Practical 3: Solutions

Please see the accompanying Stata Do file and R script file for the code to produce these solutions.

1. Load the data and explore the grade variable. Summarize the numbers and timings of CHD deaths and censorings by job grade.

Among men in Grade 1, there were 90 (7.5%) CHD deaths with 1104 censorings. In Grade 2, there were 64 (13.3%) deaths and 419 men were censored.

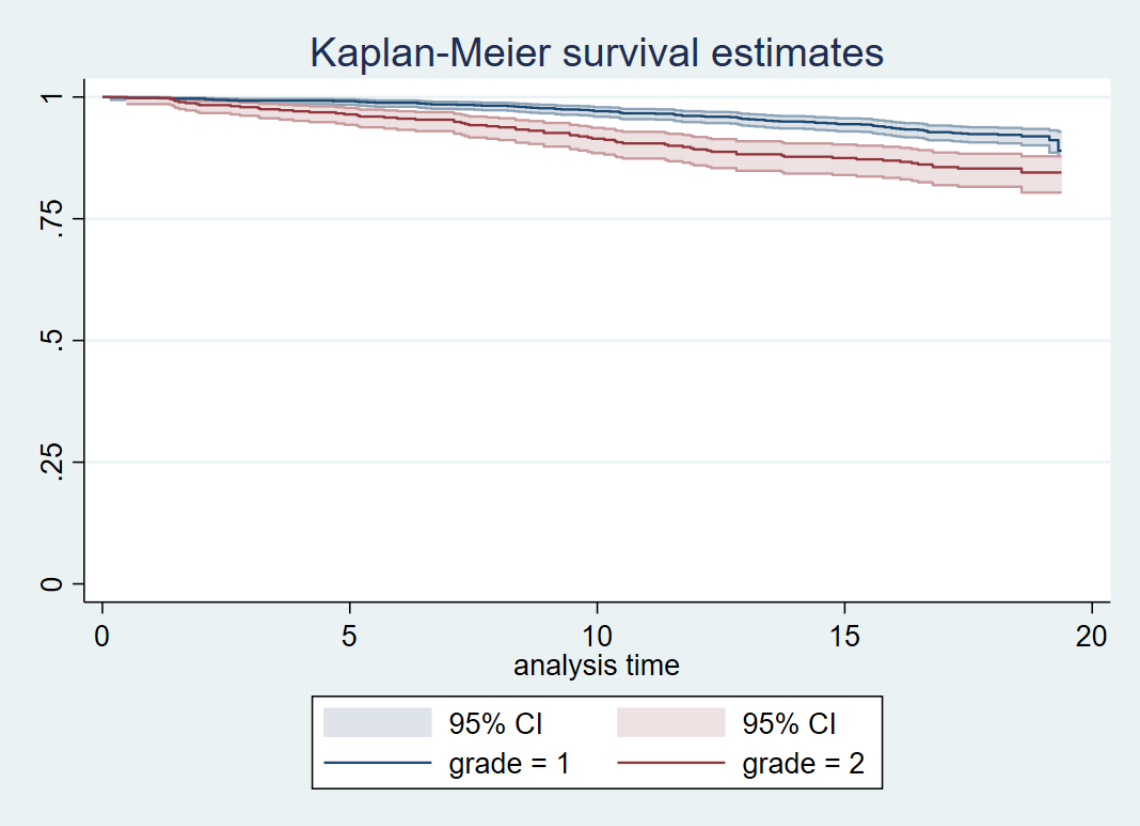
The median time to death was 11.9 years in Grade 1 and 8.8 years in Grade 2. The median time to censoring was 18.2 years and 17.9 years in Grade 1 and 2 respectively.

2. We begin by using simple methods to investigate the association between job grade and CHD.

a) Use a Kaplan-Meier plot to compare survival in the two groups, including the 95% confidence intervals. Interpret the plots.

A Kaplan-Meier plot is shown below. Those in Grade 1 had better survival compared with those in Grade 2. The confidence intervals are non-overlapping, suggesting good evidence of a true survival difference between the two groups.

