Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

Modeling Discrete Survival Time Using Genomic Feature Data

Kyle Ferber and Kellie J. Archer
Department of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA.

Abstract:
Researchers have recently shown that penalized models perform well when applied to high-throughput genomic data. Previous researchers introduced the generalized monotone incremental forward stagewise (GMIFS) method for fitting overparameterized logistic regression models. The GMIFS method was subsequently extended by others for fitting several different logit link ordinal response models to high-throughput genomic data. In this study, we further extended the GMIFS method for ordinal response modeling using a complementary log-log link, which allows one to model discrete survival data. We applied our extension to a publicly available microarray gene expression dataset (GSE53733) with a discrete survival outcome. The dataset included 70 primary glioblastoma samples from patients of the German Glioma Network with long-, intermediate-, and short-term overall survival. We tested the performance of our method by examining the prediction accuracy of the fitted model. The method has been implemented as an addition to the ordinalgmifs package in the R programming environment.

Keywords: classification, ordinal response, gene expression, survival analysis, R

Introduction
Frequently in high-throughput genomic research, we want to fit a statistical model using gene expression data in order to predict a future outcome. This has become a modern challenge for statisticians because there are far more features or variables (p) than samples (n). This is an obstacle in two regards. First, the design matrix will not be full-rank. Thus, there is an infinite number of solutions to the system of equations. Even small perturbations in the data will lead to large fluctuations in the coefficient estimates. Second, given the vast interrelatedness of genes, collinearity is likely to be a problem. Collinear predictors further contribute to the instability of the parameter estimates. Recently, penalization (also referred to as regularization) has stood out as an effective method to combat these two issues. There are several popular penalization methods, but the defining characteristic of them all is that they introduce bias into the parameter estimates in exchange for a reduction in variance. In many cases, penalization improves the model’s predictive accuracy and, relatedly, reduces the mean squared error (MSE) of the parameter estimates. In cases where model parsimony and interpretability are important, the least absolute shrinkage and selection operator (LASSO) penalization method is effective as it shrinks many parameter estimates to be exactly zero. The generalized monotone incremental forward stagewise (GMIFS) method is an algorithm that can be used in logistic regression to produce a monotone LASSO solution. The GMIFS method was subsequently extended by Archer et al for fitting several different logit link ordinal response models to high-throughput genomic data including the cumulative logit, forward continuation ratio (CR), backward CR, stereotype logit, and adjacent category models. Herein we describe the GMIFS algorithm for ordinal...
response modeling using a complementary log-log (cloglog) link, which is useful for discrete survival modeling. Therefore, in the Discrete Survival Analysis Section, we describe the model formulation for modeling a discrete survival outcome. In the GMIIFS Method for Ordinal Response Modeling section, we present the GMIIFS method for the forward CR model using the cloglog link. Next, in the Application section, we discuss the motivating dataset that examined survival in glioblastoma (GBM) patients. The Results section examines the model performance in terms of parsimony, resubstitution error, and cross-validation (CV) error. Finally, in the Conclusion we provide concluding remarks, including limitations of the study.

**Discrete Survival Analysis**

Survival analysis encompasses methods in which the outcome variable is time to event (e.g., time to death, disease relapse, etc.). The particular method used in the analysis will depend on the scale of the survival times collected. Ideally, these will be measured on a continuous scale, but sometimes for a variety of reasons, researchers only collect times on a discrete scale. For instance, for many diseases, it is impossible to record the precise date and time of relapse (i.e., a continuous measurement) because the needed data are often only collected at a physician visit. Thus, we are forced to work with discrete times. Furthermore, discrete times are used when the latent scale of the response times is discrete.

**High dimensional discrete survival data.** Assume there are \( n \) independent subjects \( (i = 1, 2, 3, \ldots, n) \) and \( p \) features per subject, where \( p >> n \). Because this design matrix will be singular, traditional statistical methods (e.g., OLS) are not applicable. The data are often presented as follows:

- Let \( Y_i \) represent the discrete survival time response variable that takes on the values \( (j = 1, 2, \ldots, K) \), where \( K \) is the largest value of \( Y \) observed.
- To facilitate the formulation of the likelihood, we define an \( n \times K \) response matrix as follows:

\[
y_{ij} = \begin{cases} 1 & \text{if } j = y_i \\ 0 & \text{otherwise} \end{cases}
\]

- A \( p \times 1 \) vector of covariates, \( x_i \), is observed for each subject.

