The Cox Proportional Hazards Model

Regression Analysis for Survival Data (from chapter 5)

Eric Delmelle

2025-10-01

Link to qmd (quarto markdown)

1 Regression in Survival Analysis

- We learned how to compare survival distributions between groups using nonparametric methods like the log-rank test.
- Now we extend these ideas to **regression analysis** for survival data.

1.1 Why Do We Need Regression for Survival Data?

When analyzing time-to-event data, we often face situations where:

- **Multiple factors** might influence survival (age, sex, treatment, biomarkers)
- We need to **adjust for confounding** variables
- Covariates are **continuous** (not just two groups)
- We want to **predict** survival for specific patient profiles
- We need to **quantify effects** while controlling for other variables

The **Cox Proportional Hazards Model** addresses all these needs while remaining flexible about the baseline hazard function.

1.2 The Proportional Hazards Concept

From Chapter 4, recall that comparing two groups can be expressed as:

$$h_1(t) = \psi \cdot h_0(t)$$

- This **proportional hazards** relationship means the hazard ratio ψ stays constant over time.
- The **Cox model** extends this idea to handle multiple covariates.

```
# Load required packages
library(survival)
library(tidyverse)
library(knitr)
library(ggplot2)
library(survminer)

# Set plotting theme
theme_set(theme_minimal(base_size = 12))

# Custom colors
col_male <- "#E74C3C"
col_female <- "#3498DB"
```

2 The Cox Proportional Hazards Model

2.1 Model Formula

For an individual with covariate vector $\mathbf{z} = (z_1, z_2, \ldots, z_p)$, the Cox model specifies:

$$\boxed{h(t|\mathbf{z}) = h_0(t) \cdot \exp(\beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p)}$$

Or more compactly:

$$h(t|\mathbf{z}) = h_0(t) \cdot \exp(\beta^T \mathbf{z})$$

Where:

- $h_0(t)$ is the **baseline hazard** function (hazard when all covariates = 0)
  - Imagine a reference population (e.g., non-smokers, or women, or age = 0 if centered).
  - The **baseline hazard** describes how the risk evolves over time in this reference group.
  - Other groups are compared to this baseline by multiplying it by a factor.

  - Example: Baseline = "non-smokers". If $\exp(\beta_1)$=2, then a "Smoker" has twice the risk at every instant.

- $\beta = (\beta_1, \ldots, \beta_p)$ are **regression coefficients** to be estimated
- $\mathbf{z} = (z_1, \ldots, z_p)$ is the **covariate vector** for an individual
- The model is **semi-parametric**: no prior assumptions about $h_0(t)$ (like Exponential, Weibull)

2.2 Why we use an Exponential function in Cox model?

Four Key Reasons for Using $\exp(\beta^T \mathbf{z})$

You might wonder: why not just use $h(t|\mathbf{z}) = h_0(t) + \beta^T \mathbf{z}$?

1. Guarantees Positive Hazard

- The hazard rate $h(t)$ can **never be negative** (it's a rate)
- A linear combination $\beta^T \mathbf{z}$ could produce negative values
- But exp(anything) > 0 always!

2. Multiplicative Interpretation (Hazard Ratios)

- We get simple, interpretable **hazard ratios**
- HR = 2 means "twice the hazard"
- HR = 0.5 means "half the hazard"

3. Proportional Hazards Property

- The effect **multiplies** the baseline hazard at all times
- This maintains the proportional hazards assumption
- The ratio $h_1(t)/h_0(t)$ is constant over time

4. Linear on Log Scale

- Taking logs: $\log h(t|\mathbf{z}) = \log h_0(t) + \beta^T \mathbf{z}$
- This is a **linear model** for log hazard
- Similar to logistic regression: $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$

2.2.1 Intuitive Analogy

Think of $h_0(t)$ as the **baseline volume** of music, and $\exp(\beta^T \mathbf{z})$ as a **volume multiplier**:

- HR = 2.0 → doubles the volume
- HR = 0.5 → halves the volume
- HR = 1.0 → no change
- **Volume never becomes negative!**

2.3 Understanding Hazard Ratios

For two individuals with covariate vectors $\mathbf{z}_i$ and $\mathbf{z}_j$:

$$\text{HR}_{i,j} = \frac{h(t|\mathbf{z}_i)}{h(t|\mathbf{z}_j)} = \frac{h_0(t)\exp(\beta^T\mathbf{z}_i)}{h_0(t)\exp(\beta^T\mathbf{z}_j)}$$

The baseline hazard cancels out:

$$\boxed{\text{HR}_{i,j} = \exp[\beta^T(\mathbf{z}_i - \mathbf{z}_j)]}$$

Key insight: The hazard ratio depends only on:

1. The **difference in covariate values** between individuals
2. The **regression coefficients** β

This is powerful: we can estimate effects **without specifying** the form of $h_0(t)$.

3 Application: Lung Cancer Survival Study

Let's apply the Cox model to real data from a lung cancer clinical trial.

3.1 The Dataset

```
# Load lung cancer data
data(lung)

# Clean and prepare the data
lung_clean <- lung %>%
  mutate(
    sex = factor(sex, levels = 1:2, labels = c("Male", "Female")),
    status_binary = status - 1,  # Convert 1/2 to 0/1
    ph.ecog = factor(ph.ecog, levels = 0:3)
  ) %>%
```

```r
  filter(complete.cases(age, sex, ph.karno, status))

# Summary statistics by sex
lung_summary <- lung_clean %>%
  group_by(sex) %>%
  summarise(
    n = n(),
    n_events = sum(status_binary),
    pct_events = round(100 * mean(status_binary), 1),
    median_time = median(time),
    mean_age = round(mean(age), 1),
    sd_age = round(sd(age), 1),
    mean_karno = round(mean(ph.karno), 1)
  )

kable(lung_summary,
      col.names = c("Sex", "N", "Events", "% Events", "Median Time",
                    "Mean Age", "SD Age", "Mean Karnofsky"),
      caption = "Table 1: Lung Cancer Dataset Summary Statistics")
```

