Tidy Survival Analysis: Applying R's Tidyverse to Survival Data

Module 1. Introduction

LU MAO

lmao@biostat.wisc.edu

Department of Biostatistics & Medical Informatics

University of Wisconsin-Madison

Aug 3, 2025

Table of contents

- Basics of Survival Analysis
- German Breast Cancer Study: A Working Example
- Standard Analysis with `survival` Package
- Summary

Basics of Survival Analysis

Time-to-Event Data

- A common type of outcome in medical and clinical studies
  - **Starting point**: Randomization, diagnosis, enrollment, birth, etc.
  - **Endpoint (Event of interest)**: Death, disease onset, hospitalization, etc.
    - **Engineering**: Failure times of machines or components (reliability)
    - **Social sciences**: Time to job change, dropout, or event occurrence

- **Right censoring:**
  - Event not observed within the follow-up period
  - Due to study ending, dropout, or loss to follow-up
  - We only know:

$$T > C$$

  where T is event time and C is censoring time

Follow-up (Swimmer) Plot

- A rat tumorigenicity study
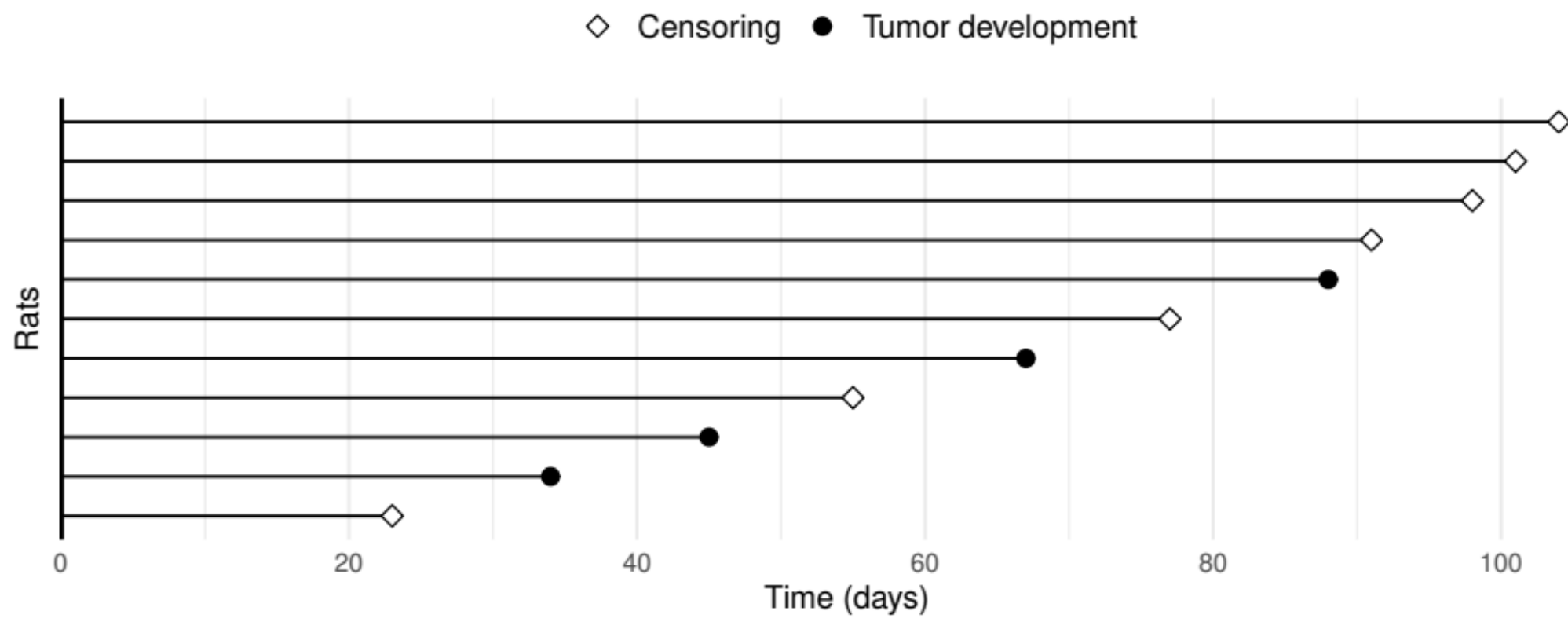
Basic Estimands

