Regularized Cox Regression on Health Evaluation and Linkage to Primary Care (HELP) Dataset

Ethan Ashby

5/07/2021

Introduction

Finding ways to practially link medical care and substance abuse treatment is a desirable goal of public health. Indeed, patients with substance abuse problems are common in general medicine practice across all demographics, and primary care physicians can play a powerful role in helping patients accept treatment (Weaver et al. 1999). Therefore, finding ways to involve primary care physicians in the rehabilition and chemical dependency treatment process could potentially lead to better adherance to treatment protocols and improved recovery.

The Health Evaluation and Linkage to Primary Care (HELP) study was a clinical trial in adult patients without a primary care physician who were undergoing in-patient detoxification treatment for alcohol, heroine, and cocaine addiction in the Boston area. Patients were randomized to recieve a multidisciplinary assessment and a brief motivational intervention or usual care, with the primary endpoint being whether the patient attended an appointment with a primary care physician within 12 months.

The inclusion/exclusion criteria for the study is as follows. Patients enrolled in the study were Spanish/English speaking adults that had reported alcohol, heroin, or cocaine as their primary or secondary drug of choice. Patients needed to reside in the proximity of the primary care clinic to which they would be referred or were homeless. Patients were excluded that had established primary care relationships, significant dementia, plans to leave the Boston area that would lead to loss of follow up, failed to provide contact information, or were pregnant.

Subjects were interviewed at baseline during their detoxification stay and follow-up interviews were undertaken every 6 months for 2 years. A variety of covariates and outcomes were measured per individual.

The Elastic Net

The elastic net (Zou & Hastie 2004) is a regularization and variable selection strategy for regression. When fitting a regression model to real world data, both generalization performance (i.e., accurate predictions on future data) and model interpretability (i.e., more sparse or parsimonious models) are highly desirable. Good performance, variable selection, and model parsimony is especially important in high-dimensional scenarios, where the number of predictors is large.

Regularization/penalization methods have been proposed to improve the performance and interpretibility of regression models. One method is ridge regression, which achieves improved generalization performance by imposing a bound on the L_2 norm of the regression coefficients (circular constraint region shown in Figure 1), decreasing the variance of the model. However, ridge regression does not produce a parsimonius model, for it retains all the predictors in the model. Another method is lasso regression, which imposes an L_1 penalty on the regression coefficients. Do to the sharpness of the L_1 constraint region (diamond-shaped constraint region shown in Figure 1), model coefficients are both shrunk and forced to 0, leading to improved generalization and model sparsity. While the lasso is a popular penalized regression method, it is limited in the case where the number of predictors (p) outnumber the number of observations (n). When $p > n$, lasso selects at most n variables. In addition, in the case of highly correlated variables, lasso tends to randomly select only one variable from the group and is empirically outperformed by ridge regression.

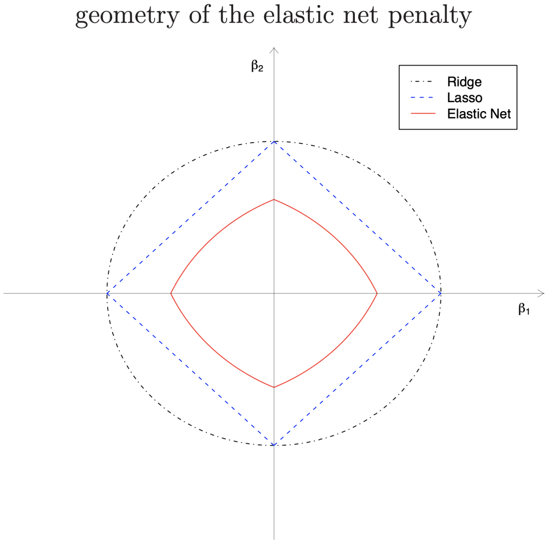

Figure 1: Figure borrowed from Zou and Hastie, 2004. Figure illustrates the constraint space of LASSO (blue diamond), Ridge (black circle), and Elastic Net (red deformed diamond). The geometry of the Elastic net constraint encourages variable shrinkage and model sparsity.

