Bayesian quantile regression and unsupervised learning methods to the US Army and Navy data

Jong-Min Kim
Division of Science and Mathematics,
University of Minnesota-Morris,
Morris, USA
Email: jongmink@morris.umn.edu

Chuwen Li
Department of Statistics,
University of Michigan,
1085 South University,
Ann Arbor, MI 48109-1107, USA
Email: chuwenli@umich.edu

Il Do Ha*
Department of Statistics,
Pukyong National University,
Busan, South Korea
Email: idha1353@pknu.ac.kr
*Corresponding author

Abstract: We apply the Bayesian quantile regression (BayesQR) model for binary response variables and the unsupervised learning methods to synthetic data (Stevens and Anderson-Cook, 2017a, 2017b), which is univariate data with a binary response of passing or failing for complex munitions generated to match age and usage rate found in US Department of Defense complex systems (Army and Navy). Instead of the generalised linear model (GLM) used in Stevens and Anderson-Cook (2017a), we propose to apply the BayesQR to predict a binary response of passing or failing for the Army and Navy data as well as the unsupervised learning methods. First, we want to find the best models for the Army and Navy through comparing statistical inference of BayesQR and GLMs and calculating their percentage correctly classified (PCC) which tests the accuracy of a prediction. The second method focuses on clustering using the k-means clustering and random forests based on the results of BayesQR. We compare models with different covariates to find the one that can best divide data into two groups: Army and Navy.

Keywords: generalised linear model; GLM; BayesQR; k-means; random forests.

Introduction

Stevens and Anderson-Cook (2017a) analysed and discussed the probability of agreement between two related populations of systems using the Army and Navy data. The Army and Navy data is about US defence applications produced by a single manufacturer that are distributed between the Army and Navy. The variables in the data are the defence application’s reliability, pass or fail (binary), each application’s age, and each application’s usage rate, which is the change in the number of annual transfers. Stevens and Anderson-Cook (2017a) fit the data using the generalised linear model (GLM) with probit link function to predict the reliability of applications in each system, and they further sought the probability of agreement in responsibility between two systems. In this paper, we suggest using a Bayesian quantile regression (BayesQR) (Yu and Moyeed, 2001) to fit the Army and Navy data and compare its results with GLM allowing for probit or logit link function. We also apply the unsupervised learning methods: the k-means clustering and random forests (RFs) to the data to see which one gives a better grouping for this specific case.

Recently, Kim et al. (2019a) proposed generalised linear mixed logit and probit models applied to US Army and Navy data. Kim et al. (2019b) proposed machine learning techniques applied to US Army and Navy data.

To propose the different method with US Army and Navy data, we propose the BayesQR model through R ‘BayesQR’ package to predict the data instead of the GLM. While ordinary least squares (OLS) can be inefficient if the errors are highly non-normal, QR is more robust to non-normal errors and outliers. QR also provides a richer characterisation of the data, allowing us to consider the impact of a covariate on
the entire distribution of $y$, not merely on its conditional mean. Different measures of central tendency and statistical dispersion can be useful to obtain a more comprehensive analysis of the relationship between variables. Standard linear regression techniques summarise the average relationship between a set of regressors and the outcome variable based on the conditional mean function $E(y \mid x)$. This provides only a partial view of the relationship, as we might be interested in describing the relationship at different points in the conditional distribution of $y$. The median estimator and other quantile estimators are influenced less by outliers in the response variables conditional on the explanatory variables (Benoit and Van den Poel, 2017). Especially, BayesQR overcomes the difficulty of inference in the traditional quantile regression when the data is clustered or censored. Traditionally, we use the R ‘quantreg’ package for the traditional quantile regression model for binary response. However, it is not easy to interpret the estimated coefficients of the parameters in quantile regression because a traditional quantile regression model establishes a relationship between the percentiles of a continuous outcome and a set of predictors Koenker and Hallock (2001). On the other hand, Bayesian modelling enables exact inference and its estimator is equivalent to the maximum likelihood solution under an asymmetric Laplace (ASL) error distribution developed by Geraci and Bottai (2007). In this paper, we choose 10%, 30%, 50%, 70%, and 90% quantile regression models, and compare them to the GLMs. Their statistical significance of parameters is tested on the 95% credible interval. Furthermore, we take GLM with logit link function into consideration. Stevens and Anderson-Cook (2017a) considered three models for the Army and Navy using GLM with probit link function. We compute the percentage correctly classified (PCC), which tests the proportion of manufacturers that are classified as good responsibility (binary response = 1) for the proposed models in this paper. Through the credible interval and PCC, we try to find the best model among BayesQR models, GLM Logit, and GLM probit in predicting the data. We also plot the beta coefficients in each BayesQR model to see how they change in different quantiles. Our results shows that most BayesQR models with odd quantile estimators with one covariate, either age or usage, contain only statistical significant parameters and provide more accurate predictions than GLM logit/probit in terms of PCC.

The second part of the paper focuses on unsupervised learning, which is a type of machine learning, to find the best model for the Army and Navy data. Machine learning is a kind of learning techniques that makes precise predictions depending on what is learned from the available information. It contains the algorithms and techniques combining ideas from statistics, probability, and optimisation Mohri et al. (2018). Based on the amount and type of supervision during the training set, machine learning can be divided into supervised learning, unsupervised learning, semi supervised learning, and reinforcement learning (Géron, 2017). For the unsupervised training, its goal is to cluster the data to see whether the data can be divided into different groups. In contrast to supervised learning that predicts and assigns data to ‘right classification’ given beforehand, unsupervised learning groups data first and we assign meanings to each group afterwards. In this paper, we apply the unsupervised learning methods to the data to see which of the GLM model covariate results (age and/or usage) partitions the data in a way that most closely resembles the way the group originally divides. We adopt two types of unsupervised learning: the k-means clustering and RFs. The k-means clustering partitions observations based on the nearest mean to the cluster while RFs is a collection of individual classification tree predictors (Seligson et al., 2005). We
compare the results of these two methods and check whether they give the same best variables in classifying the units in Army and Navy data. Further more, we analyse the results of statistical inference of BayesQR models and the results of unsupervised learning methods together to discuss common points.

