Instructor: L.-P. Chen

SS4850G/9850B Assignment #1
(Due Feb 14 2020 before 13:30)

- **Show all your works in details and provide your code for computations.**
- **In Questions 4 to 6, use the "Biweight" kernel.**
- **Write down your name, student id, and you are undergraduate/graduate students.**
- **Please submit your hard copy in class.**

1. In Section 2.1, we discussed properties of LSE $\widehat{\beta}$ and residual $e = (I_n - X(X^\top X)^{-1}X^\top)y$. Based on the notation defined in Section 2.1, please show that

   (a) the estimator $\widehat{\beta}$ is an unbiased estimator of β with $\text{var}(\widehat{\beta}|X) = \sigma_\epsilon^2 (X^\top X)^{-1}$. Moreover, $\widehat{\beta}$ is the Best Linear Unbiased Estimator (BLUE).

   (b) $E(e|X) = 0$, $\text{var}(e|X) = \sigma_\epsilon^2(I_n - H)$ with $H = X(X^\top X)^{-1}X^\top$, and $\text{cov}(e, \widehat{\beta}|X) = 0$.

   (c) Furthermore, find the unbiased estimator of σ_ϵ^2.

2. Suppose that $\{(Y_i, X_i) : i = 1, \cdots, n\}$ is a sequence of independently and identically distributed (i.i.d.) random variables with $Y_i, X_i \in \mathbb{R}$. Assume that $\text{var}(X_i) = \sigma_X^2$. We consider the simple linear regression model

$$Y_i = \beta_0 + X_i\beta_x + \epsilon_i, \tag{1}$$

   where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2)$ and ϵ_i is independent of X_i.

   (a) In some applications (e.g., when collecting data), we are not able to precisely measure X_i, but instead, we can only observe X_i^*. It is called *mismeasurement*. As a result, we usually have model (1) with X_i replaced by X_i^* in this situation. Please find the estimators of β_x based on X_i and X_i^*, and denote them by $\widehat{\beta}_x$ and $\widetilde{\beta}_x$, respectively.

   (b) We usually build up the relationship between X_i and X_i^* by the following model:

$$X_i^* = X_i + \delta_i, \tag{2}$$

   where $\delta_i \overset{i.i.d.}{\sim} N(0, \sigma_\delta^2)$ and δ_i is independent of X_i and ϵ_i. Based on (2), show that $\widehat{\beta}_x \xrightarrow{p} \omega_1\beta_x$ and $\widetilde{\beta}_x \xrightarrow{p} \omega_2\beta_x$ for some non-negative values ω_1 and ω_2 as $n \to \infty$, where "$\xrightarrow{p}$" represents *convergence in probability*. Also, you should specify the exact values of ω_1 and ω_2.

   (c) Find variances of $\widehat{\beta}_x$ and $\widetilde{\beta}_x$, i.e., $\text{var}(\widehat{\beta}_x)$ and $\text{var}(\widetilde{\beta}_x)$. Moreover, compare these two variances.

(d) (Do a simple simulation study). Consider the sample size $n = 1000$. Let $\beta_0 = \beta_x = 1$, $\sigma_\epsilon^2 = 1$. Let X_i be generated by $N(4, 1)$. Then the response Y_i can be generated by model (1). In addition, consider $\sigma_\delta^2 = 0.15, 0.55$ and 0.75, and then generate X_i^* by (2). Suppose that we run 1000 repetitions. Based on your "artificial" data, calculate numerical results for $\widehat{\beta}_x$, $\widetilde{\beta}_x$, $\mathrm{var}(\widehat{\beta}_x)$ and $\mathrm{var}(\widetilde{\beta}_x)$. Summarize your numerical results as the following table and compare with (a), (b), and (c).

Table 1: Simulation result

	$\widetilde{\beta}_x$			$\widehat{\beta}_x$
	$\sigma_\delta^2 = 0.15$	$\sigma_\delta^2 = 0.55$	$\sigma_\delta^2 = 0.75$	
Bias				
var				

Note: Bias is $\widetilde{\beta}_x - \beta_x$ and $\widehat{\beta}_x - \beta_x$.

