STATISTICAL LEARNING - MODERN INFERENCE
Data Analysis

1 Prediction intervals

This exercise provides two data sets:

1. Low-dimensional dataset
2. High-dimensional dataset

For each dataset, it is provided both a training set (x_i, y_i), $i = 1, \dots, n$ and a test set x_i^*, $i = 1, \dots, m$. The goal is to provide a prediction interval $[l_i^*, u_i^*]$ for each (unknown) y_i^* of the test set.
Your predictions will be evaluated on both

1. Coverage
$$C = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{y_i^* \in [l_i^*, u_i^*]\}$$
The target coverage is 90%. The closer to 90%, the better.
2. Average length
$$L = \frac{1}{m} \sum_{i=1}^{m} (u_i^* - l_i^*)$$
the shorter, the better.

2 Variable selection

This exercise provides two data sets:

1. Low-dimensional dataset
2. High-dimensional dataset

The goal is to select the “relevant” predictors for each dataset. Data were generated as

$$y = X\beta^0 + \varepsilon$$

with $S_0 = \{j : \beta_j^0 \neq 0\}$ and $N_0 = \{j : \beta_j^0 = 0\}$. Your selection $\hat{S}$ will be evaluated on both true positive rate and false discovery rate:

$$\text{TPR} = \frac{|S_0 \cap \hat{S}|}{|S_0|}, \qquad \text{FDR} = \frac{|N_0 \cap \hat{S}|}{|\hat{S}|}$$