To overcome the limitations of ridge regression and lasso, Zou and Hastie proposed the *elastic net*. The elastic net is essentially a weighted average of the ridge and lasso penalties, thereby combining the strengths and accounting for the weaknesses in each approach. The elastic net simultaneously performs automatic feature selection (inducing model sparsity) and variable shrinkage (reducing variance and improving generalization performance). The deformed, diamond-like geometry of the elastic net constraint region (Figure 1), imposes variable shrinkage and selection by regularizing parameter estimates to the edges and corners of the constraint region. Importantly, the elastic net penalty can be tuned to weight the L_1 and L_2 penalties differently, which would sharpen or smooth the constraint region respectively. Lastly, the elastic net is capable of identifying groups of correlated variables, creating a flexible feature selection approach like a “stretchable fishing net that retains all the big fish”.

Connecting the Elastic Net to the Cox Model

The Cox Proportional Hazards (Cox PH) model assumes a semi-parametric form for the hazard: $h_i(t) = h_0(t)e^{x_i^T \beta}$, where $h_i(t)$ denotes the hazard for patient i at time t, $h_0(t)$ is the baseline hazard at time t, and β is a vector of predictors (length p).

Typically, β is estimated in the Cox PH model by maximizing the *partial likelihood*:

$$L(\beta) = \prod_{i=1}^{m} \frac{e^{x_{j(i)}^T \beta}}{\sum_{j \in R_i} e^{x_{j(i)}^T \beta}}$$

where $(i \in \{1, \dots, m\})$ denote the observed event times, R_i is the set of indices, j, with $y_j \geq t_i$. Essentially, the partial likelihood is the product over the event times (i) of conditional probabilities of witnessing the observed failure given one failure occured among all susceptible individuals (R_i) at time t_i. Note that the partial likelihood is **not** a function of the baseline hazard, and evaluation of the partial likelihood only considers the **order/rank** of event times and ignores information between events.

Simon et al. (2011) note that maximizing the partial likelihood is equivalent to mazimizing a scaled version of the log partial likelihood (since log is a monotonic transformation),

$$\frac{2}{n}\ell(\beta) = \frac{2}{n}\left[\sum_{i=1}^{m} x_{j(i)}^T\beta - \log\left(\sum_{j\in R_i} e^{x_j^T\beta}\right)\right]$$

Next, consider the **elastic net penalty** on the vector of β:

$$\lambda P_\alpha(\beta) = \lambda\left(\alpha\sum_{k=1}^{p}|\beta_i| + \frac{1}{2}(1-\alpha)\sum_{k=1}^{p}\beta_i^2)\right)$$

where $\alpha \in [0, 1]$ denotes the relative weights of the L_1 and L_2 penalties ($\alpha = 1$ returns the lasso, $\alpha = 0$ returns ridge regession) and λ denotes the regularization strength (a tunable hyperparameter. Incorporating the elastic net penalty into the partial likelihood yields the objective function from which penalized estimates of the β values can be obtained:

$$\hat{\beta} = \text{argmax}_\beta\left[\frac{2}{n}\left(\sum_{i=1}^{m} x_{j(i)}^T\beta - \log\left(\sum_{j\in R_i} e^{x_j^T\beta}\right)\right) - \lambda P_\alpha(\beta)\right]$$

I implemented the regularized Cox model fitting using the `glmnet` package (Friedman et al. 2010) (Simon et al. 2011) as described in the Coxnet vignette (Tay et al. 2021).

Methods

This section describes the procedures used to clean the data and tune and fit the Cox PH model with elastic net penalty.

Data cleaning

The first order of business was data cleaning. I took the full dataset consisting of with 347 patients/observations (rows) and 788 features (columns) and filtered out features with greater than 10% missing values. This resulted in a filtered dataset with 347 observations and 497 features. The retained features are shown in Figure 2 below. Features with less than 10% missing values (shaded dark and to the left of the dotted red line) were retained, while features with more tha 10% missing values (shaded grey and to the right of the dotted red line) were excluded.

