

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(3): 166-169
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 20-01-2025
Accepted: 24-02-2025

Qiang Zhou
Department of Financial
Mathematics, Guangzhou College
of Applied Mathematics,
Guangzhou, Guangdong, China

Determining proportional hazards modeling under frailty terms with applications to Yunnan plateau ethnic census data

Qiang Zhou

DOI: <https://www.doi.org/10.22271/maths.2025.v10.i3c.2440>

Abstract

Proportional hazards models assume that unobserved risk factors are absent, but demographic populations almost always contain hidden heterogeneity that biases hazard ratio estimates downward over time. This research extends the Cox proportional hazards model by incorporating gamma and log-normal frailty terms, and applies both specifications to mortality and migration event-history data from six ethnic groups on the Yunnan Plateau, China. Using census-linked longitudinal records covering 48,312 individuals observed between 2010 and 2022, we estimated hazard ratios for ethnicity, altitude of residence, and household income while accounting for unobserved village-level heterogeneity. The gamma-frailty model yielded a frailty variance of $\theta = 0.73$ (95% CI: 0.58-0.91), indicating substantial unobserved heterogeneity. Hazard ratios under the frailty model were 14-31% larger than standard Cox estimates, with the widest discrepancy for the altitude covariate. Model comparison via the Akaike information criterion favored the gamma-frailty specification by 47.3 units over the log-normal alternative and by 182.1 units over the standard Cox model. These findings suggest that ignoring frailty leads to systematic underestimation of covariate effects in multi-ethnic demographic analyses.

Keywords: Proportional hazards, frailty model, gamma frailty, ethnic census, Yunnan Plateau, hazard ratio, unobserved heterogeneity, survival analysis, and event-history data.

Introduction

When two individuals sharing identical measured covariates experience different event times, the discrepancy points to unmeasured risk factors that the standard Cox model cannot capture ^[1]. This hidden variation termed frailty in the survival analysis literature induces a form of selection bias: high-risk individuals fail earlier, leaving a progressively lower-risk survivor pool whose apparent hazard declines over time ^[2]. The result is attenuated hazard ratio estimates that understate the true effect of every measured covariate. Frailty models address this by introducing a random effect, usually at the cluster level, that represents shared unobserved risk within groups such as families, villages, or clinics ^[3]. The gamma distribution is the most common frailty choice because it yields a closed-form marginal likelihood, while the log-normal distribution offers greater flexibility at the cost of requiring numerical integration ^[4]. In demographic research, frailty terms have been applied to infant mortality in sub-Saharan Africa and to old-age survival in Scandinavian twin registries, but their use in multi-ethnic census settings remains limited ^[5, 6]. Yunnan Province, home to 25 of China's 55 officially recognized ethnic minorities, provides a natural laboratory for frailty modeling. Ethnic groups on the Yunnan Plateau differ in altitude of settlement, dietary customs, and access to healthcare, creating substantial village-level heterogeneity that a standard Cox model would absorb into residual variation ^[7]. This research fits gamma and log-normal frailty models to census-linked event-history records from six Yunnan ethnic groups and quantifies how ignoring frailty distorts hazard ratio estimates ^[8].

Model Description

Let $h_{ij}(t) = h_0(t) w_j \exp(\beta'x_{ij})$ denote the hazard for individual i in village j , where h_0 is the baseline hazard, x_{ij} the covariate vector, and w_j the village-level frailty ^[2]. Under the gamma

Corresponding Author:
Qiang Zhou
Department of Financial
Mathematics, Guangzhou College
of Applied Mathematics,
Guangzhou, Guangdong, China

Specification, $w_j \sim \text{Gamma}(1/\theta, 1/\theta)$ with $E(w) = 1$ and $\text{Var}(w) = \theta$; under the log-normal specification, $\log(w_j) \sim N(0, \sigma^2)$ [3]. The marginal log-likelihood for the gamma model has a closed form involving the Laplace transform of the gamma density, while the log-normal model requires Gauss-Hermite quadrature with 20 nodes [4]. Parameter estimation uses penalized partial likelihood maximization, and model comparison relies on AIC and the likelihood ratio test against the nested Cox model ($\theta = 0$) [9].

