Supplementary material for "Bayesian Survival Modelling of University Outcomes"

C.A. Vallejos and M.F.J. Steel

This document extends the descriptive analysis of the PUC dataset and provides further details regarding the implementation of Bayesian inference and the properties of the MCMC chain. In addition, documentation for the freely available R code is provided. Finally, we discuss the proportional odds assumption for our data. Throughout, Sections and equation numbers not starting with S refer to the paper.

A. Descriptive analysis of PUC dataset

Table S1 breaks down the percentage of students satisfying the inclusion criteria (see Section 2) by program. This inclusion percentage is at least 78% for the programmes analyzed in Section 5.

Table S1: PUC dataset. Amount of students satisfying the inclusion criteria using in this study by program.

Program	No. students	% students
Acting	362	80.1
Agronomy and Forestry Engineering	2,466	85.2
Architecture	841	69.9
Art	688	76.3
Astronomy	295	88.3
Biochemistry	331	85.5
Biology	791	83.9
Business Administration and Economics	2,027	72.7
Chemistry	379	82.0
Chemistry and Pharmacy	687	85.6
Civil Construction	1,930	86.0
Design	651	65.2
Education, elementary school	1,277	81.4
Education, elementary school (Villarrica campus)	301	80.5
Education, preschool	949	83.2
Engineering	3,522	69.3
Geography	534	84.5
History	552	76.6
Journalism and Media Studies	876	76.2
Law	2,303	84.2
Literature (Spanish and English)	911	80.8
Mathematics and Statistics	598	78.0
Medicine	972	89.8
Music	161	74.5
Nursing	886	78.6
Physics	237	85.9
Psychology	801	75.9
Social Work	440	87.5
Sociology	421	74.0
Total	27,189	78.7

Figures S1 to S8 summarize a more complete descriptive analysis of the PUC dataset. These Figures confirm strong levels of heterogeneity between different programmes of the PUC.

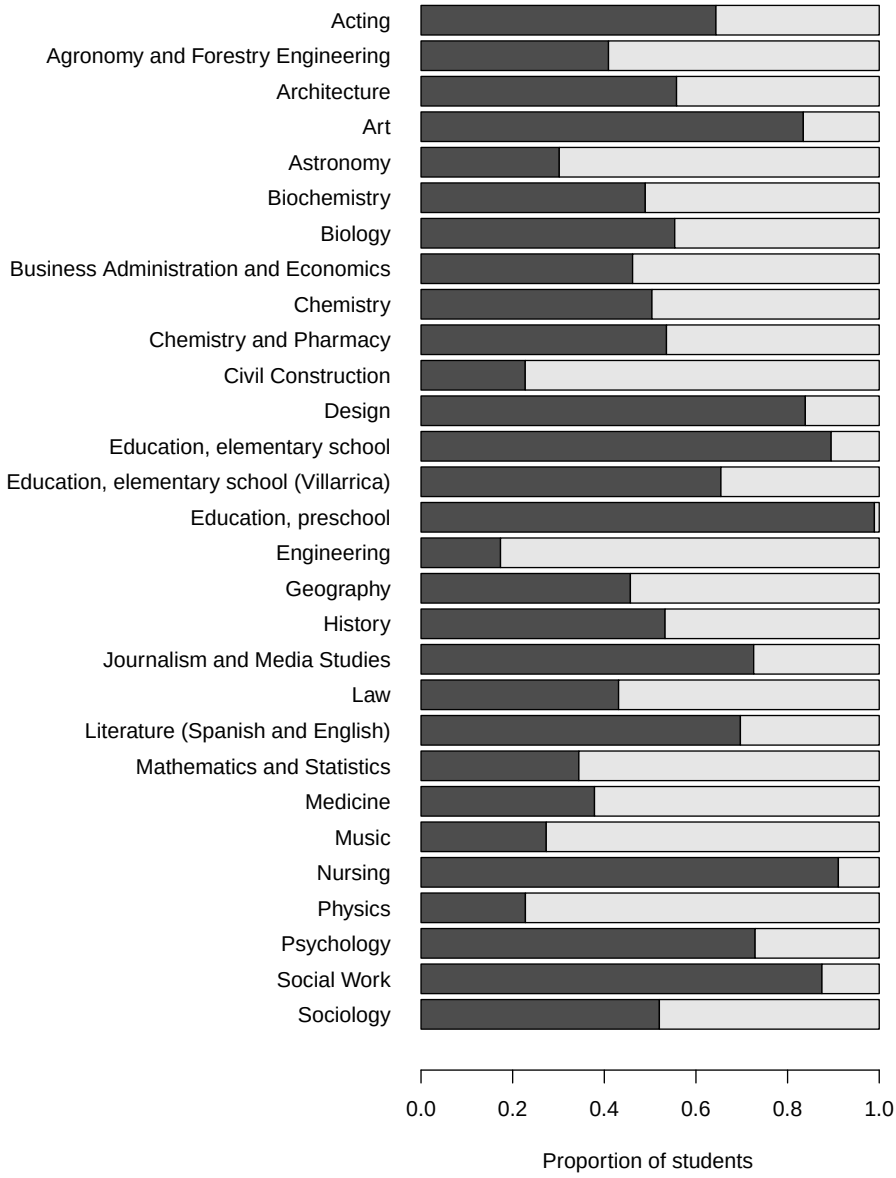

Figure S1: Distribution of students according to sex (lighter area: males).

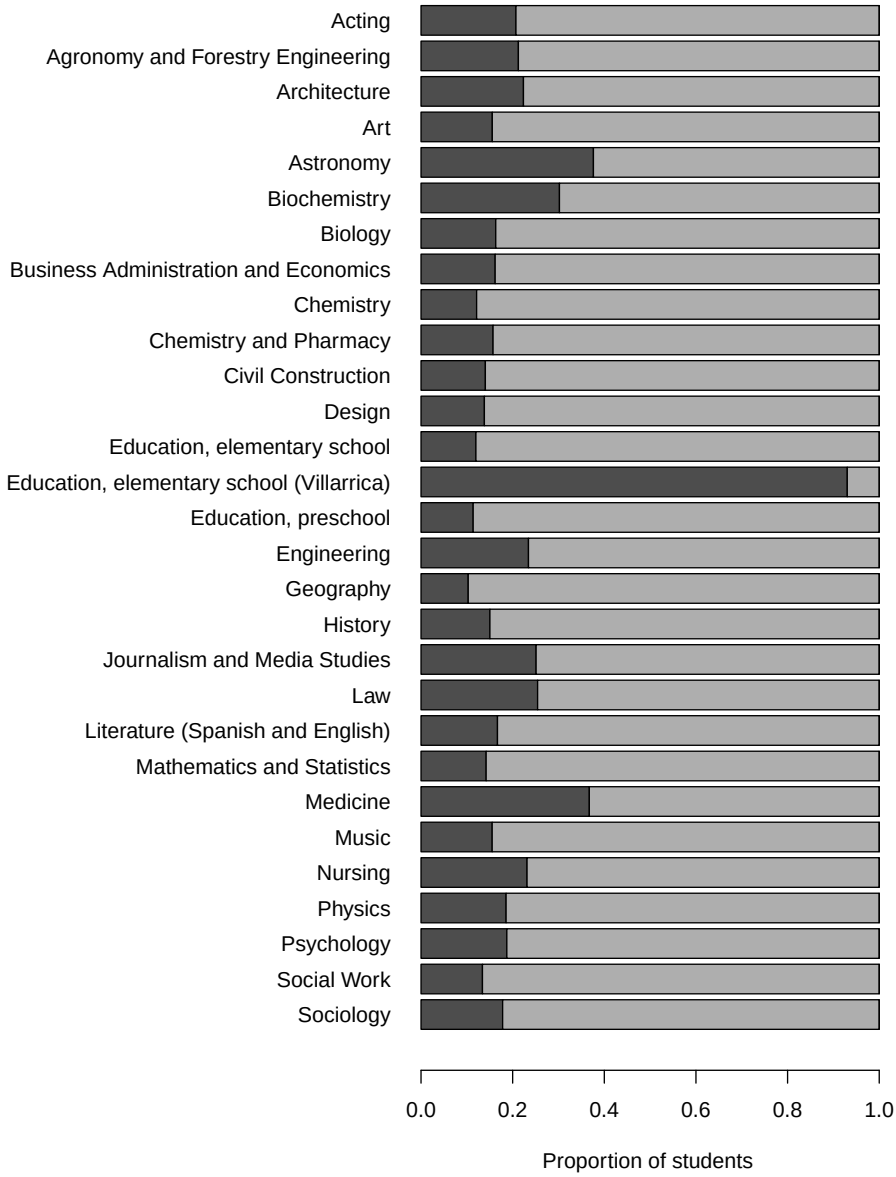

Figure S2: Distribution of students according to region of residence (lighter area: Metropolitan area).

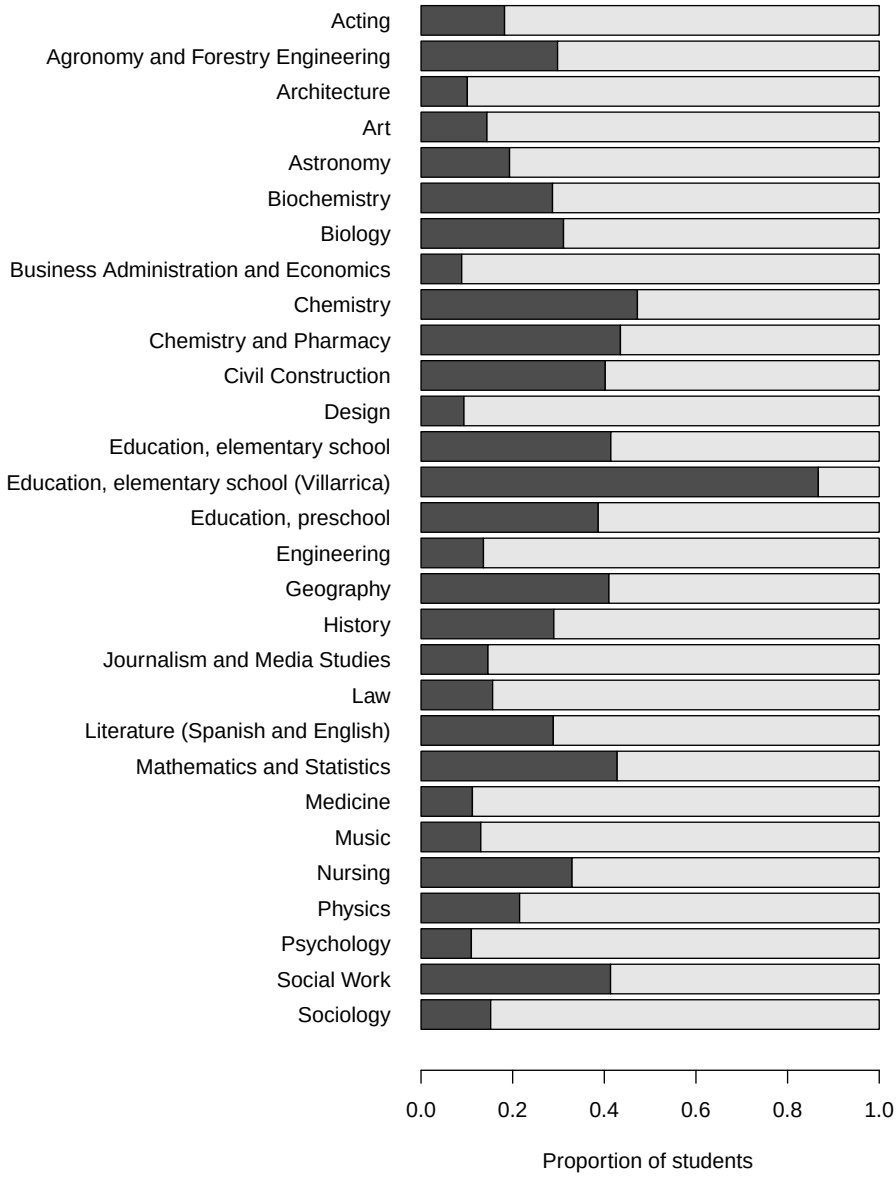

Figure S3: Distribution of students according to educational level of the parents (lighter area: students for which at least one of the parents has a university or technical degree).