**The forward CR model with a complementary log-log link function.** With discrete survival data, we are generally interested in modeling the discrete hazard rate defined as

\[
\pi_j = \pi_j(x) = P(Y_i = j | Y_i \geq j, x_i).
\]

This is also the form of a probability modeled by a forward CR model. Furthermore, if it is reasonable to assume that the data were generated by a continuous-time proportional hazards model, then we use the complementary log-log (cloglog) link function,

\[
\log[- \log(1 - \pi_j)] = \alpha_j + x_i \beta
\]

Here \( \alpha_j \) represents the intercept, or threshold, for the \( j \)th class. Notice that \( \alpha_j \) is the only component of the model that depends on time. Thus, the functions for the \( K \) time points are parallel, which implies we are assuming proportional hazards.

**Likelihood.** We define the likelihood as a product of \( n \) conditionally independent Bernoulli random variables, where \( \pi_j \) is the discrete hazard rate and \( (1 - \pi_j) \) is the conditional complement of \( \pi_j \) given by \( P(Y_i > j | Y_i = j, x) \) for the forward CR model.

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{K} \pi_j^{y_{ij}} (1 - \pi_j)^{1 - y_{ij}}
\]

Now define \( \pi_j = (\pi_{1j}, \pi_{2j}, \ldots, \pi_{nj})^T \). When using the cloglog link, the derivative of the log-likelihood is then given by

\[
\frac{\delta \log L}{\delta \beta_j} = \sum_{j=1}^{K} y_{ij} \exp\left\{-\exp\left(\alpha_j + x_i \beta + \alpha_j + x_i \beta \right) \left[\frac{y_{ij} - \sum_{i=1}^{K} y_{ij} - y_{ij}}{1 - \pi_j} \right] \right\}
\]

We use the generalized monotone incremental forward stage-wise algorithm to solve for the penalized solution:

\[
\hat{\beta} = \arg\max_{\beta} \left\{ \log L(\alpha, \beta; y, X) - \lambda \sum_{j=1}^{p} |\beta_j| \right\}
\]

The tuning parameter, \( \lambda \), controls the amount of shrinkage. As \( \lambda \) increases, the number of parameter estimates that will be shrunk to zero also increases. Using these coefficient estimates and the estimates for the \( \alpha \)'s (described later), we can recursively estimate the probability that subject \( i \) belongs to class \( j \) where

\[
P(Y_i = j | x_i) = \pi_{y_i} * P(Y_i \geq j | x_i)
\]

\[
= \left\{ \begin{array}{ll}
\pi_{y_i} & \text{for } j = 1 \\
\pi_{y_i} * \left[ 1 - \sum_{i=1}^{j-1} P(Y_i = j | x_i) \right] & \text{for } 1 < j \leq K
\end{array} \right.
\]

Subject \( i \) is then classified to the class that corresponds to the maximum class-specific probability.
GMIFS Method for Ordinal Response Modeling

The incremental forward stagewise (IFS) method is an iterative algorithm that produces a penalized solution for a linear regression model.1 The GMIFS method is an extension of IFS capable of fitting overparameterized logistic regression models.3 The GMIFS algorithm was extended by Archer et al (2014) for fitting several different logit link ordinal response models to high-throughput genomic data.4 We updated this method to allow for the use of a complementary log-log link function. The steps of the GMIFS algorithm for ordinal response modeling are as follows:

1. Enlarge the predictor space as \( \tilde{X} = [X : -X] \), where \( X \) represents the standardized predictors.
2. Initialize the \( \alpha_i \)'s to their empirical values. For the forward CR model with a clog-log link, these are initialized as \( \alpha_j = \log \left( - \frac{1}{\sum_{i=1}^{n_j} y_{ij}} \sum_{i=1}^{n_j} y_{ij} \right) \).
3. For step \( s = 0 \), initialize the components of \( \hat{\beta}^{(s)} \) as \( \hat{\beta}_1 = \hat{\beta}_2 = \ldots = \hat{\beta}_p = \hat{\beta}_{p+1} = \ldots = \hat{\beta}_{2p} = 0 \).
4. Find \( m = \arg\min_s - \delta \log L / \delta \beta^{(s)} \) at the current estimate \( \hat{\beta}^{(s)} \).
5. Update \( \hat{\beta}_{m+1}^{(s+1)} = \hat{\beta}_{m+1}^{(s)} + \epsilon \).
6. Estimate the \( \alpha_i \)'s by maximum likelihood, treating \( \hat{\beta}^{(s)} \) (from step 5) as fixed.
7. Repeat steps 4 to 6 until the difference between two successive log-likelihoods is smaller than a pre-specified tolerance, \( \tau \).

