Stat 27850/30850: Problem set 4

1. **Randomized inference.**
Suppose that you observe data where the response is modeled as a Binomial, with a parameter that depends on the features:

$$Y_i \sim \text{Binomial}(n_i, f(X_i)), \tag{1}$$

where $X_i \in \mathbb{R}^p$ is the feature vector for sample #i (treated as fixed, not random), and where $f(\cdot)$ is some link function (for example, we might have $f(x) = h(x^\top \beta)$ where h is the logistic function and β is a vector of coefficients, and we are interested in learning about β). In practice, Y_i counts the number of events that occur for sample #i, for example, each i is an person and Y_i is the number of successes out of n_i many attempts made by that person; X_i is a feature vector describing demographic information etc for person #i.

Suppose that we want to explore the data to develop some hypotheses about the form of f—for example if we assume f is of the form $f(x) = h(x^\top \beta)$ with a sparse β, then we may want to do some feature selection at the exploratory stage. We will follow a randomized inference scheme:

(Step 1) First, independently for each i, draw $Z_i \sim \text{Binomial}(Y_i, \lambda)$, where $\lambda \in (0, 1)$ is a predefined subsampling parameter. That is, for each success that occurs, it has a chance λ of being recorded for the exploratory data set.

(Step 2) Using the $\{(X_i, Z_i) : i = 1, \ldots, n\}$ data, we then develop our hypotheses or questions about the model (that is, about f—for example, if we think $f(x) = h(x^\top \beta)$ for a known transformation h and unknown coefficients β_j, we might find certain β_j's we are interested in testing).

(Step 3) Finally, we then reveal the full data $\{(X_i, Y_i)\}$ to test these hypotheses.

In this problem, we will not be developing hypotheses about f—this would be something specific to the problem or application you are working with, in practice. Instead, we will be analyzing the randomized inference framework to see how the procedure above might work in general. In order to do this, we need to develop a precise understanding of two things:

- Question (a) — how does the distribution of the (X, Z) pairs, relate to the distribution of the (X, Y) pairs? This question is important because, in practice, we are interested in asking questions/choosing hypotheses about the (X, Y) pairs, but the only data we have available in order to make these choices (at Step 2 above) are the (X, Z) pairs.
- Question (b) — if we condition on the (X, Z) data (and therefore, condition on any selection event that might have occurred after looking at the (X, Z) data), what is now the conditional distribution of the remaining data (i.e., the Y values)? This answer is needed in order to perform inference at Step 3.

(a) (Answering Question (a) above.) For a fixed feature vector X_i, the response Y_i follows the Binomial model given in (1). What is the model for Z_i given X_i?

(b) (Answering Question (b) above.) After the exploratory stage, we will reveal the true responses Y_i. What is the distribution of Y_i conditioned on the data revealed at the exploratory stage? Your answer should be related to a Binomial distribution.

(c) Explain the role of the parameter λ, and give intuition for what would happen if we chose it to be larger or smaller.

2. **Conformal prediction.**
Consider the following simplistic algorithm for fitting a regression function to predict a response Y from a feature vector X. Our data points are $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$. Here is how we fit the regression:

kNN

- For each data point (X_i, Y_i), find its k nearest neighbors in X space, i.e. the data points $i_1, \ldots, i_k \neq i$ that give the smallest distances $\|X_{i_1} - X_i\|_2, \ldots, \|X_{i_k} - X_i\|_2$.
- The fitted value is $\hat{Y}_i = \frac{1}{k} \sum_{\ell=1}^{k} Y_{i_\ell}$, i.e. the mean of the response values of the k nearest neighbors.

Suppose that you implement conformal prediction with this method above as your model fitting algorithm. Let X_{n+1} be the feature vector for the test point, and suppose that we test each potential value $y \in \mathbb{R}$ for possible inclusion into the prediction set (i.e. the set that's meant to contain Y_{n+1} with probability at least $1 - \alpha$).

Your task is to calculate the predictive interval (or predictive set, if it's not a single interval) $\mathcal{Y}$ that is the output of the conformal prediction method at level α. Your answer will be of the form

$$\mathcal{Y} = \{y \in \mathbb{R} : (y \text{ satisfies some condition that you specify})\}$$

For convenience, you can assume that all pairwise distances between X's are unique, i.e. we don't have $\|X_i - X_j\|_2 = \|X_k - X_\ell\|_2$ for two different pairs $\{i, j\}$ and $\{k, \ell\}$.