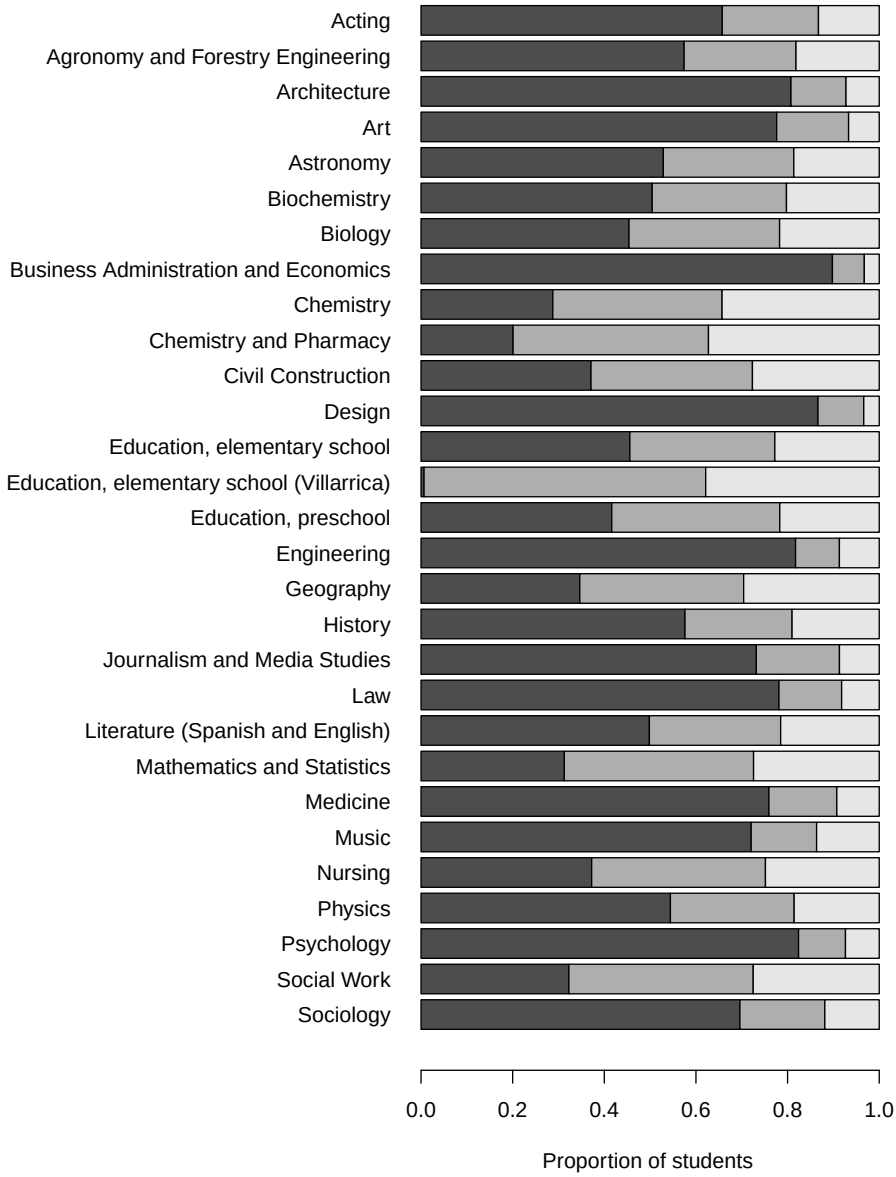

Figure S4: Distribution of students according to type of high school (from darkest to lightest, colored areas represent the proportion of students whose high school was: private, subsidized private and public, respectively).

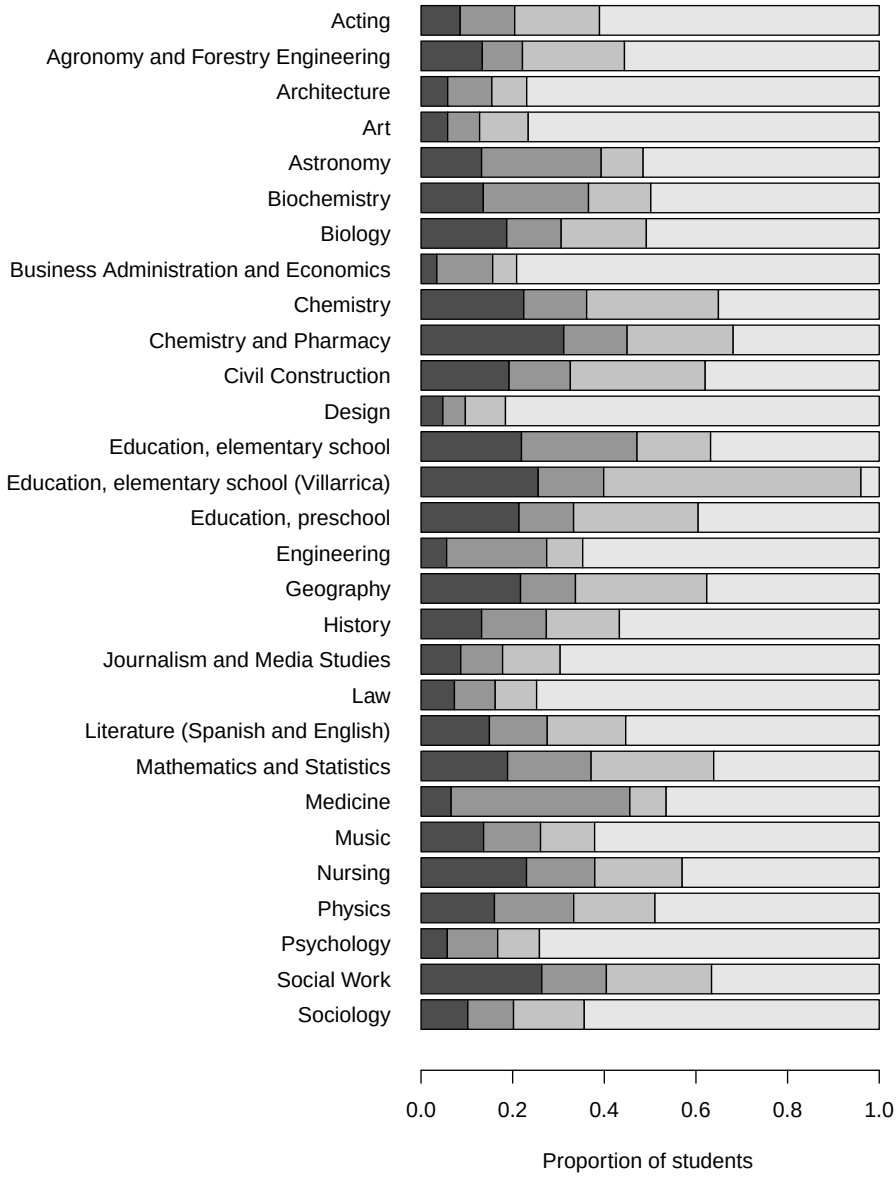

Figure S5: Distribution of students according to funding (from darkest to lightest, colored areas represent the proportion of students who have: scholarship and loan, scholarship only, loan only and no aid, respectively).

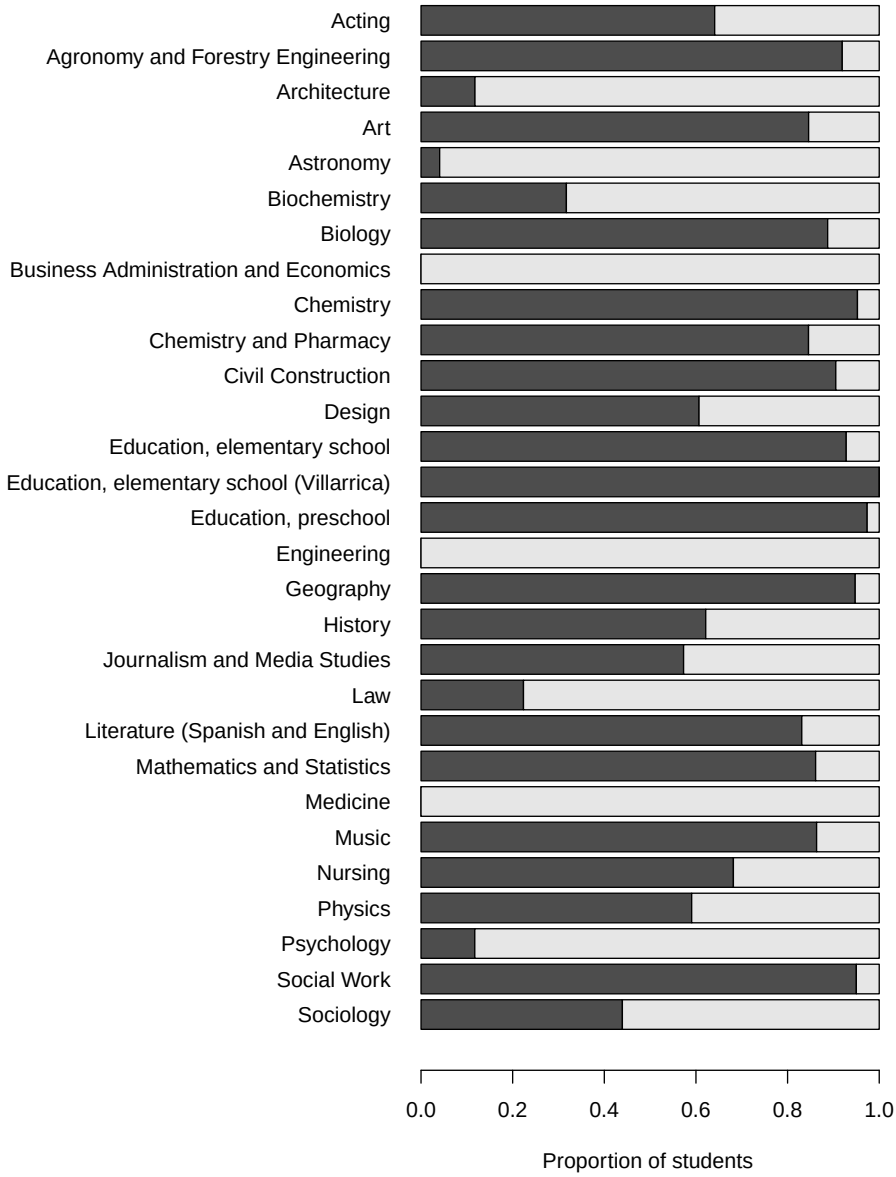

Figure S6: Distribution of students according to their selection score (lighter area: students with a selection score of 700 or more, which is typically considered a high value - the maximum possible score is 850). The minimum score required when applying to the PUC is 600 but exceptions apply for some education-related programmes.

