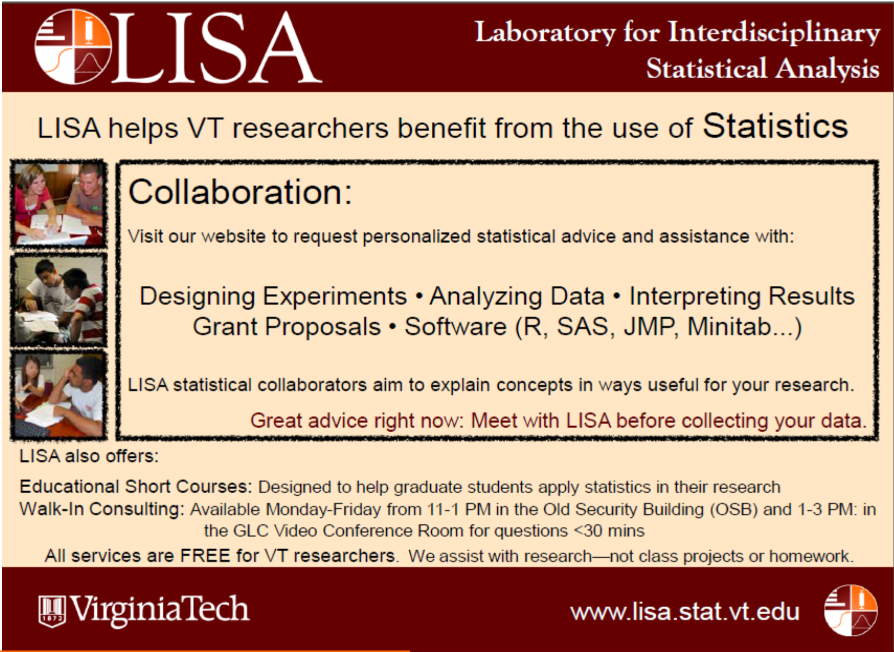
Parametric versus Semi/nonparametric Regression Models

Hamdy F. F. Mahmoud

Virginia Polytechnic Institute and State University
Department of Statistics

LISA short course series- July 23, 2014


LISA
Laboratory for Interdisciplinary Statistical Analysis
LISA helps VT researchers benefit from the use of Statistics
Collaboration:
Visit our website to request personalized statistical advice and assistance with:
Designing Experiments • Analyzing Data • Interpreting Results
Grant Proposals • Software (R, SAS, JMP, Minitab...)
LISA statistical collaborators aim to explain concepts in ways useful for your research.
Great advice right now: Meet with LISA before collecting your data.
LISA also offers:
Educational Short Courses: Designed to help graduate students apply statistics in their research
Walk-In Consulting: Available Monday-Friday from 11-1 PM in the Old Security Building (OSB) and 1-3 PM: in the GLC Video Conference Room for questions <30 mins
All services are FREE for VT researchers. We assist with research—not class projects or homework.
VirginiaTech
www.lisa.stat.vt.edu

Outline

1. What is semi/nonparametric regression?
2. When should we use semi/nonparametric regression?
3. semi/nonparametric regression estimation methods:
   a). Kernel Regression
   b). Smoothing Spline
4. Other methods in nonparametric regression models estimation.
5. Discussion and Recommendations

What is semi/nonparametric regression?

Nonparametric regression is a form of regression analysis in which NONE of the predictors take predetermined forms with the response but are constructed according to information derived from the data.

Semiparametric regression is a form of regression analysis in which a PART of the predictors do not take predetermined forms and the other part takes known forms with the response.

What is semi/nonparametric regression?