In Section 2, various statistical methods which are GLMs, quantile regression, BayesQR with binary dependent variables, and unsupervised learning will be discussed. In Section 3, we give the whole data analysis including model comparison among BayesQR with odd quantile estimators and with age and/or usage, GLM Logit with age and/or usage, and GLM probit with age and/or usage, unsupervised learning. We generate the statistical inference of BayesQR and GLMs and make the plots of Bayesian estimated coefficients of quantile regression as well as the PCC comparison between those types of models. Section 4 discusses unsupervised learning and includes the comparison of results between the k-means and RFs and the conclusion follows in Section 5.

2 Statistical methods

There are \( i = 1, 2, 3, \ldots, n \) units in each of \( j = 1, 2 \) populations, where \( j = 1 \) represents the Army population and \( j = 2 \) represents the Navy population. The binary response for the \( i \)th unit is

\[
Y_{ij} = \begin{cases} 
1 & \text{if unit } i \text{ from population } j \text{ passes} \\
0 & \text{if unit } i \text{ from population } j \text{ fails}
\end{cases}
\]

In population \( j \), we condition \( Y_{ij} \) on age and/or usage denoted by \( a_{ij} \) and \( u_{ij} \) respectively. We define the probability that the unit passes \( P(Y_{ij} = 1) \) as \( \pi_{ij} \). The GLM extends the linear mixed model to non-normal data, which implies that for GLM the response variables can come from different distributions other than Gaussian distribution.

2.1 Generalised linear models

GLMs predict data using the maximum likelihood approach and only include fixed effects in linear predictors. There are three specifications in a GLM. First, the linear predictor for the \( i \)th response in the \( j \)th population, \( Y_{ij} \), is denoted as \( \eta_{ij} \). Here \( \eta_{ij} \) is conditioned on a column vector \( x_{ij} \) of explanatory variables with regression parameters \( \beta \). That is,

\[
\eta_{ij} = x_{ij}^T \beta
\]

Next the link function \( g(\cdot) \) specifies how to convert the expected value \( \mu_{ij} = E(Y_{ij}) \) to linear predictor \( \eta_{ij} \), i.e.,

\[
\eta_{ij} = g(\mu_{ij}) = x_{ij}^T \beta.
\]

Using a logit model as an example, we have

\[
\text{logit}[P(Y_{ij} = 1|x_{ij})] = x_{ij}^T \beta,
\]
where $\beta$ is the column vector of the fixed-effects regression coefficients (the ‘betas’). For the Army and Navy data with two populations $j = 1, 2$, the equation can be written as

$$P(Y_{ij} = 1|x_{ij}) = \frac{1}{1 + e^{-x_{ij}^T\beta}} = \pi(x_{ij}).$$

The likelihood function for GLMs in our Army and Navy data is

$$L(\beta) = \prod_{j=1}^{2} \prod_{i=1}^{n_j} \pi(x_{ij})^{y_{ij}}\{1 - \pi(x_{ij})\}^{1-y_{ij}},$$

and so the log-likelihood function is given by

$$l(\beta) = \log L(\beta) = \sum_{j=1}^{2} \sum_{i=1}^{n_j} [y_{ij} \log(\pi(x_{ij})) + (1 - y_{ij}) \log\{1 - \pi(x_{ij})\}].$$

Then the maximum-likelihood estimating equation for $\beta$ has the form

$$u(\beta) = \frac{\partial l(\beta)}{\partial \beta} = \sum_{j} \sum_{i} \frac{(Y_{ij} - \pi_{ij})x_{ij}}{\text{var}(Y_{ij})} \frac{\partial \mu_{ij}}{\partial \eta_{ij}} = 0,$$

where $\mu_{ij} = E(Y_{ij}) = \pi_{ij}$ and $\pi_{ij} = \pi(x_{ij})$. This is easily solved via standard softwares (e.g., SAS or R) using the Fisher scoring or Newton-Raphson method.

In some situations GLMs with probit link function can probably give the best goodness-of-fit of the test where response variables are assumed to have normal distributions (Bolker et al., 2009).

### 2.2 Quantile regression

Koenker and Bassett (1978) first introduced quantile regression to estimate the conditional quantile function of the response variable. Benoit and Van den Poel (2017) described the quantile regression as follows: To drive the quantile regression model, we can start from the regular regression model:

$$y_i = x_i^T \beta + \varepsilon_i,$$  \hspace{1cm} (4)

where $\varepsilon_i$ is the error term, $y_i \in \mathbb{R}$, and $x_i \in \mathbb{R}^k$ and $\beta \in \mathbb{R}^k$ are $k \times 1$ column vectors of explanatory variables and regression coefficients, respectively. With the zero conditional mean assumption $E(\varepsilon|x) = 0$, we can obtain the conditional mean model,

$$\mu(x_i) = E(y_i|x_i) = x_i^T \beta.$$

From the conditional mean model, we can get the estimated $\beta$ coefficients:

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} |y_i - x_i^T \beta|.$$  \hspace{1cm} (5)
Bayesian quantile regression and unsupervised learning methods

