

Homework # 2

Due in Class Thursday, October 10

1. Consider the S-plus air data set you used in HW1 (If you type *air* in Splus, you will be able to see the data). This air data set contains the measures of ozone (OZ), solar radiation (RAD), temperature (TEMP) and wind speed (WIND) for 111 consecutive days in a city of the state of New York. Consider the nonparametric model

$$Y_i = \theta(X_i) + \epsilon_i,$$

where $Y = OZ$, $X = WIND$, $\epsilon_i \sim N(0, \sigma^2)$.

(1) Fit a linear model and a quadratic model of OZ on WIND and compare the parametric fits with a nonparametric fit using the kernel method. Comment on your results.

(2) Use the S-plus functions *ksmooth* to estimate $\theta(t)$. Try a few bandwidths and a few kernel functions and examine how the kernel estimator of $\theta(\cdot)$ is affected by the bandwidth h and the kernel function $K(u)$.

(3) Use the S-plus functions *ksmooth*, *loess.smooth* (*loess*) and *supsmu* to estimate $\theta(\cdot)$ and compare their fits. Use the *loess* function to calculate the 95% confidence interval of $\hat{\theta}(x)$.

(4) Write a function using your favorite programming language, e.g., S-plus, SAS IML, Fortran, to construct a local *linear* kernel estimate of $\theta(x)$ using the Epanechnikov kernel with bandwidth $h = 5$. Propose a method to estimate the SE of $\hat{\theta}(x; h)$ and implement it. For simplicity, you can estimate σ^2 using $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \{Y_i - \hat{\theta}(X_i)\}^2$. Plot the estimated curve and its 95% point-wise CI. Plot the estimated derivative estimate of $\theta'(x)$.

2. Consider the respiratory infection data introduced at the beginning of this semester. The data set “indon_base.dat” on the class website contains 230 Indonesian children of their respiratory infection status and covariates. The variables are ID, XERO (vitamin A deficiency status (Y/N)), AGE (centered by 36 months), SEX, HEIGHT and INFECT (Y/N).

(1) Fit linear and quadratic logistic models of respiratory infection (INFECT) on AGE and compare the parametric fits with the nonparametric *loess* logistic fit using GAM.

(2) Consider the nonparametric logistic model $\text{logit}\{E(Y_i = 1)\} = \theta(X_i)$, where $Y = \text{INFECT}$ and $X = \text{AGE}$. Use your favorite programming language to calculate the local *linear* kernel estimate of $\theta(x)$ using the Epanechnikov kernel and its 95% CI of $\theta(x)$. Plot both the curve and its 95% CI. You could try a few bandwidths and pick a good one.

3. Suppose Y_i follows the exponential family or the quasi-likelihood discussed in class with mean μ_i and variance $\phi v(\mu_i)$. Consider the generalized nonparametric model

$$g(\mu_i) = \theta(X_i),$$

where $g(\cdot)$ is a monotone differentiable link function and X_i is a continuous covariate with density $f(x)$. Consider the local *linear* kernel estimator of $\theta(x)$.

(1) Prove the asymptotic bias of $\hat{\theta}(x; h)$ is $h^2 \theta''(x)/2$.

(2) Prove the asymptotic variance of $\hat{\theta}(x; h)$ is

$$\text{var}\{\hat{\theta}(x)\} = \frac{\gamma}{nh} \phi v(\mu) \{g'(\mu)\}^2 \frac{1}{f(x)},$$

where $\gamma = \int K^2(u) du$ and $g(\mu) = \theta(x)$.

Hint:

You could do the asymptotic bias and variance calculations using the Taylor expansion given on page 12 of the GNM lecture notes.