Table 1: Table 1: Lung Cancer Dataset Summary Statistics

Sex	N	Events	% Events	Median Time	Mean Age	SD Age	Mean Karnofsky
Male	137	111	81.0	225.0	63.4	9.2	81.8
Female	90	53	58.9	292.5	61.1	8.8	82.1

Dataset description:

- **n = 227 patients** from a lung cancer trial
- **Outcome**: time to death (in days)
- **Covariates**: age, sex, ECOG performance status, Karnofsky performance score
- **Censoring**: 27.8% of observations are censored

3.2 Exploratory Survival Analysis

Before fitting the Cox model, let's visualize survival by sex:

```r
# Fit Kaplan-Meier curves
km_sex <- survfit(Surv(time, status_binary) ~ sex, data = lung_clean)
```

```r
# Plot with survminer
ggsurvplot(km_sex,
           data = lung_clean,
           pval = TRUE,
           conf.int = TRUE,
           risk.table = TRUE,
           palette = c(col_male, col_female),
           xlab = "Time (days)",
           ylab = "Survival Probability",
           title = "Kaplan-Meier Survival Curves by Sex",
           legend.labs = c("Male", "Female"),
           ggtheme = theme_minimal(base_size = 14))
```

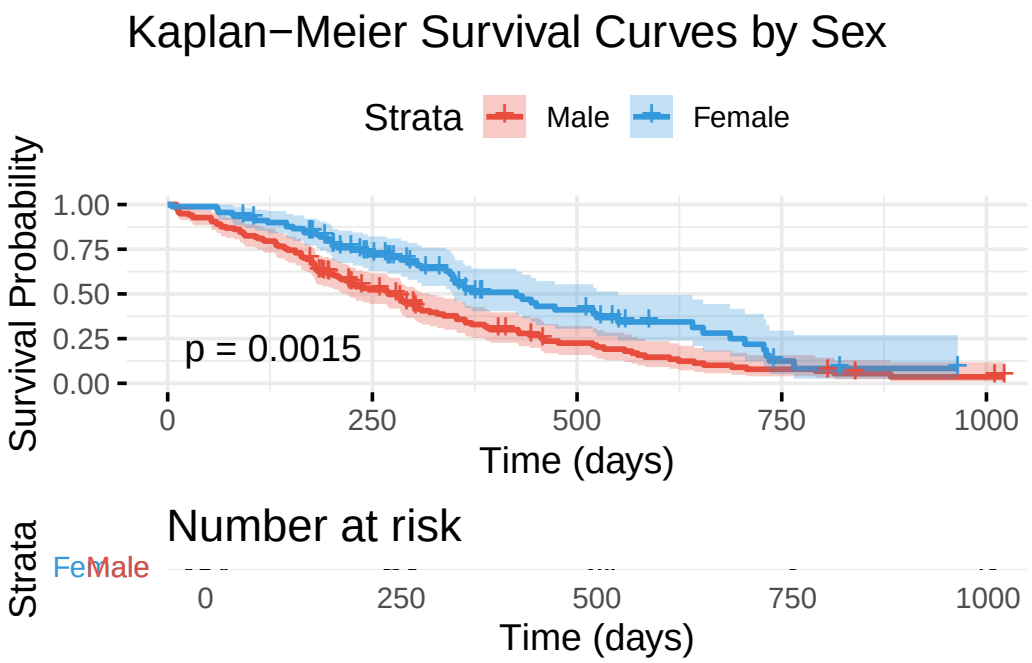

Figure 1: Figure 1: Kaplan-Meier survival curves by sex

Observation: Females appear to have better survival than males ($p < 0.05$ by log-rank test).

3.3 Fitting a Simple Cox Model

Let's start with sex as the only predictor:

```r
# Fit Cox model with sex only
fit_sex <- coxph(Surv(time, status_binary) ~ sex, data = lung_clean)

# Display summary
summary(fit_sex)
```

```
Call:
coxph(formula = Surv(time, status_binary) ~ sex, data = lung_clean)

  n= 227, number of events= 164

             coef exp(coef) se(coef)      z Pr(>|z|)
sexFemale -0.5241    0.5921   0.1674 -3.131  0.00174 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
sexFemale    0.5921      1.689    0.4265     0.822

Concordance= 0.578  (se = 0.021 )
Likelihood ratio test= 10.32  on 1 df,   p=0.001
Wald test            = 9.8  on 1 df,   p=0.002
Score (logrank) test = 10.02  on 1 df,   p=0.002
```

3.3.1 Interpreting the Results

Coefficient: $\hat{\beta}_{\text{Female}} = -0.524$

Hazard Ratio: $\exp(\hat{\beta}) = 0.592$

Interpretation:

- Females have **40.8% lower hazard** of death compared to males
- Or equivalently: males have **1.69 times the hazard** of females
- The effect is **statistically significant** ($p = 0.002$)

3.3.2 Visualizing the Effect

```r
# Create data for forest plot
hr_data <- data.frame(
  Variable = "Female vs Male",
  HR = exp(coef(fit_sex)),
  Lower = exp(confint(fit_sex)[1]),
  Upper = exp(confint(fit_sex)[2])
)

ggplot(hr_data, aes(x = HR, y = Variable)) +
  geom_vline(xintercept = 1, linetype = "dashed", color = "gray50", linewidth = 1) +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.2, linewidth = 1.5) +
  geom_point(size = 5, color = col_female) +
  geom_text(aes(label = sprintf("HR = %.2f\n95%% CI: [%.2f, %.2f]", HR, Lower, Upper)),
            vjust = -1.5, size = 4) +
  scale_x_continuous(breaks = seq(0.4, 1.2, 0.2), limits = c(0.4, 1.2)) +
  labs(title = "Hazard Ratio with 95% Confidence Interval",
       subtitle = "Females vs Males",
       x = "Hazard Ratio", y = "") +
  theme_minimal(base_size = 14) +
  theme(axis.text.y = element_text(size = 12, face = "bold"))
```