Question 2: Kaplan-Meier survival plot



b) How many individuals survived to 5, 10, 15 years of follow-up in each job grade category?

In Grade 1, the numbers are 1169, 1114, and 1045. In Grade 2 the numbers are 445, 383, and 334.

c) Use the log rank test to compare the estimated survivor functions in the two job grades.

The log rank test provides very strong evidence ($p<0.001$) against the null hypothesis that the survivor curves are the same in the two job grades.

3. We will now fit an exponential model to the Whitehall data using job grade as an explanatory variable.

a) Write down the hazard and survivor functions and the likelihood.

$$h(t|x) = \lambda e^{\beta x}$$

$$L = \prod_{i=1}^{n} \left\{\lambda e^{\beta x_i} \exp\left(-\lambda t_i e^{\beta x_i}\right)\right\}^{\delta_i} \left\{\exp\left(-\lambda t_i e^{\beta x_i}\right)\right\}^{1-\delta_i}$$

where t_i is the survival or censoring time for individual i, δ_i is the indicator of whether the person had the outcome or was censored, and x_i is an indicator for their job grade.

b) Fit the exponential model and interpret the parameter estimates. What is the association between job grade and survival?

The estimated hazard ratio (e^{β}) is 1.99, meaning that the hazard rate in grade 2 was about two times that in grade 1. The confidence interval for for HR is (1.44, 2.74), indicating strong evidence against the null hypothesis that job grade is not associated with survival. The constant term is the estimate of the hazard for individuals in job grade 1, i.e. the estimate of λ. The estimate from the model is $\lambda = 0.0044$.

c) Change to the age time scale and refit the exponential model. Compare your results with those found when using time-in-study as the timescale.

Re-running the analysis gives the same estimates. This is because the hazard for grade is constant over time in an exponential model, so the time-scale does not have any impact on the estimates. Another way to see this is that in an exponential model the hazard in each group is simply D divided by Y, where D is the total number of deaths, and Y is the total time at risk. Neither of these are altered by changing to the age time scale, as long as deferred entry is used. Hence the ratio of the hazards is unaffected.

4. Revert to the time-in-study timescale for this question and all subsequent questions. By fitting an exponential distribution we are assuming that the hazard rate does not change over time. Because this may not be a reasonable assumption we investigate

fitting a Weibull model. Fit the Weibull model. Interpret the parameters of the model. Compare your results with those from the exponential model.

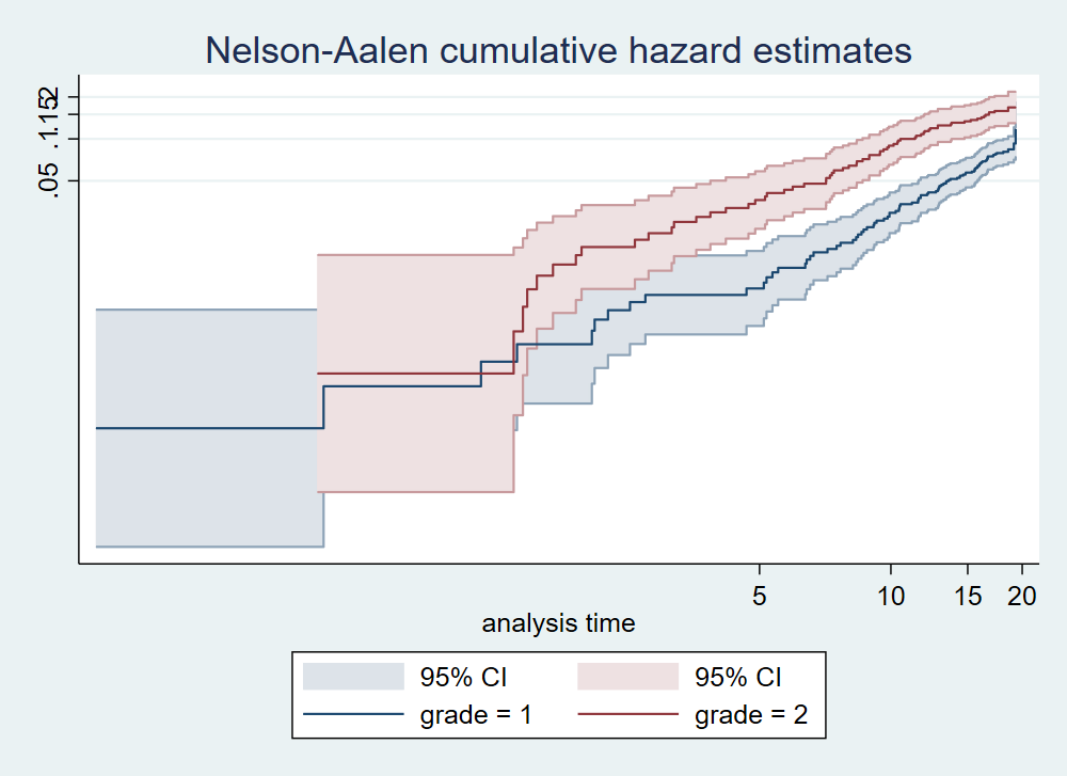
The estimate of the hazard ratio has increased slightly (compared to the exponential model) to 2.04. The 95% confidence interval (1.48 to 2.81) and p-value (p<0.001) still provide strong evidence that the risk of death due to coronary heart disease is greater among men in job grade 2 (compared to grade 1).