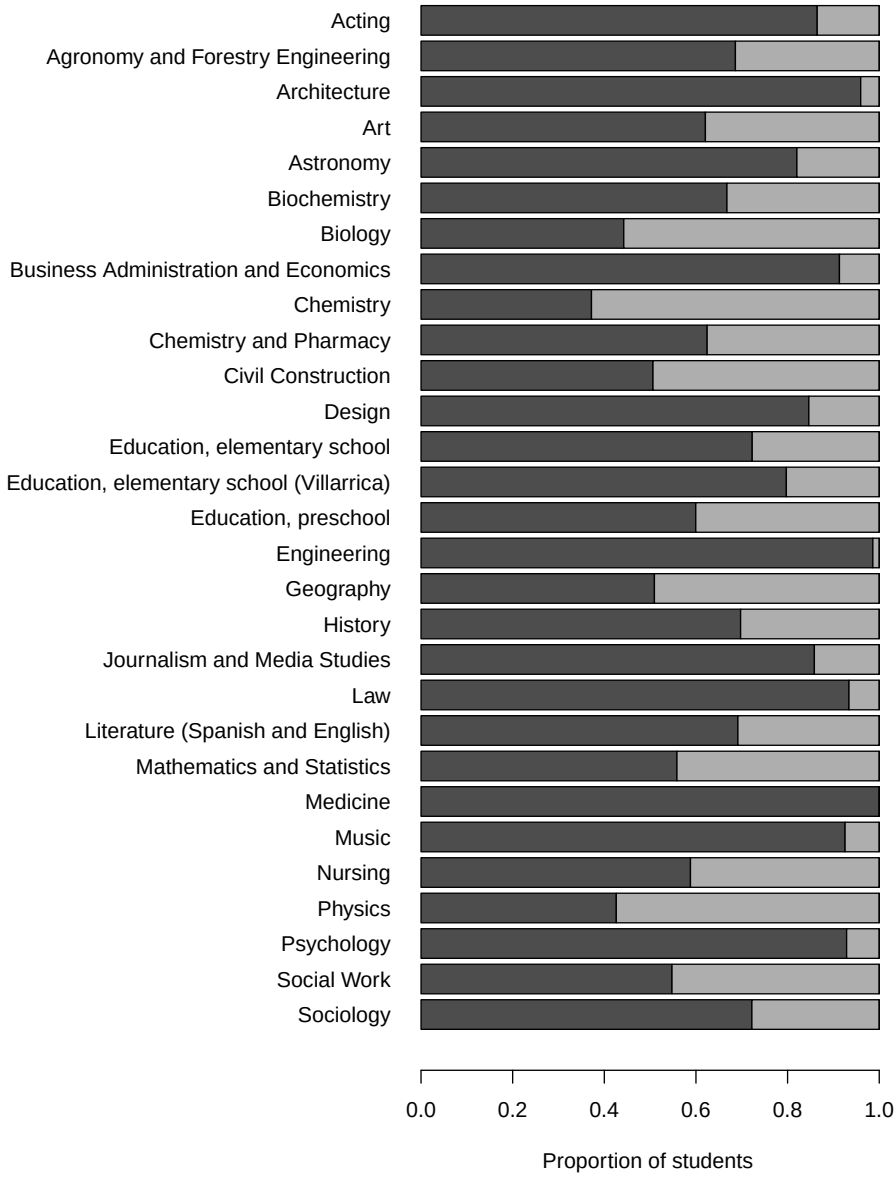

Figure S7: Distribution of students according to their application preference (lighter area: students who applied with second or lower preference to their current degree).

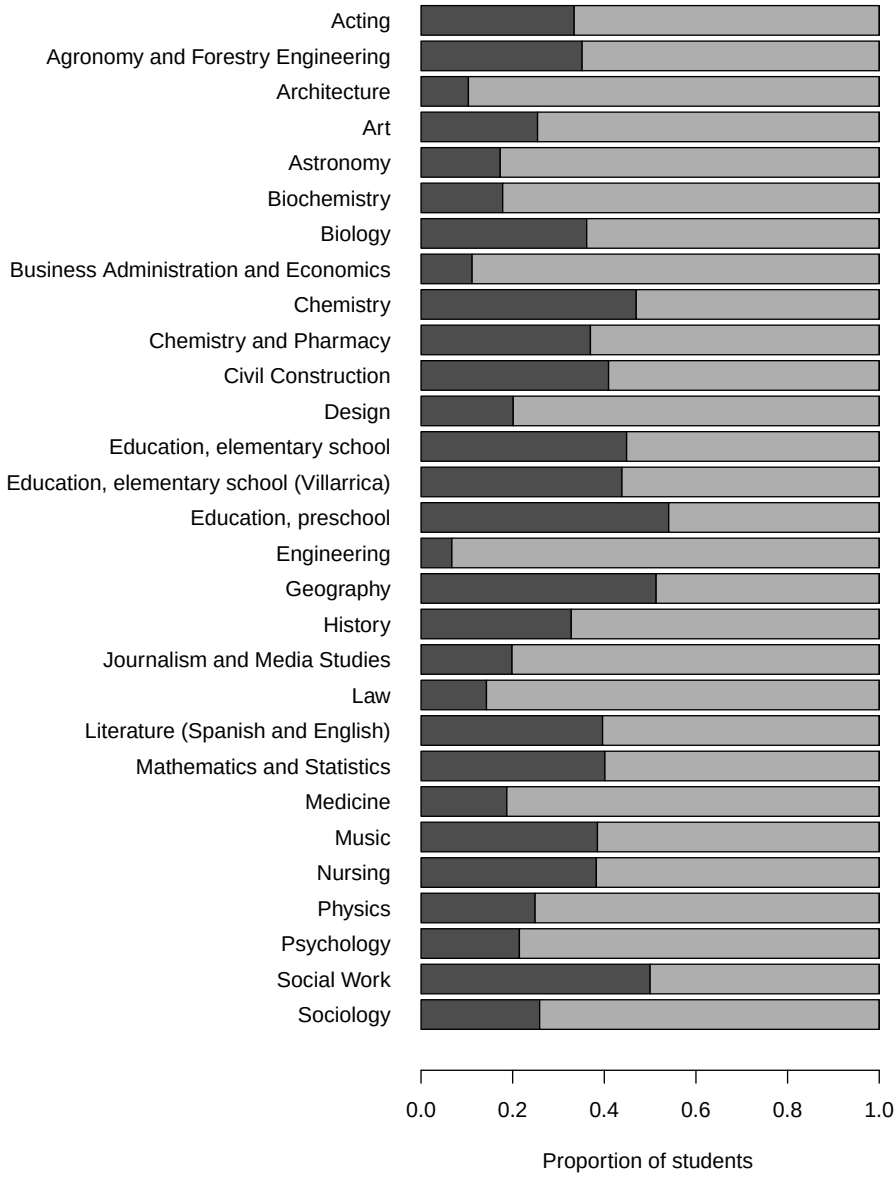

Figure S8: Distribution of students according to the gap between High School graduation and admission to PUC (lighter area: students with no gap).

B. Markov Chain Monte Carlo implementation

Bayesian inference for a multinomial (or binary) logistic regression is not straightforward. There is no conjugate prior and sampling from the posterior distribution is cumbersome [Holmes and Held, 2006]. Here we adapt the hierarchical structure proposed in Polson et al. [2013] in order to implement posterior inference for the discrete time competing risks model described in Section 3.1, under the prior described in Section 4.1.

For a binary logistic model with observations $\{y_{it} : i = 1, \ldots, n, t = 1, \ldots, t_i\}$, where $y_{it} = 1$ if the event is observed at time t for subject i, and $y_{it} = 0$ otherwise, the key result in Polson et al. [2013] is that

$$\frac{[\,e^{z_i'\beta^*}]^{y_{it}}}{e^{z_i'\beta^*}+1} \propto e^{\kappa_{it} z_i'\beta^*} \int_0^\infty \exp\{-\eta_{it}(z_i'\beta^*)^2/2\} f_{PG}(\eta_{it}|1,0)\, d\eta_{it}, \tag{S1}$$

where z_i is a vector of covariates associated with individual i, β^* is a vector of regression coefficients, $\kappa_{it} = y_{it} - 1/2$ and $f_{PG}(\cdot|a,b)$ denotes a Polya-Gamma density with parameters a and b [see Polson et al., 2013, for a description of the Polya-Gamma distribution and its properties]. In terms of the model in equation (2) of the paper, z_i includes x_i and the auxiliary binary variables linked to the δ_t's. Thus, $\beta^* = (\delta_1, \ldots, \delta_{t_0}, \beta')'$.

The result in (S1) can be used to construct a Gibbs sampling scheme for the multinomial logistic model along the lines of Holmes and Held [2006]. Now let $0, 1, \ldots, \mathcal{R}$ be the possible values for observations y_{it} associated with regression coefficients $\beta^*_{(1)}, \ldots, \beta^*_{(\mathcal{R})}$. Conditional on fixed values of $\beta^*_{(1)}, \ldots, \beta^*_{(r-1)}, \beta^*_{(r+1)}, \ldots, \beta^*_{(\mathcal{R})}$, the conditional likelihood function associated to $\beta^*_{(r)}$ is proportional to

$$\prod_{i=1}^{n}\prod_{t=1}^{t_i} \frac{\left[\exp\{z_i'\beta^*_{(r)} - C_{ir}\}\right]^{I(y_{it}=r)}}{1+\exp\{z_i'\beta^*_{(r)} - C_{ir}\}}, \quad \text{where} \quad C_{ir} = \log\left(1 + \sum_{r^*\neq r} \exp\{z_i'\beta^*_{(r^*)}\}\right). \tag{S2}$$

Assume a priori that $\beta^*_{(r)} \sim \text{Normal}_{t_0+k}(\mu_r, \Sigma_r)$, $r = 1, \ldots, \mathcal{R}$ and define $B^* = \left\{\beta^*_{(1)}, \ldots, \beta^*_{(\mathcal{R})}\right\}$. Using (S1) and (S2), a Gibbs sampler for the multinomial logistic model is defined through the following full conditionals for $r = 1, \ldots, \mathcal{R}$

$$\beta^*_{(r)}|\eta_r, \beta^*_{(1)}, \ldots, \beta^*_{(r-1)}, \beta^*_{(r+1)}, \ldots, \beta^*_{(\mathcal{R})}, y_{11} \ldots, y_{nt_n} \sim \text{Normal}_{t_0+k}(m_r, V_r), \tag{S3}$$

$$\eta_{itr}|B^* \sim \text{PG}(1, z_i'\beta^*_{(r)} - C_{ir}), \;\; t = 1, \ldots, t_i, i = 1, \ldots, n, \tag{S4}$$

defining $\mathbf{1}_t$ as a vector of t ones, $Z = (z_1 \otimes \mathbf{1}'_{t_1}, \ldots, z_n \otimes \mathbf{1}'_{t_n})'$, $\eta_r = (\eta_{11r}, \ldots, \eta_{nt_nr})'$, $D_r = \text{diag}\{\eta_r\}$, $V_r = (Z'D_rZ + \Sigma_r^{-1})^{-1}$, $m_r = V_r(Z'\kappa_r + \Sigma_r^{-1}\mu_r)$, $\kappa_r = (\kappa_{11r}, \ldots, \kappa_{nt_nr})'$ and $\kappa_{itr} = \text{I}_{\{y_{it}=r\}} - 1/2 + \eta_{itr}C_{ir}$ (where $\text{I}_A = 1$ if A is true, 0 otherwise). The previous algorithm (implemented in the `R` library `BayesLogit` by Polson et al) applies to the model in equation (6) of the paper with $\beta^*_{(r)} = (\delta'_{(r)}, \beta'_{(r)})'$, $\delta_{(r)} = (\delta_{r1}, \ldots, \delta_{rt_0})'$, $\beta_{(r)}$ being a vector of event type specific regression coefficients and defining z_i in terms of auxiliary binary variables related to the δ_{rt}'s and the observed covariates x_i. Extra steps are required to accommodate the prior adopted throughout the paper, which is a product of independent multivariate Cauchy and hyper-g prior components. Both components can be represented as a scale mixture of normal distributions (see equations (8) and (9) in the paper). Hence, conditional on $\Lambda_1, \ldots, \Lambda_{\mathcal{R}}, g_1, \ldots, g_{\mathcal{R}}$ (in (8) and (9)), the sampler above applies. To complement the sampler, at each iteration, Λ_r's and g_r's are updated using the full conditionals.

$$\Lambda_r|\delta_{(r)} \;\sim\; \text{Gamma}\left(\frac{t_0+1}{2}, \frac{\delta'_{(r)}\delta_{(r)}}{2\omega^2}\right), \qquad r = 1, \ldots, \mathcal{R}, \tag{S5}$$

$$g_r|\beta_{(r)} \;\propto\; g_r^{-k/2} \exp\left\{-\frac{\beta'_{(r)}X'X\beta_{(r)}}{2g_r}\right\} \pi(g_r), \qquad r = 1, \ldots, \mathcal{R}. \tag{S6}$$

An adaptive Metropolis-Hastings step [see Section 3 in Roberts and Rosenthal, 2009] is implemented for (S6).