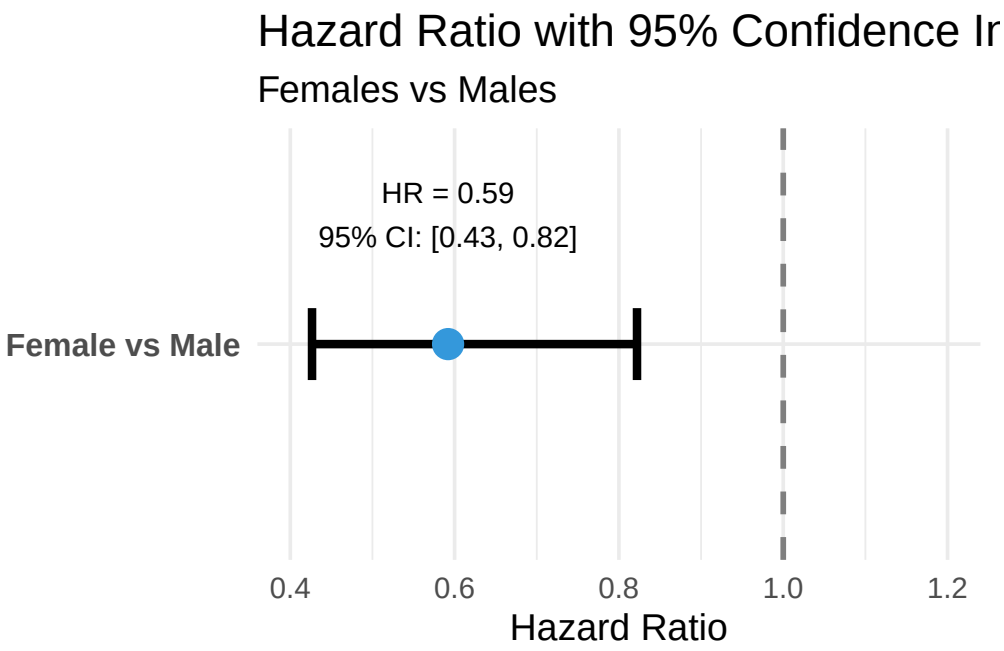

Figure 2: Figure 2: Hazard ratio for sex effect

4 Multiple Covariates Model

Now let's build a more comprehensive model that adjusts for multiple factors.

4.1 Adding Age and Performance Status

```
# Fit Cox model with multiple covariates
fit_multi <- coxph(Surv(time, status_binary) ~ age + sex + ph.karno,
                   data = lung_clean)

summary(fit_multi)
```

```
Call:
coxph(formula = Surv(time, status_binary) ~ age + sex + ph.karno,
    data = lung_clean)

  n= 227, number of events= 164

              coef exp(coef)  se(coef)      z Pr(>|z|)
```

```
age        0.012375  1.012452  0.009405  1.316  0.18821
sexFemale -0.497170  0.608249  0.167713 -2.964  0.00303 **
ph.karno  -0.013322  0.986767  0.005880 -2.266  0.02348 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef) exp(-coef) lower .95 upper .95
age          1.0125     0.9877    0.9940    1.0313
sexFemale    0.6082     1.6441    0.4378    0.8450
ph.karno     0.9868     1.0134    0.9755    0.9982

Concordance= 0.637  (se = 0.025 )
Likelihood ratio test= 18.81  on 3 df,   p=3e-04
Wald test            = 18.73  on 3 df,   p=3e-04
Score (logrank) test = 19.05  on 3 df,   p=3e-04
```

4.1.1 Interpreting Multiple Coefficients

Let's break down each covariate:

```r
# Extract coefficients and create summary table
coef_summary <- data.frame(
  Covariate = c("Age (per year)", "Sex: Female", "Karnofsky Score (per point)"),
  Beta = coef(fit_multi),
  HR = exp(coef(fit_multi)),
  Lower_CI = exp(confint(fit_multi)[,1]),
  Upper_CI = exp(confint(fit_multi)[,2]),
  P_value = summary(fit_multi)$coefficients[,5]
)

kable(coef_summary,
      digits = c(0, 3, 3, 3, 3, 4),
      col.names = c("Covariate", " ", "HR", "95% CI Lower", "95% CI Upper", "P-value"),
      caption = "Table 2: Cox Model Coefficients and Hazard Ratios")
```

Table 2: Table 2: Cox Model Coefficients and Hazard Ratios

	Covariate		HR	95% CI Lower	95% CI Upper	P-value
age	Age (per year)	0.012	1.012	0.994	1.031	0.1882
sexFemale	Sex: Female	-0.497	0.608	0.438	0.845	0.0030

	Covariate		HR	95% CI Lower	95% CI Upper	P-value
ph.karno	Karnofsky Score (per point)	-0.013	0.987	0.975	0.998	0.0235

Age:

- $\exp(\hat{\beta}_{\text{age}}) = 1.012$
- Each additional year of age increases hazard by 1.25%
- For a 10-year difference: HR = 1.13 (13% increase)

Sex:

- $\exp(\hat{\beta}_{\text{Female}}) = 0.608$
- After adjusting for age and performance score, females still have 39.2% lower hazard
- This effect remains **highly significant** ($p < 0.001$)

Karnofsky Performance Score:

- $\exp(\hat{\beta}_{\text{karno}}) = 0.987$
- Each 1-point increase in score reduces hazard by 1.3%
- For a 10-point difference: HR = 0.875 (12.5% reduction)

