Biostatistics III:
Survival analysis for epidemiologists in R

Mark Clements
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
http://www.biostat3.net/

13–21 November, 2017

http://kiwas.ki.se/katalog/katalog/kurs/2669, course 2992

Day 1: Review questions and exercises

1. What is right censoring?
2. What are the formal requirements for survival analysis for right censored data?
3. Discuss the interpretation of different time scales (exercise).
4. How is survival analysis different to a chi-square test, a t-test or linear regression?
5. What do we estimate in survival analysis?
6. What is an incidence rate? What are its properties?
7. What are the properties of the survival function?
8. How can we interpret Kaplan-Meier curves? (exercise)
9. How do we calculate the Kaplan-Meier curves?
10. What is the difference between the Kaplan-Meier estimator and the actuarial estimator for survival? When would you use each of these?
11. What formula is used to calculate the variance of the Kaplan-Meier estimator? (Advanced: on what scale do we estimate the confidence interval?)
12. Review the R commands for survival analysis.
13. Show how to calculate a log-rank test. (exercise)

Day 1: Review questions and exercises

1. What is right censoring?

Day 1: Review questions and exercises

1. What is right censoring?
2. What are the formal requirements for survival analysis for right censored data?
3. Discuss the interpretation of different time scales (exercise).

Formal requirements of time-to-event data

Three basic requirements define time-to-event measurements

a. precise definition of the start and end of follow-up time
b. unambiguous origin for the measurement of 'time'; scale of time (e.g. time since diagnosis, attained age)
c. precise definition of 'response,' or occurrence of the event of interest

Exercise: Examples of time-to-event measurements

Do the following satisfy our formal requirements for time-to-event data?

- Time from diagnosis of cancer to death due to the cancer
- Time from diagnosis of cancer to death due to any causes
- Time from diagnosis of localised cancer to metastases
- Time from randomisation to death in a cancer clinical trial
- Time from randomisation to recurrence in a cancer clinical trial
- Time from remission to relapse of leukemia
- Time to re-offending after being released from jail
- Time between two attempts to donate a unit of blood for transfusion purposes
- Time from HIV infection to AIDS
- Time to the first goal (or next goal) in a hockey game
- Time from exposure to cancer incidence in an epidemiological cohort study

Day 1: Review questions and exercises

1. What is right censoring?
2. What are the formal requirements for survival analysis for right censored data?
3. Discuss the interpretation of different time scales (exercise).
4. How is survival analysis different to a chi-square test, a t-test or linear regression?