The estimate of the shape parameter (which we call in the notes) is 1.41. For interpretation, see question 8.

5. Create a suitable non-parametric plot to investigate whether you expect the Weibull model fitted above to be appropriate.

Recall that, if the Weibull model is appropriate, we would expect that a plot of the log of the cumulative hazard against the log of time should be linear in each group, and that the two lines should be parallel. Such a plot is shown below. We can see that, after the initial period where there are few events, the lines are both straight, and are approximately parallel. This suggests the Weibull model is appropriate.

Question 5: Plots to check if the Weibull model is appropriate



6. The age at which individuals entered the study may have an important part to play in the analysis. So we will add an age variable to the Weibull model.

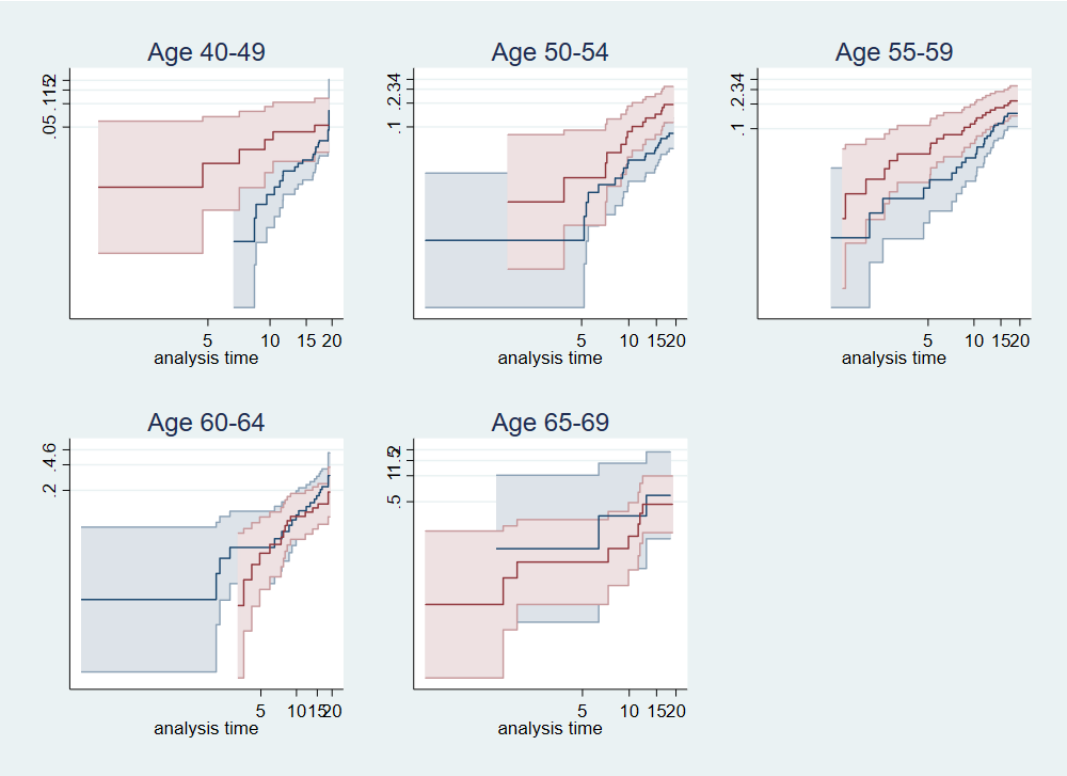
Include `agein` as an additional explanatory variable in the Weibull model fitted for job grade in Question 4. Interpret the hazard ratios for job grade and for age.