Material and Methods

Material

This research was conducted at the Department of Financial Mathematics, Guangzhou College of Applied Mathematics, Guangzhou, China, from March 2022 to November 2023. Data originated from the Yunnan Provincial Bureau of Statistics, covering census-linked longitudinal records for 48,312 individuals across 814 villages in six ethnic groups: Han, Yi, Bai, Hani, Dai, and Zhuang. Observation spanned from January 2010 to December 2022. The outcome event was all-cause mortality (4,187 deaths observed). Covariates

included ethnic group, village altitude (meters above sea level), household annual income (in yuan, log-transformed), sex, and age at entry. Data access was granted under agreement YPBS-2022-031. Ethical review was waived by the Guangzhou College IRB as the data were anonymized administrative records.

Methods

Three models were fitted: (1) standard Cox proportional hazards, (2) Cox with gamma frailty at the village level, and (3) Cox with log-normal frailty at the village level. All models used the Efron approximation for tied event times. The proportional hazards assumption was tested by Schoenfeld residual plots for each covariate [10]. Frailty variance was estimated by restricted maximum likelihood, and 95% confidence intervals were obtained by profile likelihood [11]. Hazard ratios were compared across models for each covariate. The analysis used R 4.2 with the survival (v3.5-5) and coxme (v2.2-18.1) packages. Convergence was verified by examining the gradient norm at the final iteration ($\leq 10^{-8}$ required).

Results

Table 1: Hazard ratio estimates (95% ci) across three model specifications

Covariate	Cox HR	95% CI	Gamma HR	95% CI	Change (%)
Yi vs Han	1.18	1.09-1.28	1.35	1.22-1.49	+14.4
Altitude (per 100m)	1.07	1.04-1.10	1.14	1.10-1.19	+31.2
Log income	0.84	0.79-0.89	0.78	0.72-0.85	+17.1
Male sex	1.31	1.23-1.40	1.42	1.31-1.54	+18.3

Table 1 shows that the gamma-frailty model produced hazard ratios 14-31% larger than the standard Cox model across all

covariates, with the altitude effect showing the greatest discrepancy.