Now, consider a quantile regression model. Let the $\tau^{th}$ ($0 < \tau < 1$) quantile of $\varepsilon_i$ be the value $q_\tau$ satisfying $P(\varepsilon_i < q_\tau) = \tau$. Then the $\tau^{th}$ conditional quantile model of $y_i$ given $x_i$ is defined by

$$q_\tau(y_i|x_i) = x_i^T \beta_\tau,$$

where $\beta_\tau$ is a vector of regression coefficients dependent on $\tau$. The quantile regression above extends equation (5) to different quantiles of interest by introducing the loss function $\rho_\tau$: 

$$\hat{\beta}_\tau = \arg \min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \rho_\tau[y_i - x_i^T \beta_\tau], \quad (6)$$

where $\rho_\tau(x) = x(\tau - I(x < 0))$. $I(\cdot)$ is the indicator function and $\tau \in (0, 1)$. The loss function $\rho_\tau$ assigns a weight of $\tau$ to positive residuals and a weight of $(1 - \tau)$ to negative residuals. The quantity $\hat{\beta}_\tau$ is called the $\tau^{th}$ regression quantile. When $\tau = 0.5$, we are doing the median regression by minimising the sum of absolute residuals, called the estimation of least absolute deviation (LAD).

Barrodale and Roberts (1973) developed the simplex algorithm to minimise the objective function in equation (6), which is the default method in the quantreg R package and the SAS PROC QUANTREG procedure (Benoit and Van den Poel, 2017).

2.2.1 BayesQR with binary dependent variables

Later, Koenker and Machado (1999) introduced another way to minimise the problem in equation (6) which is called independently distributed asymmetric Laplace densities (ALD). Then Yu and Moyeed (2001) further elaborated the ALD approach and initiated the take-off of BayesQR. Bayesian implementation of quantile regression with binary dependent variables begins by the standard binary regression model:

$$y_i = I(y_i^* \geq 0) = x_i^T \beta_\tau + \varepsilon_i, \quad (7)$$

where $y_i$ is the indicator of the $i^{th}$ individual’s response determined by the underlying variable $y^*$. By using the Manski’s maximum score estimator, we are able to deal with the binary quantile regression:

$$\hat{\beta}_\tau = \arg \max_{\beta \in \mathbb{R}^k} n^{-1} \sum_{i=1}^n \rho_\tau(2y_i - 1)(2y_i - 1)\text{sgn}(x_i^T \beta), \quad (8)$$

where $\text{sgn}(\cdot)$ is the signum function. Scale normalisation is necessary because the parameter $\beta$ is only identified up to a scale.

However, there are multiple problems associated with the maximum score estimator. The first one is the difficult optimisation of the multidimensional step function. Also, calculating the inference is difficult because of the complicated asymptotic distribution of the estimator. The Bayesian ALD approach was invented to avoid these problems, but at a cost of adding additional distributional assumptions on the error terms. We first let

$$P(y_i = 1|x_i, \beta) = 1 - F_{y^*}(-x_i^T \beta), \quad (9)$$

where $F_{y^*}(\cdot)$ is the distribution function of $y^*$. The Bayesian implementation of quantile regression with binary dependent variables begins by the standard binary regression model:

$$y_i = I(y_i^* \geq 0) = x_i^T \beta_\tau + \varepsilon_i, \quad (7)$$

where $y_i$ is the indicator of the $i^{th}$ individual’s response determined by the underlying variable $y^*$. By using the Manski’s maximum score estimator, we are able to deal with the binary quantile regression:

$$\hat{\beta}_\tau = \arg \max_{\beta \in \mathbb{R}^k} n^{-1} \sum_{i=1}^n \rho_\tau(2y_i - 1)(2y_i - 1)\text{sgn}(x_i^T \beta), \quad (8)$$

where $\text{sgn}(\cdot)$ is the signum function. Scale normalisation is necessary because the parameter $\beta$ is only identified up to a scale.

However, there are multiple problems associated with the maximum score estimator. The first one is the difficult optimisation of the multidimensional step function. Also, calculating the inference is difficult because of the complicated asymptotic distribution of the estimator. The Bayesian ALD approach was invented to avoid these problems, but at a cost of adding additional distributional assumptions on the error terms. We first let

$$P(y_i = 1|x_i, \beta) = 1 - F_{y^*}(-x_i^T \beta), \quad (9)$$

where $F_{y^*}(\cdot)$ is the distribution function of $y^*$. The Bayesian implementation of quantile regression with binary dependent variables begins by the standard binary regression model:

$$y_i = I(y_i^* \geq 0) = x_i^T \beta_\tau + \varepsilon_i, \quad (7)$$

where $y_i$ is the indicator of the $i^{th}$ individual’s response determined by the underlying variable $y^*$. By using the Manski’s maximum score estimator, we are able to deal with the binary quantile regression:

$$\hat{\beta}_\tau = \arg \max_{\beta \in \mathbb{R}^k} n^{-1} \sum_{i=1}^n \rho_\tau(2y_i - 1)(2y_i - 1)\text{sgn}(x_i^T \beta), \quad (8)$$

where $\text{sgn}(\cdot)$ is the signum function. Scale normalisation is necessary because the parameter $\beta$ is only identified up to a scale.