The coefficient for age is 1.11. The interpretation is that the hazard increases by 11% for each additional year of age at entry to the study. This estimate is conditional on job grade. The estimated hazard ratio for grade is 1.27, after conditioning on age at entry to the study. Before adjustment for age at entry, the hazard ratio was 2.04. The reason for the reduction in the hazard ratio is that age at entry to the study confounds the association.

7. Create non-parametric plots to investigate whether you expect the Weibull model fitted in Question 6 to be appropriate for this data. Age is a continuous variable so we could categorize the age variable for use in making (approximate) assessments of whether the Weibull model is appropriate. We recommend using the age categories: 40-49, 50-54, 55-59,..., 65-69. Example code for creating these plots is provided in the example Stata do file and R script file.

The resulting plots are shown below.There is some possible evidence that the estimated transformed survivor curves in the first age category are not parallel, however the confidence bands are also wide. Note that we could also look at similar plots of age category separately by the two job grades. In later sessions we will learn how to use alternative models which may provide a better fit to the data as well as other methods for assessing model fit.

Question 7: Plots to check the Weibull model is appropriate



8. Referring to the model fitted in question 6, perform a test of the null hypothesis that the hazard rate does not change over time. What do you conclude?

There are two ways to do this.

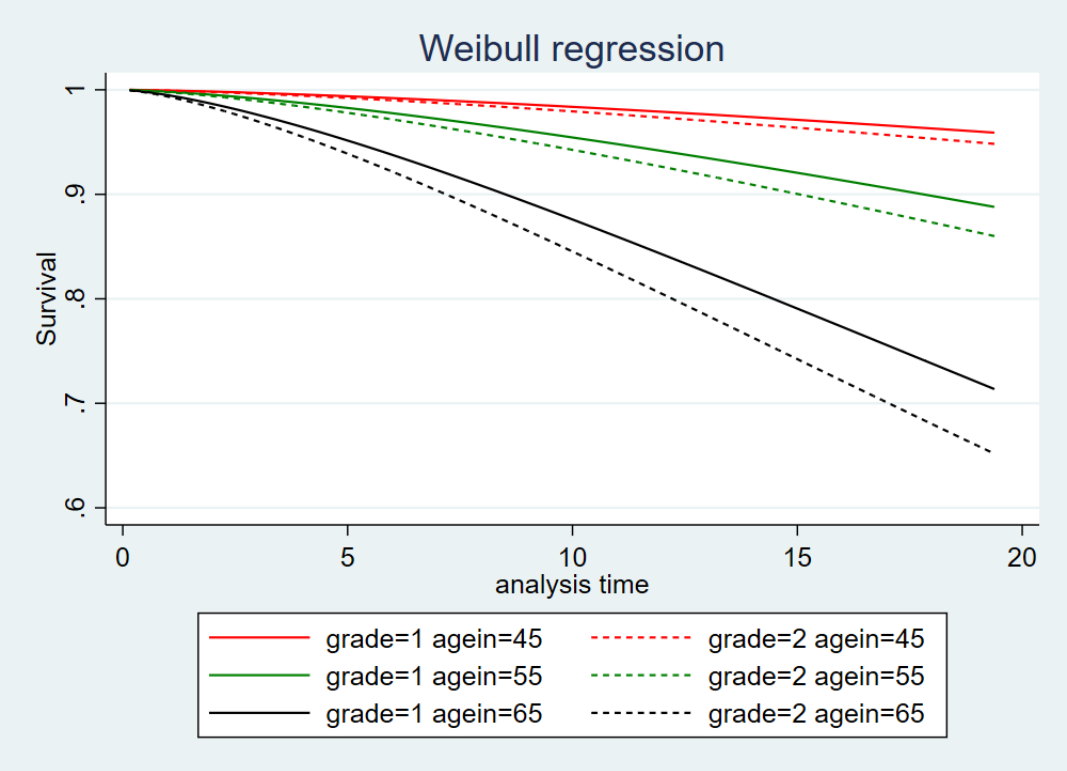
First, recall that the Weibull distribution with $\kappa = 1$ is equivalent to the exponential distribution. We can therefore test the null hypothesis that $\kappa = 1$ in order to test the null hypothesis that the hazard rate does not change over time. The estimate from the model fitted in question 6 shows the 95% confidence interval for κ, which is (1.22,1.64). This excludes the value 1, giving evidence against the null hypothesis (at the 5% level). We could equivalently look at the p-value for $\log \kappa$ (`ln_p` in the Stata output), which is very small ($p < 0.0001$).