- **Survival function**: $S(t) = \Pr(T > t)$

  - Probability subject survives beyond time t

- **Hazard function**:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

  - Instantaneous risk of failure at time t

- **Relationship**

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right)$$

  - **Cumulative hazard function**: $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$

Observed (Censored) Data

- **Notation:** (X, δ)

  - $X = \min(T, C)$: observation time (event or censoring)

  - $\delta = I(T \leq C)$: event indicator (1 for event, 0 for censoring)

- **Data format**

```
 1  # time = X, status = delta (tidy format)
 2      id    time status
 3  1    1      5     1
 4  2    2      3     0
 5  3    3      8     1
 6  4    4      2     0
 7  5    5      6     1
 8  # Alternatively
 9      id    time
10  1    1      5
11  2    2      3+
12  3    3      8
13  4    4      2+
14  5    5      6
```

German Breast Cancer Study: A Working Example

German Breast Cancer (GBC) Study

- **Study Information**

  - **Population**: 686 patients with node-positive breast cancer

  - **Objective**: Assess if tamoxifen + chemo reduces mortality/relapse

  - **Baseline info**: Age, tumor size, hormone levels, menopausal status, etc.

  - **Follow-up**: Median 44 months

    - 171 deaths $\rightarrow$ exact times known

    - 515 censored $\rightarrow$ survival time $>$ censoring time

- **Data sets**:

  - Mortality data: https://lmaowisc.github.io/tidysurv/data/gbc_mort.txt

  - Mortality + relapse: https://lmaowisc.github.io/tidysurv/data/gbc.txt

  - Download and save in a `data` folder under your root directory

Data Format (I)

- **Death only**

```
1  # Load mortality data
2  gbc_mort <- read.table("data/gbc_mort.txt", header = TRUE)
3  # Check the first few rows of the data frame
4  head(gbc_mort)
```

```
  id      time status hormone age meno size grade nodes prog estrg
1  1 74.819672      0       1  38    1   18     3     5  141   105
2  2 65.770492      0       1  52    1   20     1     1   78    14
3  3 47.737705      1       1  47    1   30     2     1  422    89
4  4  4.852459      0       1  40    1   24     1     3   25    11
5  5 61.081967      0       2  64    2   19     2     1   19     9
6  6 63.377049      0       2  49    2   56     1     3  356    64
```

```
1  # The data frame 'gbc_mort' contains:
2  #  time:   time (months) to death or censoring
3  #  status: event indicator (1 = death, 0 = censoring)
4  # hormone: Hormone therapy (1 = no, 2 = yes); age: Age at diagnosis (years);
5  # meno: Menopausal status (1 = no, 2 = yes); size: Tumor size (mm); grade: Tumor grade (1-3);
6  # nodes: Number of positive lymph nodes; prog: Progesterone receptor level (fmol/mg); estrg:
7  # Estrogen receptor level (fmol/mg).
```

Data Format (II)

- **Mortality + relapse**

```
1  # Load mortality + relapse data
2  gbc <- read.table("data/gbc.txt", header = TRUE)
3  # Check the first few rows of the data frame
4  head(gbc)
```

	id	time	status	hormone	age	meno	size	grade	nodes	prog	estrg
1	1	43.83607	1	1	38	1	18	3	5	141	105
2	1	74.81967	0	1	38	1	18	3	5	141	105
3	2	46.55738	1	1	52	1	20	1	1	78	14
4	2	65.77049	0	1	52	1	20	1	1	78	14
5	3	41.93443	1	1	47	1	30	2	1	422	89
6	3	47.73770	2	1	47	1	30	2	1	422	89

```
1  # The data frame 'gbc' contains:
2  #  time:   time (months) to death, relapse, or censoring
3  #  status: event indicator (1 = relapse, 2 = death, 0 = censoring)
4  #  other covariates the same as in gbc_mor.
```

Analysis Goals

- **Descriptive**
  - Summarize patient characteristics
  - Visualize survival distributions

- **Inferential**
  - Compare survival curves (e.g., hormone therapy vs. no hormone therapy)
  - Assess impact of covariates on survival (e.g., age, tumor size, etc.)
  - Model competing risks (e.g., relapse vs. death)

- **Predictive**
  - Develop risk prediction models
  - Evaluate model performance (e.g., concordance index, calibration)