The rationale for enlarging the predictor space is that it allows us to avoid taking the second derivative of the log-likelihood. Once the algorithm has converged, we can obtain the penalized solution by \( \hat{\beta}_m = \hat{\beta}_m - \hat{\beta}_{m+1} \) in step 5, \( \epsilon \) is a small incremental value; we used 0.001 in our analysis.

Application

Glioblastoma. Glioblastomas (GBMs) are highly malignant and aggressive tumors that arise from the supportive tissue of the brain. Among all primary brain and central nervous system (CNS) tumors, they are the second most common after meningiomas, which are predominantly benign, and the five-year survival rate for GBM patients is less than 4%.7 Aside from the aggressiveness of the tumors, one possible explanation for the low survival rate is that GBMs are rare in young people; the median age at diagnosis is 64, and the age group with the highest incidence rate is 75–84 year olds.7 Treatment involves surgical removal of as much of the tumor as is safely possible followed by radiotherapy and/or chemotherapy.8 The Cancer Genome Atlas (TCGA) Research Network revealed a subtype of GBM related to the mRNA expression and methylation of a set of genes that affects young adults and has an increased survival rate. Researchers also discovered four molecular subtypes of GBM that have unique responses to treatment and gene mutations that could lead GBMs to become resistant to therapy after a standard chemotherapy treatment.5,10 These findings highlight the importance of genomic research in the study and treatment of GBM.

Data. We downloaded the raw CEL files for GSE53733 from Gene Expression Omnibus.11 The investigators used Affymetrix HG-U133 v2.0 GeneChips to measure gene expression from patients’ tumor samples taken from their initial operation. In the dataset, there were \( n = 70 \) GBM patients, of which 16 had an overall survival (OS) of less than 12 months, 31 patients had an OS between 12 and 36 months, and 23 patients had an OS greater than 36 months.12 The patients’ survival times were reported by the investigator as short-, intermediate-, and long-term. There were \( p = 54,613 \) features per subject in the CEL files after excluding control probe sets. However, after processing the data to remove probe sets with MAS5 present calls in <30% of the subjects,12 31,744 features remained. Furthermore, a 3:5' ratio much different from 1 for the housekeeping gene glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is associated with poor cDNA and cRNA quality.14 Thus, we removed one subject with a 3:5' GAPDH ratio greater than 3, leaving us with 69 subjects. We then used the RMA method to obtain probe set expression summaries for our statistical analysis.15 Afterward, we fit a forward CR model using the cloglog link with \( \epsilon = 0.001 \) and \( \tau = 0.00001 \).