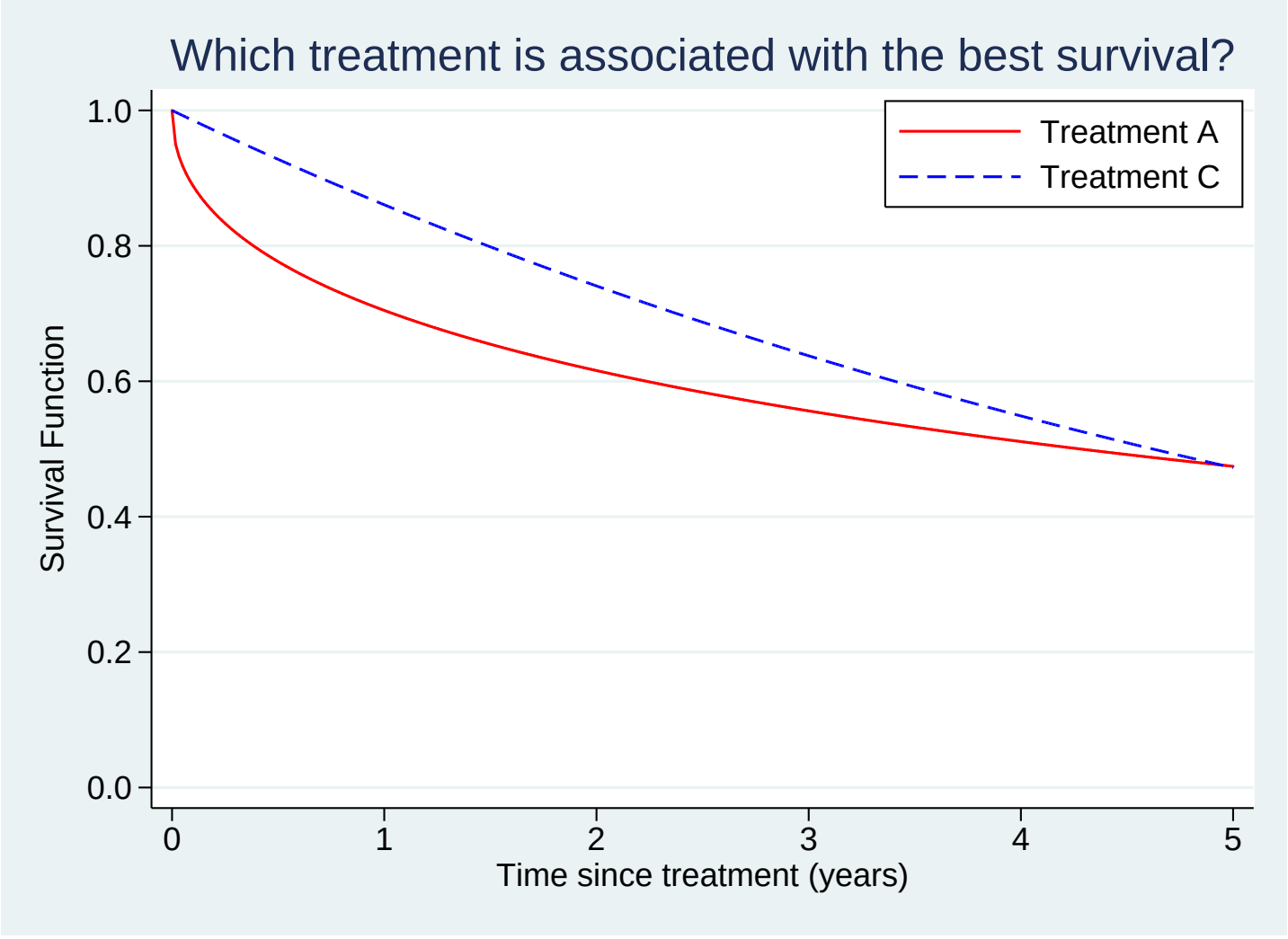
Day 1: Review questions and exercises

1. What is right censoring?
2. What are the formal requirements for survival analysis for right censored data?
3. Discuss the interpretation of different time scales (exercise).
4. How is survival analysis different to a chi-square test, a t-test or linear regression?
5. What do we estimate in survival analysis?

