Turning Hazard Models Inside Out

Non-parametric regression models are routinely used in demographic analyses to explore the dynamics shaping fertility timing, marriage and divorce, and migration outcomes, and hazard models are among the most frequently employed. Hazard models are used to analyze the time until an event occurs, allowing researchers to assess the factors that influence this duration (Box-Steffensmeier and Zorn 2001). Initially developed to examine age-specific failure rates in life tables (Cox 1972), hazard models are now widely used in demographic analysis, including applications to studies of life expectancy (Missov and Lenart 2011) and migration processes (Mehrab et al. 2024).

Such regression models may mask heterogeneity and ignore relationships between both cases and variables. In their recently published book, *Regression Inside Out,* Schoon, Melamed and Breiger (2024) outline a novel approach to regression modeling that brings cases to the fore. Among other things, Regression Inside Out (RIO) allows researchers to see each individual observation's additive contribution to the overall regression coefficient, geometrically map the regression model space, and plot the relationships among cases and variables. In these ways, RIO reveals how individual cases and clusters of cases influence components of the overall model. To date, however, RIO has been generalized only to a limited number of models.[1]

In this article, we introduce a novel generalization of RIO that allows it to be used with any model in which linear predictors are estimated, and we demonstrate its use in the case of survival, and specifically Cox proportional hazard, models. To do so, we extract the linear predictors from the model and then predict those in a separate OLS model, effectively simplifying the entire GLM into OLS for our purposes. In this way, the generalization we apply to Hazard models in this context generalizes to any model for which linear predictors are

[1] Schoon, Melamed and Breiger (2024) describe RIO in the context of OLS, logistic, Poisson, negative binomial, and random intercept regression models.

estimated. We begin by detailing the basic logic of RIO, and introduce our novel generalization of the procedures outlined by Schoon, Melamed and Breiger (2024). We also introduce a bootstrap approach for distinguishing signal from random noise in both the individual and aggregate contributions to the regression coefficients of observations. We highlight the utility of these methodological innovations through re-analyses of a published proportional hazard model that examined differential mortality risks by sex, race, and ethnicity (Yuan et al. 2025). We conclude with the implications of this generalization and potential future directions.

Generalizing Regression Inside Out

RIO (Schoon, Melamed and Breiger 2024) includes a few related methods for examining what is going on inside of a regression model. First, the *coefficient decomposition* computes each observation's additive contribution to the estimated regression coefficients. Second, the *variance decomposition* computes each observation's contribution to the error variance and can flag cases that have a disproportionate impact on the model or draw attention to parts of the sample space. Third, the *model visualization* projects the cases and the variables into the same space, maintaining the relationships embedded in the regression coefficients while illustrating where/how the cases fit into the relationships between variables. And fourth, a natural use for RIO is to inductively *identify interaction effects* or other non-linearities in the sample space. Here, we describe how to generalize RIO's coefficient decomposition to any model for which linear predictors are computed.[2] We also implement a novel bootstrapping procedure that is an alternative to existing variance decomposition techniques.

In terms of the regression *coefficient decomposition*, Equation 1 presents the familiar OLS estimator with one adjustment. Rather than post-multiply by the outcome variable as a

[2] While RIO may be used to identify interaction effects, we do not discuss this aspect of RIO in detail. However, the generalization we introduce here can be applied for the purposes of interaction detection following the same procedures described in Chapter 6 of Schoon, Melamed, and Breiger (2024).

vector, we post-multiply by the outcome as a diagonal matrix. While the traditional estimator returns a p x 1 vector of regression coefficients, where p refers to the number of estimated parameters, Eq. 1 returns a p x n matrix, where n refers to the sample size, of regression coefficient contributions whose rows sum to the estimated regression coefficients (Schoon, Melamed, and Breiger 2024). Columns of **P** may be summed to examine how subsets of cases contribute to the overall model coefficients (see Breiger and Melamed 2014).

