Homework 4 Solution Script

Ling Zhu and Songeun Emily Lee

11/27/2025

0. Loading Packages and Data

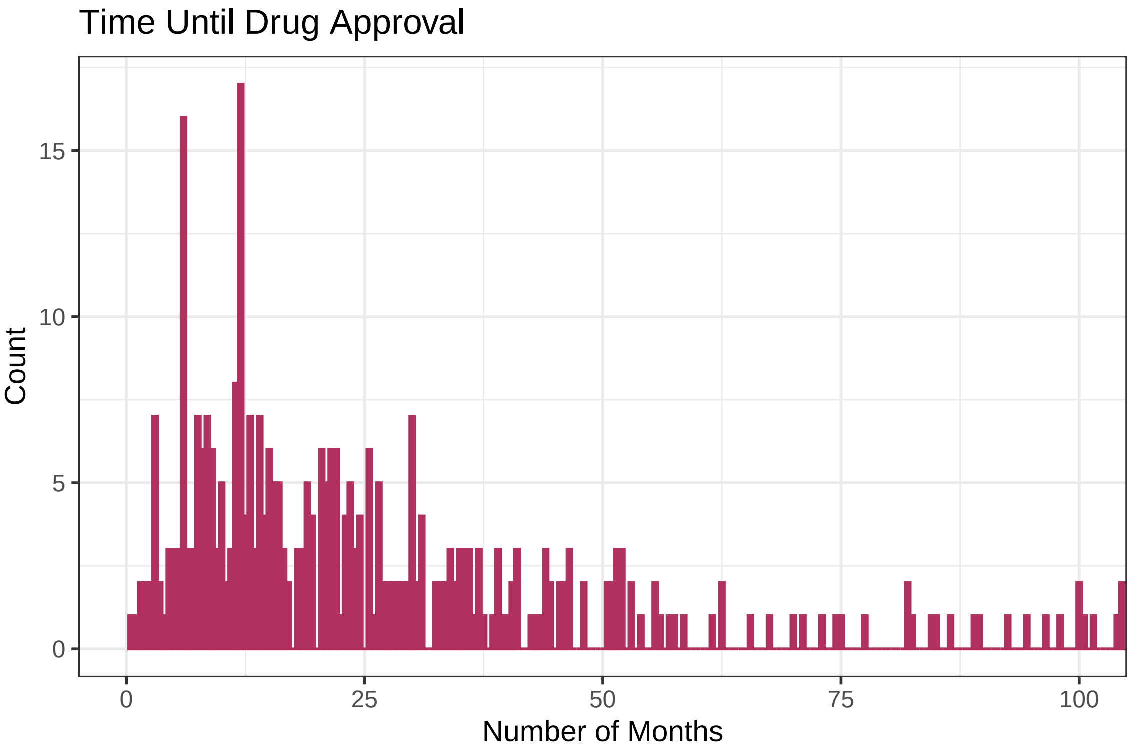
```
# rm(list = ls())

## Working directory
setwd("/Users/songeunlee/Library/CloudStorage/Dropbox/POLS6482 2015 Fall/HW Assignments/2025 HW Review
# setwd("/Users/lingzhu/Dropbox/UH Teaching/POLS6382_2025/HW Assignments/2025 HW Review and Solution Sc

## Libraries
pkgs <- c(
  "foreign", "ggplot2", "ggthemes", "dplyr", "ggpubr",
  "VGAM", "MASS", "nlme", "survival", "simPH", "KMsurv",
  "gtsummary", "ggstats", "survminer"
)

invisible(lapply(pkgs, require, character.only = TRUE))

## Data
data("CarpenterFdaData", package = "simPH")

# Optional: Taking a quick look before digging into the analysis!
#str(CarpenterFdaData)
#head(CarpenterFdaData)
```