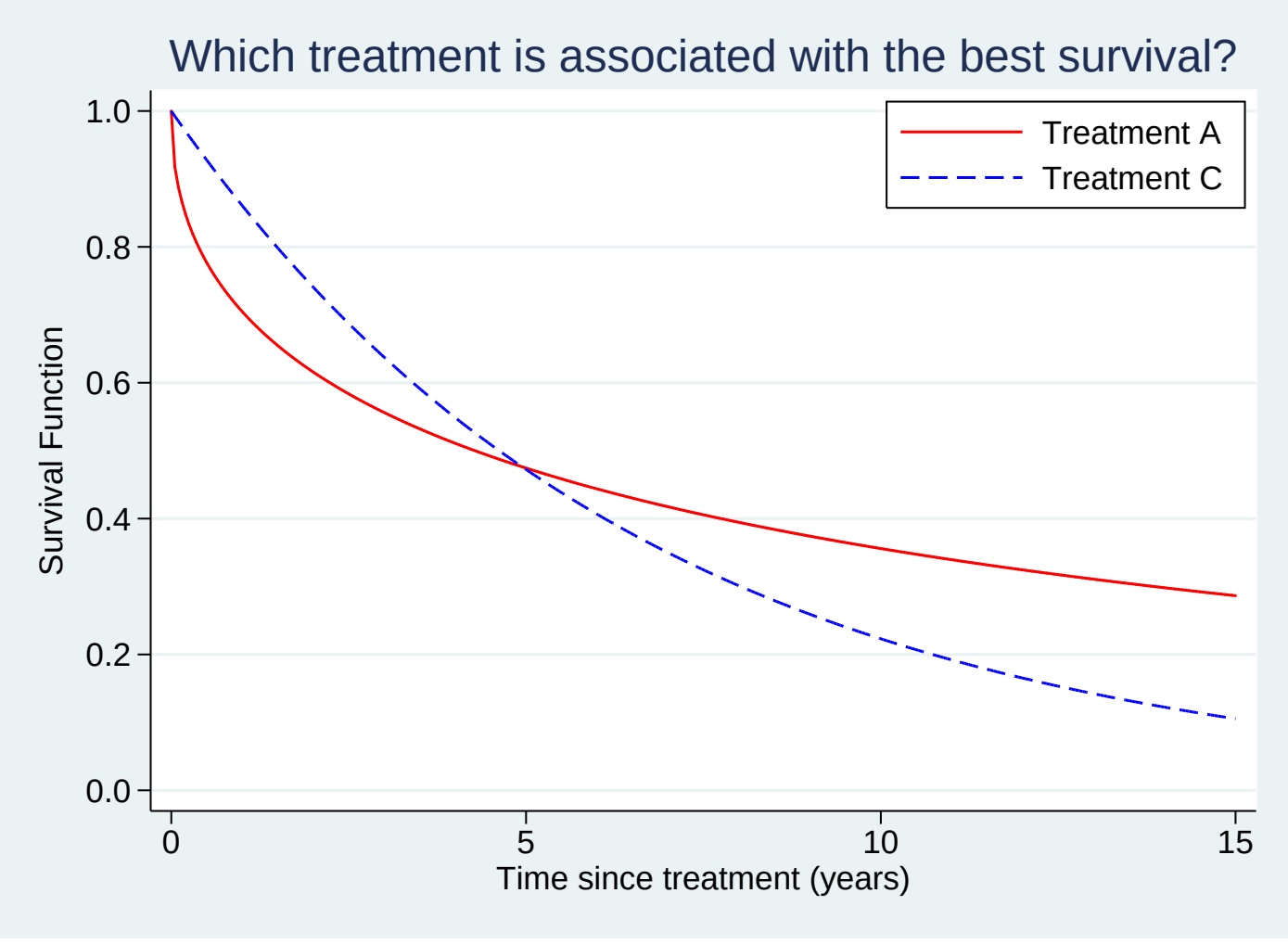
Day 1: Review questions and exercises

1. What is right censoring?
2. What are the formal requirements for survival analysis for right censored data?
3. Discuss the interpretation of different time scales (exercise).
4. How is survival analysis different to a chi-square test, a t-test or linear regression?
5. What do we estimate in survival analysis?
6. What is an incidence rate? What are its properties?
7. What are the properties of the survival function?
8. How can we interpret Kaplan-Meier curves? (exercise)