$$P=\left[\left(X^T X\right)^{-1} X^T\right]Y \tag{1}$$

Our generalization for the coefficient decomposition rests on a statistical identity, namely that the predictor variables for many classes of statistical models are linearly related to the linear predictors for the model. Often those linear predictors are transformed to some other metric, such as predicted probabilities via the logistic function in logistic regression. But the linear predictors are linearly related to the predictor variables. This forms the basis for our generalization: any model for which linear predictors exist can be decomposed by regressing the linear predictors on the independent variable matrix. This yields the same regression coefficients as the original non-OLS model and may be linearized in a way that weighted regressions (e.g., logistic or Cox Hazard models) cannot. Thus, the regression coefficients for any model with linear predictors can be decomposed into case contributions to the full regression coefficients.

In terms of *variance decomposition*, existing techniques can decompose the model error variance, but this does not translate to case-level standard errors as the variance and standard error are not linearly related (Schoon, Melamed, Breiger, and Yoon 2024). Extending the variance decomposition, here we describe a bootstrapping procedure that we adapted specifically for RIO to distinguish the signal from noise of regression contributions, and for both individual cases and subsets of them. The procedure entails sampling cases with replacement, estimating the

regression model, retaining the desired decomposition, repeating this process thousands of times (we use 10,000), and computing the confidence interval from the distribution of regression contributions. In the context of Hazard models, we extend this basic bootstrapping set up to account for the fact that cases have multiple observations through time.

The remaining aspects of RIO – *model visualization* and *detecting interaction effects* – do not need to be altered once the linear predictors have been substituted as the outcome. The model visualization applies a Singular Value Decomposition (SVD) to the predictor matrix with the linear predictors appended to it. The linear relationship between the predictors and the linear predictors is maintained across the dimensions of the SVD, and therefore the leading dimensions of the SVD may be plotted for a model visualization that is similar to a correspondence analysis except that an explicit dependent variable has been specified (Breiger et al. 2014). Likewise, the coefficient decomposition still yields vectors of effects on the outcome which can be predicted by other variables to assess moderation and inductively discover statistical interaction effects (Schoon, Melamed and Breiger 2024). Below, we illustrate these generalizations by replicating aa published example of a Cox Proportional Hazard models. In particular, we reproduce the results presented in Yuan et al. (2025) illustrating differential mortality risk by sex, race and ethnicity. Results are forthcoming and will be ready in advance of June 2026.

References:

Box-Steffensmeier, Janet M., and Christopher JW Zorn. 2001. “Duration Models and Proportional Hazards in Political Science.” *American Journal of Political Science* 972–88.

Breiger, Ronald L., and David Melamed. 2014. “The Duality of Organizations and Their Attributes: Turning Regression Modeling ‘Inside Out.’” Pp. 263–75 in *Contemporary perspectives on organizational social networks*. Emerald Group Publishing Limited.

Cox, D. R. 1972. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society Series B: Statistical Methodology* 34(2):187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.

Mehrab, Zakaria, Logan Stundal, Srinivasan Venkatramanan, Samarth Swarup, Bryan Lewis, Henning S. Mortveit, Christopher L. Barrett, Abhishek Pandey, Chad R. Wells, and Alison P. Galvani. 2024. “An Agent-Based Framework to Study Forced Migration: A Case Study of Ukraine.” *PNAS Nexus* 3(3):pgae080.

Missov, Trifon I., and Adam Lenart. 2011. “Linking Period and Cohort Life-Expectancy Linear Increases in Gompertz Proportional Hazards Models.” *Demographic Research* 24:455–68.

Schoon, Eric W., David Melamed, and Ronald L. Breiger. 2024. *Regression Inside Out*. Strategies for Social Inquiry. Cambridge: Cambridge University Press.

Yuan, Ye, Carmen R. Isasi, Tala Al-Rousan, Arnab K. Ghosh, Pricila H. Mullachery, Priya Palta, and Nour Makarem. 2025. “Associations of Concurrent Hypertension and Type 2 Diabetes With Mortality Outcomes: A Prospective Study of US Adults.” *Diabetes Care* 48(7):1241–50.