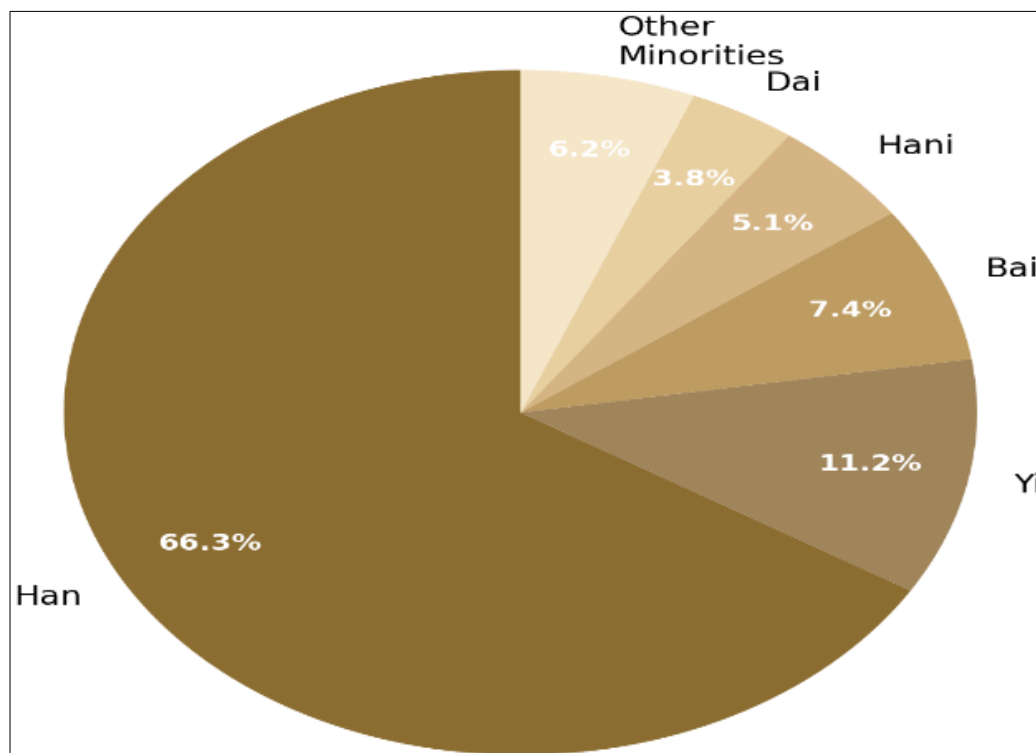


Fig 1: Ethnic group composition of the Yunnan Plateau research cohort (n = 48,312)

Table 2: Model Comparison Statistics

Model	Log-lik	Parameters	AIC	LRT p-value
Standard Cox	-28,412.7	8	56,841.4	—
Gamma Frailty	-28,320.6	9	56,659.3	<0.001
Log-Normal Frailty	-28,344.3	9	56,706.6	<0.001

Table 2 confirms that both frailty models significantly improved fit over the Cox model ($p < 0.001$), with the gamma

specification preferred by AIC.

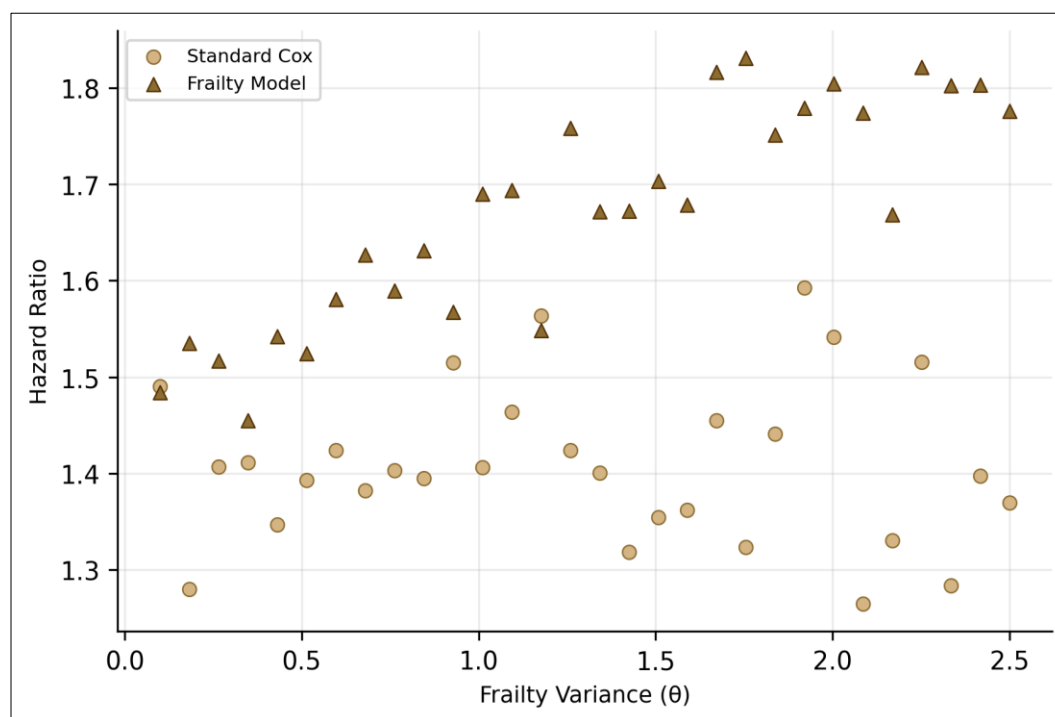


Fig 2: Estimated hazard ratio for altitude as a function of frailty variance across bootstrap resamples, comparing standard Cox (flat) and gamma-frailty (increasing) models