Next, I needed to change how variables were encoded. I looped through all the features in this dataframe. If a feature was categorical (had a minimum value of 0 or 1, a maximum value less than 16, and no decimal values), I encoded the features as a factor, being sure to add a `NA` factor level if there were missing values in the column. If a feature was continous, I imputed `NA` values by taking the mean of all the variables in the column and rounding to the nearest integer. The resulting dataframe `data_filter` was subjected to downstream model fitting.

I also exluded the `e14e` binary feature (Have you had biofeedback in the last 6 months?) from the model, because every respondent replied “No”, and a model cannot be fit to a factor with only 1 level.

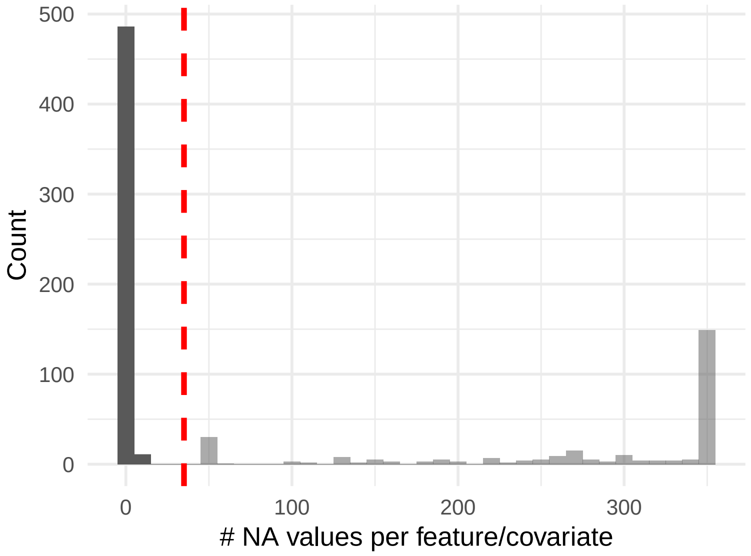

Figure 2: Histogram of the number of NA values per feature. The dotted red line denotes the 10 percent threshold of missing values. Features to the left of the threshold (shaded dark), were included in the analysis while features to the right of the threshold (shaded grey) were excluded.

Tuning and fitting the elastic net model

I needed to fit an cox model with elastic net penalty (coxnet model) to the observations in the filtered dataset. In order to do so, I needed to choose an α parameter, designating the weighting of the L_1 and L_2 penalties in the elastic net penalty. $\alpha = 0$ corresponds to a LASSO penalty (only L_1 penalty is considered) while $\alpha = 1$ corresponds to Ridge penalty (only L_2 is considered). In order to evaluate the performance of the coxnet models while varying α and λ (the regularization strength), I performed 10-fold cross validation over a logarithmically spaced grid of λ values while varying the α parameter over $\{0, 0.25, 0.5, 0.75, 1\}$ using the `cv.glmnet` function in the package `glmnet`, while specifying the `family="cox"`. Parameters were tuned to those that minimized by CV deviance by specifying `type.measure="deviance"`.

Shown in Figures 3 and 4 below, choice of α largely had no effect on the deviance profiles and C statitic profiles produced in the cross validation experiment. For all $\alpha > 0$, the curves achieved similar deviances and C-statistics in the CV datasets. The one exception was $\alpha = 0$, where the C statistic was much lower cross most $\log(\lambda)$ values surveyed. Thus, any $\alpha > 0$ would yield comparable performance in CV, meaning any variant of elastic net including ridge regression but excluding LASSO would suffice for building a Cox model that should discriminate well on future data. I elected to use the standard elastic net weighting of the LASSO and Ridge penalties ($\alpha = 0.5$) for my model.

To select the regularization strength λ, I performed 10 repeated CV experiments (with $\alpha = 0.5$) and recorded the optimal λ values that produced the smallest deviance. I ran repeated CV experiments to avoid overfitting the particular CV split structure and reduce the variance of my λ estimate. I selected the median optimal regularization strength, $\lambda = 0.1496369$, as the hyperparameter value to fit my final cox model with elastic net penalty.