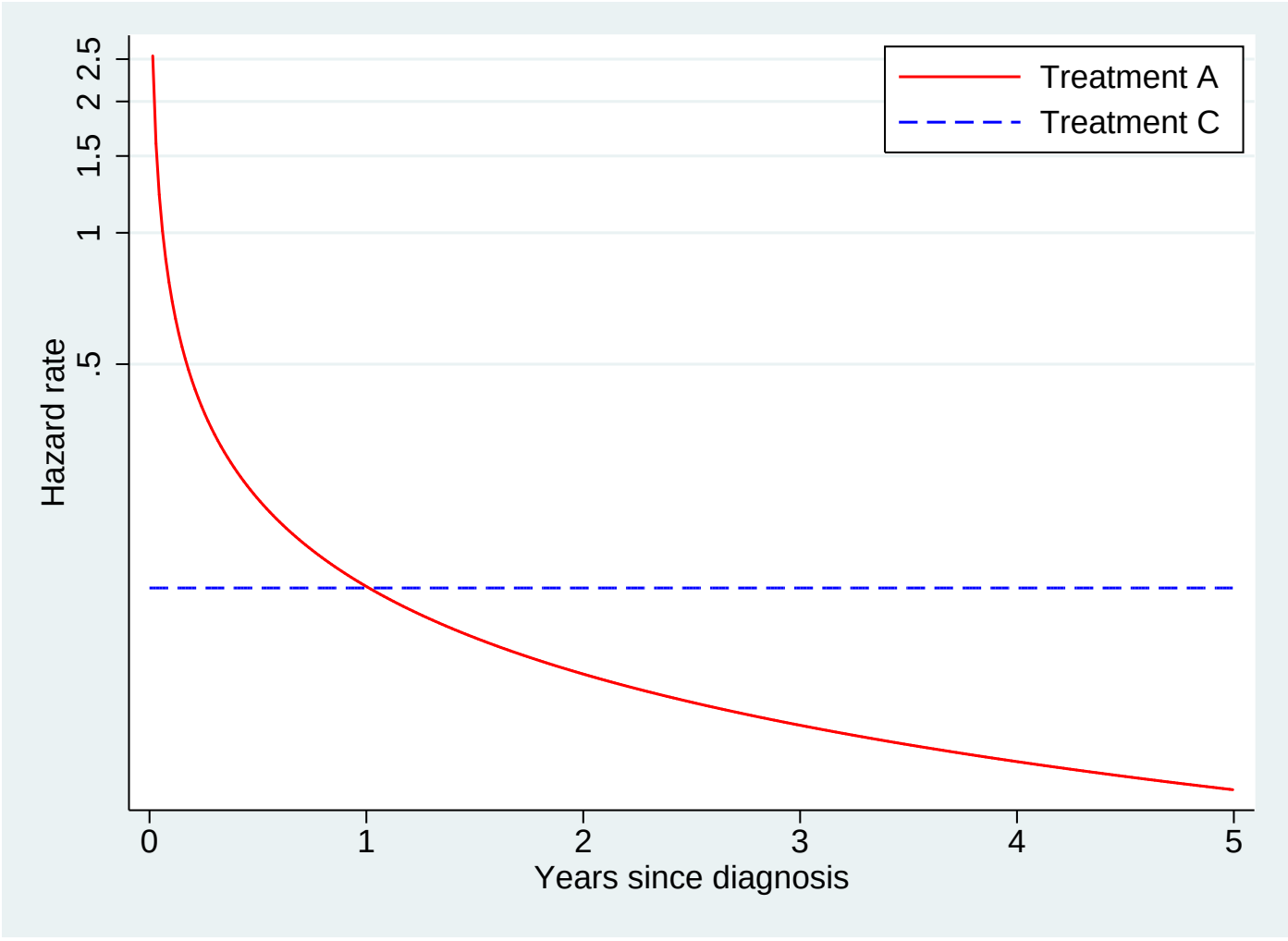
Exercise: Which group (A or C) has the best survival? I



