Simulating Duration Data for the Cox Model

Jeffrey J. Harden* Jonathan Kropko†

Online Appendix

Contents

1 Why are DGPs from known distributions problematic?

The simulation in the main text demonstrates that the random spline method is capable of generating durations for Cox model simulations. However, it does not explicitly demonstrate any advantage the method holds over the typical approach of using a known distribution to simulate durations for the Cox model. Accordingly, we next repeat the simulation with a parametric DGP and again compare the same three estimators as before.

Specifically, we use all of the same parameter values as in the previous simulation, but instead of the random spline method we draw the durations from a Weibull distribution with shape parameter $s = 5$ and scale parameter $\lambda = X\beta$. We again estimate exponential, Weibull, and Cox models and store the coefficient estimates and RMSE values. Note that the Weibull model is the correct model given this DGP. The durations come from that distribution, making the Weibull model

*Assistant Professor, Department of Political Science, University of Notre Dame, 2055 Jenkins Nanovic Halls, Notre Dame, IN, jeff.harden@nd.edu.

†Assistant Professor, Department of Politics, University of Virginia, S383 Gibson Hall, 1540 Jefferson Park Avenue, Charlottesville, VA 22904, jkropko@virginia.edu.

preferable to the Cox model; furthermore, $s \neq 1$, making it preferable to the exponential model.

Table A1 presents the results with this new simulation. Again the left columns report the means of the coefficient estimates (in proportional hazards parameterization) across the simulated datasets for each estimator. The right columns give coefficient RMSE values. Bold entries indicate the estimator with the lowest RMSE for each coefficient.

[Insert Table A1 here]

The coefficient means again indicate that the exponential model is biased. This is due to the fact that by incorrectly assuming $s = 1$, the exponential model is misspecified. As in the previous simulation the Weibull and Cox models return estimates that are, on average, close to the true values set in the DGP. However, this time the RMSEs show that the *Weibull* model performs the best of the three estimators. This is because only the Weibull correctly parameterizes the true baseline hazard function. The Cox model is not biased because it makes no attempt to parameterize the baseline hazard. However, by modeling only the duration ranks, that model discards some information, and as a result is slightly less efficient than the Weibull.

This result demonstrates a potential problem with exclusively relying on a known distribution to conduct simulations with the Cox model. Doing so means that there is an estimator that outperforms the Cox model due to a statistical artifact of the DGP. Some simulation studies compare the performance of the Cox model and other parametric survival models (e.g., Chastang, Byar, and Piantadosi 1988; Benaglia, Jackson, and Sharples 2015; Kropko and Harden 2018). Using a parametric DGP in such cases may bias the simulation in favor of one or more of the parametric models. The random spline method eliminates this problem. Using the random spline method should not be seen as biasing the simulation in favor of the Cox model: if the random spline method generates baseline hazards that are more realistic for the analyst's particular application, then a result that favors the Cox model illustrates the Cox model's inherent flexibility, *not* a bias from the simulation.

2 Mixing random spline with a parametric distribution

To further illustrate the flexibility of the random spline method, we next demonstrate how it can be combined with a parametric distribution to vary the expected performance of the Cox model and a parametric estimator. Specifically, we present a set of simulations that compare the performance of the Cox model and Weibull model when the durations are generated from mixtures of a random spline hazard and a Weibull hazard.[1] Importantly, we increase the proportion of the data generated by each hazard function in 5% increments. We begin with 0% drawn from the random spline and 100% from the Weibull, which creates a DGP in which the Weibull model should perform the best. Then we simulate again with 5% of the data drawn from the random spline hazard and 95% from the Weibull, then again with 10% random spline and 90% Weibull, etc.... Our last simulation draws all of the data from the random spline hazard. We expect the Cox model's performance to improve as more data are drawn from the random spline hazard.

Figure A1 reports the results. The x-axis graphs the proportion of the data generated from the random spline hazard. The y-axis gives the ratio of the Weibull model's RMSE to the Cox model's RMSE. Values lower than 1 on the y-axis reflect better performance by the Weibull model and values greater than 1 indicate better performance by the Cox model.

