.

$\ln wage_i = \beta_1 female_i + \beta_2 married_i + \beta_3 educ_i + \beta_4 tenure_i + g(exper_i) + \epsilon_i$

Table2: Summary of Partially Linear Model

```
Partially Linear Model
Regression data: 526 training points, in 5 variable(s)
With 4 linear parametric regressor(s), 1 nonparametric regressor(s)

                y(z)
Bandwidth(s): 2.050976

                x(z)
Bandwidth(s): 4.194368
              1.353161
              3.160555
              5.238182

                   female     married      educ     tenure
Coefficient(s): 0.2861456 -0.03833231 0.0788131 0.01616543

Kernel Regression Estimator: Local-Constant
Bandwidth Type: Fixed

Residual standard error: 0.3929321
R-squared: 0.452499

Continuous Kernel Type: Second-Order Gaussian
No. Continuous Explanatory Vars.: 1
```

Semiparametric Single-index Model

$$\ln wage_i = g(female_i + \beta_1 married_i + \beta_2 educ_i + \beta_3 exper_i + \beta_4 tenure_i) + \epsilon_i$$

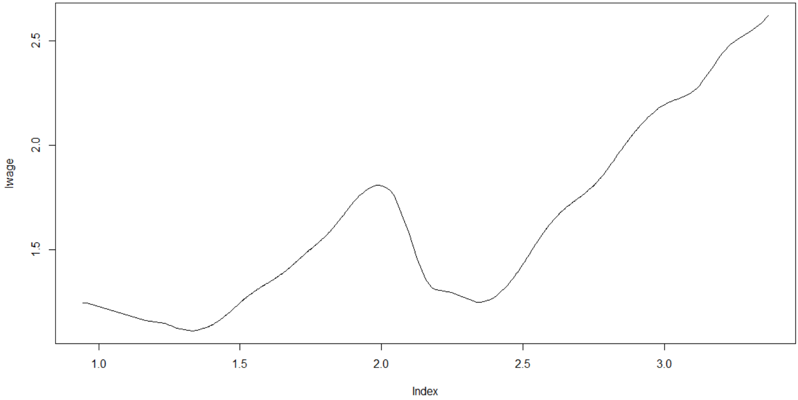

Figure3: Plot of Single-index Model

Varing Coefficients Model

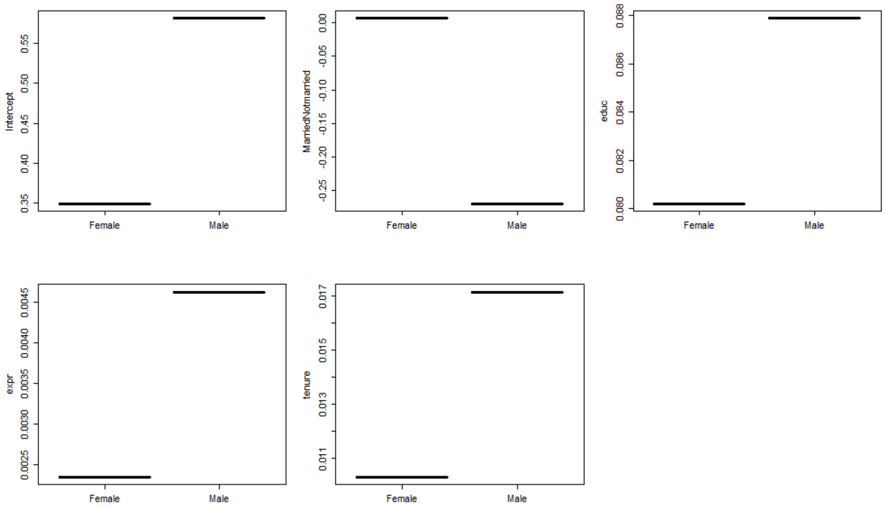

Figure4: Coefficents of the Explanatory Variables

IV. Results

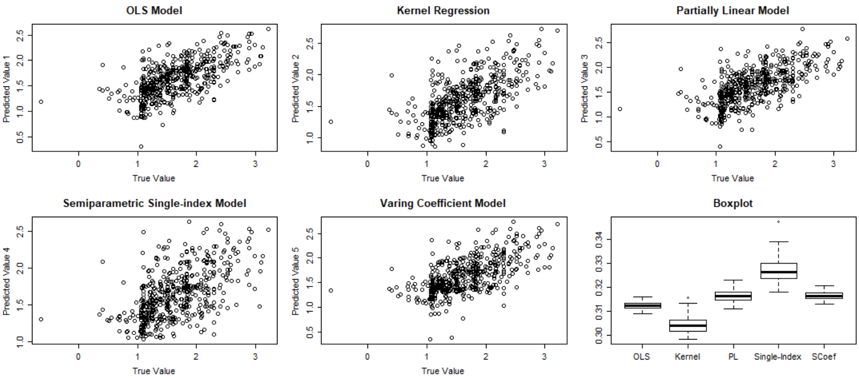

Figure5: Results of Prediction

- Nonparametric kernel regression has the best predictive ability.
- Future study: Try more combinations of variables and apply this method to analyze other datasets.