

Home assignment 2

Analysis of Survival Data, fall 2025

Criteria to get your home assignment approved (NOTE: not the same as for the previous assignment):

1. Your report must be complete and handed in on time. See the document 'Home assignment information' for further information.
2. Follow the check list (separate document) to ensure that you follow all requirements to pass (you can skip no. 14, about references, for this assignment).
3. All calculations and tables are to be produced by SAS. Graphs can be produced by SAS or, if you wish, by using other software (as long as the data is produced by SAS).
4. Include your commented SAS code in an appendix (Appendix 1).
5. Include a short statement on any use of generative AI (e.g. ChatGPT) in a second appendix (Appendix 2).

Information:

Your report (only one per group) needs to be electronically registered (pdf version) at Studium before the deadline.

All documents you need are to be found at Studium, modules "Home assignments" and "SAS examples".

This home assignment is based on the contents from lectures L1-L7.

If you pass this home assignment on the first try you get 2 bonus points to be added to the points you score on the final exam.

Data to be used:

Use the same data as for Home assignment 1.

Good luck!
/Inger

Task 1

The group differences you investigated in Home assignment 1, Task 1, have now been discussed in media.

You are hired as consultants to investigate this further.

The text below is just an example, to be adjusted to fit your specific data and results from Home assignment 1.

Previous reports have shown that the survival for leukemia patients is worse at Alfred hospital than at Ohio State, St Vincent and Hahnemann hospitals. But is the quality of care really worse at Alfred hospital than at the other hospitals?

Use the available data to answer that question. A suitable method to investigate this question is to use a **Cox regression model**, and adjust for additional factors that possibly could have an effect on survival (choose among the available variables that you described in your Data description, i.e., the two covariates you investigated in Home assignment 1, plus at least two continuous covariates).

For the final model: **check if the assumption of proportional hazards holds**, and check the **overall fit of the model**.

Include estimated survival graph(s) for the hospitals (see *SAS examples* section 4.7.1),
i) **comparing the groups without any additional covariates (here you can copy the survival plot from Home assignment 1),**
ii) **for your final model with some selected values of the covariates, i.e. show graphically the differences between the hospitals after adjusting for the covariates in the model. Tip: You can refer to “hypothetical patients”, which would help the hospital manager to understand.**

I want you to present the **“best” model** to the hospital manager (where the “best” model is chosen by statistical means).

Summarize your results in a maximum of 4 pages (excl. front page), in a way that is understandable to the hospital manager. The hospital manager has no statistical knowledge! This means that you must explain your results in a non-technical manner.

Avoid words/phrases like: model, dependent/independent, metric/nonmetric, coefficient, covariate, estimate, fit, significance level, significant (in any other way than “statistically significant”, see below).

Words/phrases that can be used but need to be explained: statistically significant, correlation, prognostic factor, explanatory variable/factor.

Include the steps below in your analysis, even if you don't include it all in your summary to the hospital manager. You don't need to describe details such as your choice of ties handling procedure, or how you have investigated the assumptions, or which variables you have investigated but not included in the final model, the manager won't understand such things and really doesn't need to know. Instead, focus on the results, and include short comments in

your SAS code to explain e.g. method choices, investigated assumptions, and chosen significance level(s).

Step 1) Investigate the data using descriptive statistics (especially graphs), focusing on the relationship between different covariates and the grouping variable (hospital).

Step 2) Identify possible covariates. Remember to investigate the correlation/multicollinearity among the possible covariates.

Step 3) Test the assumption of proportional hazards for each covariate with the time-dependent covariate test (no graphical examination needed at this stage). This way you'll know from the start if there are any covariates that you might want to stratify upon.

Step 4) Investigate which functional form to use for continuous covariates (if you have any).

Step 5) Build the “best” possible model (the hospital covariate must be included). Remember to also investigate possible interactions between the selected variables.

Step 6) Investigate the PH assumption graphically once you have decided which covariates to include in the model (and, if needed, return to step 6). Include your produced graphs in the SAS code appendix.

Step 7) Investigate the model fit for the final model, using Cox-Snell residuals and generalized R^2 . If you have any time-dependent covariates in your model, you can estimate the Cox-Snell residuals excluding the time-dependent covariates. Include your produced graph and the value of the generalized R^2 in the SAS code appendix.

Task 2

The hospital manager now asks you to clear Alfred hospital's name as far as possible, by only including explanatory variables that reduce the difference in survival between Alfred hospital and the other hospitals. What do you do? Refer to a few relevant sections of one or more codes of ethics documents and formulate a response to the manager.