However, there are multiple problems associated with the maximum score estimator. The first one is the difficult optimisation of the multidimensional step function. Also, calculating the inference is difficult because of the complicated asymptotic distribution of the estimator. The Bayesian ALD approach was invented to avoid these problems, but at a cost of adding additional distributional assumptions on the error terms. We first let

$$P(y_i = 1|x_i, \beta) = 1 - F_{y^*}(-x_i^T \beta), \quad (9)$$

where $F_{y^*}(\cdot)$ is the distribution function of $y^*$. The Bayesian implementation of quantile regression with binary dependent variables begins by the standard binary regression model:

$$y_i = I(y_i^* \geq 0) = x_i^T \beta_\tau + \varepsilon_i, \quad (7)$$

where $y_i$ is the indicator of the $i^{th}$ individual’s response determined by the underlying variable $y^*$. By using the Manski’s maximum score estimator, we are able to deal with the binary quantile regression:

$$\hat{\beta}_\tau = \arg \max_{\beta \in \mathbb{R}^k} n^{-1} \sum_{i=1}^n \rho_\tau(2y_i - 1)(2y_i - 1)\text{sgn}(x_i^T \beta), \quad (8)$$

where $\text{sgn}(\cdot)$ is the signum function. Scale normalisation is necessary because the parameter $\beta$ is only identified up to a scale.
where $F_{y^*}(\cdot)$ is the cumulative distribution function of the asymmetric Laplace variable $y^*$. Benoit and Van den Poel (2017) calculate the joint posterior distribution, $\psi(\cdot)$, of the unobservables $\beta$ and $y^*$ given the data $x$ and the quantile of interest $\tau$:

$$
\psi(\beta, y_i^* | y, x, \tau) \propto \pi(\beta) \prod_{i=1}^{n} \{I(y_i^* > 0)I(y_i = 1) + I(y_i^* \leq 0)I(y_i = 0)\} \\
\times ALD(y_i^* | x_i^T \beta, \sigma = 1, \tau), \quad (10)
$$

where $\pi(\beta)$ is the prior on the regression coefficients and $\sigma = 1$ because of identification issues. Here

$$ALD(x; \mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\{-\rho_{\tau}(x \sigma)\}$$

is the three-parameter density function of ALD.

The standard Bayesian approach for the models with a categorical dependent variable is to assume an unobserved variable determines the outcome. This can be seen from equations (7) and (10). Because $y^*$ is unknown, we need to estimate it from the data. By splitting the equation (10), we find that the fully conditional distribution of $y^*$ is given by:

$$
\psi(y_i^*|\beta, \tau, y_i, x_i) \sim ALD(y_i^* | x_i^T \beta, \sigma = 1, \tau) \\
\text{truncated at the left by 0, if } y_i = 1, \\
\psi(y_i^*|\beta, \tau, y_i, x_i) \sim ALD(y_i^* | x_i^T \beta, \sigma = 1, \tau) \\
\text{truncated at the right by 0, if } y_i = 0.
$$

(11)

In this paper, the prior is not specified so that this will result in the default dispersed prior in the R package on the model parameters. The number of MCMC iterations is set to 10,000 so that this is enough to find posterior convergence in the MCMC chain. When we summarise the outputs of the BayesQR model, we exclude the first 2,000 burn-in draws out of the 10,000 MCMC iterations.

2.3 Unsupervised learning methods

The goal of unsupervised learning is to cluster the data, so the data has no class labels or response variables at first with each vector of the same dimension.

2.3.1 The k-means

The k-means clustering is a type of unsupervised learning method. The purpose of this k-means algorithm is to find the number of groups represented by the variable $K$. Given a set of observations $x_i, i = 1, 2, \ldots, n$, where each observation is a $d$-dimensional real vector, the k-means clustering aims to group $x_i$ into $K(K \leq n)$ groups ($G = \{G_1, G_2, \ldots, G_k\}$) by minimising the within-cluster sum of squares (WCSS). The formula can be written as

$$
\arg \min_{G} \sum_{i=1}^{k} \sum_{x \in G_i} \| x - \mu_i \|^2,
$$

(12)
where $\mu_i$ is the mean of points in $G_i$. Equation (12) is equivalent to minimising the pairwise squared deviations of points:

$$
\arg\min_{G} \sum_{i=1}^{k} \frac{1}{2|G_i|} \sum_{x,y \in G_i} ||x - y||^2.
$$

### 2.3.2 Random forests

RF is a highly flexible machine learning algorithm with excellent accuracy. It integrates classification tree through ensemble learning. A RF predictor is a collection of individual classification tree predictors (Seligson et al., 2005). For every classification tree predictor, it represents a particular pattern, which gives uncorrelated prediction through top down approach under random sampling with replacement. At each split node of the classification tree, dissimilar patterns of the data are deterministically isolated by multiple tests. 99.9% individual classification predictors will give predictions that include all possible circumstances and in the end RF votes for common common classes.

RFs can be used in unsupervised learning problems starting with creating synthetic data and unknown cluster outcomes for data clustering. The null hypothesis is that there is no dependence structure in the data. Given a feature vector $x_i, i = 1, 2, \ldots, n; x_i \in \mathbb{R}^k$, the partitions are conducted at each node according to the following function:

$$
h(v, \theta_j) = \mathbb{R}^k x T,
$$

where $\theta_j$ represents the $j^{th}$ parameters associated with the $j^{th}$ tree node and $\theta_j \in T$, where $T$ is the space for the split parameters and $v$ is the incoming patterns. This equation can be reorganised as follows:

$$
\theta_j = \arg\max_{\theta_j \in T} I(\delta_j, \theta),
$$

where $I$ is the information gained by splitting samples points. We try to partition arriving data points at each node. It is obvious that the higher the information gain, the better the partition is. To maximise $I$ in equation (15), The entropy can be written as

$$
I = H(\delta) - \sum_{i \in L, R} \left| \frac{\partial^i}{\partial} H(\delta^i),
$$

where $| \cdot |$ is a determinant for the matrix. The test criteria at each weak leaner (split point) for unsupervised learning problem can be formulated as

$$
I(\delta_j, \theta) = H(\delta_j) - \sum_{i \in L, R} \left| \frac{\partial^i}{\partial} H(\delta_j^i).
$$

The entropy can be defined as

$$
H(\delta) = \frac{1}{2} log((2\pi e)^d |\Lambda(s)|).
$$

Then the information gain can be obtained as follows:

$$
I(\delta_j, \theta) = log(|\Lambda(s_j)|) - \sum_{i \in L, R} \left| \frac{\partial^i}{\partial} log(|\Lambda(s_j^i)|).
$$
3 Data analysis

To begin our data analysis and fit the data with GLMs, we want to understand the structure of the data. We constructed a table to show the basic information in the Army and Navy data.