Example:

Assume that we have a response variable Y and two explanatory variables, x_1 and x_2. In general the regression model that describes the relationship can be written as:

$$Y = f_1(x_1) + f_2(x_2) + \epsilon$$

Some parametric regression models:

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ (Multiple linear regression model)
- $Y = \beta_0 + \beta_{10} x_1 + \beta_{11} x_1^2 + \beta_{20} x_2 + \beta_{21} x_2^2 + \epsilon$ (Polynomial regression model of second order)
- $Y = \beta_0 + \beta_1 x_1 + \beta_2 e^{(\beta_3 x_2)} + \epsilon$ (Nonlinear regression model)
- $\log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (Poisson regression when Y is count)

What is semi/nonparametric regression?

- If we do not know f_1 and f_2 functions, we need to use a NONparametric regression model.
- If we do not know f_1 and know f_2, we need to use SEMIparametric regression model.

  Example:

$$Y = \beta_0 + \beta_1 x_1 + f(x_2) + \epsilon$$

When we should use semi/nonparametric regression?

The FIRST STEP in any analysis is GRAPHICAL ANALYSIS for the response (dependent) variable and the explanatory (independent) variables.

Examples:

- Boxplots
- Area plots
- Scatterplots

GO TO REAL DATA [Course R Code]

When should we use semi/nonparametric regression?

Four principal assumptions which justify the use of linear regression models for purposes of fitting and inferences:

- Linearity
- Independence
- Constant variance
- Normality

When should we use semi/nonparametric regression?

- **Violations of linearity** are extremely serious, especially when you extrapolate beyond the range of the sample data.
- **How to detect:**
  - Nonlinearity is usually most evident in a plot of residuals versus predicted values.
  - Use Goodness of fit test.
- **How to fix:**
  - Use a nonlinear transformation to the dependent and/or independent variables such as a log transformation, square root, or power transformation.
  - Add another regressors which is a nonlinear function of one of the other variables. For example, if you have regressed Y on X, it may make sense to regress Y on both X and X^2 (i.e., X-squared).
  - Use semi(non)parametric regression model.

When should we use semi/nonparametric regression?

GO to R Code file to:

- Practice on identifying the relationship (linear or not linear) using different data sets.
- For wage data set, regress log(wage) on age using linear regression model.
- Check the linearity assumption.
- Try to fix the problem.

Estimation methods

Two of the most commonly used approaches to nonparametric regression are:

1. **Kernel Regression:** estimates the conditional expectation of Y at given value x using a weighted filter to the data.
2. **Smoothing splines:** minimize the sum of squared residuals plus a term which penalizes the roughness of the fit.

Semi/nonparametric regression estimation methods.

1. Nadaraya-Watson Kernel Regression [local constant]

Nadaraya and Watson 1964 proposed a method to estimate $\hat{f}(x_0)$ at a given value x_0 as a locally weighted average of all $y's$ associated to the values around x. The Nadaraya-Watson estimator is:

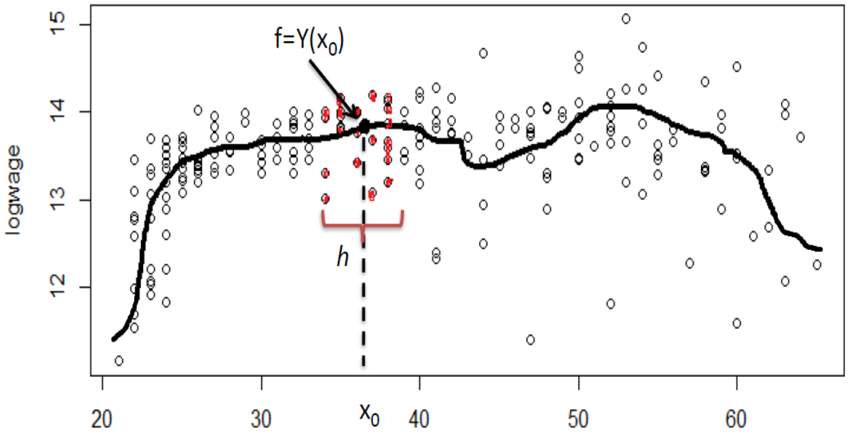
$$\widehat{f}_h(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}$$

where K is a Kernel function (weight function) with a bandwidth h.