1. The Politics of Drug Approval

Question 1a: Cox Regression Model

```
## Question 1a: Estimate a Cox regression model

## Outcome:
##  - acttime : time until drug approval (in months)
##  - censor  : event indicator (1 = approved, 0 = censored)

## Covariates required by the question:
##  - mandiz01  : drug for disease mainly affecting men (dummy)
##  - femdiz01  : drug for disease mainly affecting women (dummy)
##  - peddiz01  : drug for disease mainly affecting children (dummy)
##  - deathrt1  : death rate per 1,000
##  - lethal    : lethal condition (dummy)
##   - hhosleng  : average hospitalization length
##   - hosp01    : population of hospitalization for the disease
##   - stafcder  : FDA drug review staff (FTE)
##   - wpnoavg3  : Washington Post disease stories
```

```
## Descriptive plot of time until approval
# Create a histogram to visualize the distribution of 'acttime'
# helps us see when approvals tend to occur and how the time variable is spread

figure1 <- ggplot(CarpenterFdaData, aes(x = acttime)) +
  geom_histogram(binwidth = 0.5,
               fill = "maroon",
               color = "maroon") +
  coord_cartesian(xlim = c(0, 100)) +
  labs(
    title = "Time Until Drug Approval",
    x = "Number of Months",
    y = "Count"
  ) +
  theme_bw()

figure1
```



```
## Mean of time to approval

# A quick summary measure to see how long approval takes
mean(CarpenterFdaData$acttime, na.rm = TRUE)
```

```
[1] 45.58952
```

```
## Estimate Cox proportional hazards model
model1 <- coxph(
  Surv(acttime, censor) ~
    mandiz01 + femdiz01 + peddiz01 +
    deathrt1 + lethal + hhosleng +
    hosp01 + stafcder + wpnoavg3,
  data = CarpenterFdaData
)

summary(model1)
```

```
Call:
coxph(formula = Surv(acttime, censor) ~ mandiz01 + femdiz01 +
    peddiz01 + deathrt1 + lethal + hhosleng + hosp01 + stafcder +
    wpnoavg3, data = CarpenterFdaData)

  n= 408, number of events= 262

               coef  exp(coef)   se(coef)      z Pr(>|z|)
mandiz01  0.5353473  1.7080413  0.3510107  1.525 0.127219
femdiz01  0.7067445  2.0273803  0.2515881  2.809 0.004968 **
peddiz01 -0.4100893  0.6635910  0.3705237 -1.107 0.268388
deathrt1  0.1363942  1.1461336  0.2107326  0.647 0.517478
lethal    0.0777713  1.0808754  0.1672553  0.465 0.641942
hhosleng  0.0246447  1.0249509  0.0149205  1.652 0.098589 .
hosp01   -0.7254922  0.4840862  0.1953518 -3.714 0.000204 ***
stafcder  0.0024165  1.0024194  0.0002878  8.397  < 2e-16 ***
wpnoavg3  0.0031301  1.0031350  0.0009340  3.351 0.000805 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
mandiz01    1.7080     0.5855    0.8585    3.3984
femdiz01    2.0274     0.4932    1.2382    3.3196
peddiz01    0.6636     1.5070    0.3210    1.3718
deathrt1    1.1461     0.8725    0.7583    1.7323
lethal      1.0809     0.9252    0.7788    1.5002
hhosleng    1.0250     0.9757    0.9954    1.0554
hosp01      0.4841     2.0657    0.3301    0.7099
stafcder    1.0024     0.9976    1.0019    1.0030
wpnoavg3    1.0031     0.9969    1.0013    1.0050

Concordance= 0.739  (se = 0.015 )
Likelihood ratio test= 139  on 9 df,   p=<2e-16
Wald test            = 150.8  on 9 df,   p=<2e-16
Score (logrank) test = 172.2  on 9 df,   p=<2e-16
```

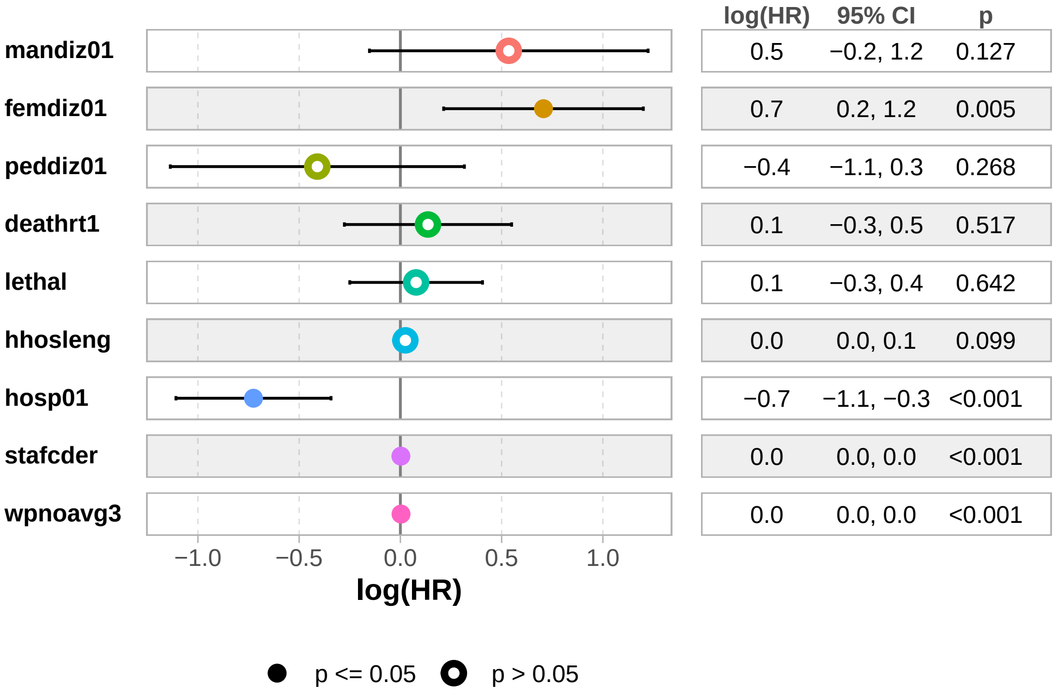
Question 1b: Interpret the Cox model results

```
tbl_regression(model1)
```