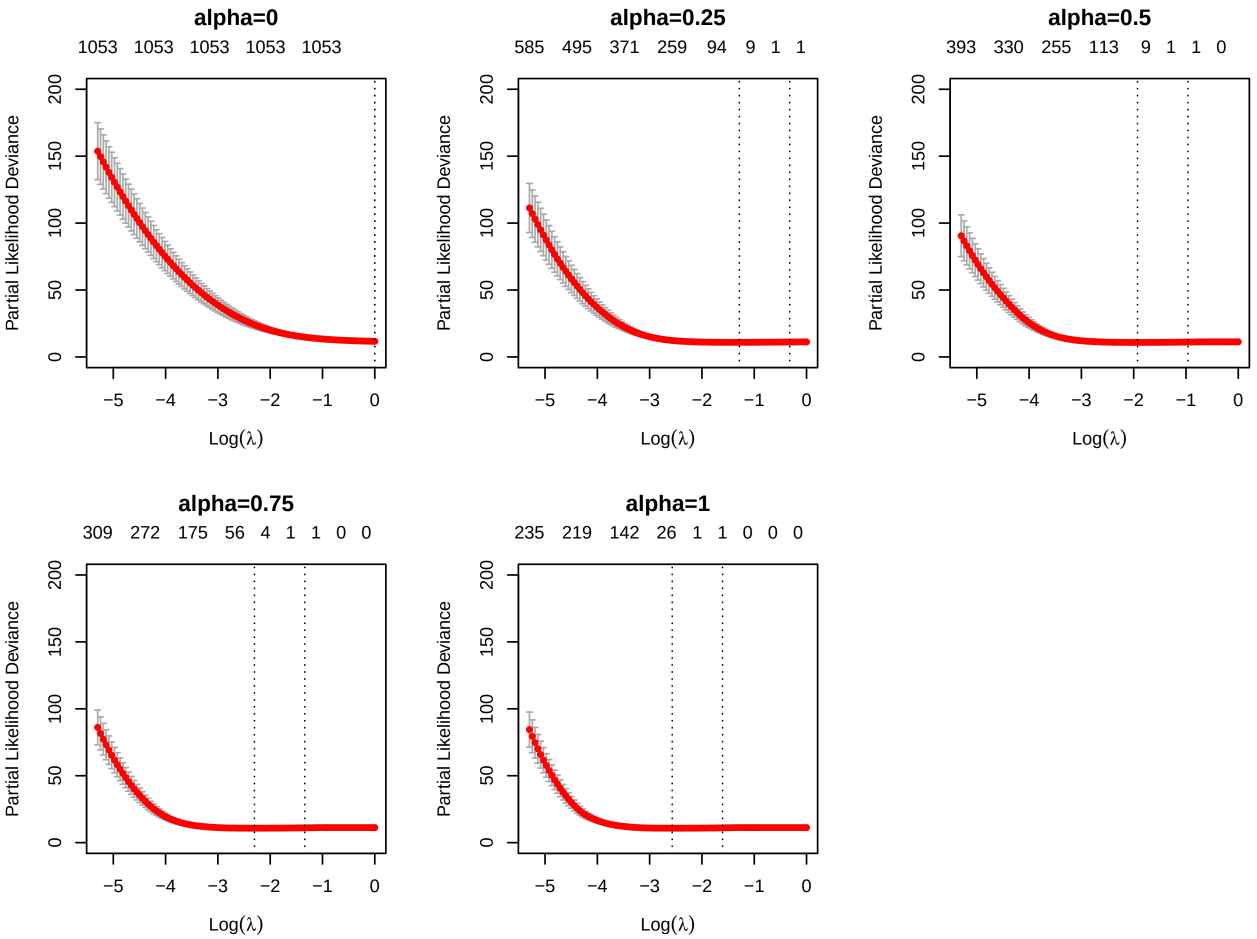

Figure 3: CV deviances for different weights L1/L2 penalization terms (alpha) and regularization strengths (lambda). Minimal differences in deviance profiles between different alphas were observed.

