Assignment 9 Solutions: Other Outcomes

Part 2: Survival Analysis

Applied Quantitative Methods II, UC3M

Spring 2026

Contents


```
library(survival)
library(survminer)
library(marginaleffects)
data(lung)
lung$dead = lung$status - 1
```

5. Kaplan-Meier survival curves

a) Explore the data:

```
nrow(lung)
```

```
## [1] 228
```

```
sum(lung$dead == 1)
```

```
## [1] 165
```

```
sum(lung$dead == 0)
```

```
## [1] 63
```

The dataset contains 228 patients. Of these, 165 experienced the event (death), and 63 were right-censored (still alive at the end of the study or lost to follow-up). The censoring proportion is 27.6%. This is a moderate censoring rate — not unusual for cancer studies. The censored patients' true survival times are unknown but at least as long as their observed times; simply dropping them or treating them as events would bias the analysis.

b) Overall Kaplan-Meier survival curve:

```
km_all = survfit(Surv(time, dead) ~ 1, data = lung)
km_all
```

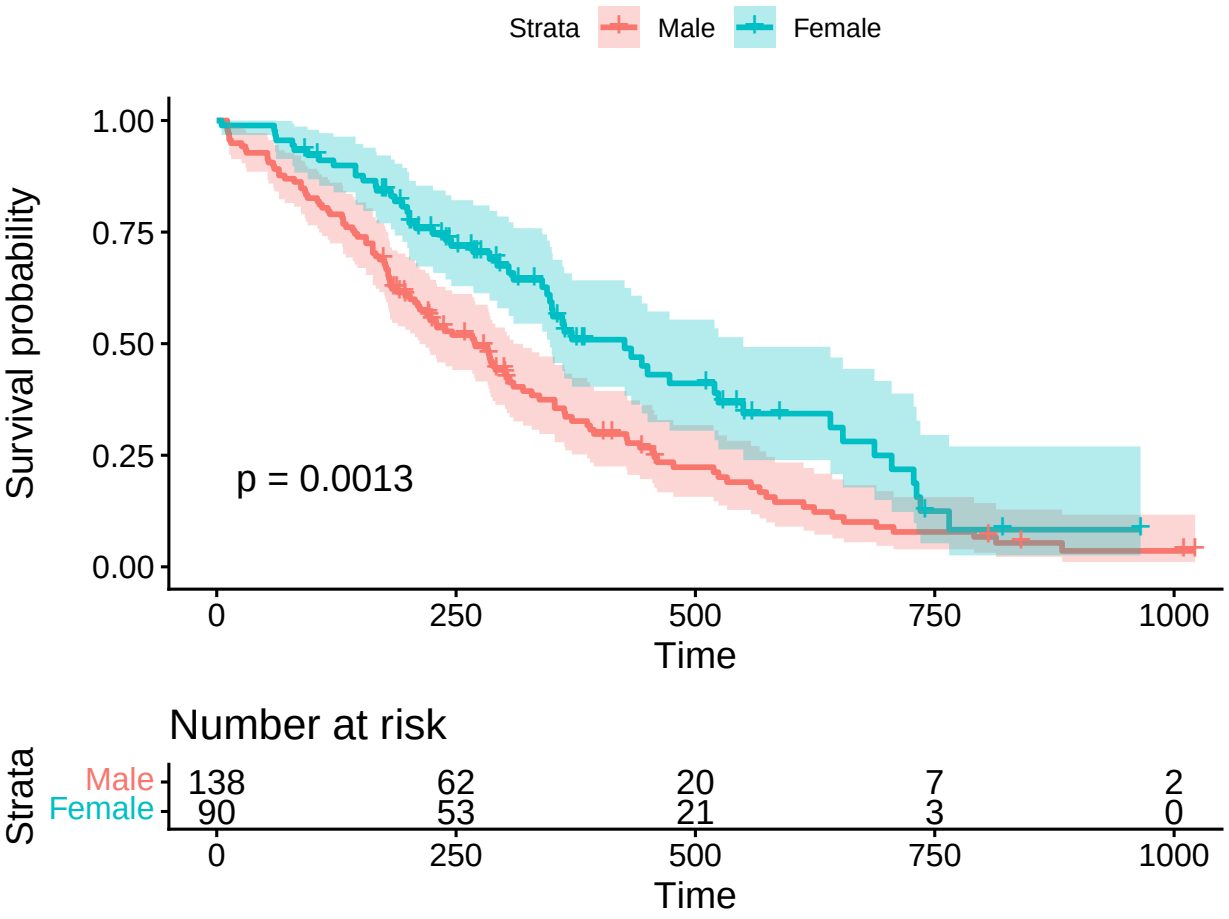
```
## Call: survfit(formula = Surv(time, dead) ~ 1, data = lung)
##
```

```
##        n events median 0.95LCL 0.95UCL
## [1,] 228    165    310     285     363
```