Characteristic	log(HR)	95% CI	p-value
mandiz01	0.54	-0.15, 1.2	0.13
femdiz01	0.71	0.21, 1.2	0.005
peddiz01	-0.41	-1.1, 0.32	0.3
deathrt1	0.14	-0.28, 0.55	0.5
lethal	0.08	-0.25, 0.41	0.6
hhosleng	0.02	0.00, 0.05	0.10
hosp01	-0.73	-1.1, -0.34	<0.001
stafcder	0.00	0.00, 0.00	<0.001
wpnoavg3	0.00	0.00, 0.00	<0.001

Abbreviations: CI = Confidence Interval, HR = Hazard Ratio

```
ggcoef_table(model1)
```



log(HR)	95% CI	p
0.5	−0.2, 1.2	0.127
0.7	0.2, 1.2	0.005
−0.4	−1.1, 0.3	0.268
0.1	−0.3, 0.5	0.517
0.1	−0.3, 0.4	0.642
0.0	0.0, 0.1	0.099
−0.7	−1.1, −0.3	<0.001
0.0	0.0, 0.0	<0.001
0.0	0.0, 0.0	<0.001

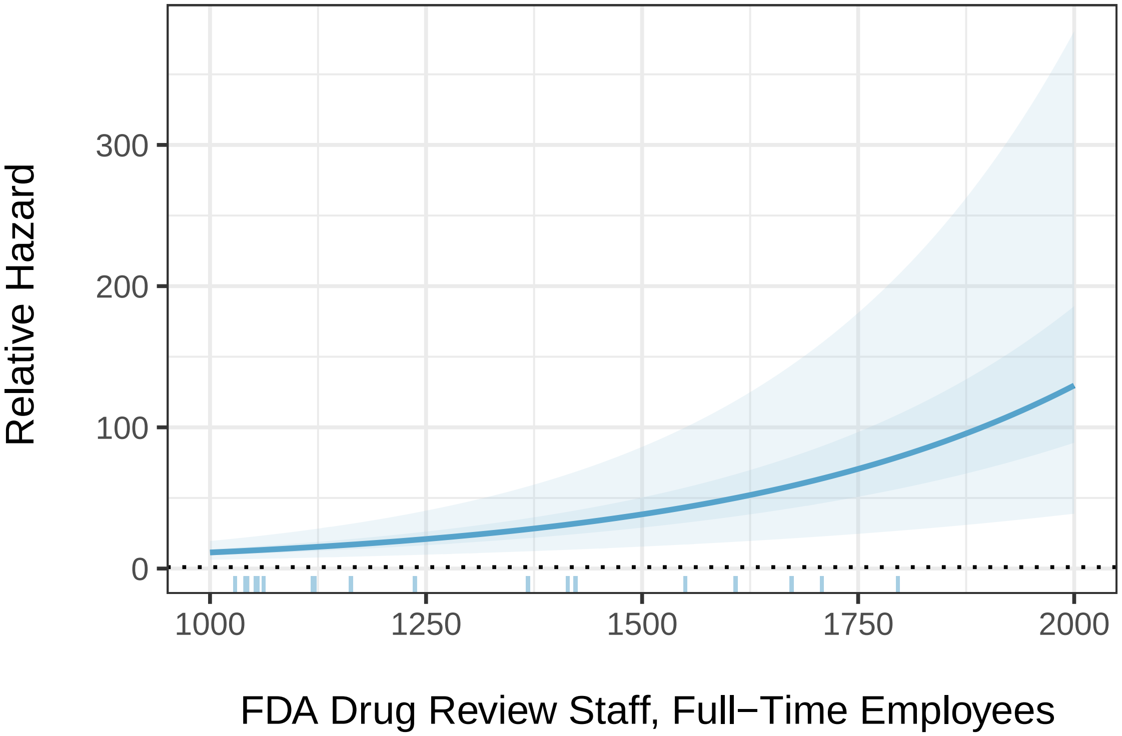
The model includes 408 cases and 262 approval events. Several predictors have statistically significant effects on approval time. Femdiz01 has a positive and significant coefficient (coef = 0.71, p = 0.005), with a hazard ratio of 2.03, meaning drugs for female-related diseases are approved much faster. Hosp01 is negative and highly significant (coef = -0.73, p < 0.001); its hazard ratio of 0.48 indicates about a 52% lower approval hazard, or slower approval. Stafcder (HR = 1.0024, p < 2e-16) and wpnoavg3 (HR = 1.0031, p = 0.0008) are also significant, showing small but meaningful increases in the hazard of approval.