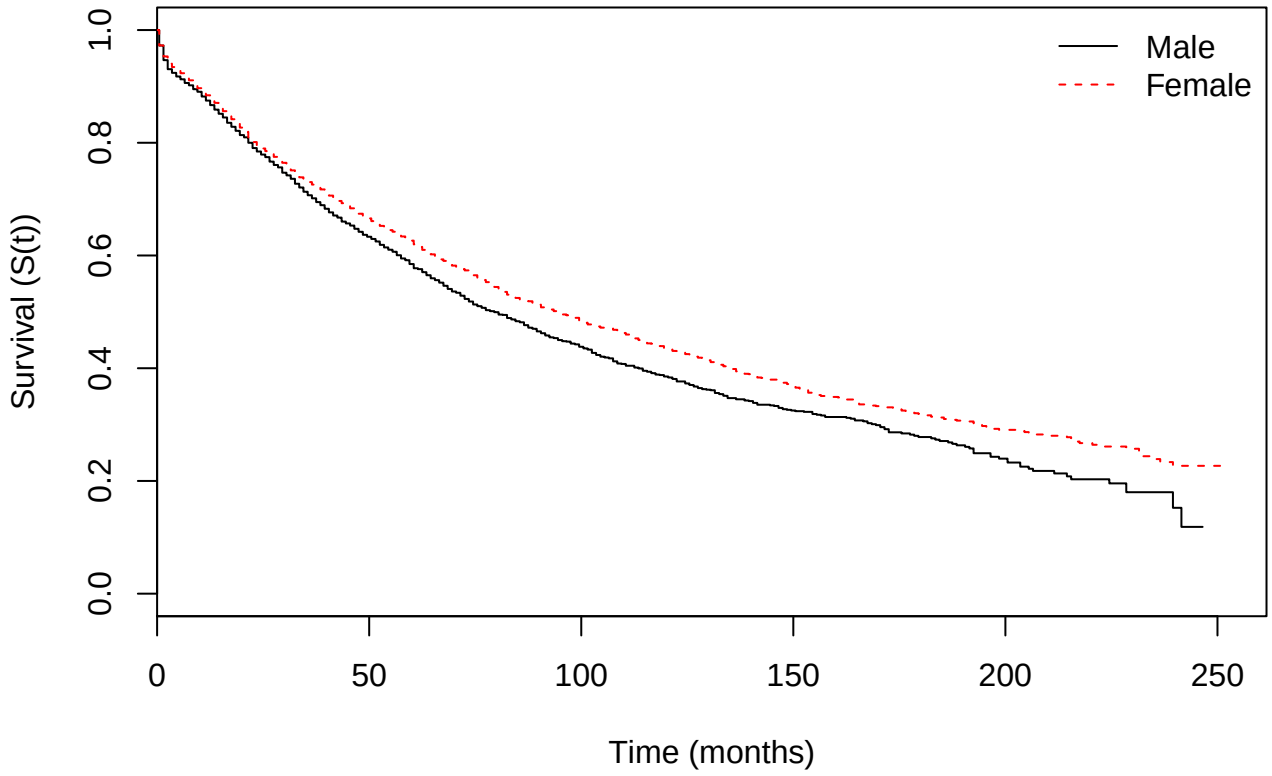
What about if we extend the follow-up?



Now plot the two hazard functions



Exercise: Survival following a diagnosis of localised colon cancer, Sweden



- What do we see? Consider the questions on the following page.

Possible questions of interest

- Which sex experiences the best survival?
- Does the group with best survival experience lower mortality throughout the follow-up?
- At what point in the follow-up is mortality the highest?

Day 1: Review questions and exercises

1. What formula is used to calculate the variance of the Kaplan-Meier estimator? (Advanced: on what scale do we estimate the confidence interval?)

Day 1: Review questions and exercises

1. What formula is used to calculate the variance of the Kaplan-Meier estimator? (Advanced: on what scale do we estimate the confidence interval?)
2. Show how to calculate a log-rank test. (exercise)

Exercise: Log rank test I

event time	males			females		
	at risk	obs	exp	at risk	obs	exp
2	19	1	1.086	16	1	0.914
3	18	1	0.545	15	0	0.455
5	17	1	0.531	15	0	0.469
7	16	1	0.516	15	0	0.484
8	15	1	0.500	15	0	0.500
9	14	0	0.483	15	1	0.517
11	14	0	0.500	14	1	0.500
19	13	0	0.520	12	1	0.480
22	13	1	0.542	11	0	0.458
27	12	0	0.545	10	1	0.455
28	11	0	0.550	9	1	0.450
32	11	2	1.158	8	0	0.842
33	9	1	0.563	7	0	0.438
43	8	0	0.615	5	1	0.385
46	8	1	0.667	4	0	0.333
102	2	0	0.500	2	1	0.500
103	2	1	0.667	1	0	0.333

Totals: $O_1 = 11$, $E_1 = 10.488$, $O_2 = 8$, $E_2 = 8.512$

Exercise: Log rank test II

What is the test statistic? The critical value for a χ^2 test with one degree of freedom is 3.84; how do we interpret the test?

Pop quiz: R

1. What is the difference between `:`, `::` and `:::`?

Pop quiz: R

1. What is the difference between `:`, `::` and `:::`?
2. What are the `base::transform`, `base::within` and `dplyr::mutate` functions? How are they similar and different? (Hint: you only need one of these.)

Pop quiz: R

1. What is the difference between `:`, `::` and `:::`?
2. What are the `base::transform`, `base::within` and `dplyr::mutate` functions? How are they similar and different? (Hint: you only need one of these.)
3. What are the `base::subset`, `dplyr::filter` and `dplyr::select` functions?