Results

After the GMIFS algorithm converged, we examined two models: (a) the model selected by minimizing the AIC criterion and (b) the model resulting from the convergence of the GMIFS algorithm (Fig. 1). Using the full dataset, the AIC-selected model classified 10 of the 69 patients, while the converged model only misclassified one patient (Tables 1 and 2). However, the AIC-selected model was more parsimonious with 25 non-zero coefficients, while the converged model contained 46 non-zero coefficients. The 25 probe sets that had non-zero coefficient estimates in the AIC-selected model are shown in Table 3. Furthermore, for each model, we examined the sensitivity and specificity for diagnosing short-term survival as well as the sensitivity and specificity for diagnosing short- or intermediate-term survival (Tables 4 and 5). Among the probe sets with non-zero coefficient estimates, the one with the largest absolute coefficient estimate in both models (among probe sets with known gene symbols) was designed to interrogate HD Domain Containing 2 (HDDC2). Long-term survivors had higher HDDC2 expression levels than short- and intermediate-term survivors (Fig. 2). This result agrees with another GBM study that showed that HDDC2 was significantly downregulated in short-term survivors compared to long-term survivors.12 There was also a clear positive relationship between Nucleoside-Triphosphatase, Cancer-Related (NTPCR) expression and survival time (Fig. 3). Interestingly, researchers have shown that NTPCR is overexpressed in neuroblastomas,16 but no study has associated NTPCR with...
GBM. Additionally, one of the probe sets that had a non-zero coefficient estimate was designed to interrogate the gene PDZ and LIM domain 4 (PDLIM4), which has previously been studied in association with gliomas. Researchers examined the expression of the gene at the protein level for patients with gliomas and discovered that the median OS for patients with high levels of the protein (PDLI4) was significantly shorter than for patients with low protein levels. We compared the mean log2 expression levels of PDLIM4 for patients with short-, intermediate-, and long-term OS using Welch’s T-test. Patients with short-term OS had significantly higher expression levels than patients with long-term OS at the Bonferroni adjusted α = 0.0017 significance level (P = 0.0019), and patients with intermediate-term OS also had significantly higher expression levels than patients with long-term OS (P < 0.0001). The difference in mean expression levels between those with short-term OS and those with intermediate-term OS was not significant. Thus, it appears that both the gene expression levels and the protein levels of the gene are lower for patients who survive longer.

A common critique of a model fitted from high-dimensional data is that the final model, even if selected by minimizing AIC, is not parsimonious. In this example, critics may say that given a sample size of 69 subjects, including 25 coefficients in the model is overfitting, and that the model performance is likely a result of chance. In response, we fit two additional models whose performances will be a result of chance alone. First, we fit a model with the same gene-expression data used in our example, but we randomly permuted the response vector. Next, we fit a model using our original response vector, but instead of using the gene expression data, we used a design matrix filled with 31,744 × 69 = 2,190,336 random variables generated from a Gaussian distribution with a mean and standard deviation equal to the corresponding sample statistics of the gene expression data. If we exclude regions of underfitting and overfitting, the model fit with the gene expression data and the original response vector had better performance than the other two models whose performances are a result of chance rather than a relationship between the features and the response (Fig. 4).

We also performed N-fold (or leave-one-out) CV to assess the generalizability of our models (where N = 69). Both the AIC-selected model and the converged model had an N-fold CV error rate of about 44.9% (Tables 6 and 7). Thus, it appears that the AIC-selected model and the model that satisfied the GMIFS convergence criterion predict discrete survival time equally well. We chose the AIC-selected model as our final model as it is more parsimonious and therefore more interpretable.

### Conclusion

GBM is a particularly dangerous tumor with a low survival rate. A specific and accurate prognosis would be very useful to both the patient and the oncologist. Thus, we were interested in predicting survival time based on a patient’s genomic feature data. We used discrete times because the investigators of this particular GBM study reported discrete times. Another case when discrete survival times would be used is when the outcome of interest (e.g., disease relapse) can only be assessed at physician visits. The GMIFS algorithm is an effective method for building a classifier for an ordinal response outcome given a high-dimensional covariate space. In this case, we fit a forward CR model with a complementary log-log link function to model discrete survival time. The model resulting from the convergence of the algorithm had only a 1.4% resubstitution error. Using N-fold CV, the model had a
Table 3. Probe sets with non-zero coefficient estimates in the AIC and converged models.

