BIOST/STAT 527
SPRING QUARTER 2024

Homework # 2
Due Via Online Submission to Canvas: Tue, Apr 30 at 11:59 PM

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, code without written answers will receive no credit. To receive credit, you must explain your answers and show your work. All plots should be appropriately labeled and legible, with axis labels, legends, etc., as needed.

Problem 1: Theory

In class, we introduced the notion of the expected risk of a function $g : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ represents the predictor space and $\mathcal{Y}$ the outcome space. Given a loss function $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, the expected risk of a function g is defined as

$$\mathbb{E}_{Y,X}[L(Y, g(X))]. \tag{1}$$

An optimal predictor g^* (sometimes referred to as a Bayes predictor) is defined as

$$g^* = \arg\min_g \mathbb{E}_{Y,X}[L(Y, g(X))]. \tag{2}$$

This represents an 'optimal prediction function' that could be achieved if we had access to an infinite amount of data, or equivalently, to the true distribution of the data. Note that this may not be unique.

Noting that

$$g^*(x) = \arg\min_g \mathbb{E}_{Y|X}[L(Y, g(X))|X = x], \qquad \forall x \in \mathcal{X}, \tag{3}$$

prove the following statements:

1. If $\mathcal{Y} = \mathbb{R}$ and $L(Y, \hat{Y}) = (Y - \hat{Y})^2$, then $g^*(x) = \mathbb{E}[Y|X = x]$.

   Hint: Prove that $\mathbb{E}[(Y - c)^2] = \mathrm{Var}(Y) + (\mathbb{E}[Y] - c)^2$, for any random variable Y and constant $c \in \mathbb{R}$.

2. If $\mathcal{Y} = \mathbb{R}$ and $L(Y, \hat{Y}) = |Y - \hat{Y}|$, then $g^*(x) = \mathrm{median}[Y|X = x]$.

3. If $\mathcal{Y} = \{0, 1\}$, $L(Y, \hat{Y}) = \mathbb{I}(Y \neq \hat{Y})$, where $\mathbb{I}(\cdot)$ is the indicator function, then $g^*(x) = \arg\max_{y' \in \{0,1\}} \mathbb{P}[Y = y'|X = x]$.

Problem 2: Methodology & Case Study

In this problem, we analyze the dataset `fev.txt` available on Canvas, which includes measures of lung function from 654 children. Lung function is assessed using forced expiratory volume (FEV), indicating the volume of air expelled from the lungs in a short period. Higher FEV values suggest better respiratory function. The dataset also includes data on each child's age, height, sex, and smoking status. For further details about this dataset, see `http://www.emersonstatistics.com/Datasets/index.asp`.

Our goal is to determine whether smoking tends to impair lung function in children.

We focus our analysis on the following variables: smoking status (denoted by $X \in \{0, 1\}$), FEV (denoted by $Y \in \mathbb{R}$), and height and age (denoted by $Z \in \mathbb{R}^2$).

1. Assume the following linear model for the data:

$$Y = \beta_1 X + \epsilon, \tag{4}$$

   where $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\epsilon^2] \leq \sigma^2 < \infty$. Estimate β_1 and report whether there is evidence of an association between the variables Y and X.

2. Discuss whether these results allow us to conclude that smoking impairs lung function in children.

3. Next, consider the following model for the data:

$$Y = \beta_1 X + g^*(Z) + \epsilon,$$

   where $g^* : \mathbb{R}^2 \to \mathbb{R}$ is an unknown and possibly non-linear function modeling the relationship between Z (the confounders) and Y (the outcome). The variable X is instead modeled parametrically. The noise term ϵ is such that $\mathbb{E}[\epsilon] = 0$ and $\epsilon \perp X, Z$. This model, known as the **partially linear model**, allows us to study the association between Y and X while controlling for Z. We aim to estimate β_1 and g^* using the tools introduced in class.