4.1.2 Forest Plot for Multiple Covariates

```
# Prepare data for forest plot (using meaningful scales)
forest_data <- data.frame(
  Variable = c("Age (10-year increase)",
               "Sex: Female vs Male",
               "Karnofsky Score (10-point increase)"),
  HR = c(exp(10 * coef(fit_multi)["age"]),
         exp(coef(fit_multi)["sexFemale"]),
         exp(10 * coef(fit_multi)["ph.karno"])),
  Lower = c(exp(10 * confint(fit_multi)["age", 1]),
            exp(confint(fit_multi)["sexFemale", 1]),
            exp(10 * confint(fit_multi)["ph.karno", 1])),
  Upper = c(exp(10 * confint(fit_multi)["age", 2]),
            exp(confint(fit_multi)["sexFemale", 2]),
            exp(10 * confint(fit_multi)["ph.karno", 2]))
)
```

```r
ggplot(forest_data, aes(y = reorder(Variable, -HR), x = HR)) +
  geom_vline(xintercept = 1, linetype = "dashed", color = "gray50", linewidth = 1) +
  geom_errorbarh(aes(xmin = Lower, xmax = Upper), height = 0.3, linewidth = 1.2) +
  geom_point(size = 5, color = "#3498DB") +
  geom_text(aes(label = sprintf("%.2f [%.2f, %.2f]", HR, Lower, Upper)),
            vjust = -2.4, hjust = .42, size = 3) +
  scale_x_log10(breaks = c(0.4, 0.6, 0.8, 1.0, 1.2, 1.5, 2.0)) +
  labs(title = "Adjusted Hazard Ratios with 95% Confidence Intervals",
       subtitle = "Cox Proportional Hazards Model",
       x = "Hazard Ratio (log scale)",
       y = "",
       caption = "HR < 1 indicates better survival; HR > 1 indicates worse survival") +
  theme_minimal(base_size = 14) +
  theme(plot.caption = element_text(hjust = 0, face = "italic", size = 10))
```

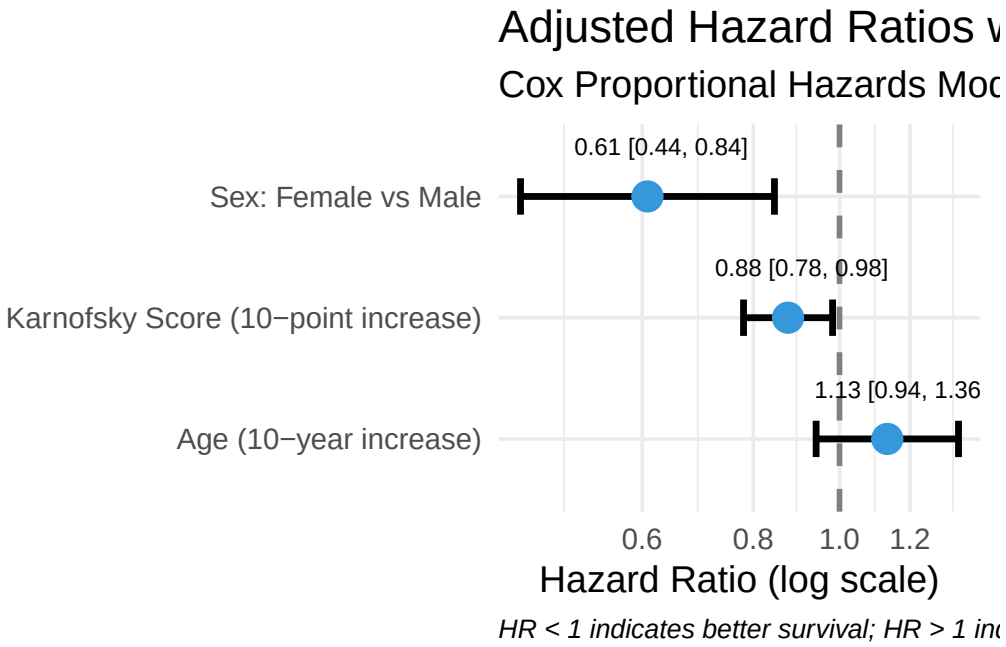

Figure 3: Figure 3: Forest plot of hazard ratios for multiple covariates

4.2 Hypothesis Testing

The Cox model output provides **three hypothesis tests** for $H_0 : \beta = \mathbf{0}$ (no effect):

4.2.1 1. Wald Test

Uses the estimated coefficient and its standard error:

$$Z_w = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \sim N(0,1)$$

Most commonly reported in tables (z-statistic and p-value).

4.2.2 2. Score Test (Log-Rank)

For a single binary covariate, equivalent to the log-rank test. Uses the score function at $\beta = 0$.

4.2.3 3. Likelihood Ratio Test

Compares the log-likelihood of the full model to the null model. Most reliable for small samples and invariant to parameter transformations.

```
# Extract test statistics
wald_stat <- summary(fit_multi)$waldtest[1]
score_stat <- summary(fit_multi)$sctest[1]
lr_stat <- summary(fit_multi)$logtest[1]

test_comparison <- data.frame(
  Test = c("Wald", "Score", "Likelihood Ratio"),
  Chi_square = c(wald_stat, score_stat, lr_stat),
  df = rep(3, 3),
  P_value = c(summary(fit_multi)$waldtest[3],
              summary(fit_multi)$sctest[3],
              summary(fit_multi)$logtest[3])
)

kable(test_comparison,
      digits = c(0, 2, 0, 5),
      col.names = c("Test", " ²", "df", "P-value"),
      caption = "Table 3: Global Tests of Model Significance")
```

Table 3: Table 3: Global Tests of Model Significance

Test	^{2}	df	P-value
Wald	18.73	3	0.00031
Score	19.05	3	0.00027
Likelihood Ratio	18.81	3	0.00030

All three tests strongly reject the null hypothesis that all coefficients are zero ($p < 0.001$).

5 Predicted Survival Curves

One of the most useful applications of the Cox model is predicting survival for individuals with specific covariate patterns.