The extension to a sampler over model space as explained in Subsection 4.3 involves drawing from the full conditionals

$$\pi(\gamma_j|\gamma_{-j},\delta,B,\Lambda,g) \propto L(\gamma) \times \left[\prod_{r=1}^{\mathcal{R}} \pi(\beta_{(r)}|g_r;X_\gamma)\right], \tag{S7}$$

with $\gamma_{-j} = \{\gamma_1 \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_{k^*}\}$, $\delta = \{\delta_{(1)}, \ldots, \delta_{(\mathcal{R})}\}$, $B = \{\beta_{(1)}, \ldots, \beta_{(\mathcal{R})}\}$, $\Lambda = \{\lambda_1, \ldots, \lambda_{\mathcal{R}}\}$ and $g = \{g_1, \ldots, g_{\mathcal{R}}\}$. In (S7), $L(\gamma)$ represents the likelihood function associated with the model in (6) and the covariate configuration induced by γ. We use a Metropolis-Hastings step with "add/remove" proposals (i.e. propose $\gamma_j = 1$ when the current value is equal to 0 and propose $\gamma_j = 0$ when the current value is equal to 1). To facilitate a faster exploration of the model space, we also use "forced moves" where, every 20 iterations, we randomly choose one of the γ_j's and directly sample its value from a Bernoulli(0.5) distribution (only one of the γ_j's is updated in this forced update).

C. Convergence of MCMC chains and prior sensitivity

In this section we display some graphical summaries to visualise the convergence of the MCMC chains used throughout in Section 5 of the manuscript. In addition, we provide results when using different priors for the regression coefficients — based on the Zellner [1986] g-prior and the Benchmark-Beta hyper prior for g used in Ley and Steel [2012]. Figures S9, S10 and S11 display cumulative estimates of MPPIs for three priors, indicating both good convergence and similarity of results between these priors. Figure S12 presents the traceplots of the regression coefficients, suggesting good mixing and convergence. Traceplots for the other events and programmes lead to the same conclusions.

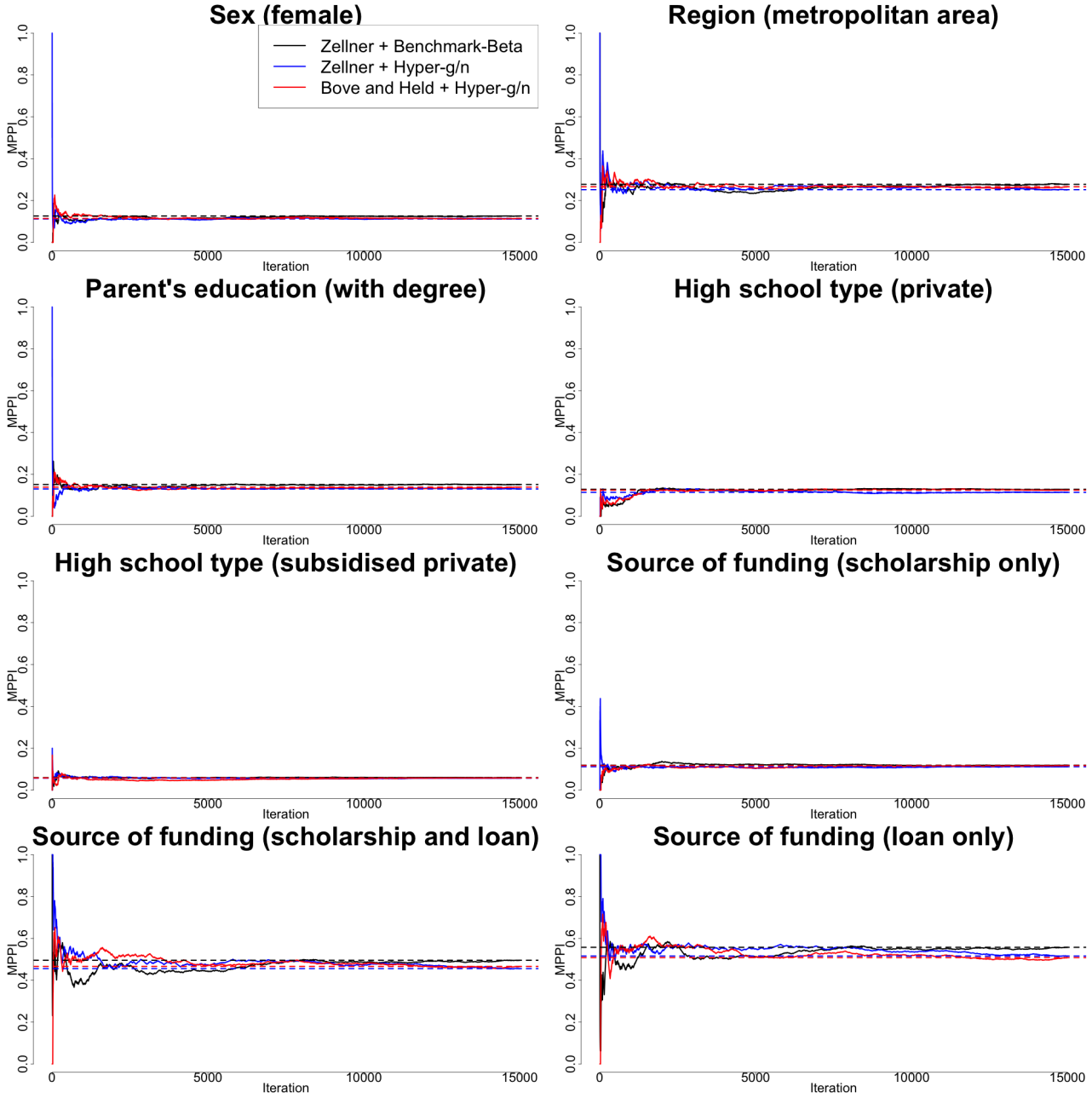

Figure S9: Chemistry students. Cumulative estimates of MPPIs under three different priors. Red lines correspond to the prior used for the results displayed in Section 5. Dotted lines indicate the final estimates.

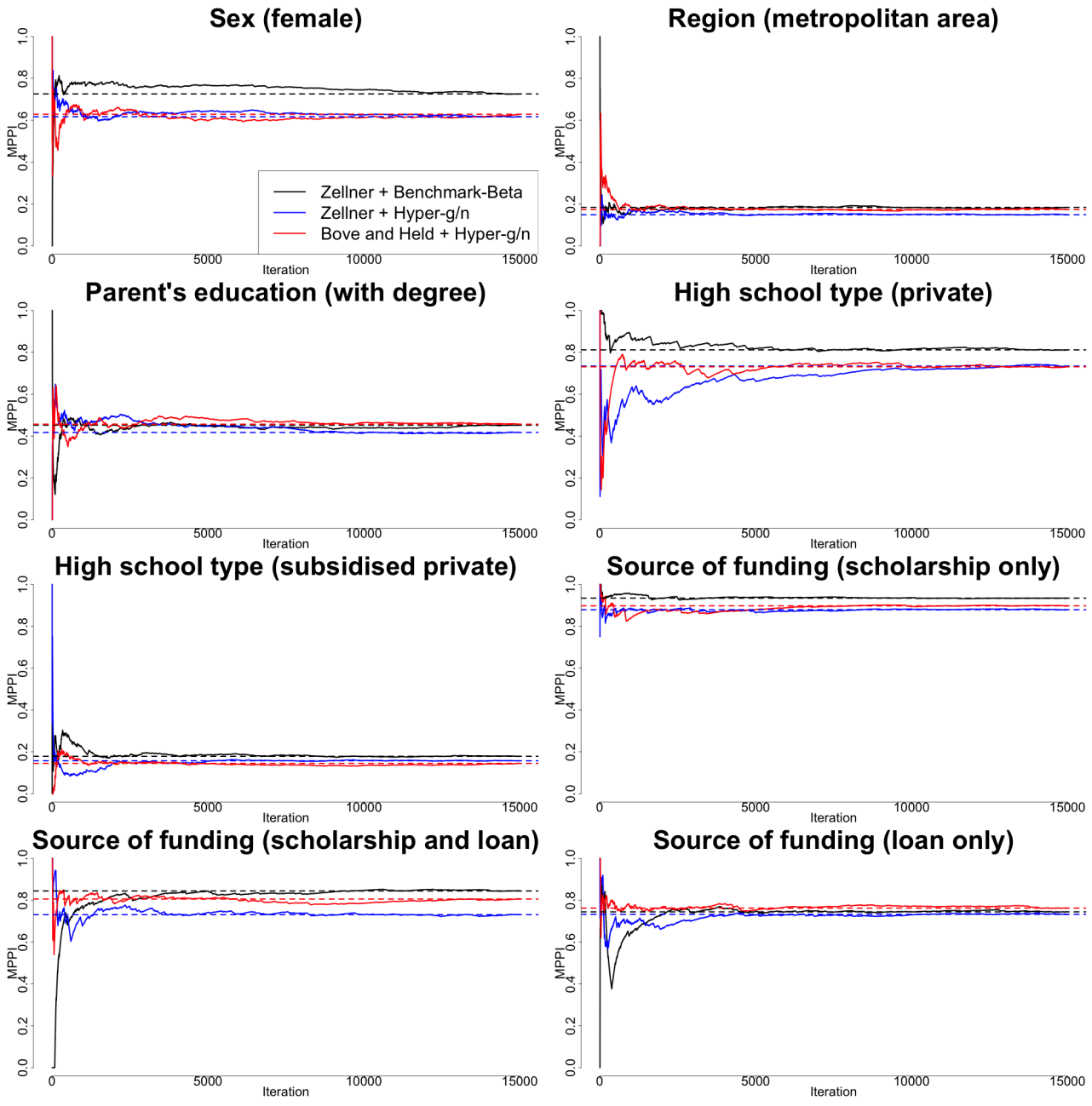

Figure S10: Mathematics and Statistics students. Cumulative estimates of MPPIs under three different priors. Red lines correspond to the prior used for the results displayed in Section 5. Dotted lines indicate the final estimates.