An alternative approach to comparing the exponential and Weibull models is to perform a likelihood ratio test (note that the `force` option is required in Stata, as otherwise it does not recognise that the exponential and Weibull models are nested).

9. Using the Weibull model fitted in question 6, plot estimated survivor curves for individuals in job grade groups 1 and 2 aged 45, 55, 65.

The plots are shown below. We can see that at each of the three ages the predicted survival is lower among men from job grade 2. The difference is smallest for men aged 45 at study entry, and is widest for men aged 65. However, even for men in grade 2 who were aged 65 at entry into the cohort, we expect that around 65% will survive for 20 years from the the time they joined the study.

Question 9: Predicted survival from a Weibull model include job grade and age category



Extra exercises

1. We used the Weibull model above to allow the hazard to change over time. A

different approach is to split the follow up time up into a few periods and fit a series of exponential models within each period. It can then be investigated whether the baseline hazard changes across the periods. To do this we need to create a record for each individual within each time period up to their event or censoring time.

a) Write down the algebraic expression for the model being fitted.

$$h(t|grade, agein, fuband) = \lambda \exp(\beta_1 x_1 + \gamma_1 a + \gamma_2 x_5 + \gamma_3 x_{10} + \gamma_4 x_{15}) \qquad (1)$$

where x_1 is an indicator variable for job grade, x_5 takes value 1 for $5 \leq t < 10$ and 0 otherwise, x_{10} takes value 1 for $10 \leq t < 15$ and 0 otherwise, and x_{15} takes value 1 for $t \geq 15$ and 0 otherwise. The variable a is the age at entry. Grade is a binary variable, agein is continuous, fuband is a vector of indicators for the follow-up period.

b) Interpret the results and compare the results from this model with those from the Weibull and exponential models fitted earlier.

The parameter estimates 1.88, 2.18, 2.92 are the estimated ratios of hazards in follow-up time band 5-10 years, 10-15 years and 15-20 years relative to 0-5 years.

The hazard ratio for grade (1.27) doesn't depend much on the model we use – we have found similar estimates using exponential and Weibull models, and using this current 'piecewise exponential' model.

However, the estimates for grade and age (1.11) under this piecewise exponential model are (very slightly) more similar to the Weibull results than they are to the exponential results. This is because the Weibull allows the baseline hazard to vary over time, and this model also does that, albeit in a more crude way.

2. We have used the exponential model to investigate the association between job grade and CHD. The exponential model is based on the assumption of a constant baseline hazard. An equivalent way of fitting this model is using Poisson regression, which should be familiar to you from earlier modules.

Fit a Poisson regression model to these data, with job grade as an explanatory variable. Check that you get the same results using Poisson regression and using an exponential model.

You should get the same results from the two models. Check the code provided if you don't.