Standard Analysis with `survival` Package

Survival Package Overview

- **Key Functions**

  - `Surv()`: Create survival object

  - `survfit()`: Fit Kaplan-Meier survival curves

  - `survdiff()`: Compare survival curves (log-rank test)

  - `coxph()`: Fit Cox proportional hazards regression models

  - `survreg()`: Fit parametric survival regression models

Kaplan-Meier Survival Curves

- **Create dataset for relapse-free survival**

```r
1  # Sort by subject id, then time
2  o <- order(gbc$id, gbc$time)
3  gbc <- gbc[o,]
4  # Keep only first row per subject => first event
5  df <- gbc[!duplicated(gbc$id), ]
6  # Convert status > 0 to 1 if it is either relapse or death
7  df$status <- ifelse(df$status > 0, 1, 0)
8  head(df)
```

```
  id      time status hormone age meno size grade nodes prog estrg
1  1 43.836066      1       1  38    1   18     3     5  141   105
3  2 46.557377      1       1  52    1   20     1     1   78    14
5  3 41.934426      1       1  47    1   30     2     1  422    89
7  4  4.852459      0       1  40    1   24     1     3   25    11
8  5 61.081967      0       2  64    2   19     2     1   19     9
9  6 63.377049      0       2  49    2   56     1     3  356    64
```

Kaplan-Meier Curves (I)

- Fit Kaplan-Meier survival curves

```r
1  library(survival)
2  # Fit KM curves by hormone treatment group
3  km_fit <- survfit(Surv(time, status) ~ hormone, data = df)
4  km_fit
```

```
Call: survfit(formula = Surv(time, status) ~ hormone, data = df)

            n events median 0.95LCL 0.95UCL
hormone=1 440    205   50.1    42.5    59.5
hormone=2 246     94   66.2    62.9      NA
```

Kaplan-Meier Curves (II)

- **Summarize survival estimates at specified time points**
    - For example, at 6, 12, 24, and 36 months

```
1  summary(km_fit, times = c(6, 12, 24, 36))
```

```
Call: survfit(formula = Surv(time, status) ~ hormone, data = df)

                hormone=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6    419       9    0.979 0.00691        0.966        0.993
   12    379      35    0.897 0.01476        0.868        0.926
   24    280      73    0.720 0.02203        0.678        0.764
   36    195      41    0.606 0.02475        0.559        0.656

                hormone=2
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6    236       4    0.983 0.00826        0.967        1.000
   12    223       8    0.950 0.01418        0.922        0.978
   24    177      38    0.785 0.02701        0.733        0.839
   36    136      16    0.708 0.03047        0.650        0.770
```
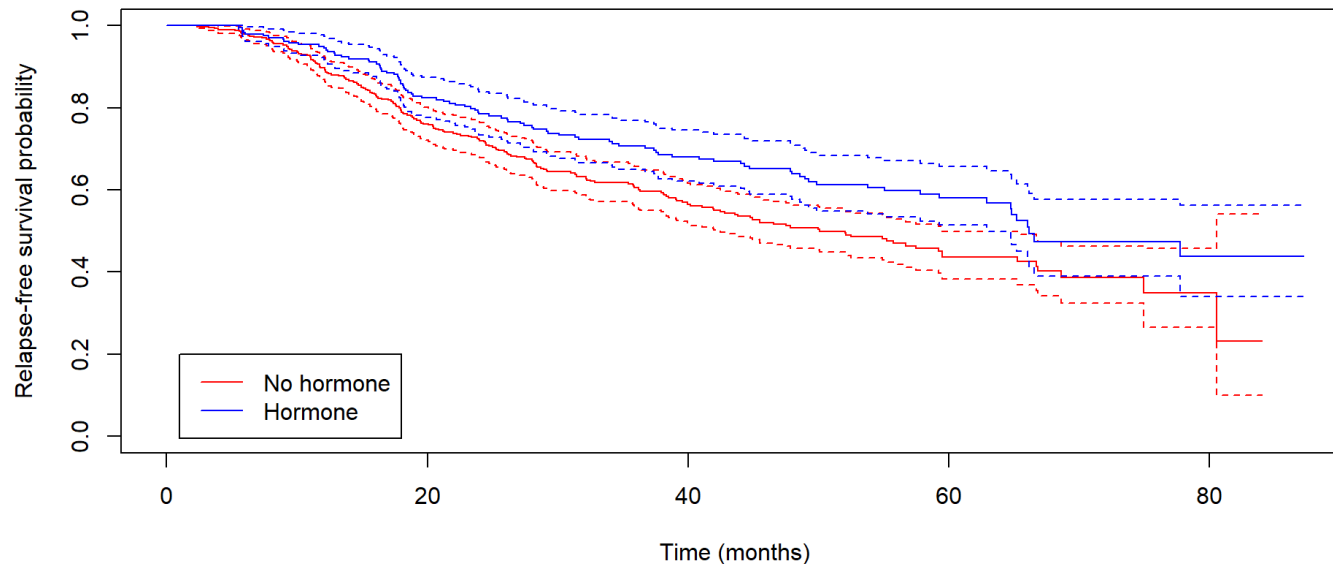
Kaplan-Meier Curves (III)