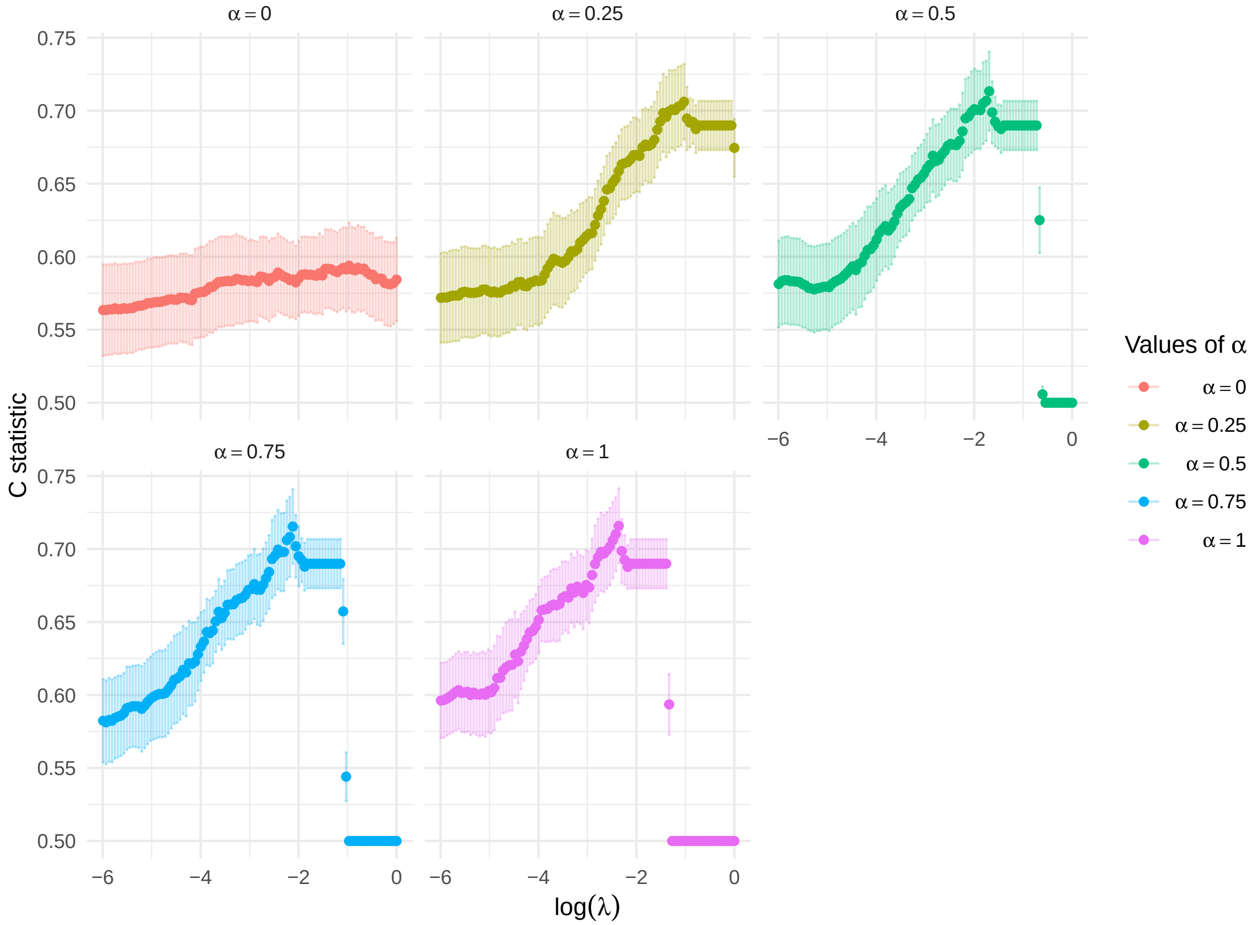

Figure 4: Illustrating the 10-fold CV C-statistics associated with different regularization strengths (lambda) and different elastic net penalty weights (alpha). Lasso (alpha=0) demonstrated the lowest C-statistic for all models assayed, while any alpha>0 produced similar profiles of CV C-statistic.