Other variables—mandiz01, peddiz01, deathrt1, and lethal—are not statistically significant. Hhosleng is marginally significant (HR = 1.025, p = 0.099), suggesting a slight increase in approval hazard.

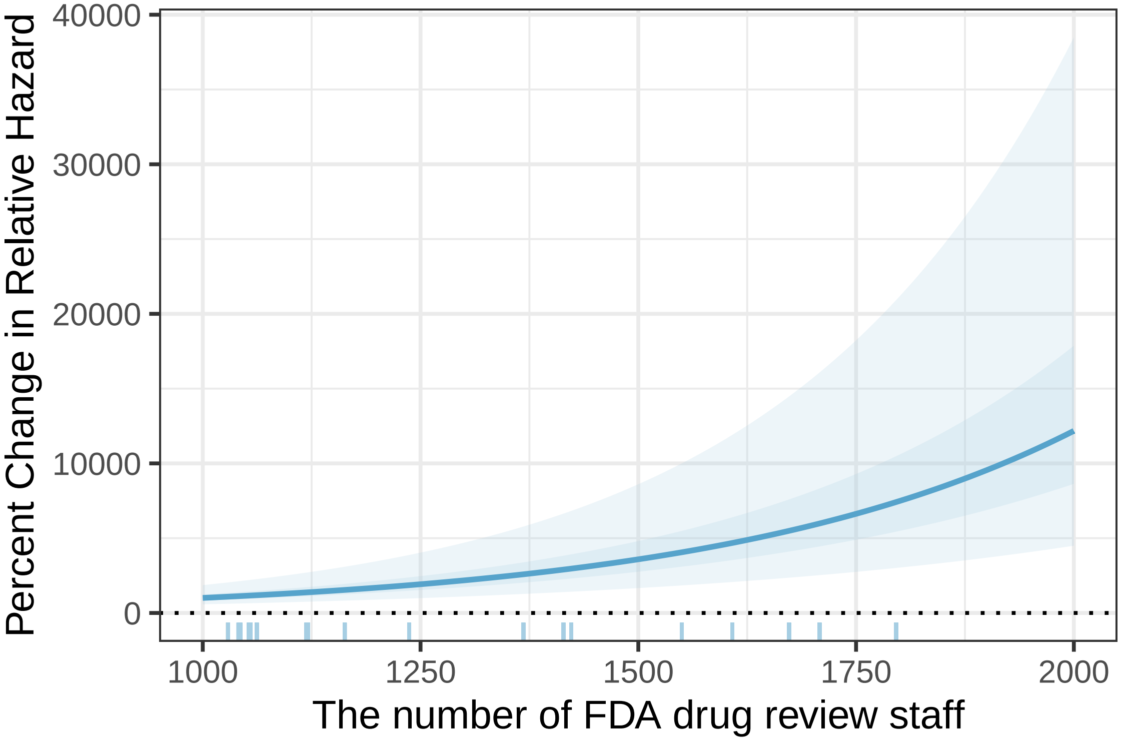
The likelihood-ratio, Wald, and Score tests all strongly reject the null that all coefficients equal zero, indicating that the model overall explains approval time well.