- **Plot Kaplan-Meier survival curves by group**

```r
1  plot(km_fit, ylim = c(0,1), xlab = "Time (months)", ylab = "Relapse-free survival probability",
2       col = c("red", "blue"), conf.int = TRUE)
3  # Add legend
4  legend(1, 0.2, col=c("red", "blue"), lty = 1,
5         c("No hormone", "Hormone")) # Legend text
```

Log-Rank Test

- ## Compare survival curves between groups

```r
1  lgr_obj <- survdiff(Surv(time, status) ~ hormone, data = df)
2  lgr_obj # Print log-rank test results
```

```
Call:
survdiff(formula = Surv(time, status) ~ hormone, data = df)

          N Observed Expected (O-E)^2/E (O-E)^2/V
hormone=1 440      205      180      3.37      8.56
hormone=2 246       94      119      5.12      8.56

 Chisq= 8.6  on 1 degrees of freedom, p= 0.003
```

```r
1  lgr_obj$pvalue # Extract p-value
```

```
[1] 0.003427282
```

> **Exercise**
>
> Perform a log-rank test on treatment stratified by patient menopausal status `meno`.
>
> ▶ Solution

Cox Model - Model Specification

- **Cox proportional hazards model**

$$\lambda(t \mid Z) = \lambda_0(t) \exp(\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p)$$

  - $\lambda_0(t)$: baseline hazard function

  - $Z = (Z_1, \ldots, Z_p)^{\mathrm{T}}$: covariates (e.g., hormone therapy, age, tumor size)

  - $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$: regression coefficients

  - $\exp(\beta_j)$: hazard ratio for covariate Z_j

- **Proportional hazards (PH) assumption**

$$\frac{\lambda(t \mid Z)}{\lambda(t \mid Z^*)} = \exp\{\beta^{\mathrm{T}}(Z - Z^*)\}$$

  - HR constant over time, i.e., $\beta(t) \equiv \beta$ (for each covariate)

Cox Model - Model Fitting (I)

- **Model fitting**: `survival::coxph()`

```r
1  cox_fit <- coxph(Surv(time, status) ~ hormone + meno + age + grade + size + prog + estrg,
2                   data = df)
3  summary(cox_fit) # Print model summary
```

```
Call:
coxph(formula = Surv(time, status) ~ hormone + meno + age + grade +
    size + prog + estrg, data = df)

  n= 686, number of events= 299


             coef  exp(coef)   se(coef)       z Pr(>|z|)
hormone -0.3422139  0.7101963  0.1290669  -2.651  0.00801 **
meno     0.2765637  1.3185909  0.1837781   1.505  0.13236
age     -0.0087813  0.9912572  0.0093375  -0.940  0.34700
grade    0.2785797  1.3212519  0.1051531   2.649  0.00807 **
size     0.0152793  1.0153966  0.0036877   4.143 3.42e-05 ***
prog    -0.0023307  0.9976720  0.0005803  -4.016 5.91e-05 ***
estrg    0.0001678  1.0001679  0.0004669   0.359  0.71923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

        exp(coef) exp(-coef) lower .95 upper .95
```

Cox Model - Model Fitting (II)