Table 1 US Army and Navy data

<table>
<thead>
<tr>
<th>System</th>
<th>Characteristics</th>
<th>Data observed</th>
<th>Passes (rate)</th>
<th>Failures (rate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Army</td>
<td>Age, usage</td>
<td>380</td>
<td>336 (88.4%)</td>
<td>44 (11.6%)</td>
</tr>
<tr>
<td>Navy</td>
<td>Age, usage</td>
<td>470</td>
<td>465 (98.9%)</td>
<td>5 (1.0%)</td>
</tr>
</tbody>
</table>

Figure 1 Density curves of age comparison for the Army and Navy (see online version for colours)

In Table 1, we can see that the Army and Navy data includes two populations: Army and Navy. The total number of the Army systems in the data have 380 units of complex munitions, which is 90 units fewer than the total number of the Navy systems in the data. Out of 380 units, the 44 units of complex munitions fail to pass the reliability test in the Army and only the 5 units out of 470 munitions in the Navy systems are considered unreliable. Here, the passing rate and failure rate for each system is important. For the Army, 88.4% of the units are considered to be responsible and it is 98.9% for the Navy. These rates are the main standard for later decisions on which model gives a better prediction as well as on which method in unsupervised learning partitions the data closer to the true division. We also constructed density plots for both
systems to analyse how Army and Navy data is distributed as a function of age and usage. As shown in Figure 1, the Navy skews farther to the right than the Army, which implies that munitions in the Navy typically have a longer age than units in the Army. In Figure 2, munitions in the Army experience more transfers (larger usage rate) than most of the units in the Navy. Thus, we can summarise that most munitions in the Army were transferred more times than in the Navy in a shorter period of time.

Figure 2 Density curves of usage comparison for the Army and Navy (see online version for colours)

3.1 Bayesian quantile regression

It is obvious to see from the density plots that the data is non-normal. Since BayesQR is superior to GLM in dealing with non-normal data, we believe that BayesQR will provide a more precise prediction. We want to check the statistical significance of QR coefficients over the quantiles. We first got the 95% credible interval for BayesQR with odd quantile estimators ($\tau = 0.1, 0.3, 0.5, 0.7, 0.9$), GLM logit, and GLM probit for both Army and Navy in Table 2. We can see that for the Army, all parameters in models with one covariate, age or usage, are statistically significant, while for models with two covariates, age and usage, the coefficient for the age variable $\beta_1$ is statistically insignificant. This holds true for both BayesQR and GLMs. For the Navy, things are getting interesting. Still, all $\beta_o$ are statistically significant over quantiles in BayesQR, while $\beta_1$ and $\beta_2$ are statistically insignificant over $\tau = 0.7$ and $\tau = 0.9$ in BayesQR with age and/or usage covariate. Both BayesQR and GLMs fit poorly with two covariates except for BayesQR ($\tau = 0.9$) in which all parameters are statistically significant.
Overall, models with only one covariate outperform models with two covariates for both BayesQR and GLMs.

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Model</th>
<th>$\hat{\beta}_0$ (CI)</th>
<th>$\hat{\beta}_1$ (CI)</th>
<th>$\hat{\beta}_2$ (CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age only</td>
<td>BayesQR ($\tau = 0.1$)</td>
<td>8.31 (4.75, 12.17)</td>
<td>$-0.67$ ($-0.95, -0.41$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.3$)</td>
<td>7.80 (5.65, 10.12)</td>
<td>$-0.48$ ($-0.63, -0.34$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.5$)</td>
<td>9.80 (7.66, 12.10)</td>
<td>$-0.52$ ($-0.66, -0.38$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.7$)</td>
<td>16.52 (13.73, 19.32)</td>
<td>$-0.79$ ($-0.96, -0.63$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.9$)</td>
<td>41.25 (35.80, 47.21)</td>
<td>$-1.74$ ($-2.13, -1.35$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM logit</td>
<td>6.69 (5.18, 8.53)</td>
<td>$-0.35$ ($-0.47, -0.25$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM probit</td>
<td>3.64 (2.86, 4.57)</td>
<td>$-0.19$ ($-0.25, -0.14$)</td>
<td>-</td>
</tr>
<tr>
<td>Usage only</td>
<td>BayesQR ($\tau = 0.1$)</td>
<td>11.33 (7.00, 16.31)</td>
<td>$-0.13$ ($-0.18, -0.08$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.3$)</td>
<td>10.29 (7.49, 13.41)</td>
<td>$-0.09$ ($-0.12, -0.06$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.5$)</td>
<td>12.83 (10.29, 15.50)</td>
<td>$-0.10$ ($-0.12, -0.08$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.7$)</td>
<td>18.93 (15.46, 23.00)</td>
<td>$-0.13$ ($-0.17, -0.10$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.9$)</td>
<td>43.89 (36.20, 53.43)</td>
<td>$-0.26$ ($-0.33, -0.21$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM logit</td>
<td>7.94 (6.20, 10.16)</td>
<td>$-0.06$ ($-0.08, -0.05$)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM probit</td>
<td>4.27 (3.41, 5.31)</td>
<td>$-0.03$ ($-0.04, -0.02$)</td>
<td>-</td>
</tr>
<tr>
<td>Age and usage</td>
<td>BayesQR ($\tau = 0.1$)</td>
<td>11.47 (7.11, 17.37)</td>
<td>$0.04$ ($-0.38, 0.42$)</td>
<td>$-0.13$ ($-0.18, -0.09$)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.3$)</td>
<td>10.32 (7.28, 14.72)</td>
<td>$-0.00$ ($-0.28, 0.25$)</td>
<td>$-0.09$ ($-0.12, -0.06$)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.5$)</td>
<td>13.05 (9.73, 17.85)</td>
<td>$-0.07$ ($-0.30, 0.20$)</td>
<td>$-0.09$ ($-0.13, -0.06$)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.7$)</td>
<td>20.52 (14.98, 26.21)</td>
<td>$-0.10$ ($-0.48, 0.31$)</td>
<td>$-0.13$ ($-0.18, -0.09$)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.9$)</td>
<td>43.31 (35.89, 52.20)</td>
<td>$0.05$ ($-0.67, 0.75$)</td>
<td>$-0.27$ ($-0.33, -0.20$)</td>
</tr>
<tr>
<td></td>
<td>GLM logit</td>
<td>8.19 (6.08, 10.92)</td>
<td>$-0.03$ ($-0.21, 0.14$)</td>
<td>$-0.06$ ($-0.08, -0.04$)</td>
</tr>
<tr>
<td></td>
<td>GLM probit</td>
<td>4.29 (3.27, 5.57)</td>
<td>$-0.00$ ($-0.10, 0.09$)</td>
<td>$-0.03$ ($-0.04, -0.02$)</td>
</tr>
</tbody>
</table>