Table 1: Variables that were retained in the Cox model after application of the elastic net penalty. Each term is paired with its fitted beta estimate and its associated Hazard Ratio (HR).

term	estimate	HR
b9c3	0.1160157	1.1230135
d43	0.0739853	1.0767909
e2a1	-0.0126033	0.9874758
m241	0.0034333	1.0034392
m441	0.0453162	1.0463587
n2d7	0.0507930	1.0521051
n2n7	0.1017245	1.1070784
q1a1	-0.0269062	0.9734525
q162	0.1679884	1.1829229
q202	-0.0960402	0.9084275
group1	0.8945788	2.4463052
hs_grad1	-0.0639636	0.9380391
epi_sum1	0.0860250	1.0898335
h3_prb1	-0.1231054	0.8841705
phys5	-0.0347928	0.9658055
inter10	0.0118460	1.0119165
sr2	-0.2558515	0.7742569

Results

The elastic-net-regularized cox model produced a relatively parsimonious final model with 20 predictors: `a12b` (categorical variable denoting Hollingshead category or professional status), `b9c` (a categorical variable denoting whether someone felt nothing could cheer them in last 4 weeks), `d2` (binary variable denoting whether patient takes prescription medication regularly for physical problem), `d4` (a categorical variable denoting how bothered an indivdual felt by a medical problem in the last 30 days),`e2a` (binary variable denoting whether enrolled in a alcohol or drug detox program over last 6 months), `m24` (binary variable denoting whether physical health was harmed by alcohol/drug), `m44` (binary variable denoting whether they were suspended/fired/left job or school due to substance abuse), `n2d` (binary variable denoting whether confiding in family members made their family members uncomfortable), `n2n` (binary variable denoting whether wished family were much different), `q1a` (binary variable denoting whether they ever injected drugs), `q16` (categorical variable denoting how frequently the individual was paid for sex over the last 6 months), `q20` (categorical variable denoting whether the individual believed that safer sex was always their responsibility), `group` (a binary variable specifying assignment to either the treatment group of control group), `birthplc` (binary variable denoting whether born in US or foreign nation), `hs_grad` (binary variable denoting whether the individual graduated high school), `epi_sum` (categorical variable denoting sum of episodic events), `h3_prb` (a binary variable specifying whether an individual has a substance problem related to heroin), `phys` (categorical variable denoting physical consequences of drug use), `inter` (categorical variable denoting interpersonal consequences of drug use), and `sr` (categorical variable denoting social responsibility consequences of drug use). The regularized estimates and hazard ratios are shown in Table 1.

The alluvial plot in Figure 5 above illustrates the importance of the `group` and `sr` variable in determining the time-to-event variable `dayslink`. The first axis and color of each line denotes the `group` assignment to the control (0) or treatment (1) arm of the clinical trial. The second axis denotes the `sr` variable, denoting the percieved social responsibility harms associated with the individual's addiction (1-7). The third axis denotes the time-to-event variable `dayslink` broken into discrete 50 day bins (<50, <100, <150, <200, <250, <300, and <350). Right censored observations are binned into 300+, 350+, and 400+ categories. A key insight from this graphic is that the majority of individuals from the treatment (`group=1`) group show small time-to-events, with a good fraction of individuals connecting with a primary physician after less than 50 days after treatment. However, a miority of people in the control (`group=0`) group ever connect with a primary