<table>
<thead>
<tr>
<th>PROBE SET</th>
<th>ENTREZ ID</th>
<th>GENE SYMBOL</th>
<th>CHROMOSOME</th>
<th>$\hat{\beta}_{AIC}$</th>
<th>$\hat{\beta}_{CONVERGED}$</th>
<th>CANCER ASSOCIATIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>203260_at</td>
<td>51020</td>
<td>HDCC2</td>
<td>6</td>
<td>0.309</td>
<td>0.309</td>
<td>Glioblastoma</td>
</tr>
<tr>
<td>1557883_a_at</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>0.268</td>
<td>0.268</td>
<td></td>
</tr>
<tr>
<td>206565_x_at</td>
<td>11039</td>
<td>SMA4</td>
<td>5</td>
<td>0.267</td>
<td>0.267</td>
<td></td>
</tr>
<tr>
<td>1558723_at</td>
<td>284014</td>
<td>LOC284014</td>
<td>17</td>
<td>0.178</td>
<td>0.226</td>
<td></td>
</tr>
<tr>
<td>202447_at</td>
<td>1666</td>
<td>DECR1</td>
<td>8</td>
<td>0.163</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td>226813_at</td>
<td>84284</td>
<td>NTPCR</td>
<td>1</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>209078_s_at</td>
<td>25828</td>
<td>TXN2</td>
<td>22</td>
<td>0.133</td>
<td>0.133</td>
<td></td>
</tr>
<tr>
<td>230581_at</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>0.203</td>
<td>0.203</td>
<td></td>
</tr>
<tr>
<td>215962_at</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>0.267</td>
<td>0.267</td>
<td></td>
</tr>
<tr>
<td>155100_s_at</td>
<td>25831</td>
<td>HECTD1</td>
<td>14</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>242333_at</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>0.267</td>
<td>0.267</td>
<td></td>
</tr>
<tr>
<td>206697_s_at</td>
<td>3240</td>
<td>HP</td>
<td>16</td>
<td>0.267</td>
<td>0.267</td>
<td></td>
</tr>
<tr>
<td>222992_s_at</td>
<td>4715</td>
<td>NDUFB9</td>
<td>8</td>
<td>0.203</td>
<td>0.203</td>
<td></td>
</tr>
<tr>
<td>219221_at</td>
<td>253461</td>
<td>ZBTB38</td>
<td>3</td>
<td>0.163</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td>230353_at</td>
<td>284112</td>
<td>LOC284112</td>
<td>17</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>243957_at</td>
<td>400464</td>
<td>LOC400464</td>
<td>15</td>
<td>0.163</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td>231773_at</td>
<td>9068</td>
<td>ANGPTL1</td>
<td>1</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>211564_s_at</td>
<td>8572</td>
<td>PDLIM4</td>
<td>5</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>218669_at</td>
<td>57826</td>
<td>RAP2C</td>
<td>X</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>1561759_at</td>
<td>645513</td>
<td>LOC645513</td>
<td>4</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>1559283_a_at</td>
<td>285888</td>
<td>CNPY1</td>
<td>7</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>221900_at</td>
<td>1296</td>
<td>COL8A2</td>
<td>1</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>203184_at</td>
<td>2201</td>
<td>FB2N2</td>
<td>5 9</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>234547_at</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>&lt;NA&gt;</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
<tr>
<td>229146_at</td>
<td>136895</td>
<td>C7orf31</td>
<td>7</td>
<td>0.194</td>
<td>0.194</td>
<td></td>
</tr>
</tbody>
</table>

Table 4. AIC-selected model sensitivity and specificity for predicting short-term survival and for predicting short- or intermediate-term survival.

<table>
<thead>
<tr>
<th>OUTCOME</th>
<th>SENSITIVITY</th>
<th>SPECIFICITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term survival</td>
<td>63</td>
<td>100</td>
</tr>
<tr>
<td>Short- or intermediate-term survival</td>
<td>100</td>
<td>82</td>
</tr>
</tbody>
</table>

Table 5. Converged model sensitivity and specificity for predicting short-term survival and for predicting short- or intermediate-term survival.

<table>
<thead>
<tr>
<th>OUTCOME</th>
<th>SENSITIVITY</th>
<th>SPECIFICITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term survival</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Short- or intermediate-term survival</td>
<td>100</td>
<td>95</td>
</tr>
</tbody>
</table>
they could have been effective unpenalized predictors as well (despite the fact that these two variables were not significantly different across the three classes in this study). Additionally, a specific month of death would have provided more information than a range of months. However, for each discrete time value, we would need enough subjects with that response to fit a reliable model. As the price of microarray experiments decreases, we will have greater access to datasets with a larger number of subjects, making this a reasonable expectation for the future.

Author Contributions
Conceived and designed the methods: KF, KJA. Analyzed the data: KF. Wrote the first draft of the manuscript: KF. Contributed to the writing of the manuscript: KJA. Agree with manuscript results and conclusions: KF, KJA. Jointly developed the structure and arguments for the paper: KF, KJA. Made critical revisions and approved final version: KF, KJA. Both authors reviewed and approved of the final manuscript.

REFERENCES