Pop quiz: R

1. What is the difference between `:`, `::` and `:::`?
2. What are the `base::transform`, `base::within` and `dplyr::mutate` functions? How are they similar and different? (Hint: you only need one of these.)
3. What are the `base::subset`, `dplyr::filter` and `dplyr::select` functions?
4. What is the difference between `x=list(a=1:2,b=2)` and `data.frame(a=1:2,b=2)`?

Pop quiz: R

1. What is the difference between `:`, `::` and `:::`?
2. What are the `base::transform`, `base::within` and `dplyr::mutate` functions? How are they similar and different? (Hint: you only need one of these.)
3. What are the `base::subset`, `dplyr::filter` and `dplyr::select` functions?
4. What is the difference between `x=list(a=1:2,b=2)` and `data.frame(a=1:2,b=2)`?
5. What are three ways to get the `b` element from `x`?

Exercises for Monday afternoon

1a. Hand calculation: Kaplan-Meier estimates of cause-specific survival (35 patients)

1b. Kaplan-Meier estimates of cause-specific survival using R (35 patients)

2. Melanoma: Comparing survival proportions and mortality rates according to stage

3. Localised melanoma: Comparing estimates of cause-specific survival between periods; first graphically and then using the log rank test

4. Localised melanoma: Comparing various approaches to estimating the 10-year survival proportion

Organisation: Each group should prepare a short presentation of their answers for the next morning's teaching occasion. The presentation should include brief computer code, output and summary findings. (The content of the presentation is intended to mirror the examination requirements.)

Day 2: Review questions and exercises

- What are the possible time scales for the following and discuss which is the most appropriate: (i) time to first myocardial infarction (MI); (ii) survival following admission for first MI; (iii) survival for a man with screen-detected localised prostate cancer.

Day 2: Review questions and exercises

- What are the possible time scales for the following and discuss which is the most appropriate: (i) time to first myocardial infarction (MI); (ii) survival following admission for first MI; (iii) survival for a man with screen-detected localised prostate cancer.
- What is a mathematical definition of the hazard?
- What is the relationship between the hazard function, the cumulative hazard function and survival?

Day 2: Review questions and exercises

- What distribution is used to model event counts? What regression model is used to model event rates? What is the "offset" and how does that affect the regression model interpretation?
- How can we calculate an incidence rate ratio and its 95% confidence interval from a regression coefficient from a Poisson regression?

Day 2: Review questions and exercises

- What distribution is used to model event counts? What regression model is used to model event rates? What is the "offset" and how does that affect the regression model interpretation?
- How can we calculate an incidence rate ratio and its 95% confidence interval from a regression coefficient from a Poisson regression?
- How can we model (i.e. parameterise) for (i) a continuous covariate with a linear effect, (ii) a continuous covariate with a quadratic effect, (iii) an ordinal variable with three levels and (iv) a nominal variable with three levels? How many parameters are there for each?

Day 2: Review questions and exercises

- What distribution is used to model event counts? What regression model is used to model event rates? What is the "offset" and how does that affect the regression model interpretation?
- How can we calculate an incidence rate ratio and its 95% confidence interval from a regression coefficient from a Poisson regression?
- How can we model (i.e. parameterise) for (i) a continuous covariate with a linear effect, (ii) a continuous covariate with a quadratic effect, (iii) an ordinal variable with three levels and (iv) a nominal variable with three levels? How many parameters are there for each?
- What is the difference between a main effects model and an interaction model?
- Explain the difference between
  1. `glm(chd  hieng+job+hieng:job+offset(log(y)), data=diet, family=poisson)`
  2. `glm(chd  hieng+hieng:job+offset(log(y)), data=diet, family=poisson)`
- Can we use the parameters in (1) to estimate the parameters in (2)? If so, which R functions could we use?

Day 2: Review questions and exercises

- Describe two ways to assess whether an exposure adds significantly to a Poisson regression model.

