The 3 Regression Modes implemented in tornado

- **Poisson regression mode**
- **Negative binomial regression mode**
- **QPD (Quantile-parameterized distribution) regression mode**

Each mode has its own characteristics. It is necessary to select the appropriate mode according to the problem to be applied.
The following are examples of the results obtained when each mode is applied to data.

Brief introduction of each mode

- Poisson regression mode

The Poisson regression mode assumes that the objective variable follows a Poisson distribution. Therefore, accuracy is often better when the objective variable follows a Poisson distribution. The Poisson distribution is also used to predict probability distributions.

- Negative binomial regression mode

Negative binomial regression mode also assumes that the objective variable follows a Poisson distribution. Therefore, accuracy is often better when the objective variable follows a Poisson distribution. The difference with Poisson regression is that when forecasting probability distributions (as opposed to forecasting expected values (mean values)), forecasts are made using a negative binomial distribution. Negative binomial distribution often has better accuracy because it predicts the probability distribution taking into account the variance, whereas Poisson distribution predicts the probability distribution based only on the estimated expected value.

- QPD (Quantile-parameterized distribution) regression mode

The QPD regression mode does not assume any specific distribution that the objective variable follows. It analyzes the true distribution based on information from the center and both ends of the distribution of the observed data. Even if the true distribution cannot be known a priori, it can be fitted flexibly to both lineally symmetric and skewed distributions. QPD regression mode has more settable parameters than the others. There is `quantile` to set the quantile position to be estimated, `bound` to set whether the supported target range is limited or not, and `lower` and `upper` to specify the range. The `bound` can be `"U"` (unbound) to not limit the target range, `"S"` (semi-bound) to set the lower

bound of the target range, or `"B"` (bound) to set the upper or lower bound of the target range. Settings that do not match the target range can cause errors.

Performance comparison among regression modes

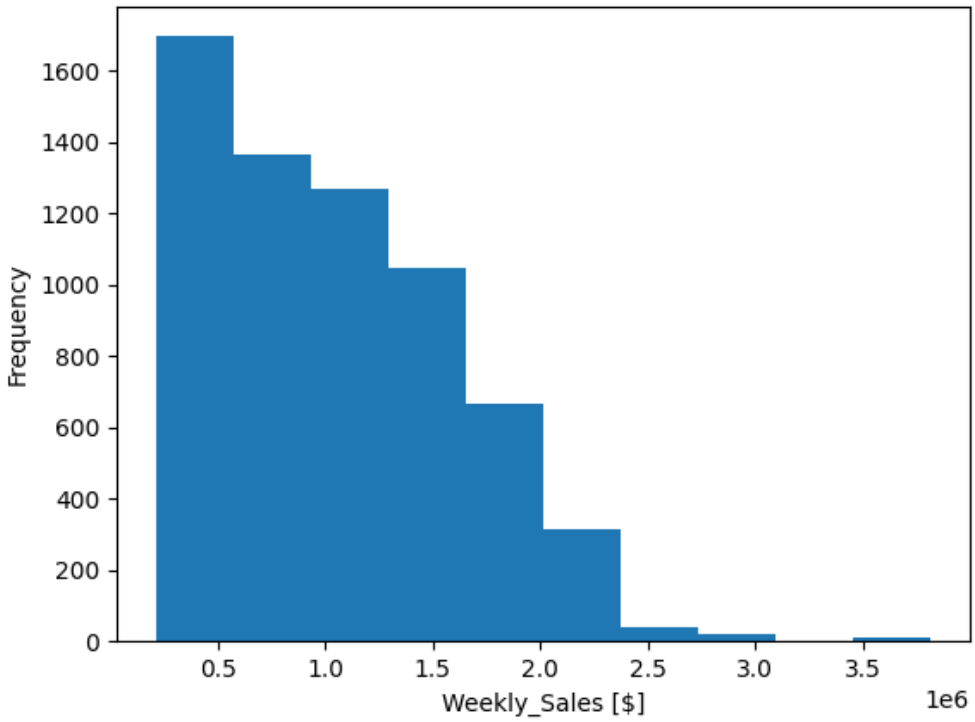
Apply estimation by the 3 regression modes to the 3 datasets and compare the results.

Datasets

- Walmart Dataset

Weekly sales data per store of Walmart, a major U.S. retailer
Walmart Dataset (kaggle.com)

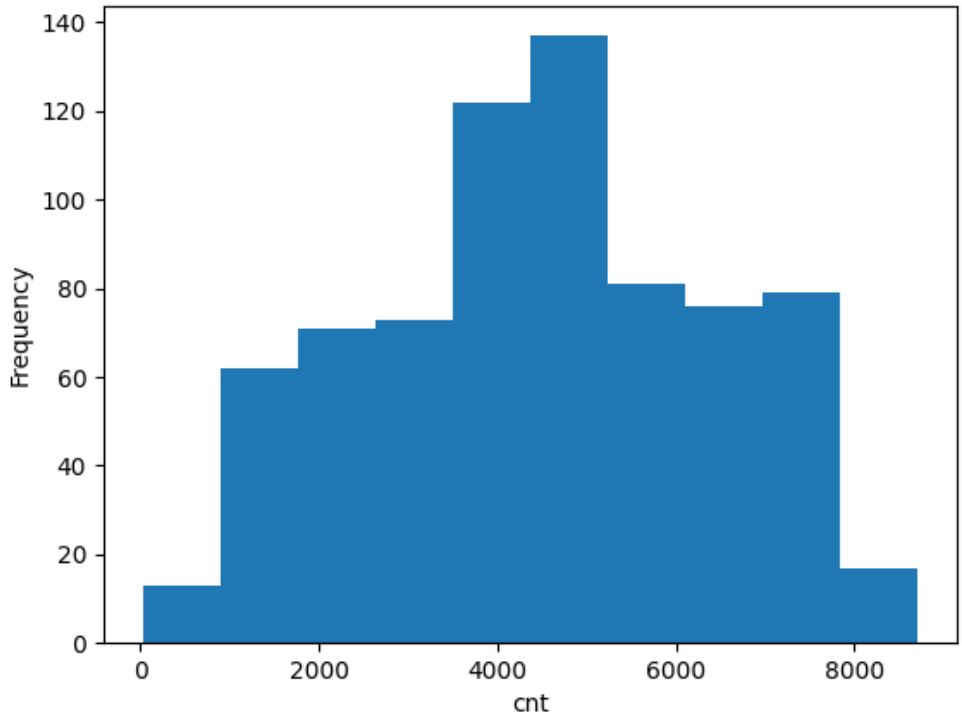
Records: 6435
Distribution of object variable:



- Daily Bike Sharing Dataset

Daily usage count data of the Capital Bikeshare system in Washington, D.C., U.S.
Bike Sharing Dataset (kaggle.com)

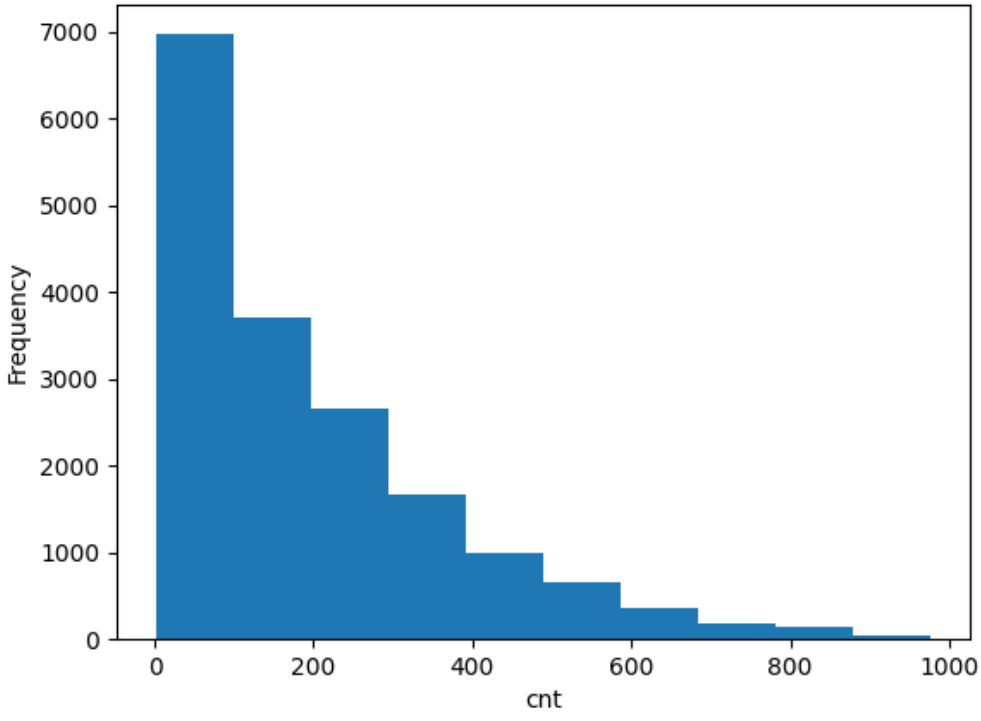
Records: 731
Distribution of object variable:



- Hourly Bike Sharing Dataset

Hourly usage count data of the Capital Bikeshare system in Washington, D.C., U.S.
Bike Sharing Dataset (kaggle.com)

Records: 17379
Distribution of object variable:



Evaluation

RMSE (Root Mean Squared Error)

dataset \ mode	Poisson	N-Binomial	QPD
Walmart	48827.96	48827.96	130686.01
Daily Bike Sharing	681.85	681.85	965.39
Hourly Bike Sharing	33.92	33.92	50.45

MAE (Mean Absolute Error)

dataset \ mode	Poisson	N-Binomial	QPD
Walmart	34706.14	34706.14	78088.17
Daily Bike Sharing	480.55	480.55	602.12
Hourly Bike Sharing	21.42	21.42	30.84

MAPE (Mean Absolute Percentage Error) [%]

dataset \ mode	Poisson	N-Binomial	QPD
Walmart	3.65	3.65	7.80
Daily Bike Sharing	88.39	88.39	153.41
Hourly Bike Sharing	17.85	17.85	25.91

Accuracy of Probability Distribution Prediction

Accuracy of the probability distribution calculated based on Wasserstein distance between the cumulative distribution function (CDF) of the predicted probability distribution at each observed value and the uniform distribution. The value range is from 0 to 1. The closer to 1, the better the accuracy. For more details, visit here.

dataset \ mode	Poisson	N-Binomial	QPD
Walmart	0.51	0.96	0.64
Daily Bike Sharing	0.61	0.92	0.71
Hourly Bike Sharing	0.84	0.98	0.71

Processing Time [s]

dataset \ mode	Poisson	N-Binomial	QPD
Walmart	24.69	24.71	251.81
Daily Bike Sharing	8.68	8.80	35.11
Hourly Bike Sharing	118.69	113.23	384.41

The QPD regression mode has a longer processing time than the others for the three types of data sets. Here, `bound` ="S" (semi-bound) and the lower limit is set to 0.0 to account for the characteristics of the target variable in the datasets.
Mean prediction is the same for Poisson regression mode and negative binomial regression mode, but negative binomial regression mode is more accurate in predicting the probability distribution because it takes into account the variance of the probability distribution.
Considering the characteristics of each mode with reference to the above, it is necessary to select the mode that best suits the purpose of the problem.