(e) Summarize your findings in (a) - (d).

3. Suppose $f(y)$ is a probability density function (pdf). Let

$$R(f^{(r)}) = \int_{-\infty}^{\infty} \left\{ f^{(r)}(y) \right\}^2 dy,$$

where $f^{(r)}(y)$ is the rth derivative of $f(y)$.

(a) When f is pdf of $N(\mu, \sigma^2)$, please find $R(f^{(2)})$ so that we are able to obtain the bandwidth based on normal scale rule.

(b) Show that under some conditions,

$$R(f^{(r)}) = (-1)^r \int_{-\infty}^{\infty} f^{(2r)}(y) f(y) dy.$$

Which kinds of conditions do we need here? Does standard normal distribution satisfy these conditions?

4. Consider the wool prices data set (*wool.txt*) that reports the wool prices at weekly markets. The response of interest is the log price difference between the price of a particular wool 19 μm (cents per kilogram clean) and the floor wool price (cents per kilogram clean) at markets:

$$y_t = \log(19\ \mu m \text{ price/floor price}),$$

and the covariate x_t is the time in weeks since January 1, 1976.

(a) Fit the data by a simple linear regression model and a polynomial model of order 10. Give scatterplot of the data and add the two fitted lines, one for simple linear model and one for polynomial model. Put clear and proper legends on it.

(b) Fit the data by local constant kernel estimator and local linear kernel estimator. Choose the bandwidths in these two estimators by the CV method. Give scatterplot of the data and add the two fitted lines. Put clear and proper legends on it.

(c) Fit the data by local linear kernel estimator. Choose the bandwidths by the CV and direct plug-in methods. Give scatterplot of the data and add the two fitted lines. Put clear and proper legends on it.

(d) Let $\widehat{y}_t$ denote the fitted values determined by methods in (a) to (c). Compute $\sum_{t=1}^{n}(\widehat{y}_t - y_t)^2$. Finally, summarize your findings.

5. Consider the data from undergoing corrective spinal surgery. The objective was to determine important risk factors for kyphosis. The response (y) is the presence or absence of the kyphosis. The risk factor (x) is the age in weeks. The data is in *kyphosis.txt*.

We consider the following model:

$$y_i \sim Bernoulli\Big(\pi(x_i)\Big) \text{ with } \log\left\{\frac{\pi(x_i)}{1-\pi(x_i)}\right\} = f(x_i).$$

(a) Suppose that we use the local constant and local linear kernel methods to estimate $\pi(x)$. Choose the bandwidths h by the CV method, and use h to calculate the estimators $\widehat{\pi}(x_i)$;

(b) Assume that $f(x)$ is a linear function of x. We estimate $\pi(x)$ under this assumption. Plot this estimator and local constant / linear kernel estimators in the same graph. Put clear and proper legends on it. Does the linear assumption for $f(x)$ look reasonable here?

6. This problem refers to data from a study of nesting horseshoe crabs. Each female horseshoe crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. The response outcome y for each female crab is her number of satellites. The covariate (x) is the female crab's carapace width. The data is in *crab.txt*.

We consider the following model:

$$y_i \sim Poisson\Big(\mu(x_i)\Big) \text{ with } \log\{\mu(x_i)\} = f(x_i).$$

You are asked to answer the following questions:

(a) By the similar idea in Section 3.4, please write down the formulations of $\ell_i(\beta_0, \beta_1)$ and the performance measure. Also, determine the CV score.

(b) Choose the bandwidth h by the CV score based on local constant and local linear kernel estimators, and then calculate $\widehat{\mu}(x_i)$ by local constant and local linear kernel estimators.

(c) Assume that $f(x)$ is a linear function of x, estimate $\mu(x)$ under this assumption and plot it and local constant and linear kernel estimators in the same graph. Put clear and proper legends on it. Does the linear assumption look reasonable here?