The estimated median survival time is days. This means that, according to the Kaplan-Meier estimator, 50% of patients survived beyond days from enrollment. The KM estimator handles censored observations correctly: when a patient is censored, they drop out of the risk set but do not count as an event, so the survival probability is not artificially deflated.

c) Kaplan-Meier curves by sex:

```
km_sex = survfit(Surv(time, dead) ~ sex, data = lung)
p = ggsurvplot(km_sex, data = lung, conf.int = TRUE,
               pval = TRUE, risk.table = TRUE,
               legend.labs = c("Male", "Female"))
print(p)
```



The Kaplan-Meier curves show clear separation: females survive longer than males throughout the study period. The female survival curve lies consistently above the male curve, and the confidence intervals show limited overlap in the early and middle portions of follow-up. The log-rank test p-value (displayed on the plot) is statistically significant, indicating that the survival distributions differ significantly between the two groups. The log-rank test evaluates the null hypothesis that the survival curves are identical across groups — a significant result rejects this null and supports the conclusion that sex is associated with survival in this lung cancer cohort.

6. Cox proportional hazards model

a) Fit the Cox PH model:

```
cox_fit = coxph(Surv(time, dead) ~ age + sex + ph.ecog, data = lung)
summary(cox_fit)
```

```
## Call:
## coxph(formula = Surv(time, dead) ~ age + sex + ph.ecog, data = lung)
##
##   n= 227, number of events= 164
##    (1 observation deleted due to missingness)
##
##              coef exp(coef)  se(coef)      z  Pr(>|z|)
## age      0.011067  1.011128  0.009267  1.194  0.232416
## sex     -0.552612  0.575445  0.167739 -3.294  0.000986 ***
## ph.ecog  0.463728  1.589991  0.113577  4.083 0.0000445 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age        1.0111     0.9890    0.9929    1.0297
## sex        0.5754     1.7378    0.4142    0.7994
## ph.ecog    1.5900     0.6289    1.2727    1.9864
##
## Concordance= 0.637  (se = 0.025 )
## Likelihood ratio test= 30.5  on 3 df,   p=0.000001
## Wald test            = 29.93  on 3 df,   p=0.000001
## Score (logrank) test = 30.5  on 3 df,   p=0.000001
```

The hazard ratio for `sex` is 0.575. Since sex is coded as 1 = male and 2 = female, a one-unit increase in `sex` corresponds to comparing females to males. A hazard ratio of 0.575 means that females have approximately 42% lower hazard of death compared to males, conditional on age and ECOG performance score. Put differently, at any given time point, a female patient's instantaneous risk of death is about 42% lower than a comparable male patient's. The effect is statistically significant ($p < 0.01$), consistent with the Kaplan-Meier analysis.

b) Interpret the hazard ratio for `ph.ecog`:

```
exp(coef(cox_fit)["ph.ecog"])
```

```
##  ph.ecog
## 1.589991
```

The hazard ratio for `ph.ecog` is 1.59. A one-unit increase in ECOG performance score (i.e., moving from fully active toward more impaired functioning) is associated with approximately 59% higher hazard of death. This is a substantial effect and is highly statistically significant ($p < 0.001$). Physical functioning is the strongest predictor of survival in this model: patients with worse baseline performance status face considerably higher risk.

c) Test the proportional hazards assumption:

```
cox.zph(cox_fit)
```

```
##         chisq df    p
```

```
## age     0.188  1 0.66
## sex     2.305  1 0.13
## ph.ecog 2.054  1 0.15
## GLOBAL  4.464  3 0.22
```

The `cox.zph()` function tests whether each covariate's effect is constant over time — the proportional hazards assumption. A significant p-value indicates that the hazard ratio for that variable changes over the course of follow-up, violating the PH assumption. In this case, none of the individual tests nor the global test are significant at the 5% level, suggesting that the proportional hazards assumption is reasonable for all three covariates. If, hypothetically, the test for `ph.ecog` were significant, it would mean that the effect of physical functioning on hazard changes over time — perhaps being strongly predictive early on but less so among long-term survivors.

d) Summary:

The Kaplan-Meier analysis reveals a clear and statistically significant survival advantage for female lung cancer patients compared to males, with the log-rank test rejecting the null of equal survival distributions. The Cox proportional hazards model confirms this finding after adjusting for age and physical functioning: females have approximately 42% lower hazard of death (HR = 0.58, $p < 0.01$). ECOG performance score is the strongest predictor, with each unit increase (worse functioning) associated with 59% higher hazard (HR = 1.59, $p < 0.001$). Age is not significantly associated with survival after controlling for sex and performance status. The proportional hazards assumption is supported by the `cox.zph()` test (all $p > 0.05$). The key substantive finding is that baseline physical functioning — not age — is the dominant predictor of lung cancer survival in this cohort, highlighting the clinical importance of performance status in prognosis.