[Insert Figure A1 here]

The results show that the Cox model's performance relative to the Weibull model improves as the proportion of data drawn from the random spline hazard increases. In fact, the Cox model yields a lower RMSE with as little as 10% of the data drawn from the random spline hazard. The Weibull model only produces smaller RMSE values when the durations are all from a Weibull distribution or from a mixture of 5% random spline/95% Weibull. This finding underscores the importance of the baseline hazard assumption and the potential benefits of the Cox model. Even relatively small deviations from the Weibull hazard render the Weibull model an inferior estimator compared to the Cox model.

[1] The other parameters and aspects of these simulations are the same as the simulation presented in the main text (see the replication materials for more details).

3 Controlling the random baseline hazard's shape

In some situations the analyst may wish to exert some control over the shape of the baseline hazard function while still using the random spline method. For example, he or she may want the hazard to monotonically increase or decrease, but still be drawn from random points rather than from a Weibull distribution. Such an approach may be more realistic for some contexts. We illustrate an example with a monotonically increasing function here.

To accomplish this objective we first follow the steps described in the main text to generate the failure CDF. Then, instead of using the resulting function as the CDF and computing the baseline functions as described in the main text, we use that function as the hazard function itself. This approach yields a monotonically increasing function that does not follow a specific distribution.[2] Figure A2 illustrates the process. The graph in panel (a) plots the baseline hazard function. Panel (b) gives a histogram of a sample of durations simulated from that hazard function.

[Insert Figure A2 here]

After generating the baseline hazard with this restriction, the rest of the simulation occurs just as before. Table A2 presents results from a simulation identical to the one presented in the main text, but with the monotonically increasing hazard function from Figure A2.

[Insert Table A2 here]

Table A2 show the same relative performance between the three estimators as the simulation from the main text. In particular, the Cox model yields unbiased estimates with the lowest RMSE. Additionally, the absolute performance of the exponential model and Cox model remain essentially unchanged compared to the main text. The major difference is in the performance of the Weibull model, which is better here than in the simulation from the main text.

[2] Similarly, if the analyst wanted a monotonically decreasing hazard, he or she could simply insert the survivor function from the original method directly as the hazard.

4 Time-varying covariates

The random spline method can easily be adapted to include TVC, as we show in another simulation. Specifically, an option in the R function employs the permutation algorithm (PA) described by Sylvestre and Abrahamowicz (2008) to link the durations generated via the random spline method to user-controlled covariates and coefficient values.

The first step in this process is to generate durations using the random spline method. We then generate three new covariates in a $NT \times p$ matrix such that there are T rows for each of the N observations. The first two covariates are TVC; we generate X_1 from a standard normal distribution and X_2 from a Poisson distribution with a mean of 2. X_1 is a binary variable that varies across observations but is static over time within each observation. We also generate censoring times (on the same scale as the actual durations) for each observation as random draws from a uniform distribution on the interval $[1, T]$.

Next, the PA matches each of the N observed durations with the T rows of the covariate matrix for each observation. This matching is done according to permutation probabilities derived from the Cox model's partial likelihood computation (see Sylvestre and Abrahamowicz 2008, 2620–2621). In brief, this procedure follows the steps listed below.

- First, the algorithm defines a variable t^* to be the lesser of (1) the time until failure and (2) the censoring time for each of the generated durations. Then it sorts the values of t^* in ascending order.
- Next, the algorithm splits the covariate matrix into N individual matrices, one for each observation, where each matrix has T rows for each of the T time points. After drawing simulated coefficients, it uses these coefficients to generate the linear predictors from these covariates, which are then transformed into the Cox model probabilities that a particular observation still at risk will be the next to fail.
- Finally, the algorithm connects the covariates to the outcomes. If the observation is not censored, the probability of assignment for each observation i is defined from the Cox model

probabilities. If the observation is censored, assignment occurs as a draw with uniform probability based on the size of the observation's risk set.

The result is a dataset with TVC that follows the DGP defined via the random spline hazard function, the analyst's covariates and the analyst's true coefficient values. We repeat this process 1,000 times, monitoring the coefficient estimates of the Cox model and their RMSEs.

Table A3 reports the means of the Cox model coefficient estimates and RMSE values for the simulations with TVC. We complete the simulation with N = 100, 500, and 1,000. The true coefficient values in this simulation are $\beta_1 = 0.50$, $\beta_2 = 0.25$, and $\beta_3 = 0.75$.

