# Survival Analysis in R

## Censoring-aware summaries, Kaplan–Meier curves and Cox models

Rverse Analytics

**Time and event must travel together.** `Surv(time, status)` represents follow-up with right censoring; conventionally `status = 1` means the event occurred and `0` means censored.

## Start with the estimand

| Question | Method | Function | Main output |
|---|---|---|---|
| Survival over time | Kaplan–Meier | `survfit()` | S(t), median survival |
| Compare entire curves | Log-rank test | `survdiff()` | $\chi^2$ test |
| Adjusted relative hazard | Cox PH | `coxph()` | hazard ratio |
| Absolute survival at time *t* | standardized survival | `survfit(cox_model, newdata = ...)` | adjusted S(t) |
| Competing event prevents outcome | cumulative incidence | `tidycmprsk::cuminc()` | cause-specific risk |
| Covariate changes over time | time-varying Cox | `coxph(... + tt(x))` | time-varying HR |

## Kaplan–Meier workflow

```r
library(survival)
library(gtsummary)

# Verify coding and follow-up before fitting
with(d, table(status, useNA = "ifany"))
summary(d$time)

km <- survfit(Surv(time, status) ~ treatment, data = d)
summary(km, times = c(30, 90, 180, 365))
quantile(km, probs = 0.5, conf.int = TRUE)

# Overall log-rank comparison
survdiff(Surv(time, status) ~ treatment, data = d)

# Clean table at prespecified times
tbl_survfit(km, times = c(90, 180, 365),
            label_header = "**{time} days**") |>
  add_p()

# Base R curve; keep confidence bands and censoring marks
plot(km, conf.int = TRUE, mark.time = TRUE,
     xlab = "Days", ylab = "Survival probability")
legend("bottomleft", legend = levels(d$treatment), lty = seq_along(levels(d$treatment)))
```

## Cox proportional hazards model

```r
d$treatment <- relevel(factor(d$treatment), ref = "Control")

cox <- coxph(
  Surv(time, status) ~ treatment + age + stage,
  data = d,
  ties = "efron",
  x = TRUE, y = TRUE
)

summary(cox)                         # coefficients, HRs, tests
broom::tidy(cox, exponentiate = TRUE, conf.int = TRUE)
tbl_regression(cox, exponentiate = TRUE) |>
  add_global_p() |>
  bold_labels()
```

**Read the HR carefully:** `HR = 0.70` is a 30% lower instantaneous event rate under the model. It is not “30% lower risk by day 365” and does not give an absolute effect.

## Proportional-hazards check

```
ph <- cox.zph(cox, transform = "km")
print(ph)  # variable-specific and global tests
plot(ph)   # scaled Schoenfeld residuals over time

# If treatment is non-proportional, choose a scientifically defensible response
cox_strata <- coxph(Surv(time, status) ~ age + stage + strata(treatment), data = d)
cox_time   <- coxph(Surv(time, status) ~ treatment + tt(treatment) + age,
                    data = d, tt = function(x, t, ...) x * log(t + 1))
```

| Pattern | Response | Consequence |
|---|---|---|
| Baseline factor violates PH, effect not needed | `strata(factor)` | no HR for factor |
| Effect changes smoothly | time interaction / `tt()` | report HR at meaningful times |
| Clinically distinct periods | piecewise effect | prespecify cut points |
| HR is a poor summary | RMST / survival at fixed times | report absolute time-scale effect |

## Data structures beyond one row per patient

```
# Delayed entry (left truncation)
Surv(entry_time, exit_time, status)

# Start-stop rows for time-varying covariates
coxph(Surv(tstart, tstop, status) ~ biomarker + cluster(id), data = long_d)

# Robust SE for clustered observations
coxph(Surv(time, status) ~ treatment + age + cluster(center), data = d)
```

## Review checklist

- Define time origin, event, censoring, competing events and analysis horizon.
- Show numbers at risk under KM curves; report censoring marks and confidence intervals.
- Do not treat deaths from other causes as ordinary censoring when estimating cause-specific risk.
- Check functional form of continuous predictors and PH assumptions; show the checks.
- Pair relative effects with absolute survival or risk at clinically meaningful times.

---

*Rverse Analytics · rverseanalytics.com · Original reference sheet*