Survival model for responsive follow-up in a 12-week survey

Margherita Zuppardo, Statistics Iceland, margherita.zuppardo@hagstofa.is

Abstract.

Time-to-event modelling, widely used in biostatistics and epidemiology, can also be useful during survey fieldwork.
This paper shows how a Cox proportional hazards model [1] was used in the Icelandic 2026 European Health Interview Survey (EHIS) [2] to forecast response by a fixed closing date.

The survey was organised in12 weekly invitation waves. The model used early fieldwork outcomes and register-based auxiliary variables to estimate each individual's probability of completing the questionnaire before closure. These predictions were used to prioritise limited phone follow-up, focusing on cases with low predicted response probability.

The approach provides a practical example of responsive survey design in official statistics and shows how survival modelling can support fieldwork monitoring when response time is unequal across sample members

Keywords: survey design, nonresponse, forecasting, R

1. EHIS and the 2026 implementation in Iceland

The European Health Interview Survey (EHIS) is conducted every six years across Europe. The survey provides harmonised statistics on health status, health behaviour, and the use and availability of health services. For the 2025 Icelandic implementation, data collection has been organised in 12 weekly invitation waves, as shown in Fig. 1. Data are collected first by web questionnaire (CAWI), supported by reminder one week later and, when necessary, follow-up by telephone (CATI).

A practical complication is that, since the survey closed on a fixed date for everyone selected, individuals invited in later waves have less time available to respond. In addition, previous survey experience suggests that response propensity differs across demographic groups.

Since the capacity for follow-up by phone is limited, a prioritisation was required. We built a model to answer the operational question: which individuals are less likely to respond before the fixed closing date, under the standard fieldwork process? These should be prioritised for phone follow-up.

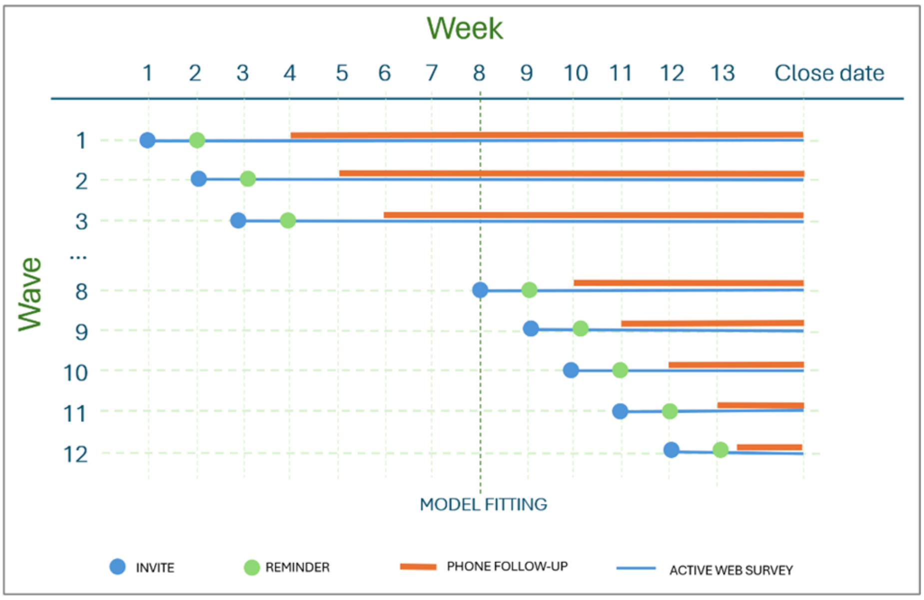

Figure 1: Fieldwork timeline for the 2026 Icelandic EHIS. Individuals were invited in 12 weekly waves. Each invitation was followed by a reminder one week later. Cases remained open for web response until the common closing date, while phone follow-up was conducted during the later part of fieldwork. The response model was fitted after wave 8 and used to support follow-up prioritisation.