- Extracting $\hat{\beta}$ and $\hat{\text{var}}(\hat{\beta})$

```
1  beta <- cox_fit$coefficients # Estimated coefficients
2  vbeta <- vcov(cox_fit) # Estimated variance-covariance matrix
3  # Extract regression table (as data frame)
4  coef(summary(cox_fit))
```

```
                coef exp(coef)     se(coef)          z     Pr(>|z|)
hormone -0.3422138547 0.7101963 0.1290669350 -2.6514448 8.014821e-03
meno     0.2765636858 1.3185909 0.1837780595  1.5048787 1.323553e-01
age     -0.0087812621 0.9912572 0.0093375120 -0.9404285 3.469978e-01
grade    0.2785796730 1.3212519 0.1051531448  2.6492757 8.066449e-03
size     0.0152793172 1.0153966 0.0036877471  4.1432660 3.423945e-05
prog    -0.0023307288 0.9976720 0.0005803186 -4.0162919 5.912102e-05
estrg    0.0001678465 1.0001679 0.0004669057  0.3594870 7.192308e-01
```

- **Conclusion**
    - Hormone therapy significantly reduces the risk of relapse or death by $1 - 0.710 = 29\%$ ($p = 0.008$)

Cox Model - Prediction (I)

- **Predicted survival function**

$$\hat{S}(t \mid z) = \exp\left\{ -\exp(\hat{\beta}^{\mathrm{T}} z)\hat{\Lambda}_0(t) \right\}$$

- Prepare new data for prediction

```
1  # Create new data for prediction
2  # specify all covariate values
3  new_data <- data.frame(hormone = 1, meno = 1,
4                         age = 45, grade = 2,
5                         size = 20, prog = 100,
6                         estrg = 100)
7  new_data
```

```
  hormone meno age grade size prog estrg
1       1    1  45     2   20  100   100
```

Cox Model - Prediction (II)

- **Predict survival probabilities at specified time points**

```
1  # Predict survival probabilities at 6, 12, 24, 26 months
2  predicted_survival <- survfit(cox_fit, newdata = new_data[1, ], times = c(6, 12, 24, 36))
3  summary(predicted_survival, times = c(6, 12, 24, 36))
```

```
Call: survfit(formula = cox_fit, newdata = new_data[1, ], times = c(6,
    12, 24, 36))

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6    655      13    0.985 0.00441        0.976        0.994
   12    602      43    0.933 0.01059        0.913        0.954
   24    457     111    0.786 0.02304        0.743        0.833
   36    331      57    0.696 0.02925        0.641        0.755
```

Cox Model - Prediction (III)

- **Plot predicted survival function**

```r
# Plot predicted survival function
plot(predicted_survival, ylim = c(0, 1), xlab = "Time (months)",
     ylab = "Relapsep-free survival probability", conf.int = TRUE)
```
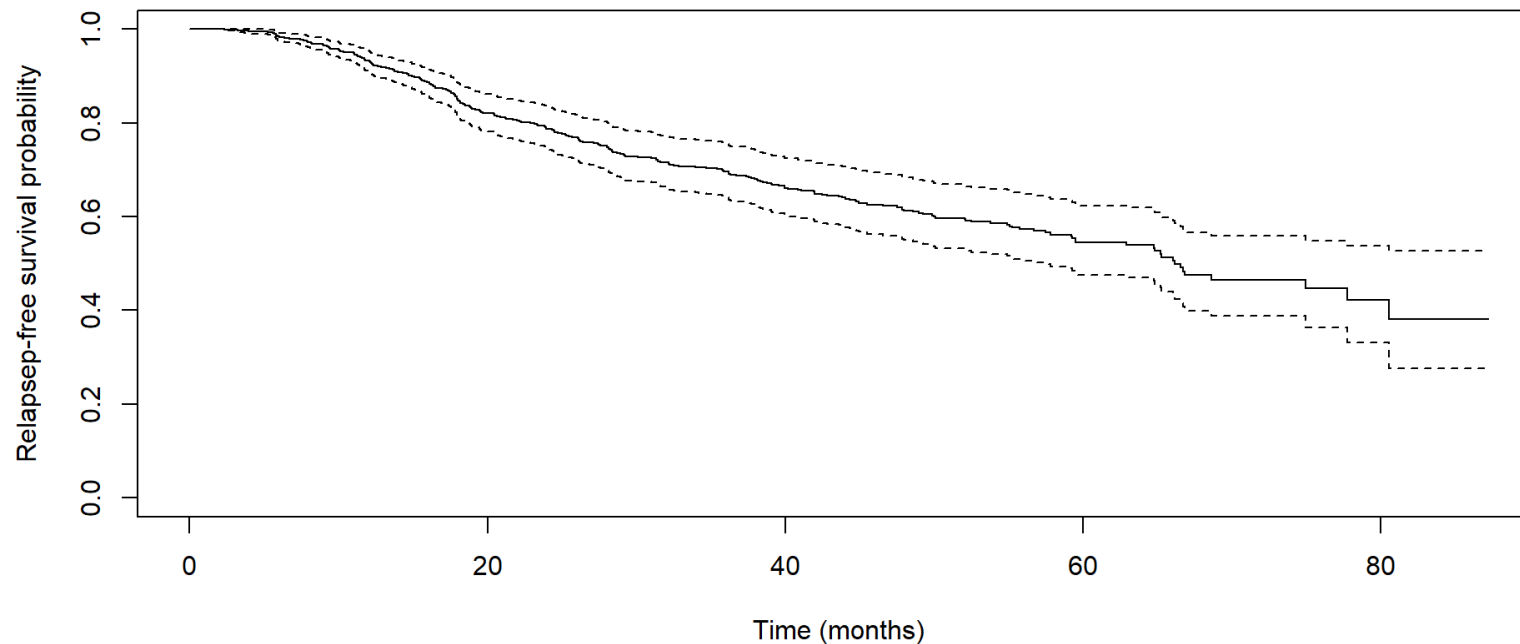
Cox Model - Check PH Assumptions (I)