[Insert Table A3 here]

The results show that the Cox model recovers the true DGP, with coefficient means close to the true values. Additionally, the estimates generally improve (coefficient means closer to the true values and smaller RMSEs) with increased sample sizes. Overall, these results indicate that researchers conducting simulation studies can use the random spline method for simulations and include TVC in their designs.

5 Violations to the proportional hazards assumption

The proportional hazards assumption is a critical aspect of the Cox model. Accordingly, researchers may be interested in investigating the consequences of assuming proportional hazards in model estimation with and without proportional hazards in the true DGP. Here we demonstrate that violations to the proportional hazards assumption can be included in the DGP with the random spline method. We start with the original simulation from the main text, then modify it to vary the effect of one independent variable depending on time. In this simulation we hold constant the effects of the last two independent variables (X_2 and X_3), but vary the effect of X_1 by multiplying its coefficient (β_1) by the log of the time index. This can be done by setting `type = ''tvbeta''` in the R function.[3]

[3] See Hendry (2014) for more on simulating proportional hazards assumption violations.

Table A4 demonstrates that this procedure produces violations of the proportional hazards assumption. The first row in the table summarizes a simulation in which the DGP includes proportional hazards and the second row gives results of the DGP without proportional hazards for β_1. Columns report the means of all three coefficient estimates across the simulation as well as the mean p-values from a Schoenfeld residuals test with a log time transformation (Grambsch and Therneau 1994). This test computes a chi-squared test statistic for each variable as well as a global statistic. The null hypothesis is no violation; thus, a statistically significant test statistic indicates a violation of the proportional hazards assumption. If the DGP summarized in the second row of the table correctly generates non-proportional hazards in β_1, the test should produce statistically significant test statistics associated with X_1 and the global test, but not with X_2 and X_3.

[Insert Table A4 here]

The first row of results correctly shows no evidence of proportional hazards violations. The coefficient means fall near the true values and the mean Schoenfeld residuals test p-values are far from conventional significance thresholds. In contrast, the second row of results correctly shows evidence of a proportional hazards violation for X_1. The average estimate of β_1 is far from the true value set in the DGP. The average estimates for the other coefficients are closer to their true values, but still display bias.[4] Furthermore, the p-values also show evidence of a proportional hazards violation. The mean p-value does not reach conventional significance levels for X_2 and X_3, but is less than 0.01 for X_1 and the global test.

Another means of verifying that the second DGP produces data with non-proportional hazards is a graphical examination of Schoenfeld residuals against time. Figure A3 provides such graphs for one iteration of each simulation for X_1. The time index is presented on the x-axes and the y-axes plot scaled Schoenfeld residuals for X_1. Panel (a) presents results from the DGP with proportional hazards and panel (b) presents results without proportional hazards for that variable.

[Insert Figure A3 here]

[4]This bias suggests that failing to correct a proportional hazards violation may have problematic implications for the entire regression model.

The graph in panel (a) demonstrates that in the first DGP the Schoenfeld residuals are virtually uncorrelated with time, correctly indicating no violation to the proportional hazards assumption. In contrast, panel (b) shows that the Schoenfeld residuals are relatively small initially, then increase in average magnitude. In other words, not accounting for the non-proportional hazards means the model overestimates the effect of X_1 initially, then underestimates it for later time periods. This is consistent with the DGP, in which the true effect of X_1 increases across time.

In sum, like the TVC example, this simulation demonstrates that researchers can use the random spline method and include violations to the proportional hazards assumption in their designs. Both examples show that the random spline method is flexible enough to allow researchers considerable control over how they use simulations to evaluate the Cox model's performance.

References

Benaglia, Tatiana, Christopher H. Jackson, and Linda D. Sharples. 2015. "Survival Extrapolation in the Presence of Cause Specific Hazards." *Statistics in Medicine* 34(5): 796–811.

Chastang, Claude, David Byar, and Steven Piantadosi. 1988. "A Quantitative Study of the Bias in Estimating the Treatment Effect Caused by Omitting a Balanced Covariate in Survival Models." *Statistics in Medicine* 7(12): 1243–1255.