Survival model for response forecasting

The problem described in Section 1 can be expressed as a time-to-event problem. For each sampled individual, the event of interest is the completion of the EHIS questionnaire. The time origin is the date of invitation for each individual, and the event time is the number of days from invitation to questionnaire completion.

A key feature of this problem is that non-response is not known until the survey has closed. If an individual has not completed the questionnaire by time *t*, this does not mean that they will never respond. It only means that the event has not yet been observed. Such observations are therefore right-censored: the individual is known not to have responded up to a certain time, but their final outcome remains unknown at that point. Right-censoring is common in medical and epidemiological studies (often describing events such as death, recovery, relapse, or disease onset), where individuals may leave a study or remain event-free at the end of follow-up.

2.1 Cox model

A Cox proportional hazards model [1]:, which for a given sampled individual *i* is expressed as

$$\boldsymbol{h_i(t) = h_0(t)\exp(\beta^T X_i)}$$

was fitted using the *survival* package in R. where $\boldsymbol{h_i(t)}$is the hazard of response at time $\boldsymbol{t}$, conditional on not having responded before $\boldsymbol{t}$. The term $\boldsymbol{h_0(t)}$ is the baseline hazard. The vector $\mathbf{X}_{\boldsymbol{i}}$contains individual-level auxiliary variables, while $\boldsymbol{\beta}$contains the regression coefficients estimated by the model.

The model used response outcomes observed after the first seven invitation waves, treating completed questionnaires as the event, and outstanding cases as right-censored observations.

The estimated hazard ratios exp(β) are shown in the following table:

Covariate	Hazard ratio	95% CI	p-value
Icelandic citizenship	2.21	2.00–2.43	<0.001
Annual income, per 5 million ISK	1.03	1.01–1.06	0.009
No university education	1.29	1.19–1.41	<0.001
University education	1.80	1.65–1.96	<0.001
Age, per 10 years	1.19	1.17–1.21	<0.001

Hazard ratios above 1 indicate higher response hazard at a given point in time, relative to the reference group, while values below 1 would represent lower instant hazard. The largest differences in response timing were associated with citizenship, education and age. Other variables, such as gender, were included in the model first, but discarded because they were not found to be statistically significant.

2.2 Parametric baseline hazard

The Cox model estimates the covariate effects, but leaves the baseline hazard unspecified. The goal of our model is to predict the probability of response by a fixed future date. For this purpose, a smooth parametric approximation to the baseline cumulative hazard was fitted.
The empirical baseline hazard, as shown in Fig.2, showed a clear shape: a strong initial response, peaking shortly after invitation, followed by a smaller response peak after the scheduled reminder, decaying into a low background response rate. To represent this pattern, the cumulative baseline hazard was fitted as:

$$\boldsymbol{H_0(t) = ct + B_1\Phi\left(\frac{t-\mu_1}{s_1}\right) + A_2 b_2\left(1 - e^{-(t-\tau)/b_2}\right)\mathbb{I}(t \geq \tau)},$$

where $\boldsymbol{t}$ is time in hours since invitation. The term $\boldsymbol{ct}$ represents a low background response rate. The second gaussian component represents the initial response peak after invitation, with location $\boldsymbol{\mu_1}$ and standard deviation $\boldsymbol{s_1}$. The exponential component represents a peak starting at time $\boldsymbol{\tau}$ and decaying with time constant $\boldsymbol{b_2}$. The indicator function $\mathbb{I}(\boldsymbol{t} \geq \boldsymbol{\tau})$ means that this component only contributes after the scheduled reminder period.