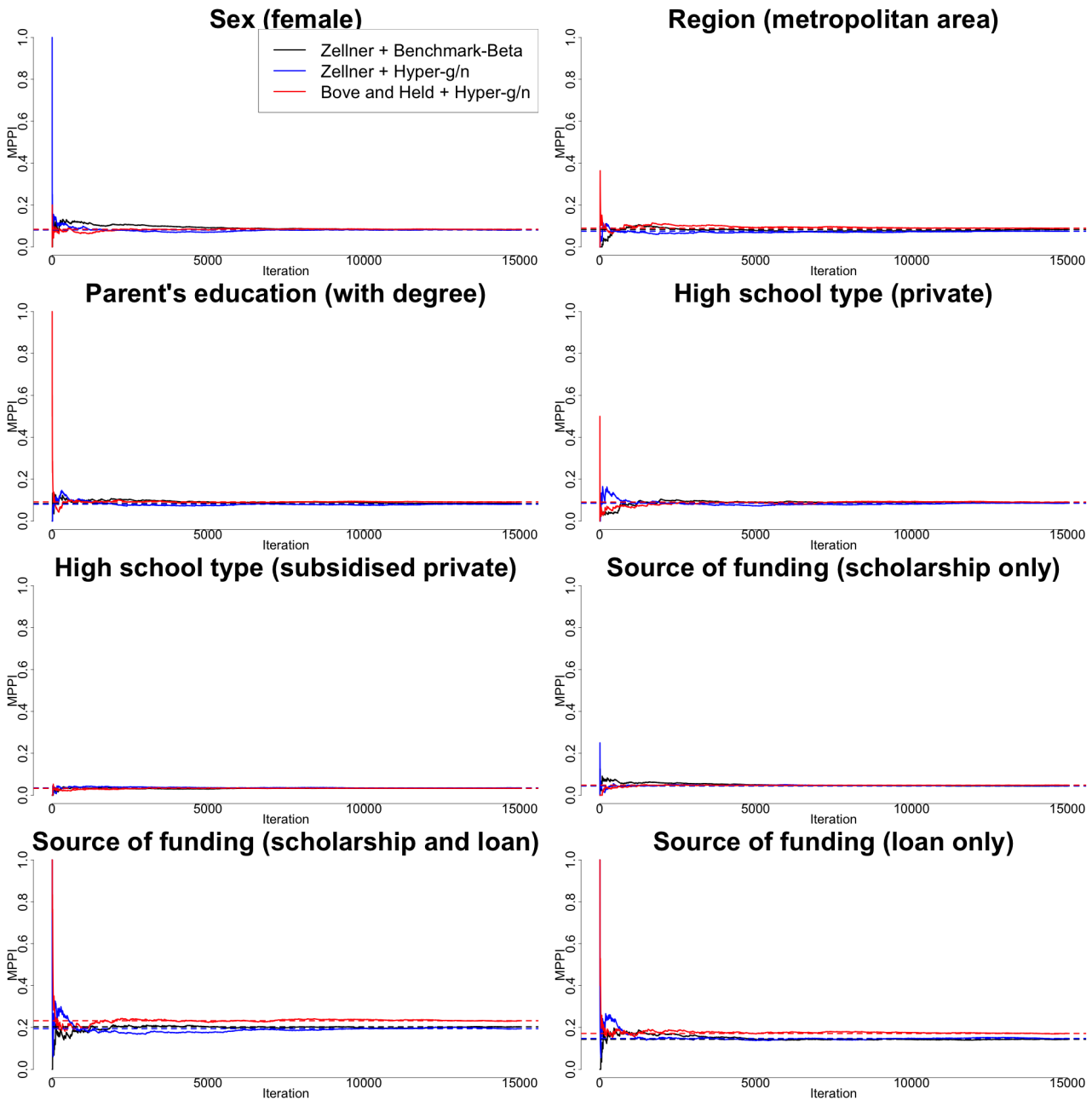

Figure S11: Physics students. Cumulative estimates of MPPIs under three different priors. Red lines correspond to the prior used for the results displayed in Section 5. Dotted lines indicate the final estimates.

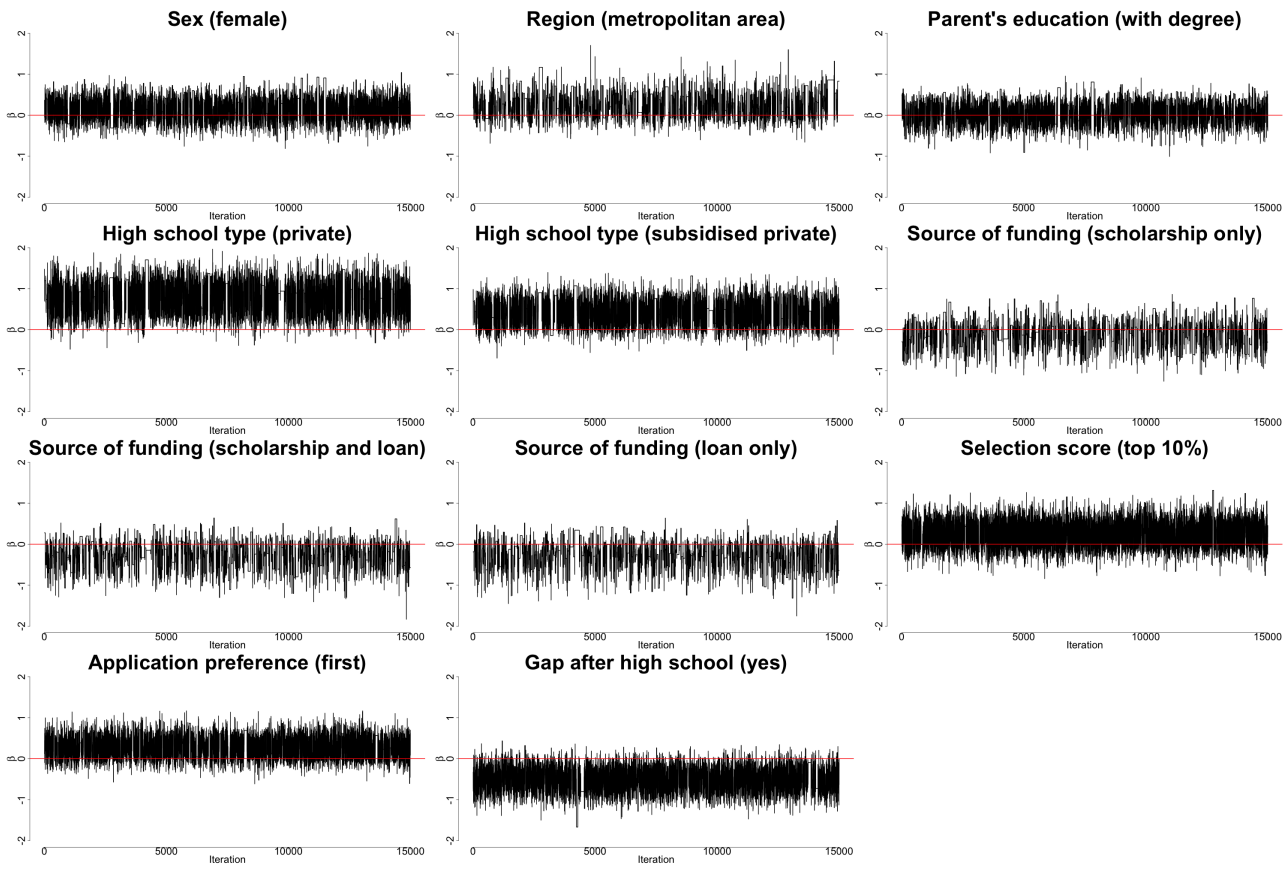

Figure S12: Mathematics and Statistics students, graduation event. Traceplots for regression coefficients under the prior described in Section 4.

D. Documentation for the R code

Bayesian inference is implemented through the Markov chain Monte Carlo (MCMC) sampler and priors described in Section 4. Inference was implemented in R[1] version 3.0.1. The code is freely available at
http://www.warwick.ac.uk/go/msteel/steel_homepage/software/university_codes.zip

This includes the MCMC algorithm and the Bayesian variable selection methods described in the paper. Before using this code, the following libraries must be installed in R: BayesLogit, MASS, mvtnorm, Matrix and compiler. All of these are freely available from standard R repositories and are loaded in R when "Internal_Codes.R" is executed. The last two libraries speed up matrix calculations and the "for" loops, respectively. Table S2 explains the notation used throughout the code. The implementation was based on three-dimensional arrays, with the third dimension representing the event type.

Table S2: Notation used throughout the R code

Variable name	Description
CATEGORIES	Number of possible outcomes, excluding censoring (equal to 3 for the PUC dataset)
n	Total number of students
nt	Total number of multinomial outcomes (i.e. $\sum t_i$ across all students)
t0	Number of period-specific baseline log-odds coefficients δ_{rt} (for each cause)
k	Number of effects (t0 + number of covariate effects)
Y	Vector of outcomes. Dimension: n $\times 1$
X	Design matrix, including the binary indicators (denoted by Z in the paper). Dimension: n $\times$ k
X.Period	Design matrix related to period-specific baseline log-odds coefficients δ_{rt}'s only. Dimension nt $\times$ t0
inc	Vector containing covariate indicators $\gamma_1, \ldots, \gamma_{k*}$
beta	β^* (period-specific baseline log-odds and covariates effects for all event types)
mean.beta	Prior mean for $\{\beta_1^*, \ldots, \beta_{\mathcal{R}}^*\}$. Dimension: $1\times$ k $\times$ CATEGORIES
prec.delta	Precision matrix for $(\delta_{r1}, \ldots, \delta_{rt_0})'$. Dimension: t0 $\times$ t0
df.delta	Degrees of freedom for prior of $(\delta_{r1}, \ldots, \delta_{rt_0})'$. Default value: 1
fix.g	If TRUE, $g_1, \ldots, g_{\mathcal{R}}$ are fixed. Default value: FALSE
prior	Choice of hyper prior for g_r: *(i)* Benchmark-Beta or *(ii)* Hyper-g/n [see Ley and Steel, 2012]
N	Total number of MCMC iterations
thin	Thinning period for MCMC algorithm
burn	Burn-in period for MCMC algorithm
beta0	Starting value for $\{\beta_1^*, \ldots, \beta_{\mathcal{R}}^*\}$. Dimension: $1\times$ k $\times$ CATEGORIES
logg0	Starting value (log-scale) of $\{g_1, \ldots, g_{\mathcal{R}}\}$. Dimension: $1\times$ CATEGORIES
ls.g0	Starting value (log-scale) of the adaptive proposal variance used in Metropolis-Hastings updates of $\log(g_1), \ldots, \log(g_{\mathcal{R}})$. Dimension: $1\times$ CATEGORIES
ar	Optimal acceptance rate for the adaptive Metropolis-Hastings updates. Default value: 0.44
ncov	Indicates how many potential covariates are included in the design matrix (might not match the number of columns due to categorical covariates with more than two levels) Default value: 8 (as in the PUC dataset)
include	Vector indicating which covariates from the design matrix are to be included in the model If missing a sampler over the model space will be run. Default: NULL
gamma	γ (covariate inclusion indicators)

The code is separated into two files. The file "Internal_Codes.R" contains functions that are required for the implementation but the user is not expected to directly interact with these. These functions must be loaded in R before doing any calculations. The main function — used to run the MCMC algorithm — is contained in

[1]Copyright (C) The R Foundation for Statistical Computing.

the file "User_Codes.R". In the following, a short description of this function is provided. Its use is illustrated in the file "Example.R" using a simulated dataset.

- `MCMC.MLOG`. Adaptive Metropolis-within-Gibbs algorithm [Roberts and Rosenthal, 2009] for the competing risks Proportional Odds model used throughout the paper. If not fixed, univariate Gaussian random walk proposals are implemented for $\log(g_1), \ldots, \log(g_{\mathcal{R}})$. Arguments: `N`, `thin`, `Y`, `X`, `t0`, `beta0`, `mean.beta`, `prec.delta`, `df.delta`, `logg0`, `ls.g0`, `prior`, `ar`, `fix.g`, `ncov` and `include`. The output is a list containing the following elements: `beta` MCMC sample of β^* (array of dimension (`N/thin+1`) $\times$ `k` $\times$ `CATEGORIES`), `gamma` MCMC sample of γ (matrix of dimension (`N/thin+1`) $\times$ `k`, `logg` MCMC sample of $\log(g_1), \ldots, \log(g_{\mathcal{R}})$ (dimension (`N/thin+1`) $\times$ `CATEGORIES`), `ls.g` stored values for the logarithm of the proposal variances for $\log(g_1), \ldots, \log(g_{\mathcal{R}})$ (dimension (`N/thin+1`) $\times$ `CATEGORIES`) and `lambda` MCMC sample of $\lambda_1, \ldots, \lambda_{\mathcal{R}})$, which are defined in equation (9) in the paper (dimension (`N/thin+1`) $\times$ `CATEGORIES`). Recording `ls.g` allows the user to evaluate if the adaptive variances have been stabilized. Overall acceptance rates are printed in the `R` console (if appropriate). This value should be close to the optimal acceptance rate `ar`.

E. Assessment of proportional odds assumption

Our modelling approach uses the proportional odds (PO) assumption. To approximately assess this assumption, this Section displays non-parametric estimates of the ratio between the cause-specific hazard rates and the hazard associated with no event, stratified by the levels of each covariate (irrespective of the value of the remaining covariates). As defined by equation (4), if the PO assumption holds, these should be proportional. While the proportional odds assumption does not seem too unreasonable for some covariates and degree programmes (e.g. preference and gap covariates for the Mathematics and Statistics data), this is less obvious in other cases. This is perhaps not that critical for those covariates that are not robustly associated with the analyzed outcomes (e.g. for Physics students, where the null model concentrates more than 50% of the posterior probability). In addition, this simple stratification per covariate does not account for imbalance in the other covariates (and we would need a very large data set indeed to be able to resolve this), which could substantially affect the results.

