STA4042 24-25T1 Homework 1: Regressions

Inquery: 223040237@link.cuhk.edu.cn

September 9, 2024

We will use Python with the packages `ipykernel,numpy,pandas,scikit-learn,matplotlib,(scipy)`

Problem 1

In this section, you will be given some synthetic data points $\{x_i, y_i\}_i$ with simple underlying relationship $y_i = \alpha x_i + \epsilon_i$, where ϵ is a random noise. We follow the assumption that ϵ is drown from a Gaussian distribution. Therefore we work with OLS (Ordinary Least Squares) regression.

Questions (5+5+5+5+20+5+5 = 50 points)

1. Given some data points, derive the closed-form OLS solution and print your estimation.

From now on, you will work with data with outliers.

2. Given some data points with outliers, derive the closed-form OLS solution and print your estimation.

3. Display your estimated function together with the underlying true function.

4. Make some simple comments: Is the outcome satisfactory? Why?

5. A technique called "Huber Regression" allows to put less weight to the outliers solving this minimization problem:

$$\min_{\beta,\sigma} \sum_{i=1}^{n} \left(\sigma + H_\delta \left(\frac{\beta^T x_i - y_i}{\sigma} \right) \sigma \right) + \alpha \|\beta\|^2, \qquad \text{where: } H_\delta = \begin{cases} z^2 & \text{if } |z| < \delta \\ 2\delta|z| - \delta^2 & \text{otherwise.} \end{cases}$$

   Implement this method from scratch with some well chosen parameters α, δ (no sklearn models allowed) and print your estimation.

6. Display your Huber result, as well as the sklearn Huber (given in the code) result and the previous regression result.

7. Make some simple comments: How do you interpret the different outcomes?

Problem 2

Inn this section, you will deliver Poisson Regression to fit on a car insurance dataset. Please read `https://scikit-learn.org/stable/auto_examples/linear_model/plot_poisson_regression_non_normal_loss.html` to understand the dataset and the task.
In the end, you will perform your own Poisson Regression without the sklearn models. You are free to use any code from the site (e.g. the data splitting, preprocessing, framework for display, etc.). You might also use your own data preprocessing.

Expected output

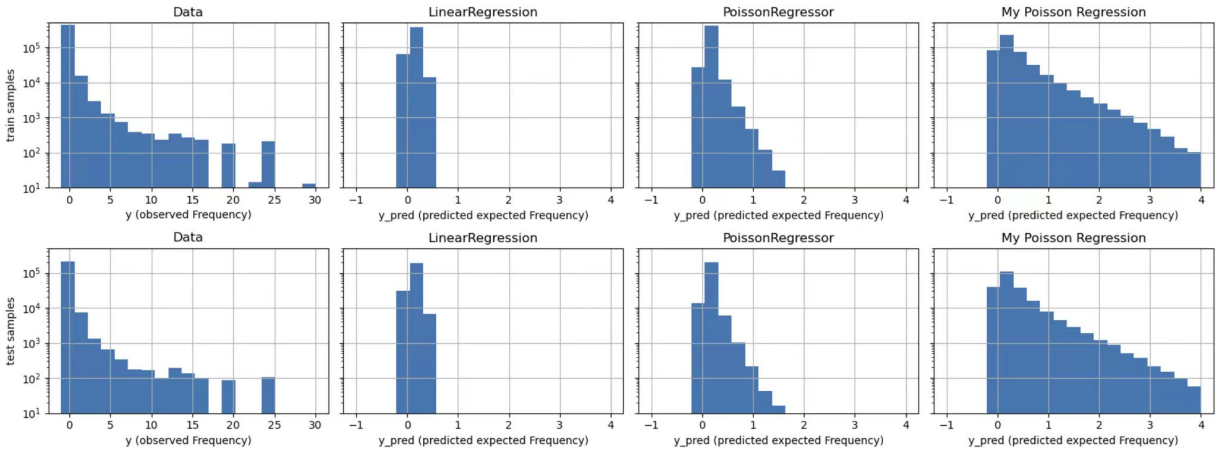

Figure 1: The two rows represent train and test data. Each column represents respectively:

- Data histogram
- Prediction histogram from Linear Regression
- Prediction histogram from Sklearn PoissonRegressor
- Prediction histogram from your own Poisson Regression

Questions (45+5 = 50 points)

8. Fit the models and show the desired 8 plots. [**Hint**]: If you do no see how to do it, here are some questions to be thought through: What is the likelihood for the data? What is the MLE in this case? Why is it impossible to derive the closed form solution directly as we did for OLS? What does the problem become now? You might use some solvers like *scipy.optimize.minimize*.

9. Comments: Do you think this method works well on the data? Give some short comments or any thinking that occurred to you during this task.