Jaehwan Han

Survival Analysis Project
Study of comparing treatment effects on reducing drug abuse constructing a Cox PH model

ABSTRACT

The Cox Proportional-hazards regression model has been widely used in analyzing survival or failure time data. The model can incorporate explanatory covariates, including a mix of numeric and categorical variables. Our main objectives are to compare treatment effects on reducing drug abuse adjusting for all covariates. We assumed that the effect of variables on hazard ratio is constant over time and there are several approaches to assess the model assumption. If this assumption is violated then, either piecewise or stratified remedy must be made to account for disproportionality. Besides checking the model assumption, diagnostics methods for assessing goodness-of-fit for Cox's regression model are implemented by means of Cox-Snell, Martingale residual and Score residual test.

Method

In this project, we are interested to study how treatment affects return to drug use, after adjusting for few important risk factors, using survival analysis.

We started our analysis by going through our data in a univariate way, checking for the distributions of each individual variable and looking for any obvious errors in the data or missing values. Then, we conducted univariate analysis on every variable of interest, using Kaplan-Meier estimates for categorical variables and Cox Proportional Hazard (Cox PH) Models for continuous variables. We decided to only include variables with p-values of Log-rank tests and Wald's tests less than or equal to 0.2, for further feature selection. We adopted forward selection, backward selection and stepwise selection criteria for variable selection to decide on a consensus model of covariates of interest. Afterwards, a Cox PH model was built using those selected variates.

We used Martingale residual plots with LOESS smooth curves to determine the best functional forms of the continuous covariates. To assess the PH assumption of our model, we used the log interaction approach as well as Standardized score residual plots. We plotted Cox-Snell residuals to assess the overall model fitting. Finally, we used Deviance residuals and DFBETAS to detect points of high influence.

Result

The summary statistics of each individual variable did not report any obvious data errors but did show that a few of the observations were missing, the highest count being 33 for the Beck depression inventory scores. As the number was not too many, considering the large sample size we had (n=628), we proceeded with our analyses without dropping any observation.

After univariate analysis, we included all the covariates, present in our dataset, in the preliminary model because all of them met our criterion of p-value≤0.2 (Table 1). Model selection using all three methods resulted in the same set of covariates (treatment, age, number of prior drug treatments and IV drug use history) and thus, we went forward with model diagnostics using these set of covariates in our model.

The LOESS smooth curves in our Martingale residual plots of age, the number of prior drug treatments and IV drug use history were quite linear (Figure 1a, b), suggesting that we didn't require any transformations of these continuous covariates. The log interaction approach for checking PH assumption did not show any violations of the PH assumption, but the Standardized score residual plots showed that treatment covariate violated the PH assumption (p-value = 0.0180) (Figure 1d). As treatment was our primary risk factor of interest, the remedy for this was to fit a piecewise model, but before doing that, we went ahead with the rest of the diagnostics.

The Cox-Snell residuals followed the diagonal line quite well, suggesting that the overall fit of our model was pretty good (Figure 1c). Examining the data for influential points using the Deviance residual plot did not show us any observation very far away from the rest of the data which can be of a major concern. The DFBETA plots also did not reveal any observation with a high influence on our estimation of our betas, the highest value being even less than 0.02. Thus, we concluded that there were no major influential points need to be addressed in our data.

Finally, as all the other diagnostics were done and everything was okay, we proceeded with fitting the final piecewise model. We decided to use the "change point" as the median time to return to drug use, which turned out to be 166 days. We dichotomized the time using this change point and fitted our final piecewise Cox PH model (Table 2).

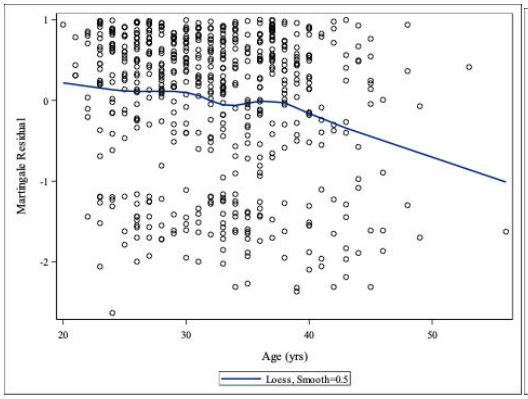
The final piecewise model suggests that till 166 days, the people in the six-month treatment group has 0.656 times the hazard of returning to drug use than the people in the three month treatment group (95% CI 0.520-0.829, p=0.0004), holding other risk factors constant. But, after 166 days, the people in the six-month treatment group had 1.082 times the hazard of returning to drug use than the people in the three-month treatment group (95% CI 0.810-1.445, p=0.5935), holding other risk factors constant. Thus, being in the six-month treatment group is significantly protective till 166 days, while it is more hazardous after that, though not significant.