Potential reasons for deviations from PO are unobserved confounders and time-varying covariate effects (e.g. if some of the variables recorded at admission might have a diminishing effect throughout time). In such cases, possible solutions would be to keep the proportional odds specification but to add an interaction effect between time and covariates (e.g. different effect magnitudes during the first year of admission) and to incorporate random effects in order to account for unobserved sources of heterogeneity. This will be investigated in future follow-up analyses. However, it should be borne in mind that inference in such more general models might well be challenging with the available sample sizes.

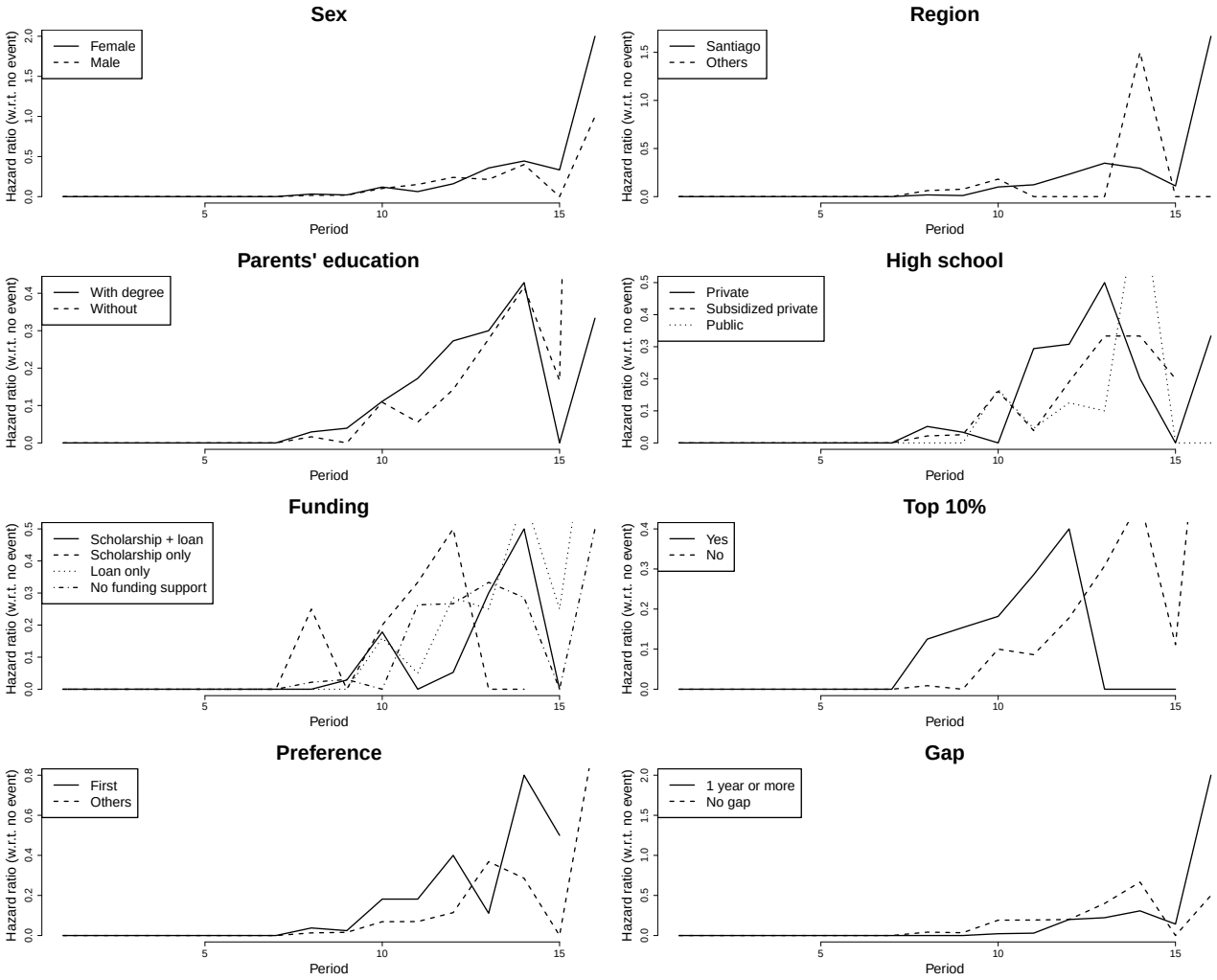

Figure S13: Chemistry students, graduation events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

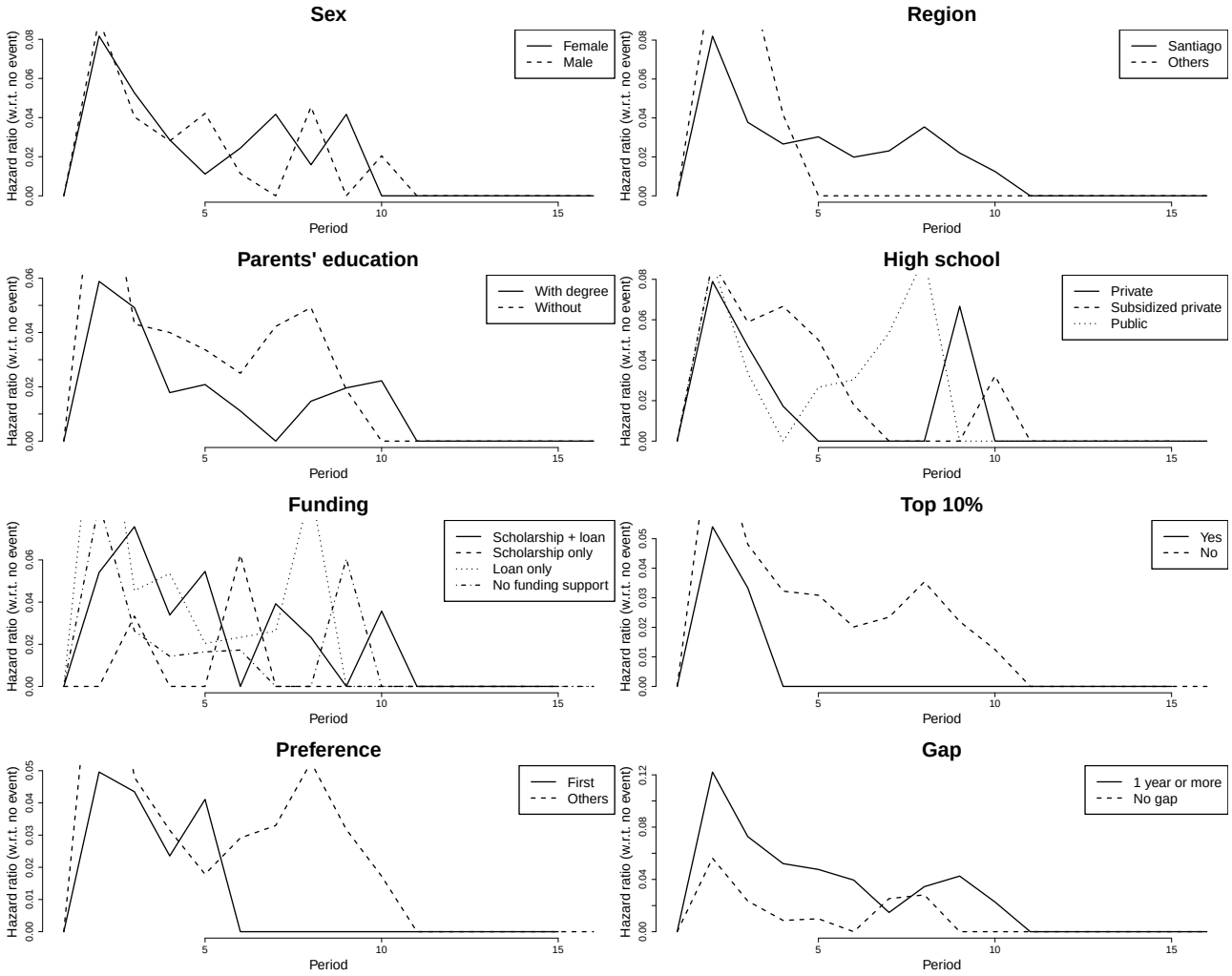

Figure S14: Chemistry students, involuntary dropout events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

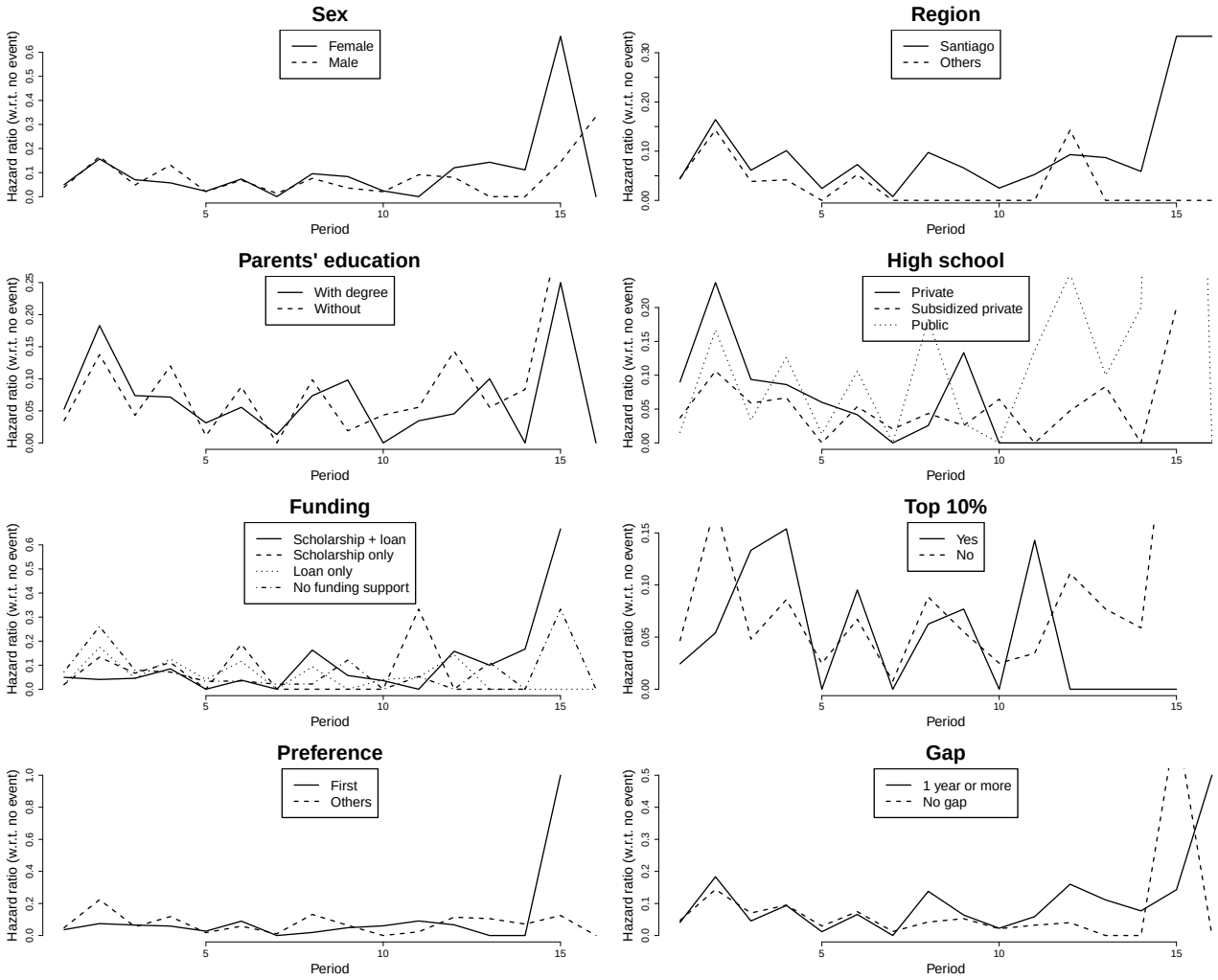

Figure S15: Chemistry students, voluntary dropout events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