5.1 Creating Patient Profiles

```
# Define typical patient profiles
patient_profiles <- data.frame(
  Profile = c("Young Male, Good Status",
              "Young Female, Good Status",
              "Older Male, Poor Status",
              "Older Female, Poor Status"),
  age = c(50, 50, 70, 70),
  sex = factor(c("Male", "Female", "Male", "Female")),
  ph.karno = c(90, 90, 70, 70)
)

kable(patient_profiles,
      caption = "Table 4: Patient Profiles for Prediction")
```

Table 4: Table 4: Patient Profiles for Prediction

Profile	age	sex	ph.karno
Young Male, Good Status	50	Male	90
Young Female, Good Status	50	Female	90
Older Male, Poor Status	70	Male	70
Older Female, Poor Status	70	Female	70

5.2 Computing Survival Curves

```r
# Generate survival curves for each profile
surv_pred <- survfit(fit_multi, newdata = patient_profiles)

# Plot with ggsurvplot
ggsurvplot(surv_pred,
           data = patient_profiles,
           conf.int = FALSE,
           legend.labs = patient_profiles$Profile,
           palette = c("#E74C3C", "#3498DB", "#C0392B", "#2980B9"),
           xlab = "Time (days)",
           ylab = "Survival Probability",
           title = "Predicted Survival Curves by Patient Profile",
           ggtheme = theme_minimal(base_size = 14),
           legend = "right")
```

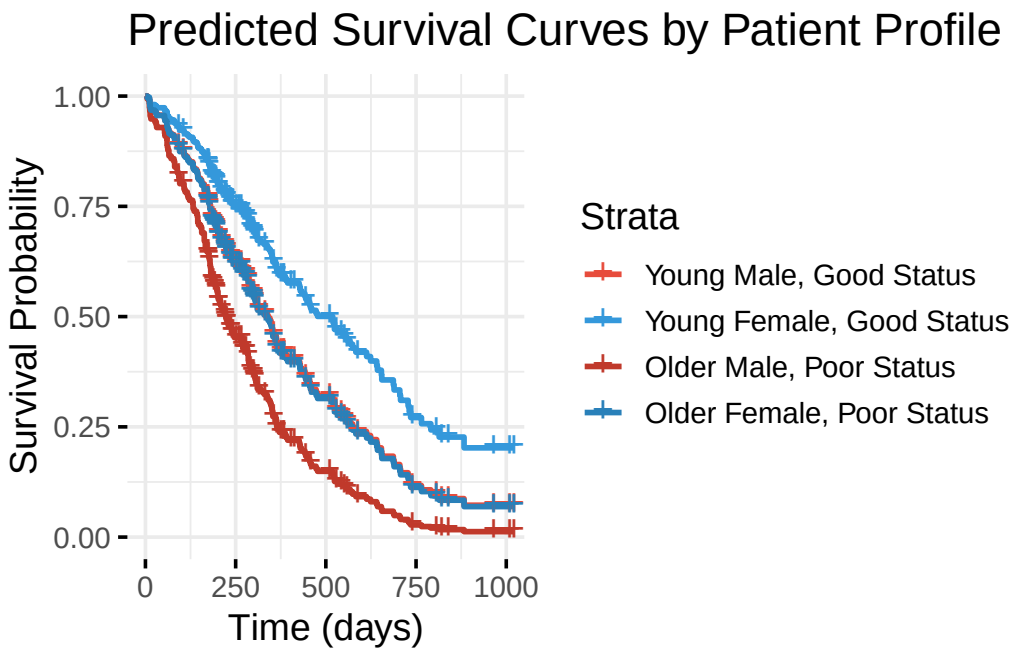

Figure 4: Figure 4: Predicted survival curves for different patient profiles

5.2.1 Predicted Median Survival Times

```
# Extract median survival times
median_survival <- summary(surv_pred)$table[, "median"]

median_table <- data.frame(
  Profile = patient_profiles$Profile,
  Predicted_Median_Survival = round(median_survival, 0)
)

kable(median_table,
      col.names = c("Patient Profile", "Predicted Median Survival (days)"),
      caption = "Table 5: Predicted Median Survival Times")
```

Table 5: Table 5: Predicted Median Survival Times

Patient Profile	Predicted Median Survival (days)
Young Male, Good Status	340
Young Female, Good Status	519
Older Male, Poor Status	223
Older Female, Poor Status	337

Key observations:

- **Sex effect**: Females consistently have better survival than males with similar characteristics
- **Age effect**: Older patients have shorter survival times
- **Performance status**: Better Karnofsky scores dramatically improve survival
- **Combined effects**: A 70-year-old male with poor performance has substantially worse prognosis than a 50-year-old female with good performance

6 Model Diagnostics and Assumptions

The Cox model makes an important assumption: **proportional hazards**. Let's verify this and check other model assumptions.

6.1 Testing the Proportional Hazards Assumption

The proportional hazards assumption states that hazard ratios are **constant over time**. We can test this using **Schoenfeld residuals**.

```
# Test proportional hazards assumption
ph_test <- cox.zph(fit_multi)
print(ph_test)
```

```
           chisq df      p
age        0.478  1 0.4892
sex        3.085  1 0.0790
ph.karno   8.017  1 0.0046
GLOBAL    10.359  3 0.0157
```

Interpretation:

- **p-value $>$ 0.05**: No evidence against PH assumption
- **p-value $<$ 0.05**: PH assumption may be violated

For our model:

- All individual covariates have p $>$ 0.05
- Global test also shows p $>$ 0.05
- **Conclusion**: PH assumption appears reasonable

6.1.1 Visualizing Schoenfeld Residuals

```
# Create plots for each covariate
par(mfrow = c(2, 2))
plot(ph_test)
par(mfrow = c(1, 1))
```