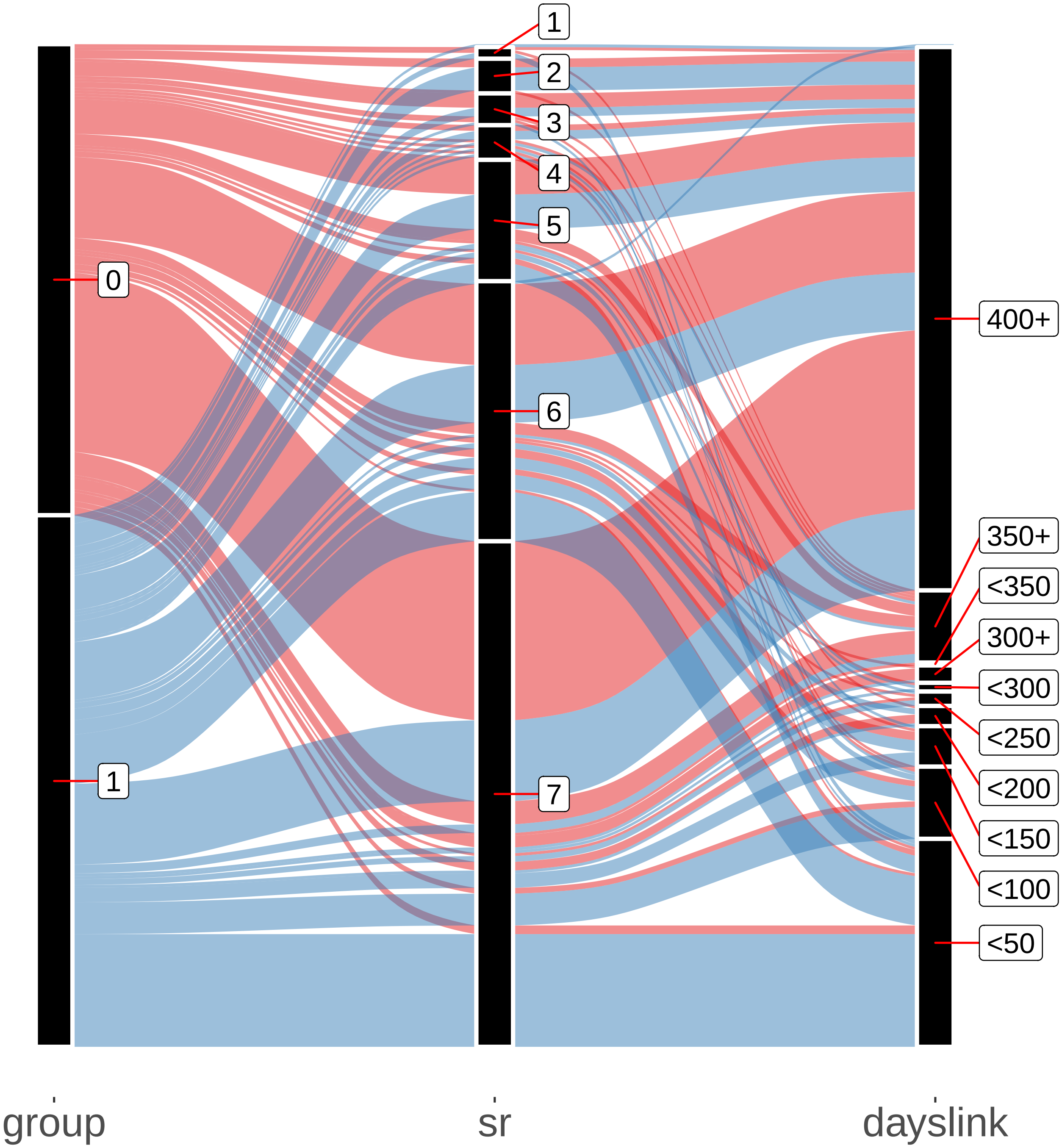

Figure 5: Alluvial plot of 2 most important features (group and sr) and outcome categories. Red lines denote patients in the control group while blue lines denote patients in the treatment group. The first axis denotes group membership. The second axis denotes the percieved consequences of alcohol on social responsibility. The last axis denotes the outcome of the individual into complete or censored time-to-event categories.

physician, as the majority of those individuals were right censored observations that hadn't connected with primary care after 400 days. And individuals with high percieved social responsibility harms (6 & 7) from both treatment groups tend to deflect to lower event times, while individuals with low responsibility harms (1 & 2) tend to have long event times or censored observations.

Discussion

The Health Evaluation and Linkage to Primary Care (HELP) study was a clinical trial that sought to test interventions and identify factors related to linking drug detoxification programs to primary medical care. Patients were randomized to recieve a multidisciplinary assessment and a brief motivational intervention or usual care, with the primary endpoint being whether the patient attended an appointment with a primary care physician within 12 months. The response variable (time until connecting with a primary care physician) represented time-to-event data, making survival analysis methods (like Cox PH regression) applicable to the problem at hand.