- **Schoenfeld residuals**
  - Difference between observed and expected covariate values at each event time
 - Use `cox.zph()` to test PH assumption

```
1  ph_test <- cox.zph(cox_fit)
2  ph_test # Print test results
```

```
        chisq df      p
hormone  0.272  1 0.6017
meno     5.514  1 0.0189
age      9.430  1 0.0021
grade    8.490  1 0.0036
size     0.872  1 0.3505
prog     4.881  1 0.0272
estrg    5.403  1 0.0201
GLOBAL  20.636  7 0.0043
```

Cox Model - Check PH Assumptions (II)

- **Graphical check of PH assumptions**
    - Plot Schoenfeld residuals against time

```
1  par(mfrow= c(2, 4)) # Set up 2x4 plotting area for 7 covariates
2  plot(ph_test, se = TRUE, col = "blue", lwd = 2) # Plot Schoenfeld residuals
```
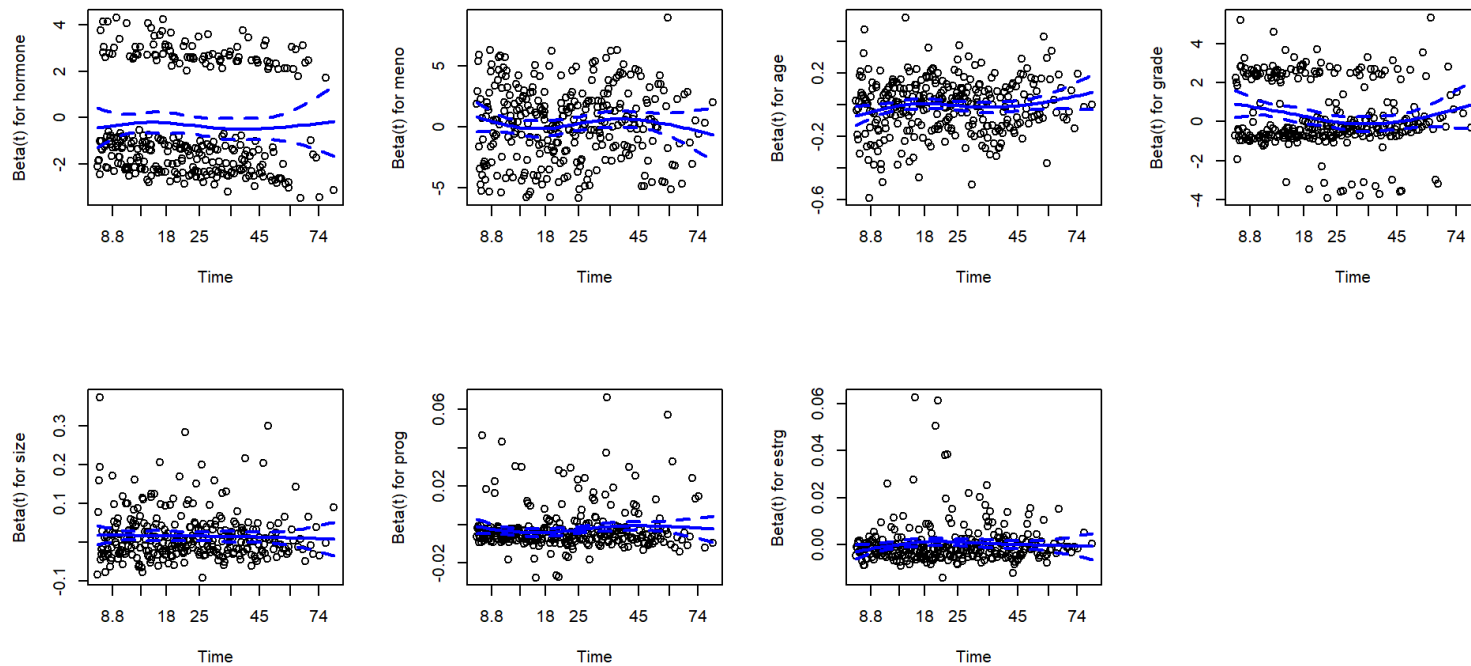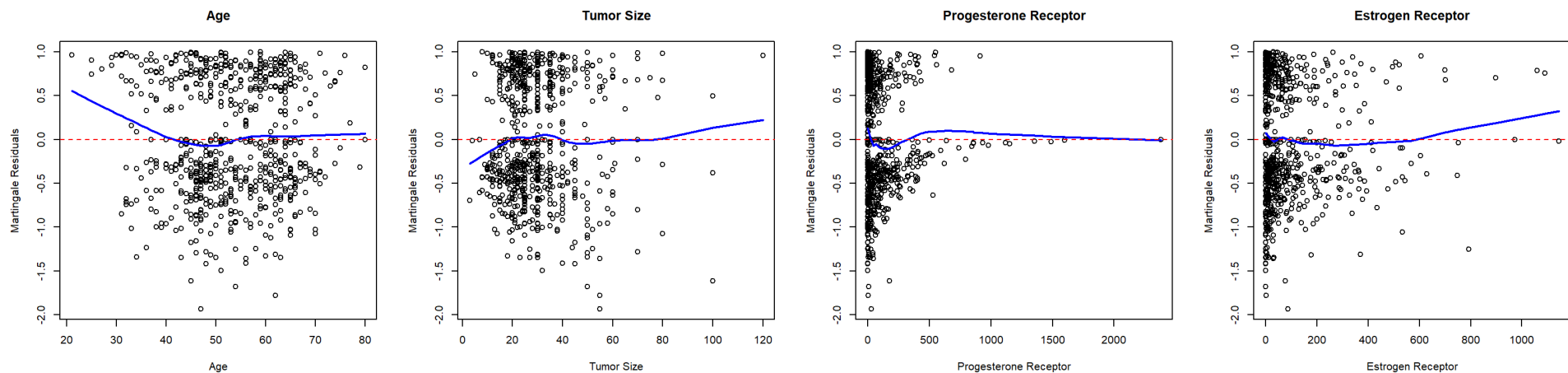
Cox Model - Check Covariate Forms

- **Check linearity of covariate effects**
  - Plot martingale residuals against (quantitative) covariates

```
1  # Extract martingale residuals
2  mart_resid <- residuals(cox_fit, type = 'martingale')
```

▶ Plotting

Coding Exercise

Exercise

Residual analyses show that the proportional hazards assumption is violated for tumor grade, and that the effect of age is not linear.

Fit a different model to address these issues.

▶ Sample solution

Summary

Key Takeaways

- Survival analysis is essential for (often censored) time-to-event data

- Key estimands: survival function, hazard function, cumulative hazard

- Standard analysis tools

  - Kaplan-Meier curves (`survfit()`)

  - Log-rank test (`survdiff()`)

  - Cox proportional hazards model (`coxph()`)

Open Questions

- Efficient/effective presentation of survival probabilities

  - Point estimates, confidence intervals

- Customizable survival curves

  - Add at risk table below graph

- Presentation of regression results

  - Hazard ratios, confidence intervals, p-values

  - Visualize regression results (e.g., forest plots)