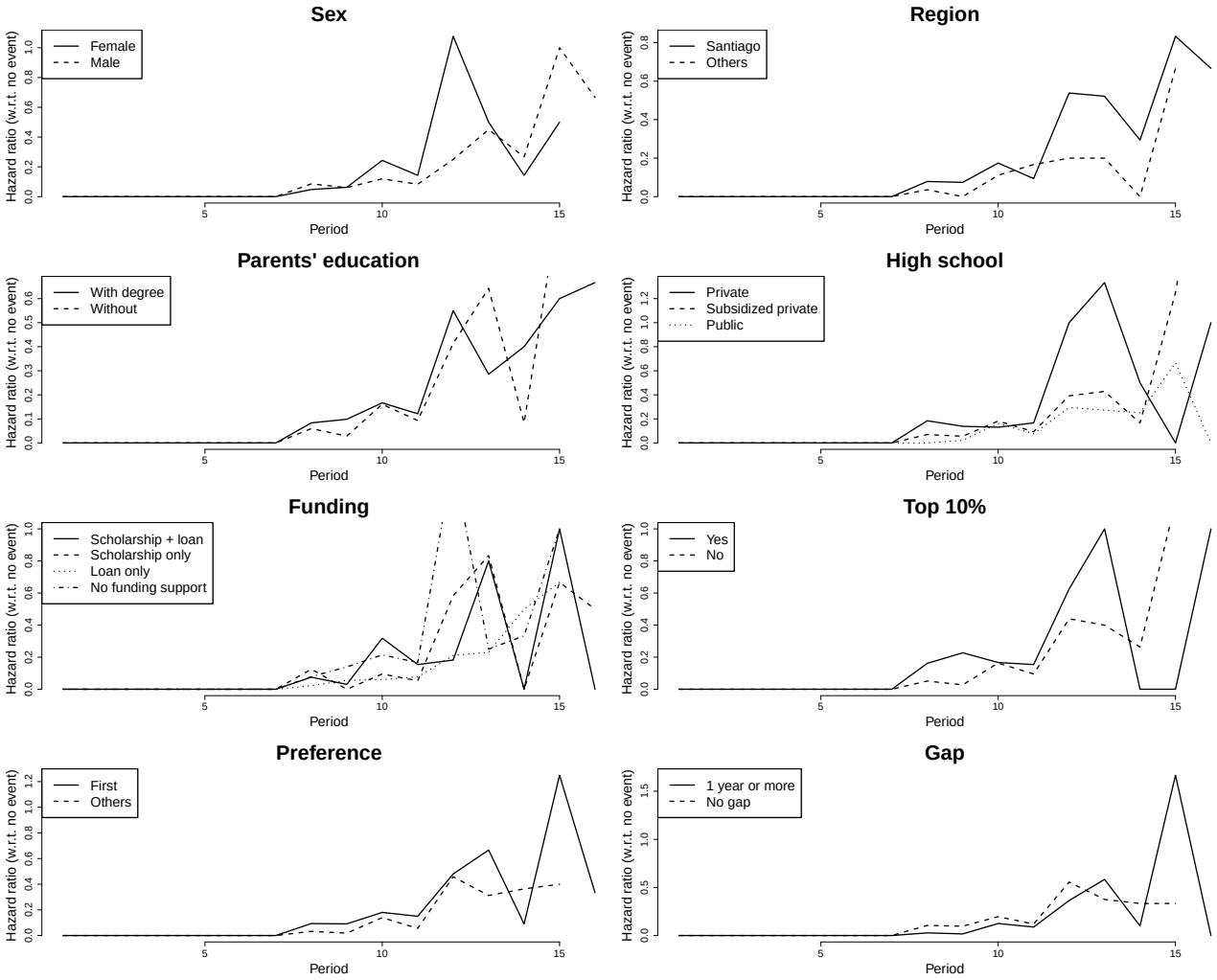

Figure S16: Mathematics and Statistics, graduation events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

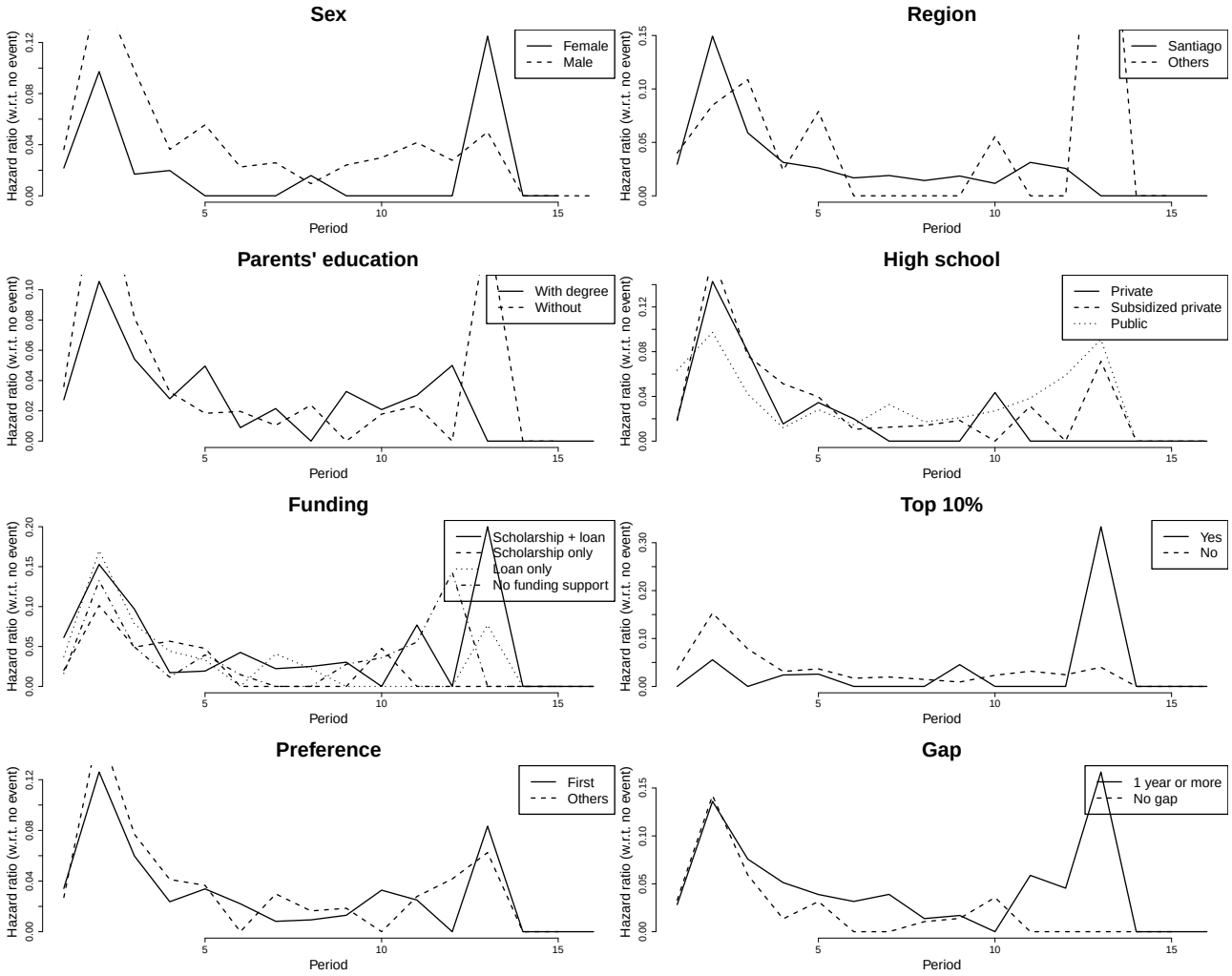

Figure S17: Mathematics and Statistics, involuntary dropout events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

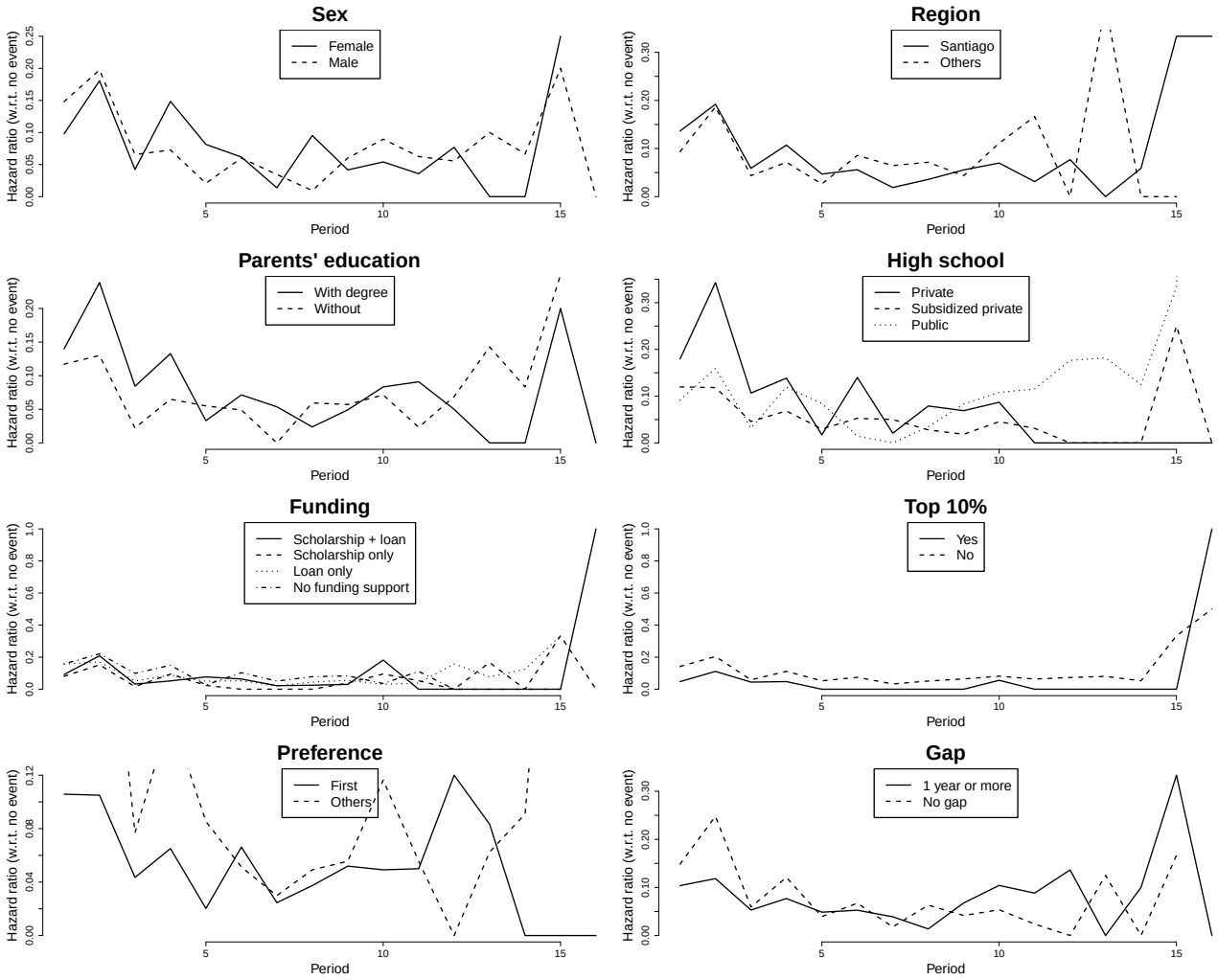

Figure S18: Mathematics and Statistics students, voluntary dropout events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