The final model suggested that age and six-month treatment in early time were protective to drug use. Increase in age also turns out to be protective, with every 10yr increase in age, the hazard of returning to drug use decreases by 28% (HR = 0.7217, 95% CI 0.6170-0.8442, p<0.0001), holding other covariates constant. On the contrary, when holding other risk factors constant, each unit increase in the number of prior drug treatments resulted in a hazard ratio of 1.030 (95% CI 1.014-1.047, p=0.0002). Compared to the subjects who never had a history of IV drug use, people who have a previous history of IV drug use has a 24.5% increased risk of returning to drug use (HR = 1.245, 95% CI 0.957-1.619, p=0.1021), and people who have a recent history of IV drug use has 51% increased risk of returning to drug use (HR = 1.510, 95% CI 1.211-1.882, p=0.0002), holding other risk factors constant.

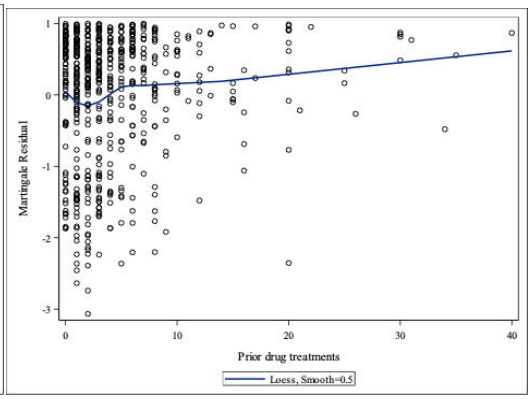
Figure & Table

Table 1: Univariate analysis on candidate variables.

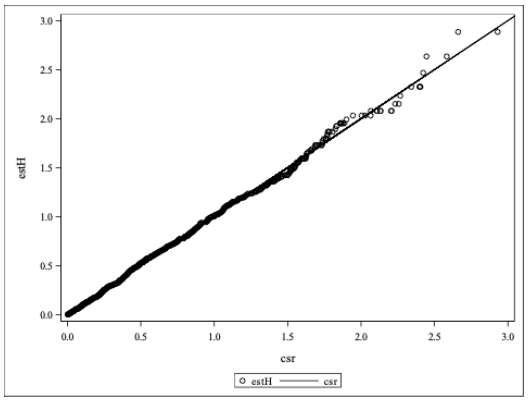
Univariate Analysis				
Variable	**Test used**	**Test statistic**	**P-value**	**Label**
Treat	Log-rank	6.7979	0.0091 *	Treat
Race	Log-rank	7.2836	0.0070 *	Race
Site	Log-rank	2.3658	0.1240	Treatment site
hercoc	Log-rank	7.9536	0.0470	Heroin/cocaine use
ivhx	Log-rank	14.7196	0.0006 *	IV drug use history
Age	Wald	3.2022	0.0735	Age
ndrugtx	Wald	15.3470	<0.0001 *	Number of prior drug treatments
Beck	Wald	5.4045	0.0201	Beck Depression Inventory score



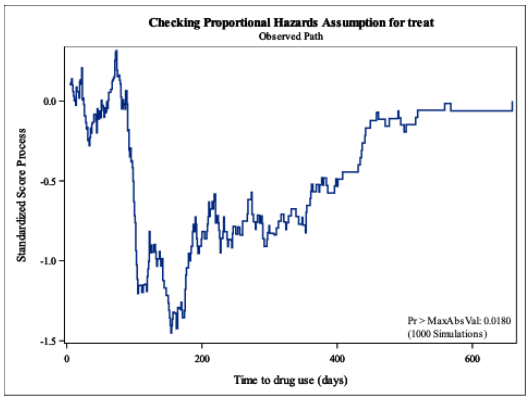
(a). Martingale residual vs. age



(b). Martingale residual vs. number of prior drug treatment



(c). Cox-Snell residual



(d) Standardized score residual vs treat

Figure 1: Types of diagnosis. Graphs of the Martingale residuals and their LOESS smooth curves for the covariates: (a) age and (b) number of prior drug treatment. (c) Cumulative hazard plot of the Cox-Snell residual of the proportional hazards Cox regression model in Table 2. 45 degree line through the origin is drawn for reference. (d) Standardized score process for treat covariate to check satisfy PH assumption.

Table 2: Results of the multivariable proportional hazards Cox regression model containing the main and piecewise effect.

Analysis of Maximum Likelihood Estimates										
Parameter		**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**95% Hazard Ratio Confidence Limits**		**Label**
age		1	-0.03261	0.00800	16.6221	<.0001	0.968	0.953	0.983	Age (yrs)
ndrugtx		1	0.02986	0.00803	13.8231	0.0002	1.030	1.014	1.047	Prior drug treatments
ivhx	**2**	1	0.21925	0.13411	2.6727	0.1021	1.245	0.957	1.619	IV drug use 2
ivhx	**3**	1	0.41203	0.11245	13.4259	0.0002	1.510	1.211	1.882	IV drug use 3
treat_bef		1	-0.42133	0.11907	12.5215	0.0004	0.656	0.520	0.829	Treat <166 days
treat_after		1	0.07887	0.14775	0.2849	0.5935	1.082	0.810	1.445	Treat >166 days