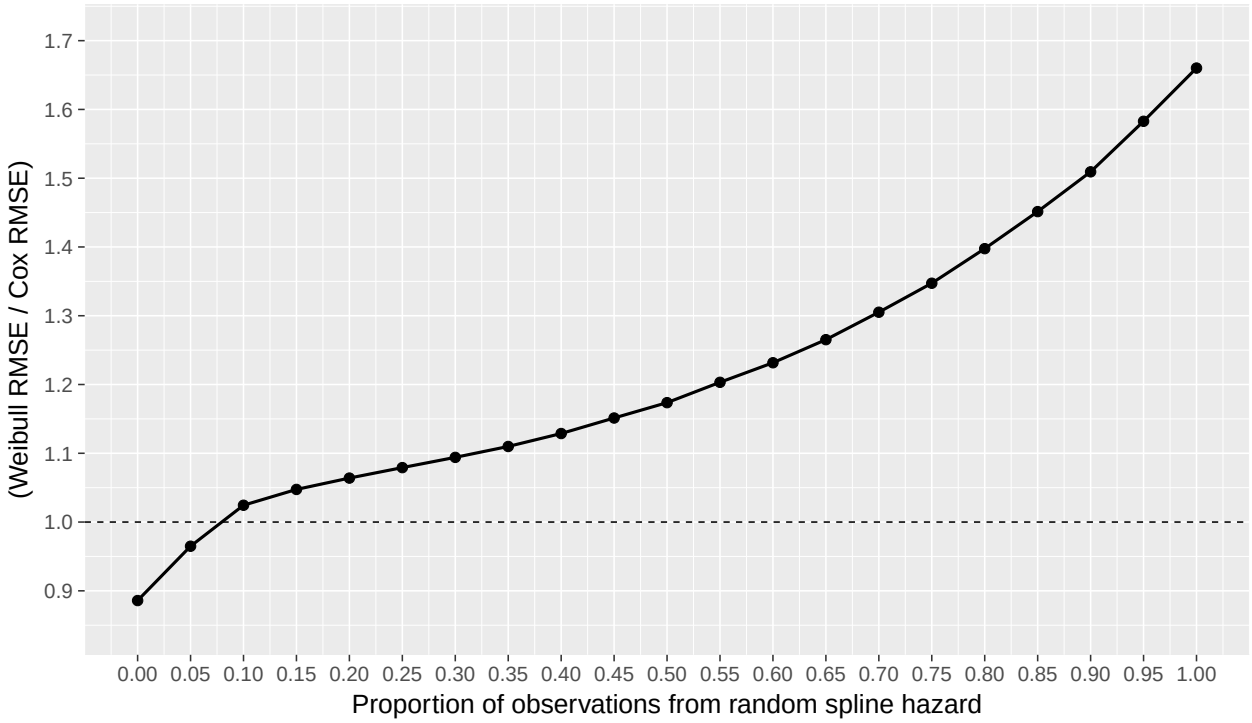
Grambsch, Patricia M., and Terry M. Therneau. 1994. "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals." *Biometrika* 81(3): 515–526.

Hendry, David J. 2014. "Data Generation for the Cox Proportional Hazards Model with Time-Dependent Covariates: A Method for Medical Researchers." *Statistics in Medicine* 33(3): 436–454.

Kropko, Jonathan, and Jeffrey J. Harden. 2018. "Beyond the Hazard Ratio: Generating Expected Durations from the Cox Proportional Hazards Model." Forthcoming, *British Journal of Political Science*. https://doi.org/10.1017/S000712341700045X.