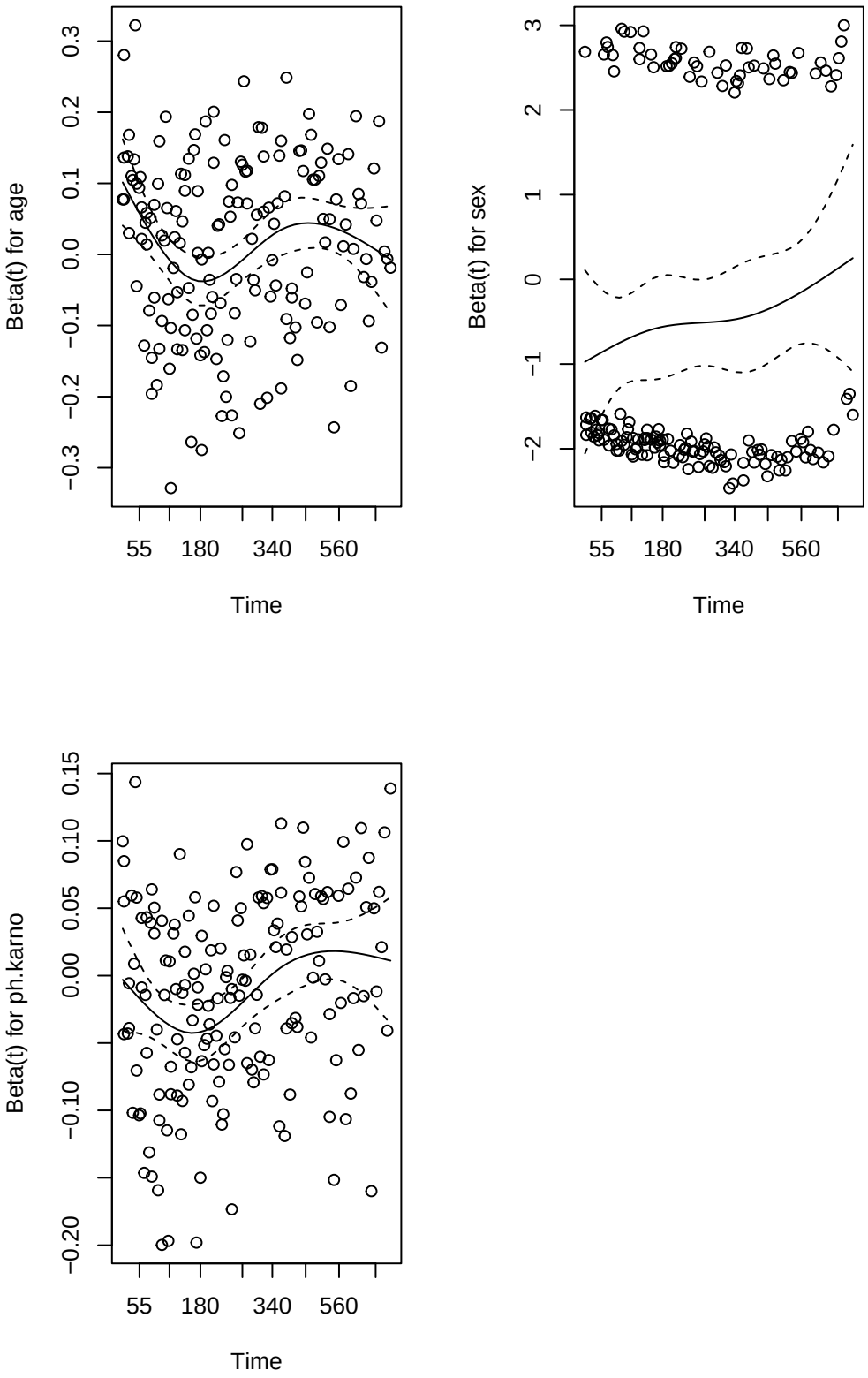

Figure 5: Figure 5: Schoenfeld residual plots for assessing proportional hazards

What to look for:

- **Flat trend** (horizontal line) → PH assumption holds
- **Clear pattern** (upward/downward trend) → potential PH violation
- The smooth curve should be relatively flat around zero

6.2 Checking Functional Form: Martingale Residuals

For continuous covariates (like age), we should verify that the **linear relationship** is appropriate.

```
# Calculate martingale residuals
mart_resid <- residuals(fit_multi, type = "martingale")

# Plot against age
lung_clean %>%
  mutate(mart_resid = mart_resid) %>%
  ggplot(aes(x = age, y = mart_resid)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "loess", color = "#E74C3C", se = TRUE, linewidth = 1.5) +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Martingale Residuals vs. Age",
       subtitle = "Checking linearity assumption for age effect",
       x = "Age (years)",
       y = "Martingale Residuals") +
  theme_minimal(base_size = 14)
```