Day 2: Review questions and exercises

- Describe two ways to assess whether an exposure adds significantly to a Poisson regression model.
- Is time a potential confounder in an analysis of event rates? If so, how can we adjust for time with Poisson regression?
- Draw a schematic to represent how time splitting works for Poisson regression. Show the events and person-time for several individuals who have been split by (i) time since study entry and (ii) attained age.
- For Poisson regression, how can we assess whether a rate ratio for an exposure varies across time?

Day 2: Review questions and exercises

- Describe two ways to assess whether an exposure adds significantly to a Poisson regression model.
- Is time a potential confounder in an analysis of event rates? If so, how can we adjust for time with Poisson regression?
- Draw a schematic to represent how time splitting works for Poisson regression. Show the events and person-time for several individuals who have been split by (i) time since study entry and (ii) attained age.
- For Poisson regression, how can we assess whether a rate ratio for an exposure varies across time?
- Assuming the same model, how does Poisson regression differ between individual-level data and collapsed data? Which form of data should be used to assess goodness of fit?

Pop quiz: R

1. Explain `factor` variables. What do `relevel` and `unclass` do?

Pop quiz: R

1. Explain `factor` variables. What do `relevel` and `unclass` do?
2. Describe the use of the following generic functions: `print`, `plot`, `summary`, `anova`, `coef`, `vcov`.

Pop quiz: R

1. Explain `factor` variables. What do `relevel` and `unclass` do?
2. Describe the use of the following generic functions: `print`, `plot`, `summary`, `anova`, `coef`, `vcov`.
3. Describe the use of the following base graphics plot arguments: `main`, `xlab`, `ylab`, `xlim`, `ylim`, `lty`, `lwd`, `col`.

Regression equation formulations

We can represent regression equations using a number of different formulations (or representations). As examples

$$\begin{aligned}
\log(\text{rate}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_4 x_1 x_2 \\
\text{rate} &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_4 x_1 x_2\right) \\
E(\text{count}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_4 x_1 x_2\right) \text{PersonTime} \\
E(\text{count}) &= \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_3 + \beta_4 x_1 x_2 + \log(\text{PersonTime})\right) \\
\log(\textit{rate}) &= \beta_0 + \beta_1 \text{age} + \beta_2 I(\text{sex}=\text{"Male"}) + \beta_4 \text{age}\ I(\text{sex}=\text{"Male"})
\end{aligned}$$

where x_1 and x_2 are indicator or continuous variables.

Regression equation re-parameterisation

The reading material for Day 2 often uses Greek symbols to represent the interaction models. We can re-express these effects using indicators. For example, for slides 149–150

$$\begin{aligned}\log(\text{rate}) = \log(\lambda_0) &+ \beta_1 I(\text{hieng="high"}) + \\ &\beta_2 I(\text{job="conductor"}) + \beta_3 I(\text{job="bank"}) + \\ &\beta_4 I(\text{hieng="high" \& job="conductor"}) + \\ &\beta_5 I(\text{hieng="high" \& job="bank"})\end{aligned}$$

and for slides 154–155

$$\begin{aligned}\log(\text{rate}) = \log(\lambda_0) &+ \\ &\beta_1 I(\text{job="conductor"}) + \beta_2 I(\text{job="bank"}) + \\ &\beta_3 I(\text{hieng="high" \& job="driver"}) + \\ &\beta_4 I(\text{hieng="high" \& job="conductor"}) + \\ &\beta_5 I(\text{hieng="high" \& job="bank"})\end{aligned}$$

(Now: slide 176 ff)

Review questions for Day 3

- What are the hazard and survival functions for an exponential distribution?
- What is the connection between Poisson regression and the exponential distribution?
- In epidemiology, very few time-to-event analyses assume a Weibull, Gamma or log-normal parametric distribution. Why is this?

Review questions for Day 3

- What are the hazard and survival functions for an exponential distribution?
- What is the connection between Poisson regression and the exponential distribution?
- In epidemiology, very few time-to-event analyses assume a Weibull, Gamma or log-normal parametric distribution. Why is this?
- What is the standard model in survival analysis for estimating hazard ratios?
- What is the hazard model for a Cox regression model? What does the Cox model assume about the form of the baseline hazard? Why is there no explicit intercept term?