Remark: K function should give us weights decline as one moves away from the target value.

Semi/nonparametric regression estimation methods.

Kernel Regression (local constant)



Popular choices of weight function are:

- Epanechnikov: $K(\cdot) = \frac{3}{4}(1 - d^2)$, $d^2 < 1$, 0 otherwise,
- Minimum var: $K(\cdot) = \frac{3}{8}(1 - 5d^2)$, $d^2 < 1$, 0 otherwise,
- Gaussian density: $exp(-\frac{x - x_i}{h})$
- Tricube function: $W(z) = (1 - |z|^3)^3$ for $|z| < 1$ and 0 otherwise.

Problem: local constant has one difficulty is that a kernel smoother still exhibits bias at the end points.
Solution: Use local linear kernel regression

Semi/nonparametric regression estimation methods.

How to choose the bandwidth?

- Rule of thumb: If we use Gaussian then it can be shown that the optimal choice for h is

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma} n^{-1/5},$$

where $\hat{\sigma}$ is the standard deviation of the samples.

Semi/nonparametric regression estimation methods.

GO to R Code file to:

1. Apply Kernel regression on wage data
2. Study the effect of bandwidth h on estimation
3. Compare between Kernel regression (nonparametric) and second order polynomial regression (parametric) in terms of fitting and prediction.

Semi/nonparametric regression estimation methods.

2. Spline Smoothing

- A spline is a piecewise polynomial with pieces defined by a sequence of knots

$$\theta_1 < \theta_2 < ..... < \theta_K$$

such that the pieces join smoothly at the knots.

- A spline of degree p can be represented as a power series:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ... + \beta_p x^p + \sum_{k=1}^{K} \beta_{1k}(x - \theta_k)_+^p,$$

where $(x - \theta_k)_+ = x - \theta_k, x > \theta_k$ and 0 otherwise

Example: $f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} \beta_{1k}(x - \theta_k)_+$ (linear spline)

Semi/nonparametric regression estimation methods.

- How many knots need to be used?
- Where those knots should be located?
- Number of parameters is $1+p+K$ that we need big number of observations.

Possible solution: use penalized spline smoothing

Consider fitting a spline with knots of every data point, so it could fit perfectly, but estimate its parameters by minimizing the usual sum of squares plus a roughness penalty. A suitable penalty is to integrate the squared second derivative, leading to penalized sum of squares criterion:

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx$$

where λ is a tunning parameter controls smoothness.

Semi/nonparametric regression estimation methods.

GO to R Code file to:

1. Apply spline regression on prestige data, and
2. Study the effect of λ on smoothing

Semi/nonparametric regression estimation methods.

What if we have more than one explanatory (independent) variable?

1. Kernel Regression
2. Spline Regression

GO to R Code file

Discussion and Recommendations

Pros:

- It is flexible.
- Better in fitting the data than parametric regression models.

Cons:

- Nonparametric regression requires larger sample sizes than regression based on parametric models because the data must supply the model structure as well as the model estimates. limitations

Discussion and Recommendations

Steps of modeling:

- Graphical Analysis
- If you have a nonlinear and unknown relationship between a response and an explanatory variable:
  - use transformation
  - add a new variable in the model to capture the relationship.
- If transformation does not work, use nonparametric regression.

References

1. D. Ruppert, M. P. Wand, and R. J. Carrol (2003), "Semiparametric Regression", Cambridge University Press, NY.
2. J. S. Simonoff (1996) "Smoothing methods in statistics", New York : Springer.
3. Y. Wang (2011) "Smoothing splines : methods and applications" Boca Raton, FL: CRC Press.
4. M.P. Wand and M.C. Jones (1995) "Kernel smoothing", London; New York: Chapman and Hall .