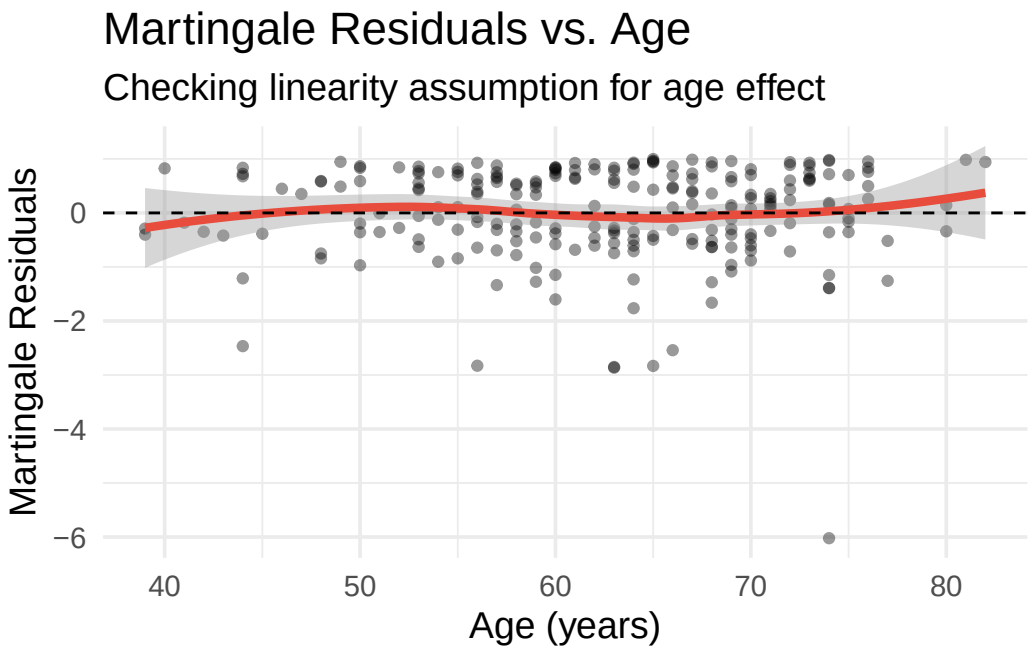

Figure 6: Figure 6: Martingale residuals vs. age

Interpretation:

- If the smooth curve is approximately **linear**, the linear form is appropriate
- **Non-linear patterns** suggest we might need transformations (e.g., age^2, log(age), splines)
- Our plot shows a reasonably linear relationship

6.3 Identifying Influential Observations

Deviance residuals help identify outliers or influential observations.

```
# Calculate deviance residuals
dev_resid <- residuals(fit_multi, type = "deviance")

# Create index plot
lung_clean %>%
  mutate(
    dev_resid = dev_resid,
    obs_id = row_number(),
    outlier = abs(dev_resid) > 3
  ) %>%
```

```r
ggplot(aes(x = obs_id, y = dev_resid, color = outlier)) +
geom_point(alpha = 0.6, size = 2) +
geom_hline(yintercept = c(-3, -2, 0, 2, 3),
           linetype = c("dashed", "dotted", "solid", "dotted", "dashed"),
           color = c("red", "orange", "gray50", "orange", "red")) +
scale_color_manual(values = c("FALSE" = "#3498DB", "TRUE" = "#E74C3C")) +
labs(title = "Deviance Residuals: Identifying Influential Observations",
     subtitle = "Points beyond ±3 may be outliers",
     x = "Observation Index",
     y = "Deviance Residuals") +
theme_minimal(base_size = 14) +
theme(legend.position = "none")
```

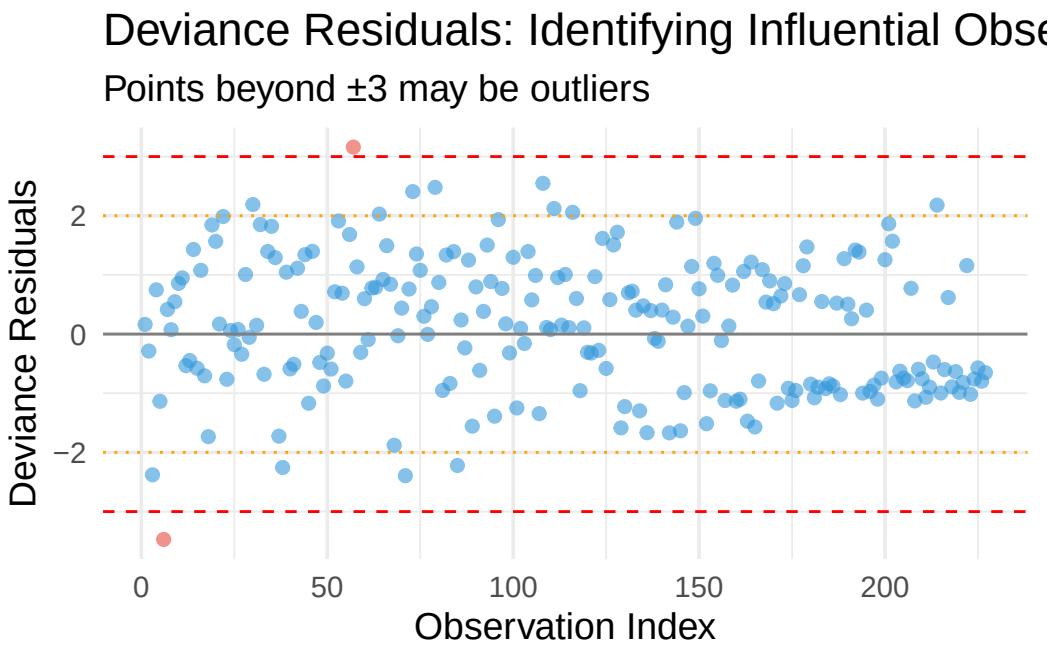

Figure 7: Figure 7: Deviance residuals for identifying outliers

Guidelines:

- Residuals between $\pm$**2**: typical
- Residuals beyond $\pm$**2**: potentially unusual
- Residuals beyond $\pm$**3**: likely outliers (investigate further)

```r
# Identify extreme residuals
n_outliers <- sum(abs(dev_resid) > 3)
pct_outliers <- round(100 * n_outliers / length(dev_resid), 1)

cat(sprintf("Number of potential outliers (|residual| > 3): %d (%.1f%%)\n",
            n_outliers, pct_outliers))
```

```
Number of potential outliers (|residual| > 3): 2 (0.9%)
```