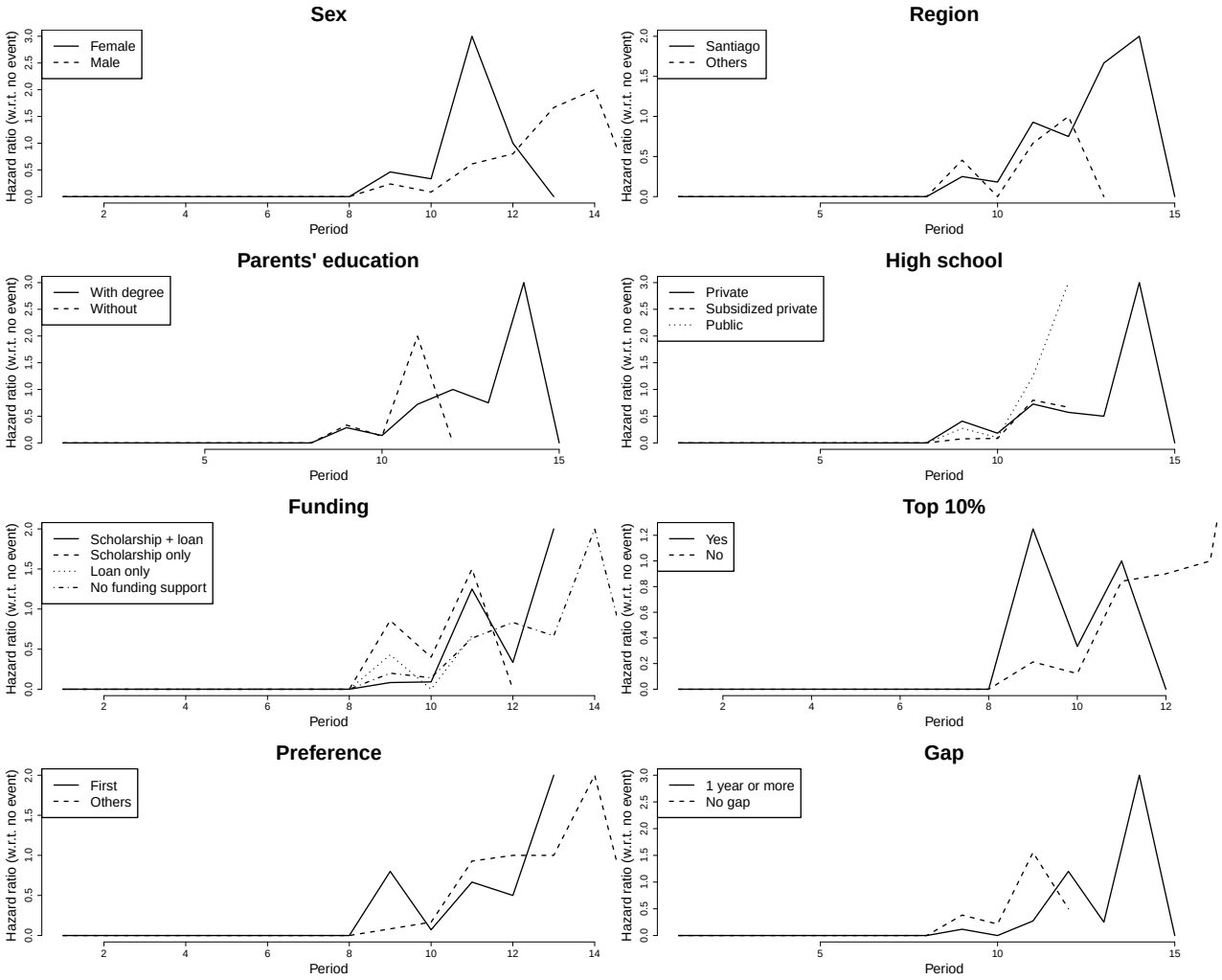

Figure S19: Physics, graduation events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

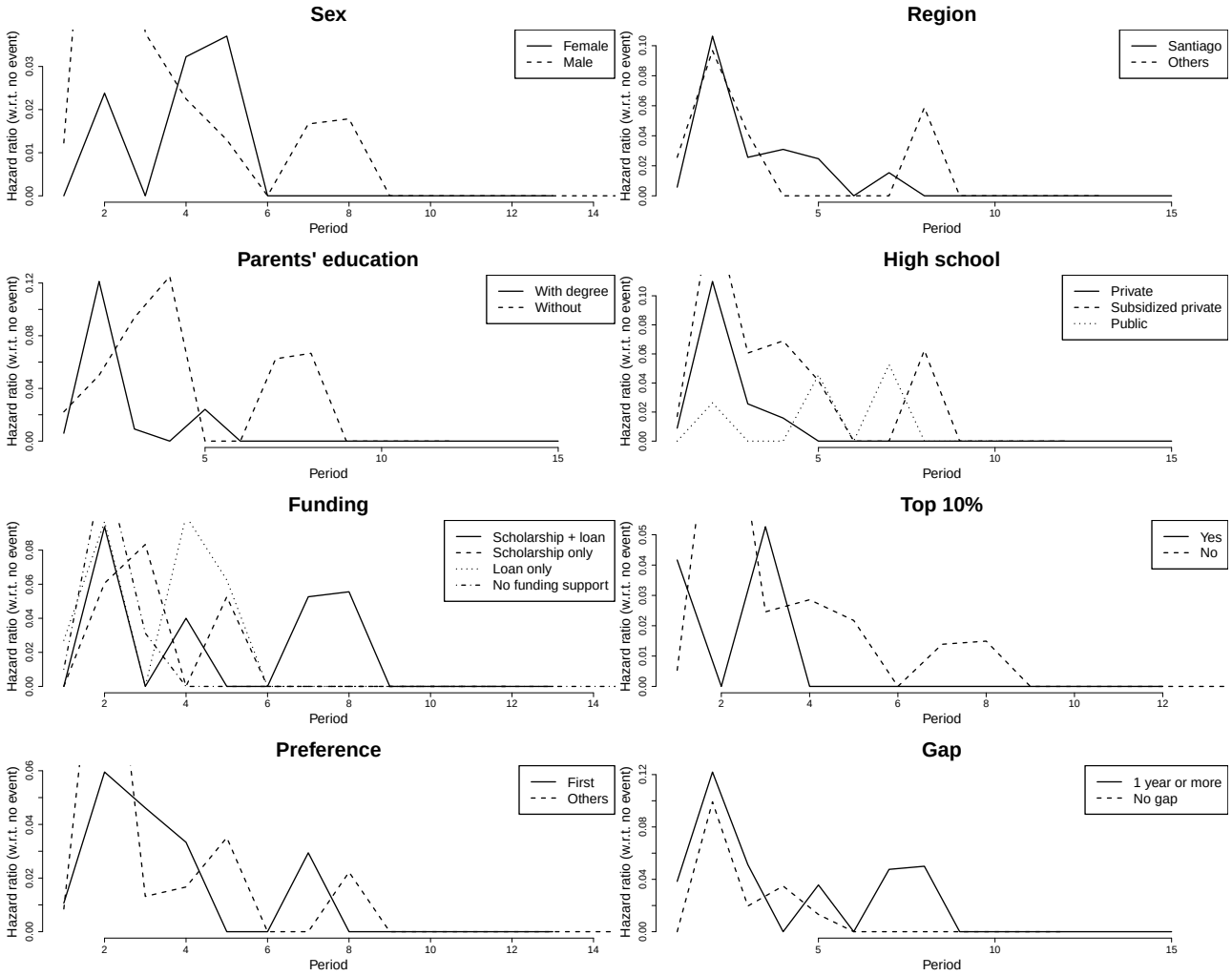

Figure S20: Physics, involuntary dropout events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

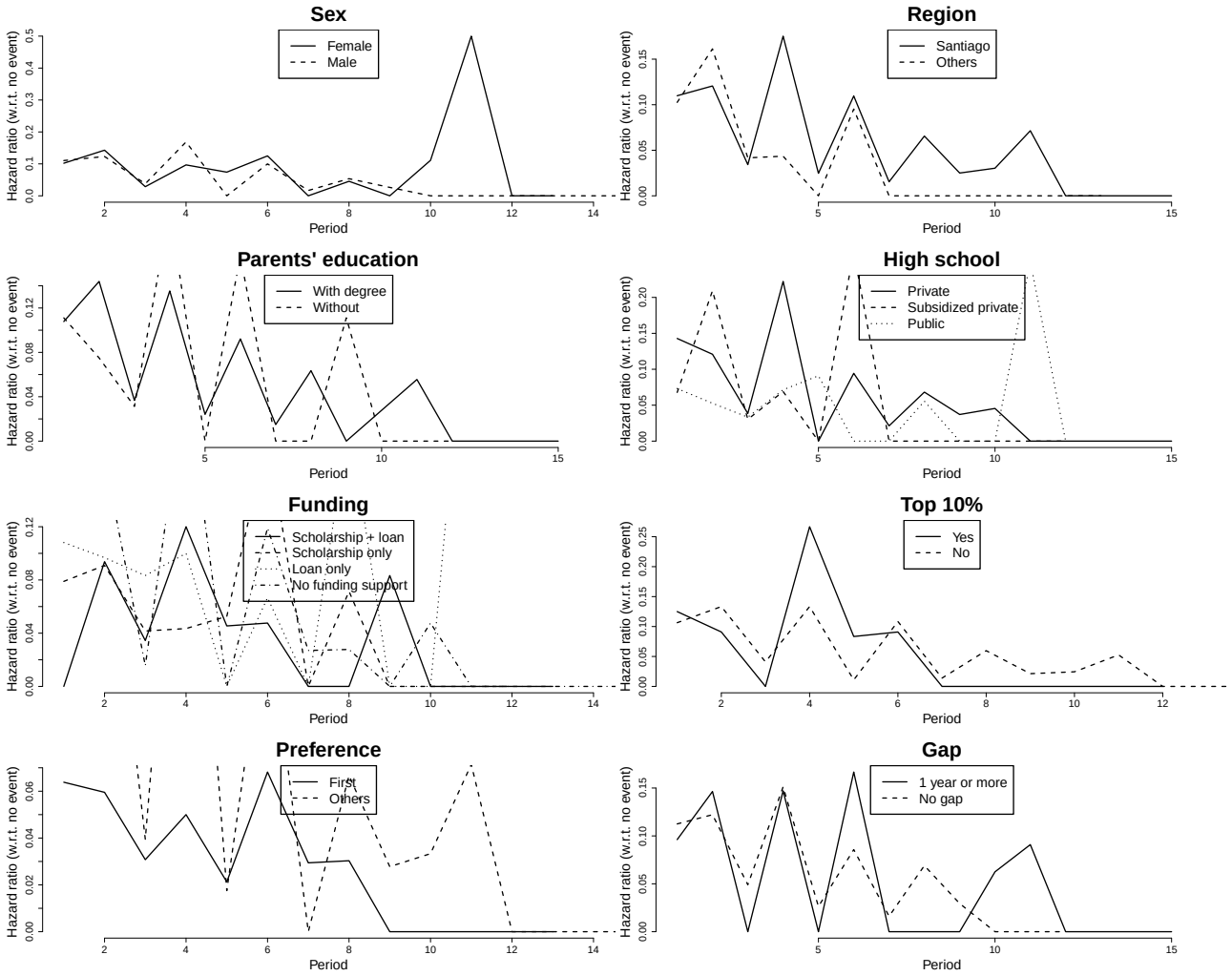

Figure S21: Physics, voluntary dropout events. Non-parametric estimates for hazard ratio with respect to no event ($h(r,t)/h(0,t)$) stratified according to the levels of the available covariates.

References

H. Cho, J. G. Ibrahim, D. Sinha, and H. Zhu. Bayesian case influence diagnostics for survival models. *Biometrics*, 65:116–124, 2009.

S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153–160, 1979.

C.C. Holmes and L. Held. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168, 2006.

E. Ley and M.F.J. Steel. Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics*, 171:251–266, 2012.

X.L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11: 552–586, 2002.

X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica*, 6:831–860, 1996.

N. Polson, J. Scott, and J. Windle. Bayesian inference for logistic models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108:1339–1349, 2013.

G.O. Roberts and J.S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367, 2009.

D.J. Spiegelhalter, N.G. Best, B.P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B*, 64:583–640, 2002.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, pages 233–243, North-Holland: Amsterdam, 1986.