Notes: $\tau$ is a weight to residuals. The quantity $\hat{\beta}$ is called the $\tau^\text{th}$ regression quantile.

Next, we plot the parameters in BayesQR against changing quantile estimators for the Army and Navy in Figures 3 and 4 to visually see the overall trend of the coefficients over quantiles including both odd and even quantile estimators. The shaded area around the line suggests the credible intervals for each $\beta$ against the corresponding $\tau$. For the Army, plots show that $\beta$ have quite a strong relationship with $\tau$: As $\tau$ increases in all BayesQR models, $\hat{\beta}_0$ increases as well and the increasing speed accelerates. The narrowing shaded area suggests that the credible interval for $\hat{\beta}$ narrows as $\tau$ increases. As $\tau$ increases in BayesQR models with one covariate, $\hat{\beta}_1$ and $\hat{\beta}_2$ decrease and their decreasing speed accelerates. When the BayesQR model includes two covariates, only $\hat{\beta}_0$ has a positive relationship and $\hat{\beta}_2$ has a negative relationship with $\tau$ while $\hat{\beta}_1$ shows no correlation with $\tau$ on the plot. For the Navy, $\hat{\beta}_0$ in all BayesQR models also have a positive correlation with $\tau$, but their credible intervals are much wider than those in the Army. The situations for $\hat{\beta}_1$ and $\hat{\beta}_2$ are completely different in the Navy compared with those in the Army. It seems that neither $\hat{\beta}_1$ nor $\hat{\beta}_2$ has a strong correlation with $\tau$ in any BayesQR models. In BayesQR with age covariate, there might be a slight upward trend around $\tau = 0.7$ that as $\tau$ increases, $\hat{\beta}_1$ increases as well. In BayesQR with age covariate, there are two upward trends between $\tau$ and $\hat{\beta}_2$: one is around $\tau = 0.4$ and the other is around $\tau = 0.6$. In BayesQR with both age and usage covariates, there is an upward trend between $\tau$ and $\hat{\beta}_1$ around $\tau = 0.6$, while no relationship is shown between
Bayesian quantile regression and unsupervised learning methods

The Navy shows a smaller correlation between $\tau$ and $\beta_2$ than for the Army and it shows larger credible intervals for $\beta$ given the same $\tau$ in the Army. These make sense in Tables 2 and 3 where fewer coefficients of covariates in BayesQR models for the Navy are statistically significant over quantiles compared to coefficients in BayesQR models for the Army.