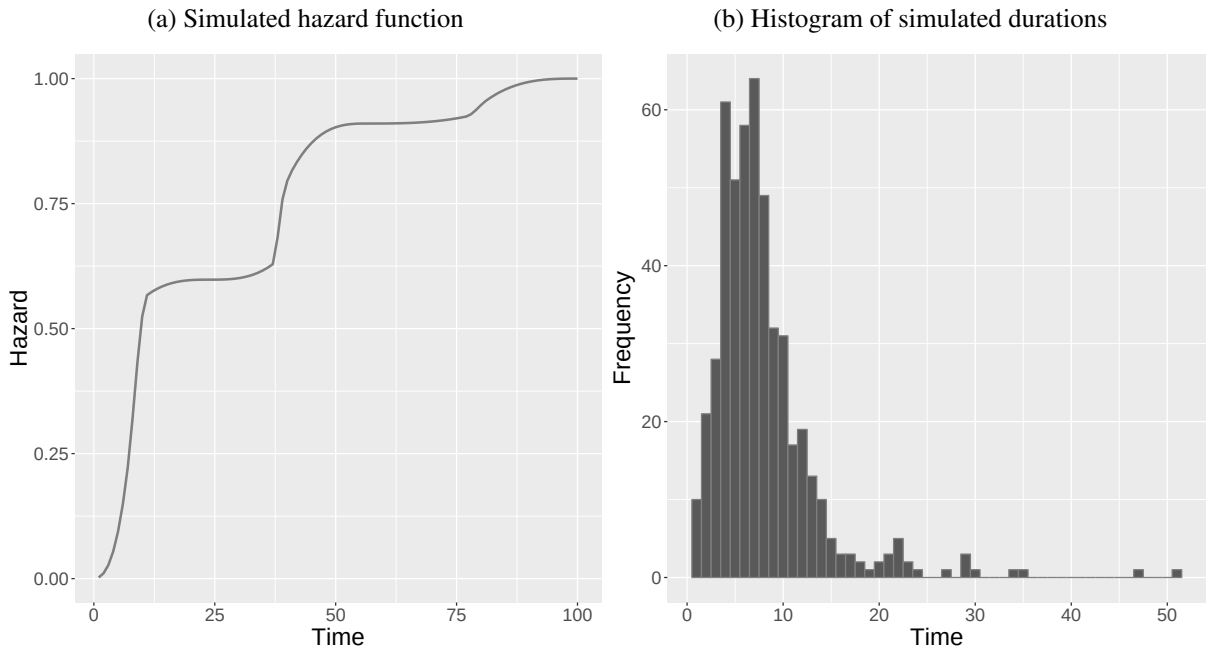
Sylvestre, Marie-Pierre, and Michal Abrahamowicz. 2008. "Comparison of Algorithms to Generate Event Times Conditional on Time-Dependent Covariates." *Statistics in Medicine* 27(14): 2618–2634.

Figure A1: Cox model and Weibull model performance in the mixture simulations



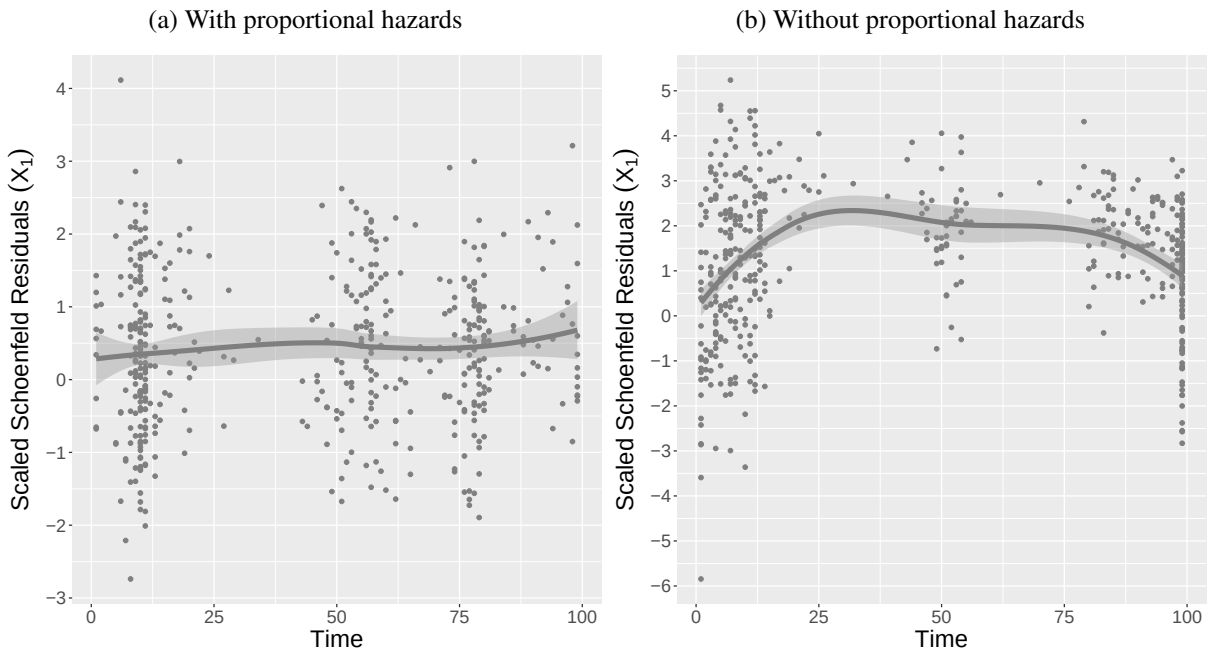
Note: The graph presents results from the mixture simulations. The x-axis graphs the proportion of the data generated from the random spline hazard. The y-axis gives the ratio of the Weibull model's RMSE to the Cox model's RMSE. Values lower than 1 on the y-axis reflect better performance by the Weibull model and values greater than 1 indicate better performance by the Cox model.

