Poisson Binomial Distribution Model (PBD)

Tianyang Xie, Qie He

November 2019

Our goal of the analytical module is to provide prediction for the total number of uncovered work in each combination of day, garage, duty category and time part. Instead of the individual response (which piece of work will be uncovered), the aggregated response (the number of uncovered works) is what we try to predict in the end. Therefore we introduce PBD, developed for modeling the relationship between features of individual pieces of work and aggregated responses. The PBD model has the following advantages comparing to the regular analytic tools.

- Unlike the binomial distribution, the PBD model does not require the parameters of individual Bernoulli distributions to be the same.
- The theory of the PBD model is well defined. The estimation, inference and prediction techniques follow the classical statistical theory.
- The PBD model not only provides point prediction, but also reveals distributions. Overestimation for prediction can be easily achieved.

To our knowledge there are no well-developed R or Python packages for the PBD model. The UMN team built our own modules for estimation, selection, and prediction with the PBD model in R. Different modules of the PBD model will be discussed in later sections.

1 Formulation

Assume that there are N days of duty fulfillment data and each data point represents a piece of work along with a set of features. Assume that there are n_k pieces of work in the kth day, for $k = 1, \dots, N$. Let $X_{k,i} \in R^p$ be the feature vector of the ith piece of work in the kth day, for $i = 1, \dots, n_k$ and $k = 1, \dots, N$. Let $p_{k,i}$ be the probability of whether the ith piece of work in the kth day will be uncovered. Let $D_k \in R$ be the aggregated observation (the total number of uncover work) in the kth day. Let $\theta \in R^p$ be the coefficient vector of the PBD model. The PBD model can be described as follows.

$$\begin{aligned}
p_{k,i} &= \frac{1}{1+e^{-X_{k,i}^T\theta}}, && i=1,\dots,n_k, k=1,\dots,N,\\
y_{k,i} &\sim^{ind} \text{Bernoulli}(p_{k,i}), && i=1,\dots,n_k, k=1,\dots,N,\\
D_k &= \sum_{i=1}^{n_k} y_{k,i}, && k=1,\dots,N.
\end{aligned}$$

It is difficult to derive the gradient or Hessian matrix directly from the exact distribution of D_k. Therefore, employing the idea from E. Rosenman & N. Viswanathan in 2018 [1], we use a normal distribution to approximate the exact distribution, so that the gradient and Hessian could be computed in an easier way.

From Lyapunov CLT, assuming each $p_{k,i}$ is bounded below from 0 and above from 1, we have the following conclusion:

$$D_k \xrightarrow{d} N(\sum_{i=1}^{n_k} p_{k,i}, \sum_{i=1}^{n_k} p_{k,i}(1-p_{k,i})), \quad \text{as } n_k \to \infty.$$

2 Estimation

Based on the approximated distribution for the aggregated responses, we can estimate the unknown coefficients θ by maximum likelihood estimation.

The log-likelihood function, the gradient and the Hessian matrix with respect to the coefficients θ for each group can be derived as the following (group index k is omitted for clarity, and all summation here is in group only):

The log-likelihood function for each group:

$$l = -\frac{(D-\sum p_i)^2}{2\phi^2} - \frac{1}{2}\log\phi^2, \text{where} \quad \phi^2 = \sum p_i(1-p_i).$$

The gradient vector for each group is calculated as follows:

$$\nabla_\theta l = \underbrace{\frac{1}{2}\left[\frac{1}{\phi^4}(D-\sum p_i)^2 - \frac{1}{\phi^2}\right]}_{①} \cdot \underbrace{\left[\sum(1-2p_i)p_i(1-p_i)X_i\right]}_{②} + \underbrace{\frac{1}{\phi^2}(D-\sum p_i)\left[\sum p_i(1-p_i)X_i\right]}_{③}.$$

To derive the Hessian matrix for each group, we first derive the derivative of each components in the gradient vector:

$$
\begin{aligned}
\nabla_\theta \textcircled{1} &= \frac{1}{2}\left[\frac{1}{\phi^4} - \frac{2}{\phi^6}(D - \sum p_i)^2\right]\left[\sum (1-2p_i)p_i(1-p_i)X_i\right] - \frac{1}{\phi^4}(D - \sum p_i)\left[\sum p_i(1-p_i)X_i\right] \\
\nabla_\theta \textcircled{2} &= \sum \left[X_i^T(1 - 6p_i + 6p_i^2)p_i(1-p_i)X_i\right] \\
\nabla_\theta \textcircled{3} &= -\frac{1}{\phi^4}(D - \sum p_i)\left[\sum p_i(1-p_i)X_i\right]^T\left[\sum (1-2p_i)p_i(1-p_i)X_i\right] \\
&\quad + \frac{1}{\phi^2}\left\{\left[\sum p_i(1-p_i)X_i\right]^T\left[-\sum p_i(1-p_i)X_i\right] + (D - \sum p_i)\left[\sum X_i^T(1-2p_i)p_i(1-p_i)X_i\right]\right\}.
\end{aligned}
$$

Therefore, the Hessian matrix for each group is:

$$\nabla_\theta^2 l = \nabla_\theta \textcircled{1} \cdot \textcircled{2}^T + \textcircled{1} \cdot \nabla_\theta \textcircled{2} + \nabla_\theta \textcircled{3}.$$

After we derived the gradient and Hessian matrix for coefficients θ, we can utilize first-order algorithms such as gradient descent, pseudo-second-order algorithms such as BFGS, or second-order methods such as Newton algorithm, to solve the maximum likelihood estimator of θ. In our training module, we utilized the BFGS algorithm for the estimation. The training process takes less than 5 minutes.

3 Prediction & Inference

After we compute the estimator $\hat{\theta}$ for true coefficients θ, we can perform the expectation prediction as:

$$\hat{D}_k = \sum_{i=1}^{n_k} p_{\hat{k},i} = \sum_{i=1}^{n_k} \frac{1}{1 + e^{-X_{k,i}^T \hat{\theta}}}.$$

Since PBD model not only provide point prediction, but also reveals the distribution of D_k, we can also present overestimated prediction by upper tail confidence bound, controlled by a confidence level. For example, if one wants to have predictions that has 90% probability to cover the true value, he can select the confidence level to be 90% to perform overestimated prediction:

$$\hat{D}_{k,90\%} = \hat{D}_k + Z_{90\%} * \hat{\phi}_k^2.$$

In the equation, $\hat{D}_{k,90\%}$ the overestimated prediction, $\hat{D}_k$ is the expectation prediction, $Z_{90\%}$ is the upper tail bound for standard normal distribution of 90% confidence level, $\hat{\phi}_k^2$ is the estimated variance defined previously.

Please note that expectation prediction is actually a special case of overestimated prediction: when the confidence level is set to be 50%, the two prediction technique provide the same prediction results.

For inference, we utilize the asymptotic property of maximum likelihood estimator, presented as the following:

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, I^{-1}(\theta))$$

In the theory above, $\hat{\theta}$ is the MLE estimator for θ, and $I^{-1}(\theta)$ is the inverse matrix of Hessian matrix. N refers to the normal distribution. In a word, when sample size approach infinity, the MLE estimator will converge to the true coefficients, and the error of them will converge to a normal distribution.

By measuring the difference between estimated coefficients and 0 on the normal distribution with designated variance, we can detect the significance of each feature. The concept of this inference tool is the same as logistic regression.

References

[1] Evan Rosenman and Nitin Viswanathan. Using poisson binomial GLMs to reveal voter preferences. *arXiv preprint arXiv:1802.01053*, 2018.