6.4 Model Comparison: Which Model is Better?

Let's compare our multiple covariate model to simpler alternatives using the **Akaike Information Criterion (AIC)**.

```r
# Fit alternative models
fit_null <- coxph(Surv(time, status_binary) ~ 1, data = lung_clean)
fit_sex_only <- coxph(Surv(time, status_binary) ~ sex, data = lung_clean)
fit_age_sex <- coxph(Surv(time, status_binary) ~ age + sex, data = lung_clean)
fit_full <- fit_multi

# Compare AIC values
model_comp <- data.frame(
  Model = c("Null (no predictors)",
            "Sex only",
            "Age + Sex",
            "Age + Sex + Karnofsky"),
  AIC = c(AIC(fit_null), AIC(fit_sex_only), AIC(fit_age_sex), AIC(fit_full)),
  n_parameters = c(0, 1, 2, 3)
) %>%
  arrange(AIC) %>%
  mutate(Delta_AIC = AIC - min(AIC))

kable(model_comp,
      digits = 1,
      caption = "Table 6: Model Comparison Using AIC")
```

Table 6: Table 6: Model Comparison Using AIC

Model	AIC	n_parameters	Delta_AIC
Age + Sex + Karnofsky	1476.2	3	0.0

Model	AIC	n_parameters	Delta_AIC
Age + Sex	1479.1	2	3.0
Sex only	1480.7	1	4.5
Null (no predictors)	1489.0	0	12.8

Interpretation:

- **Lower AIC is better** (indicates better model fit while penalizing complexity)
- Models with $\mathbf{\Delta AIC < 2}$ are considered equivalent
- The full model (Age + Sex + Karnofsky) has the **lowest AIC**, suggesting it's the best model

7 Advanced Topics

7.1 Stratified Cox Model

When the proportional hazards assumption is violated for a categorical variable, we can **stratify** by that variable:

```
# Example: stratify by sex if PH assumption fails
fit_stratified <- coxph(Surv(time, status_binary) ~ age + ph.karno + strata(sex),
                        data = lung_clean)
```

This allows **different baseline hazards** for each sex while assuming age and Karnofsky score have the **same effects** across sexes.

7.2 Time-Varying Covariates

When covariates change over time (e.g., treatment changes, biomarker values), we can model them as **time-dependent**:

$$h(t|\mathbf{z}(t)) = h_0(t) \exp[\beta^T \mathbf{z}(t)]$$

This requires special data formatting (multiple rows per subject).

8 Summary

Key Takeaways

1. Cox Model Formula

$$h(t|\mathbf{z}) = h_0(t) \cdot \exp(\beta^T \mathbf{z})$$

2. Why Exponential Function?

- Guarantees positive hazard
- Multiplicative interpretation (hazard ratios)
- Maintains proportional hazards
- Linear on log scale

3. Hazard Ratio Interpretation

$$\text{HR} = \exp(\beta^T \Delta\mathbf{z})$$

- HR > 1: increased hazard (worse survival)
- HR < 1: decreased hazard (better survival)
- HR $= 1$: no effect

4. Model Assumptions

- **Proportional hazards**: test with Schoenfeld residuals
- **Linearity**: check with martingale residuals
- **No influential outliers**: examine deviance residuals

5. Multiple Covariates

- Adjust for confounding
- Quantify independent effects
- Make predictions for specific profiles

6. Practical Modeling Steps

1. Explore data with Kaplan-Meier curves
2. Fit Cox model with relevant covariates
3. Interpret hazard ratios and p-values
4. Check model assumptions
5. Compare alternative models
6. Generate predicted survival curves

9 Exercises

9.1 Exercise 1: Basic Cox Model

Using the `aml` dataset from the `survival` package:

```
data(aml)
head(aml)
```

```
  time status          x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained
```

Tasks:

a. Fit a Cox model comparing maintained vs. non-maintained groups
b. Interpret the hazard ratio
c. Is the treatment effect statistically significant?
d. Plot the predicted survival curves for both groups

9.2 Exercise 2: Multiple Covariates

Using the `lung` dataset:

Tasks:

a. Fit a model including: age, sex, ECOG performance status (`ph.ecog`), and weight loss (`wt.loss`)
b. Which variables are statistically significant?
c. Calculate the hazard ratio for a 10-year age difference
d. Create a forest plot of all hazard ratios

9.3 Exercise 3: Model Diagnostics

For your model from Exercise 2:

Tasks:

a. Test the proportional hazards assumption
b. Create Schoenfeld residual plots
c. Check for outliers using deviance residuals
d. Is there evidence of any assumption violations?

9.4 Exercise 4: Prediction

Using your model from Exercise 2:

Tasks:

a. Create three patient profiles:
    - Low risk: 50 years, female, ECOG=0, no weight loss
    - Medium risk: 65 years, male, ECOG=1, 5 lbs loss
    - High risk: 75 years, male, ECOG=2, 15 lbs loss
b. Plot predicted survival curves for these profiles
c. Calculate predicted median survival for each
d. What is the ratio of hazards between high-risk and low-risk patients?

9.5 Exercise 5: Model Comparison

Tasks:

a. Fit three models: (1) sex only, (2) age + sex, (3) age + sex + ECOG + weight loss
b. Compare them using AIC
c. Which model would you choose and why?
d. Does adding ECOG and weight loss significantly improve the model? (Hint: use likelihood ratio test)

10 References

Primary Sources:

- Cox, D.R. (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B*, 34(2), 187-202.
- Cox, D.R. (1975). "Partial likelihood." *Biometrika*, 62(2), 269-276.

Textbooks:

- Therneau, T.M. & Grambsch, P.M. (2000). *Modeling Survival Data: Extending the Cox Model*. Springer.
- Klein, J.P. & Moeschberger, M.L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer.
- Moore, D.F. (2016). *Applied Survival Analysis Using R*. Springer.
- Collett, D. (2015). *Modelling Survival Data in Medical Research* (3rd ed.). Chapman & Hall/CRC.

R Packages:

- Therneau, T. (2023). *survival*: Survival Analysis. R package.
- Kassambara, A. (2021). *survminer*: Drawing Survival Curves using 'ggplot2'. R package.

Statistical Theory:

- Andersen, P.K., Borgan, O., Gill, R.D., & Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer.
- Fleming, T.R. & Harrington, D.P. (2011). *Counting Processes and Survival Analysis*. Wiley.

← Return to Course Materials