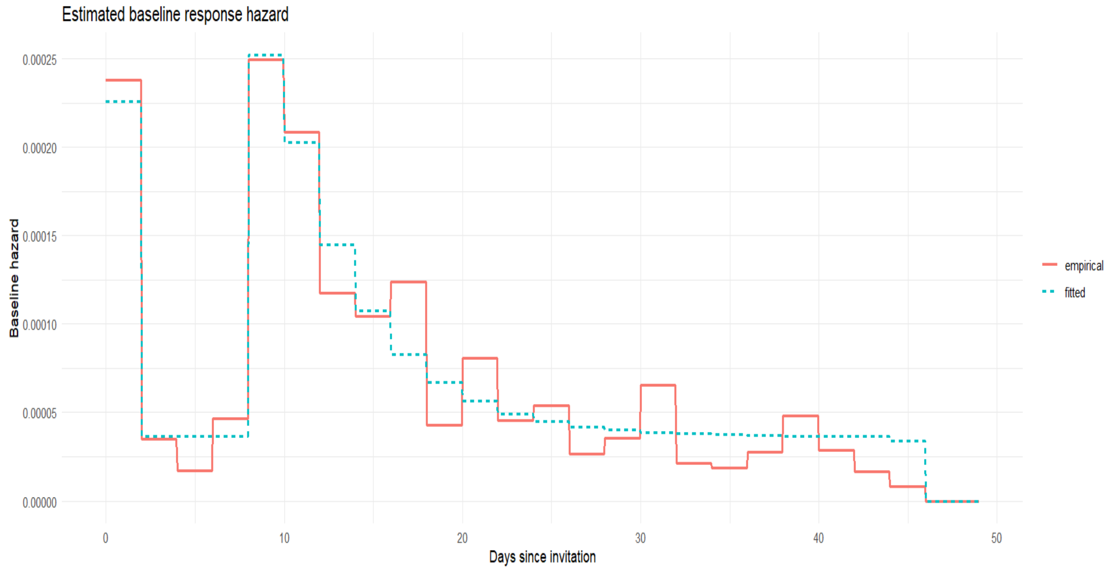

Figure 2: comparison between the empirical baseline hazard estimated from the Cox model with the fitted parametric baseline. In order to make the comparison easier and remove the noise, the empirical component was smoothed over 48 hours intervals.

From this model, the final probability of response was calculated for each indiviudal according to the formula

$$P(T_i \leq t \mid X_i) = 1 - \exp\left(-\exp(\beta^T X_i)\, H_0(t)\right),$$

by fixing the final time t to the expected close date of the survey (2 December 2025.)

3 Operational implementation

The fitted survival model was applied to the full sample to estimate each individual's probability of completing the questionnaire before the common closing date.

Individuals with lower predicted response probability were assigned higher priority for telephone follow-up, subject to practical constraints such as limited interviewer capacity.

The following table shows the resulting probabilities by subgroup

Variable	Subgroup	n	Predicted response (%)	Observed response (%)
Citizenship	FOR	2005	22	22,2
Citizenship	IS	6595	45,5	44,5
Education	elementary	2869	29,6	30
Education	no_university	3112	39,3	38,6
Education	university	2620	52,1	50,4
Age group	35-54	2786	42,5	41,5
Age group	55+	2798	50,3	49,8
Age group	<35	3017	28,1	27,6
Income	Above median income	4301	44,5	44,2
Income	Below median income	4300	35,6	34,4

4 Conclusions

Our survival-based model for the response probability uses the timing of responses directly, as opposed to a binary response model such as a logistic regression.

This approach has some underlying assumptions. One is that composition of the first seven waves used for training is similar to that of the last waves. Otherwise, the projection would be biased.

Also, the Cox model assumes proportional hazards, meaning that the covariates are assumed to multiply the baseline hazard. This is an approximation, because in reality, the baseline hazard may also be different for some groups in reality. Finally, our parametric projection is also a simplification of the actual process.

However, the model performed well enough for its intended purpose. The aggregate predicted response probability was close to the final observed one, and the predictions for individuals provided guidance for prioritisation of CATI follow-up.

5. Appendix: List of references

[1] Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B, 34(2), 187–220.

[2] Eurostat. (2024). European Health Interview Survey wave 4: Methodological manual. Publications Office of the European Union.