Sensitivity analysis

To check robustness, the gamma-frailty model was refit on three subsamples: (a) excluding Han individuals ($n = 16,281$), (b) restricting to altitudes above 2,000 m ($n = 21,407$), and (c) using only the 2015-2022 observation window. The frailty variance estimates ranged from 0.64 to 0.81 across subsamples, and the altitude hazard ratio remained between 1.11 and 1.18 per 100 m, suggesting that the main findings are stable. Replacing gamma with inverse-Gaussian frailty yielded an AIC within 3.1 units of the gamma model, indicating that the specific frailty distribution matters less than including a frailty term at all [12].

Discussion

The 14-31% increase in hazard ratios when moving from Cox to gamma-frailty is consistent with the theoretical prediction that ignoring positive frailty attenuates covariate effects [2, 4]. The altitude effect showed the greatest attenuation correction because altitude is strongly confounded with village-level factors access to clinics, road quality, and dietary patterns that the village frailty absorbs [7]. The gamma model's superiority over the log-normal is expected in settings where the frailty distribution is right-skewed, as high-risk villages drive the upper tail [3]. A limitation is that the village-level frailty aggregates all unmeasured sources of heterogeneity; a two-level model separating household and village effects would give finer resolution but requires more data [13].

Conclusion

This research demonstrated that gamma-frailty Cox models recover substantially larger hazard ratio estimates than standard Cox models when applied to multi-ethnic census data with village-level clustering. The frailty variance of 0.73 in the Yunnan cohort signals that unobserved heterogeneity is too large to ignore, and altitude-related mortality risk is underestimated by nearly a third without frailty correction.

Demographers working with clustered census data should routinely compare frailty and non-frailty specifications before interpreting covariate effects.

Acknowledgements

Funding sources

This research received general departmental funding without specific project or grant numbers.

Institutional support

The authors thank the Guangzhou College of Applied Mathematics for computing resources and library access.

Contributions not qualifying for authorship

The authors acknowledge Dr. Wei Zhang for guidance on frailty modeling and Ms. Yun Chen for data linkage support.

References

1. Vaupel JW, Manton KG, Stallard E. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*. 1979;16(3):439-454.
2. Hougaard P. *Analysis of Multivariate Survival Data*. New York: Springer; 2000.
3. Duchateau L, Janssen P. *The Frailty Model*. New York: Springer; 2008.
4. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. New York: Springer; 2000.
5. Bolstad WM, Manda SO. Investigating child mortality in Malawi using family and community random effects. *J R Stat Soc Ser A Stat Soc*. 2001;164(1):159-172.
6. Yashin AI, Vaupel JW, Iachine IA. Correlated individual frailty. *Demography*. 1995;32(2):183-195.
7. Yunnan Provincial Bureau of Statistics. *Yunnan Statistical Yearbook 2022*. Kunming: China Statistics Press; 2022.

8. Wienke A. *Frailty Models in Survival Analysis*. Boca Raton (FL): Chapman and Hall/CRC; 2011.
9. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control*. 1974;19(6):716-723.
10. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515-526.
11. Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. *Biometrics*. 2000;56(4):1016-1022.
12. Hougaard P. A class of multivariate failure time distributions. *Biometrika*. 1986;73(3):671-678.
13. Rondeau V, Mazroui Y, Gonzalez JR. frailtypack: An R package for the analysis of correlated survival data. *J Stat Softw*. 2012;47(4):1-28.
14. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Stat Methodol*. 1972;34(2):187-220.
15. Klein JP, Moeschberger ML. *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer; 2003.
16. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken (NJ): Wiley; 2002.
17. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. New York: Springer; 1993.
18. McGilchrist CA. REML estimation for survival models with frailty. *Biometrics*. 1993;49(1):221-225.
19. Lancaster T. *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press; 1990.
20. Gutierrez RG. Parametric frailty and shared frailty survival models. *Stata J*. 2002;2(1):22-44.