Spring 1998 John Rust
Economics 551b 37 Hillhouse, Rm. 27

PROBLEM SET 0
Applied Parametric and Nonparametric Regression

QUESTION 1 Extract data in file `data1.asc` in the

`pub/John_Rust/courses/econ551/regression/`

directory on `gemini.econ.yale.edu` (either ftp to `gemini.econ.yale.edu` and login as "anonymous" and `cd pub/John_Rust/courses/econ551/regression` and `get data1.asc` or click on the hyperlink in the html version of this document). This data file contains $n = 1500$ *IID* observations (y_i, x_i) that I generated on the computer from the nonlinear regression

$$y = \exp\{\beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3\} + \epsilon$$

where ϵ is normally distributed with mean zero, and the independent variable is a scalar random variable x which is also normally distributed. The sorted (y, x) observations are graphed in the file `data1.eps`, also available by clicking on the hyperlink: to

`http://gemini.econ.yale.edu/jrust/econ551/exams/98/ps0/data_ex1.eps`

1. Using the data in `data1.asc` Comput the maximum likelihood estimates of the parameter vector $\theta = (\beta, \sigma^2, \mu, \delta^2)$, where β is the (4×1) vector of regression coefficients, σ^2 is the variance of ϵ, and μ is the mean of the x distribution and δ^2 is its variance. Show theoretically that the asymptotic covariance between the (μ, δ^2) parameters and the (β, σ^2) parameters is zero. Is zero also the sample estimate of this covariance from your estimation algorithm?

2. Compute White misspecification-consistent estimates of the standard errors for your parameters and compare them to the standard estimates from an estimate of the inverse of the estimated information matrix. Are there big discrepancies between these two different estimates that would lead you to be concerned about possible misspecification of your model?

3. What happens if you try to estimate (β, σ) by simple OLS with the log-linear specification:

$$\log(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \ ?$$

   Compare the OLS (or MLE) estimates of this log-linear model to those you obtained in step 1. Can you come up with a theoretical argument that the probability limits for (β, σ) are the same for the two different estimation methods? If so, write down a proof, otherwise provide an argument of why the probability limits are different.

4. Examine the estimated residuals from the nonlinear regression model in part 1 for evidence of heteroscedasticity. What kinds of statistics could you think of to provide evidence of the possibility of heteroscedasticity? Can you think of a simple way to test the hypothesis of no heteroscedasticity, i.e. homoscedasticity?

5. One way to test for homoscedasticity is to nest the model in part 1 in a larger model that allows for heteroscedasticity and then test the null of homoscedasticity via a likelihood ratio or Wald test (topics we will cover later in Econ 551). Restimate the model in part 1 but now allow for heteroscedasticity of the following form:

$$\sigma^2(x) = \exp\{\gamma_0 + \gamma_1 x + \gamma_2 x^2\}$$

Now what are your maximum likelihood estimates of $\theta = (\beta, \gamma, \mu, \delta^2)$? Is the asymptotic covariance between the β and γ parameters zero? Why or why not? Can you reject the hypothesis of homoscedasticity via likelihood ratio or Wald test at the 5% significance level?

6. Now load the data in `data2.asc`. These are data from an unspecified linear or nonlinear regression function

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

where $f(x) = E\{y|x\}$ is the unknown regression function to be estimated. Estimate f using your favorite non-parametric regression method (i.e. kernel regression, series estimator, nearest neighbor, neural networks, splines, etc.), and plot the estimate over the $[0, 1]$ interval (the range of the x's). Describe clearly the method you used and your choices for auxiliary smoothing parameters. Discuss the sensitivity of your estimates to these smoothing parameters or other aspects of your estimation procedure. Compare your non-parametric estimate of f to the results you might get from a simple OLS with coefficients on a series expansion to f, i.e.

$$f(x) = \theta_0 + \theta_1 x + \theta_2 x^2 \cdots$$

and to a nonlinear regression of the form

$$f(x|\theta) = \exp\{\theta_0 + \theta_1 x + \theta_2 x^2 \cdots\}$$

using the software you developed to answer parts 1 to 5 above.

7. Now compute the estimated residuals from your nonparametric regression:

$$\hat{e}_i = y_i - \hat{f}(x_i), \quad i = 1, \ldots, N$$

Describe how you might use these residuals as data for a nonparametric regression to uncover the form of unknown heteroscedasticity, i.e. to estimate the unknown function

$$\sigma^2(x) = \mathrm{var}(\epsilon|x)$$

8. Repeat steps 6 and 7 with the `data1.asc` data. Do a plot of the parametric and non-parametric estimates of $f(x)$ and $\sigma^2(x)$ over the interval $[-3, 3]$. How close are the two sets of estimates over this interval? Which ones do you trust more? Do you have any reason to believe that I mislead you in suggesting the particular functional forms for $f(x)$ in Part 1 and $\sigma^2(x)$ in part 5?