Figure A2: Example output from the random spline method with a monotonically increasing baseline hazard



Note: The graphs illustrate output from one iteration of the random spline method with the baseline hazard restricted to increase monotonically. Panel (a) shows the hazard function and panel (b) shows a sample of durations generated from the hazard function.

Figure A3: Schoenfeld residuals plots with and without proportional hazards in X_1



Note: The graphs present scaled Schoenfeld residuals for X_3 on the y-axes and the time index on the x-axes. The solid lines and shading indicate LOESS fits and their 95% confidence intervals. Panel (a) presents results from the DGP in which the proportional hazards assumption holds. Panel (b) presents results from the DGP with a violation to the proportional hazards assumption in the effect of X_3.

Table A1: Coefficient estimate means and root mean squared error from data simulated via the Weibull distribution

	Coefficient Means			RMSE		
Estimator	β_1	β_2	β_3	β_1	β_2	β_3
Exponential	0.100	−0.099	0.151	0.401	0.402	0.600
Weibull	0.500	−0.501	0.759	**0.094**	**0.096**	**0.099**
Cox	0.500	−0.501	0.759	0.095	0.097	0.100

Note: Cell entries report coefficient estimate means and RMSE for the exponential, Weibull, and Cox model estimates (in proportional hazards parameterization). True coefficient values are $\beta_1 = 0.50$, $\beta_2 = -0.50$, and $\beta_3 = 0.75$. Bold entries indicate the estimator with the lowest RMSE for each coefficient. $N = 500$.

Table A2: Coefficient estimate means and root mean squared error from data simulated via the random spline method with a monotonically increasing baseline hazard function

	Coefficient Means			RMSE		
Estimator	β_1	β_2	β_3	β_1	β_2	β_3
Exponential	0.200	−0.201	0.299	0.301	0.300	0.451
Weibull	0.522	−0.522	0.783	0.058	0.058	0.064
Cox	0.490	−0.492	0.735	**0.051**	**0.050**	**0.056**

Note: Cell entries report coefficient estimate means and RMSE for the exponential, Weibull, and Cox model estimates (in proportional hazards parameterization). True coefficient values are $\beta_1 = 0.50$, $\beta_2 = -0.50$, and $\beta_3 = 0.75$. Bold entries indicate the estimator with the lowest RMSE for each coefficient. $N = 500$ in all simulations.

Table A3: Cox model coefficient estimate means and root mean squared error from data simulated via the random spline method with time-varying covariates

	Coefficient Means			RMSE		
N	β_1	β_2	β_3	β_1	β_2	β_3
100	0.520	0.263	0.786	0.058	0.039	0.200
500	0.513	0.258	0.764	0.026	0.017	0.077
1,000	0.507	0.254	0.758	0.018	0.011	0.053

Note: Cell entries report coefficient estimate means and RMSE for the Cox model estimates with N = 100, 500, and 1,000. True coefficient values are $\beta_1 = 0.50$, $\beta_2 = 0.25$, and $\beta_3 = 0.75$.

Table A4: Cox model coefficient estimate means and mean Schoenfeld residuals test p-values from data simulated via the random spline method with non-proportional hazards

	Coefficient Means			Mean p-value			
Condition	β_1	β_2	β_3	X_1	X_2	X_3	Global
With PH	0.492	−0.494	0.739	0.498	0.496	0.497	0.496
Without PH	1.300	−0.363	0.542	0.011	0.270	0.143	0.001

Note: Cell entries report results with (row 1) and without (row 2) proportional hazards in the β_1 coefficient. The first three columns give coefficient means and the next four columns give the mean p-values from a Schoenfeld residuals test for non-proportional hazards (individual variables and global test). True coefficient values for the coefficients that remain constant across both simulations are $\beta_2 = -0.50$ and $\beta_3 = 0.75$. $N = 500$ in both simulations.