Review questions for Day 3

- What are the hazard and survival functions for an exponential distribution?
- What is the connection between Poisson regression and the exponential distribution?
- In epidemiology, very few time-to-event analyses assume a Weibull, Gamma or log-normal parametric distribution. Why is this?
- What is the standard model in survival analysis for estimating hazard ratios?
- What is the hazard model for a Cox regression model? What does the Cox model assume about the form of the baseline hazard? Why is there no explicit intercept term?
- [Slides 228–236]

Review questions for Day 3

- What are the hazard and survival functions for an exponential distribution?
- What is the connection between Poisson regression and the exponential distribution?
- In epidemiology, very few time-to-event analyses assume a Weibull, Gamma or log-normal parametric distribution. Why is this?
- What is the standard model in survival analysis for estimating hazard ratios?
- What is the hazard model for a Cox regression model? What does the Cox model assume about the form of the baseline hazard? Why is there no explicit intercept term?
- [Slides 228–236]
- For the diet data, how does the modelled rate vary across time for: `glm(chd~hieng+offset(log(y)),data=diet,family=poisson)`

Review questions for Day 3

- What are the hazard and survival functions for an exponential distribution?
- What is the connection between Poisson regression and the exponential distribution?
- In epidemiology, very few time-to-event analyses assume a Weibull, Gamma or log-normal parametric distribution. Why is this?
- What is the standard model in survival analysis for estimating hazard ratios?
- What is the hazard model for a Cox regression model? What does the Cox model assume about the form of the baseline hazard? Why is there no explicit intercept term?
- [Slides 228–236]
- For the diet data, how does the modelled rate vary across time for: `glm(chd~hieng+offset(log(y)),data=diet,family=poisson)`
- We could model (i) using time since study entry as the primary time scale with age at study entry as a covariate or (ii) using attained age as the primary time scale. How are these two models different? When would you use each?

Review questions for Day 3

- What is the mathematical relationship between covariate-specific survival and baseline survival (i.e. when covariates=0)?
- Can we estimate the baseline cumulative hazard function after a Cox regression model fit?

Review questions for Day 3

- What is the mathematical relationship between covariate-specific survival and baseline survival (i.e. when covariates=0)?
- Can we estimate the baseline cumulative hazard function after a Cox regression model fit?
- What is the relationship between Cox regression and Poisson regression? When would we prefer Poisson regression over Cox regression?
- Describe each of the three recommended approaches to assess non-proportionality for a Cox model.

Review questions for Day 3

- What is the mathematical relationship between covariate-specific survival and baseline survival (i.e. when covariates=0)?
- Can we estimate the baseline cumulative hazard function after a Cox regression model fit?
- What is the relationship between Cox regression and Poisson regression? When would we prefer Poisson regression over Cox regression?
- Describe each of the three recommended approaches to assess non-proportionality for a Cox model.
- [Slides 246 ff.]

Review questions for Day 3

- What is the mathematical relationship between covariate-specific survival and baseline survival (i.e. when covariates=0)?
- Can we estimate the baseline cumulative hazard function after a Cox regression model fit?
- What is the relationship between Cox regression and Poisson regression? When would we prefer Poisson regression over Cox regression?
- Describe each of the three recommended approaches to assess non-proportionality for a Cox model.
- [Slides 246 ff.]
- What is the regression equation for a stratified Cox model? How can this model be used to account for non-proportionality?

Review questions for Day 3

- What is the mathematical relationship between covariate-specific survival and baseline survival (i.e. when covariates=0)?
- Can we estimate the baseline cumulative hazard function after a Cox regression model fit?
- What is the relationship between Cox regression and Poisson regression? When would we prefer Poisson regression over Cox regression?
- Describe each of the three recommended approaches to assess non-proportionality for a Cox model.
- [Slides 246 ff.]
- What is the regression equation for a stratified Cox model? How can this model be used to account for non-proportionality?
- [Slides 246 ff.]

Review questions for Day 4 I

- What are the main differences between the generalised survival models (GSMs) and Cox regression? What are the principal advantages of each? When would you use each model class?
- What are other names for "generalised survival models"?
- Cox regression and nested case control studies both can be thought of on terms of risk sets. How do each select each risk set?