Table 3 Statistical significance of BayesQR model and fixed model for the Navy

<table>
<thead>
<tr>
<th>Variable(s)</th>
<th>Model</th>
<th>$\hat{\beta}_0$ (CI)</th>
<th>$\hat{\beta}_1$ (CI)</th>
<th>$\hat{\beta}_2$ (CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age only</td>
<td>BayesQR ($\tau = 0.1$)</td>
<td>6.31 (2.95, 11.65)</td>
<td>-0.25 (–0.55, –0.05)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.3$)</td>
<td>9.36 (4.76, 15.27)</td>
<td>-0.31 (–0.63, –0.02)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.5$)</td>
<td>13.23 (7.68, 18.84)</td>
<td>-0.38 (–0.68, –0.03)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.7$)</td>
<td>20.77 (13.29, 30.03)</td>
<td>-0.45 (–0.97, 0.06)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.9$)</td>
<td>38.77 (28.80, 46.97)</td>
<td>-0.02 (–0.79, 0.99)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM logit</td>
<td>8.00 (4.89, 12.82)</td>
<td>-0.24 (–0.49, –0.04)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM probit</td>
<td>3.71 (2.48, 5.73)</td>
<td>-0.10 (–0.21, –0.02)</td>
<td>-</td>
</tr>
<tr>
<td>Usage only</td>
<td>BayesQR ($\tau = 0.1$)</td>
<td>5.00 (2.46, 8.78)</td>
<td>-</td>
<td>7.17 (4.88, 10.11)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.3$)</td>
<td>7.69 (5.10, 10.62)</td>
<td>-0.04 (–0.08, –0.00)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.5$)</td>
<td>15.97 (11.92, 20.38)</td>
<td>-0.07 (–0.12, –0.02)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.7$)</td>
<td>16.44 (11.10, 22.25)</td>
<td>-0.07 (–0.15, 0.03)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.9$)</td>
<td>45.03 (33.94, 55.31)</td>
<td>-0.12 (–0.32, 0.08)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM logit</td>
<td>6.53 (4.63, 9.27)</td>
<td>-0.04 (–0.08, –0.01)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>GLM probit</td>
<td>3.18 (2.40, 4.28)</td>
<td>-0.02 (–0.03, 0.00)</td>
<td>-</td>
</tr>
<tr>
<td>Age and usage</td>
<td>BayesQR ($\tau = 0.1$)</td>
<td>7.30 (2.91, 14.62)</td>
<td>-0.19 (–0.65, 0.15)</td>
<td>-0.03 (–0.09, 0.02)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.3$)</td>
<td>8.55 (4.94, 13.01)</td>
<td>-0.13 (–0.47, 0.19)</td>
<td>-0.03 (–0.09, 0.02)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.5$)</td>
<td>13.89 (8.26, 19.91)</td>
<td>-0.33 (–0.82, 0.19)</td>
<td>-0.04 (–0.16, 0.04)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.7$)</td>
<td>20.08 (14.48, 26.15)</td>
<td>-0.18 (–0.81, 0.41)</td>
<td>-0.07 (–0.18, 0.06)</td>
</tr>
<tr>
<td></td>
<td>BayesQR ($\tau = 0.9$)</td>
<td>40.67 (34.28, 47.99)</td>
<td>1.98 (1.11, 2.71)</td>
<td>-0.51 (–0.71, –0.28)</td>
</tr>
<tr>
<td></td>
<td>GLM logit</td>
<td>7.84 (4.80,12.57)</td>
<td>-0.15 (–0.45, 0.14)</td>
<td>-0.02 (–0.07, 0.02)</td>
</tr>
<tr>
<td></td>
<td>GLM probit</td>
<td>3.69 (2.46, 5.71)</td>
<td>-0.06 (–0.19, 0.06)</td>
<td>-0.01 (–0.03, 0.01)</td>
</tr>
</tbody>
</table>

Notes: $\tau$ is a weight to residuals. The quantity $\beta$ is called the $\tau$th regression quantile.

After we checked the credible interval of BayesQR and GLMs, we found overall that models with one covariate are better than those with two covariates and BayesQR and GLMs are not much different in terms of statistical inference for the Army and Navy. From the plots about how $\beta$ coefficients in BayesQR change against the quantile estimator, we found out that $\beta$ coefficients have an obvious relationship with the quantile estimator for the Army but the relationship is less obvious for the Navy. Next, we started to analyse the accuracy of prediction of BayesQR and GLMs using the PCC generated in Table 4. PPC is calculated based on the predicted values of all BayesQR ($\tau = 0.1, 0.2, ..., 0.9$). We need to compare PCC to the actual passing rate for the Army and Navy in Table 1. We know that for the Army, the rate of responsible units is 0.884 (88.4%) and for the Navy, the rate is 0.989 (98.9%). If the PCC of a model is close to 0.884 for the Army and close to 0.989 for the Navy, this model is considered to have a high accuracy of prediction. From Table 4, we can see that for the Navy, the PCC of all models is 1, which makes it hard to differentiate models; For the Army, BayesQR with usage covariate and GLM Logit also with usage covariate are closest to 0.884, so they have the best prediction among all models. When it comes to models with age
and models with age and usage for the Army, we can see that PCC of BayesQR care less than or equal to PCC of GLMs. Together with the results of statistical inference, we conclude that for the Army, BayesQR ($\tau = 0.1, 0.3, 0.5, 0.7, 0.9$) and GLMs with one covariate contain only statistically significant parameters and for the Navy, models with one covariate also perform better in statistical inference than models with two, except for BayesQR ($\tau = 0.7, 0.9$). However, BayesQR gives slightly more predictions than GLMs that the difference of PCC is within 1% while GLM Logit provides better prediction than GLM probit.

Figure 3 The rate of change of parameters with respect to quantile for Army

For Army with Age Covariance

For Army with Usage Covariance

For Army with Age and Usage Covariance

Table 4 Percentage correctly classified

<table>
<thead>
<tr>
<th></th>
<th>$A1$</th>
<th>$U1$</th>
<th>$A1$ and $U1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Army</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BayesQR</td>
<td>0.9474</td>
<td>0.9237</td>
<td>0.9263</td>
</tr>
<tr>
<td>GLM logit</td>
<td>0.9526</td>
<td>0.9237</td>
<td>0.9263</td>
</tr>
<tr>
<td>GLM probit</td>
<td>0.9579</td>
<td>0.9289</td>
<td>0.9289</td>
</tr>
<tr>
<td><strong>Navy</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BayesQR</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GLM logit</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GLM probit</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

3.2 Unsupervised learning

This section focuses on using cluster analysis in unsupervised learning to group the Army and Navy data and find which variable(s), age and/or usage, gives the best
Bayesian quantile regression and unsupervised learning methods

The two types of unsupervised learning of cluster analysis used are the k-means clustering and RFs, which are used to find groups that are not explicitly labelled in the data. In this paper, we already know our data groups as passes and failures for the Army and Navy, respectively, so the k-means and RFs are used to correctly assign data in the Army or Navy in different groups.