The HELP study measured a variety of covariates and outcomes per individual. In fact, the number of predictors (788) in the HELP study exceeded the number of observations (347). Fitting a Cox model to all covariates in the study would yield a very complicated model, overfit to the training data. Highly complex and overfit models suffer from poor interpretability and generalizability to future datasets.

To ensure model interpretability and good generalization performance, I imposed an elastic net penalty on the partial likelihood used to estimate the effect sizes of the covariates: $\hat{\beta}$. The elastic net penalty is a weighted average of the lasso (L_1) and ridge (L_2) penalties, which enjoys the advantages of model sparsity and retaining all the important variables regardless of the correlation structure among the covariates.

I employed a cross validation approaches to tune the α and λ hyperparameters, specifying the weight of the L_1/L_2 penalties and specifying the regularization strength respectively. When I fit a Cox PH model with the tuned hyperparameters, I obtained a parsimonious model with 20 covariates, representing a nearly 25-fold reduction in model complexity relative to the full model.

One key limitation of this study was that fitting regularized Cox models through the `glmnet` package does not support diagnostic plots to test the proportional hazards assumption. To indirectly address this question, I fit a full (unregularized) Cox model and used the `cox.zph` function to test for significant deviations from the proportional hazards assumptions. No variable showed a significant departure from PH at the $p<0.05$ level (data not shown). `group` was the second most significant variable (p-val: 0.129), but log-log diagnostic plots of a model built using the `group` variable (Figure 6) showed no major departure from the proportional hazards assumption. Since `group` evidently did not violate the PH assumption, and `group` was one of the most significant variables, I conclude that no features included in the final model (`final_fit`) significantly deviated from the PH assumption.

In summary, using a cox model with an elastic net penalization term, I identified a parsimonious survival model for drug rehab patients connecting with primary physicians in the Health Evaluation and Linkage to Primary Care study. The final model, `final_fit`, contains 20 features, displays good fit to the data (illustrated by low CV deviance), and good discrimination ability (illustrated by high CV C statistic).

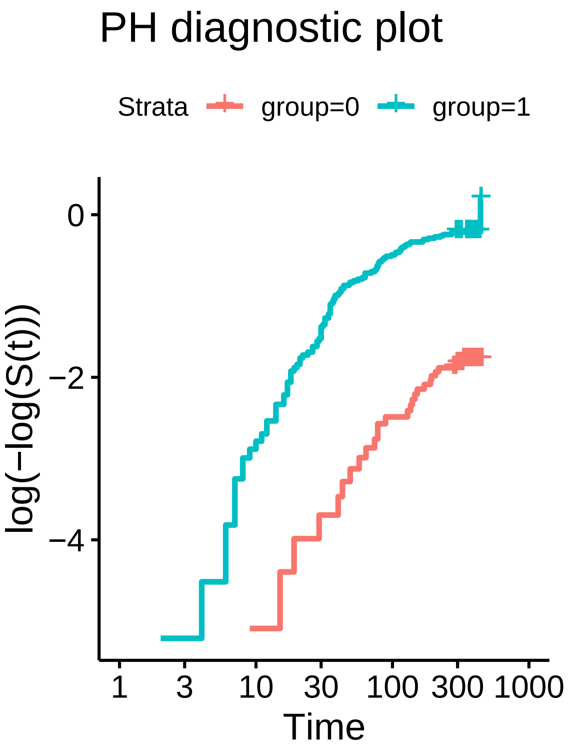

Figure 6: log-log diagnostic plot of unregularized Cox model fit to the group variable demonstrates that PH assumption likely holds.

References

1. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J. Stat. Softw. 39, (2011).

2. Tay, K., Simon, N., Friedman, J., Hastie, T., Tibshirani, R., & Narasimhan, B. Regularized Cox Regression. (2021).

3. Weaver, M. F., Jarvis, M. A. E. & Schnoll, S. H. Role of the Primary Care Physician in Problems of Substance Abuse. JAMA Intern. Med. 159, 913–924 (1999).

4. Zou, H. & Hastie, T. Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. 67, 301–320 (2004).

5. Friedman, J., Hastie, T., Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33, 1-22 (2010).