   Show that

$$\mathbb{E}(Y|Z) = \beta_1 \mathbb{E}(X|Z) + g^*(Z),$$

   and that

$$Y - \mathbb{E}(Y|Z) = \beta_1 (X - \mathbb{E}(X|Z)) + \epsilon.$$

   This is known as the Robinson transformation/decomposition, proposed in Robinson (1988). This shows that β_1 can be interpreted as the coefficient in a linear model where the outcome variable is $(Y - E(Y|Z))$ and the predictor variable is $(X - E(X|Z))$.

4. The above decomposition suggests the following procedure, which uses only tools we have introduced in class:

   - Estimate $\mathbb{E}(Y|Z)$ using a nonparametric regression model (e.g., local linear regression) and compute the residuals $Y - \mathbb{E}(Y|Z)$;
   - Estimate $\mathbb{E}(X|Z)$ using a nonparametric regression model (e.g., local linear regression) and compute the residuals $X - \mathbb{E}(X|Z)$;

- Estimate β_1 by regressing the residuals $Y - \mathbb{E}(Y|Z)$ on $X - \mathbb{E}(X|Z)$ using linear regression.

Implement this procedure, which we refer to as Robinson procedure, using the `npregbw` and `npreg` commands from the `np` library, and `lm`. Use cross-validation to select bandwidths (this is implemented in `npregbw`).

5. Plot the estimate of the regression function $\mathbb{E}(Y|Z)$ and that of $\mathbb{E}(X|Z)$. Comment on the results.

6. Report the estimate of β_1 and the p-value associated with the hypothesis test $\beta_1 = 0$. Compare this result with that from the model in equation (4) and comment on it.

7. Attach your **commented** code.

Problem 3: Simulations

In this problem, we design a simulation to validate the procedure used in the application described above and to examine the behavior of local regression models with respect to the choice of hyperparameters.

- **Nonparametric model:** Generate 100 observations from the following model:

$$Y = g^*(Z) + \epsilon, \tag{5}$$

where $Z \sim \text{Unif}(0, 3)$, $\epsilon \sim \mathcal{N}(0, 0.05^2)$, and $g^*(z) = \sin(z)^2$.

  - Fit both a Nadaraya-Watson and a local linear regression for various choices of the bandwidth using the `npregbw` and `npreg` commands from the `np` library. Comment on the results.
  - Regenerate the data with $g^*(Z) = Z$ and fit both a Nadaraya-Watson and a local linear regression for various choices of the bandwidth. Comment on the results.

- **Partially linear model:** Generate 100 observations from the following model:

$$Y = \beta_1 \cdot X + g^*(Z) + \epsilon, \tag{6}$$

where $\beta_1 = 0$, $X \sim \text{Unif}(0, 3)$, $\epsilon \sim \mathcal{N}(0, 0.05^2)$, and $g^*(z) = \sin(z)^2$.

  - Apply Robinson's procedure from Problem 2 to estimate β_1. Report the p-value and comment on the result.
  - Regenerate the data 200 times and each time apply Robinson's procedure. Count how many times you reject the null hypothesis at a significance level of $\alpha = 0.05$. Discuss whether this number is close to your expectations.
  - Set $\beta_1 = 1$, regenerate the data 200 times, and each time repeat Robinson's procedure, counting how many times you reject the null hypothesis at a significance level of $\alpha = 0.05$. Comment on the results.

- **[Extra credit] Partially linear model with dependence:** Generate 100 observations from the following model:

$$Y = \beta_1 \cdot X + g^*(Z) + \epsilon, \tag{7}$$

where $\beta_1 = 0$, $g^*(z) = \sin(z)^2$, but **enforce some association** between X and Z.

  - Regenerate the data 200 times and each time repeat Robinson's procedure. Count how many times you reject the null hypothesis at a significance level of $\alpha = 0.05$. Discuss whether this number is close to your expectations.
  - For each value of $\beta_1 \neq 0$, on a suitable grid of positive values, regenerate the data 200 times, and each time repeat Robinson's procedure, counting how many times you reject the null hypothesis at a significance level of $\alpha = 0.05$. Plot the power curve and comment on the results.