Figure 4  The rate of change of parameters with respect to quantile for Navy

The k-means clustering starts with \( k \) randomly selected centroids and then it follows two steps: data assignment step and centroid update step. Data assignment step is to assign each data point to its nearest centroid based on the calculation of squared Euclidean distance. After all the data points have their groups, the centroid update step will recompute the \( k \) centroids based on the mean of all data points in each centroid cluster. Then the data assignment step will be repeated according to the new centroids. The k-means clustering will be done until no data points can change the group it belongs to. There are many ways to determine the number \( k \). In this paper, we set \( k = 3 \).

RF algorithms generate a proximity matrix, which gives a rough estimate of distance between samples based on division of tomes. The proximity matrix is converted to a matrix which is later input to the hierarchal clustering algorithm.

The results of the k-means and RFs are given in Table 5. Using Table 1 as the reference, we can see that for the Army, using PF1 (denoted in Table 5) and U1 based on the random forests method gives the perfect classification: 44 units out of 380 are in the failure group and 336 units are in the passing group; for the Navy, neither method gives the perfect classifications. Also, if we only use PF1 or PF2 (denoted in Table 5) with one variable, RFs gives more accurate classification than the k-means. If, on the other hand, we use PF1 or PF2 along with two variables, the k-means groups units better than RFs for both the Army and Navy. Referring to the statistical inference in Tables 2 and 3, better grouping for the k-means using PF1 or PF2 together with two variables suggests that the k-means is superior than random forecasts only when at least one of two covariates is not statistically significant to the response variable. We found
that when most variables are not statistically significant in a regression model, using the parameters in that regression model will hardly give us better classification results under unsupervised learning. As a result, for Army and Navy data, PF1 or PF2 with one variable is the outperforming model under unsupervised learning.

Table 5 The k-means and RFs for Army and Navy

<table>
<thead>
<tr>
<th>Army</th>
<th>k-means</th>
<th>Random forests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PF1 = 0</td>
<td>PF1 = 1</td>
</tr>
<tr>
<td>PF1 and A1</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PF1 and U1</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PF1 and A1, U1</td>
<td>1</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Navy</th>
<th>k-means</th>
<th>Random forests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PF2 = 0</td>
<td>PF2 = 1</td>
</tr>
<tr>
<td>PF2 and A2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>PF2 and U2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>PF2 and A2, U2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: PF1 is the binary response of reliability of units in the Army: failure = 0 and pass = 1; PF2 is the binary response of reliability of units in the Navy: failure = 0 and pass = 1.

4 Conclusions

Stevens and Anderson-Cook (2017a) only used the GLM with probit link function to fit the data. We proposed two different methods to determine the best model to fit the Army and Navy data. The first method is to use BayesQR with odd quantile estimators to fit the data. We compared the statistical inferences among BayesQR, GLM probit, and GLM Logit in Tables 2 and 3 in Section 3. There is not much difference regarding statistical inference between BayesQR and GLMs except that coefficients in models with one covariate have better statistical inference than those in models with two. When checking the statistical inference for BayesQR over quantiles, we can see that BayesQR gives pretty good statistical inference for the Army, while for the Navy, when \( \tau \) equals 0.7, 0.9, \( \beta_1 \) and/or \( \beta_2 \) in BayesQR are not significant at all. However, it is noticeable that BayesQR \( \tau = 0.9 \) contains all statistical significant parameters for the Navy with two covariates when none of the other models do well in the inference with two covariates. We then compared the PCC, which is a type of accuracy measure, among all three types of models and the results are in Table 4. It shows that all three types of models give PCC of 1 for the Navy while for the Army, GLM Logit and BayesQR with usage covariate give slightly better results of PCC (0.9237) compared to the rest because their
PCC is closer to the actual passing rate for the Army (0.884). Overall, BayesQR gives more accurate predictions than GLMs with age and/or usage.

The second method we used is cluster analysis in unsupervised learning, which are the k-means clustering and RFs. Their results are included in Table 5. They do not give very different results regarding grouping for the Army and Navy, but the RFs method gives the best clustering for the Army using PF1 and U1 and for the Navy using PF2 and U2. Together with the statistical inference results in Tables 2 and 3, we found that the k-means classifies units slightly better than RFs only when using variables in the models where at least one of two covariates is not statically significant. Comparing classification results between the Army and Navy, we can see that by selecting corresponding variables from models that are doing well in statistical inference, both the k-means and RFs will improve their classification results.

We showed that BayesQR gives a more accurate prediction overall than GLMs for Army and Navy data in Stevens and Anderson-Cook (2017a). BayesQR with usage covariate gives the best accuracy of prediction. When considering cluster analysis on Army and Navy data, RFs are better than k-means in terms of differentiation between passes and failures in the Army and Navy. Usage is a more important input variable than Age when doing clustering analysis and using one variable is better than using two variables. We suggest that BayesQR could be applied to Army and Navy data if concerned about regression with a quantile other than mean since BayesQR gives good statistical inference as well as the same or more accurate predictions than GLMs. We also suggest using RFs rather than the k-means on Army and Navy data. We have illustrated that RFs gives better clustering than the k-means, especially for the Navy. Since the probability of agreement proposed in Stevens and Anderson-Cook (2017b) can be used to compare the reliability of related populations, which is going to have a prospective application in quality management, we suggest using BayesQR instead of GLM in the probability of agreement to ensure more accurate predictions when facing the data with non-normal errors and outliers or interesting in distributions in various conditional distribution of the outcome variable.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2017R1E1A1A03070747).

References