Question 1c: Visualize Hazard Ratios

```
# Figure 1
Sim1 <- coxsimLinear(model1, b = "stafcder",
                     Xj = seq(1000, 2000, by = 10))
simGG(Sim1, xlab = "\n FDA Drug Review Staff, Full-Time Employees")
```



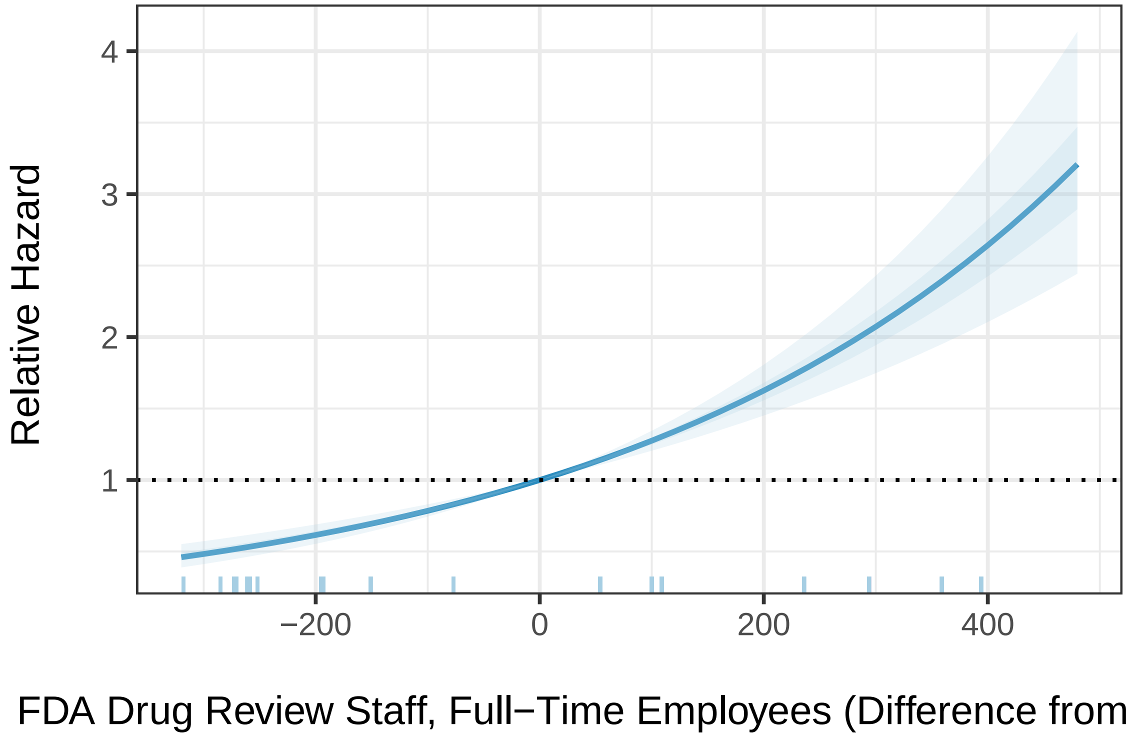
```
# Figure 2
Sim2 <- coxsimLinear(model1, b = "stafcder", qi = "First Difference", Xj = seq(1000, 2000, by = 10))

simGG(Sim2, xlab = "The number of FDA drug review staff", ylab="Percent Change in Relative Hazard")
```



```
# Mean-centering variable "stafcder"
summary(CarpenterFdaData$stafcder)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    996    1055    1237    1314    1608    1796 
```

```
CarpenterFdaData$staff<-CarpenterFdaData$stafcder-1314
model3 <- coxph(Surv(acttime, censor) ~ mandiz01 + femdiz01 + peddiz01 + deathrt1 + lethal + hhosleng +
Sim3 <- coxsimLinear(model3, b = "staff",
                     Xj = seq(-320, 480, by = 20))
simGG(Sim3, xlab = "\n FDA Drug Review Staff, Full-Time Employees (Difference from 1,314)")
```

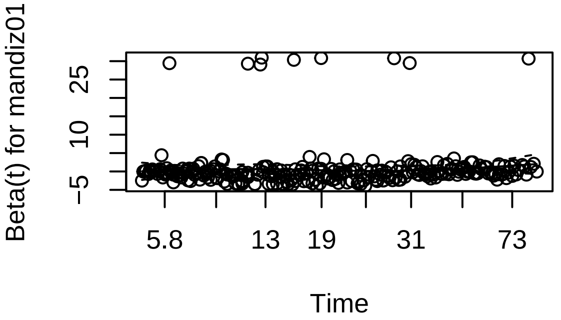


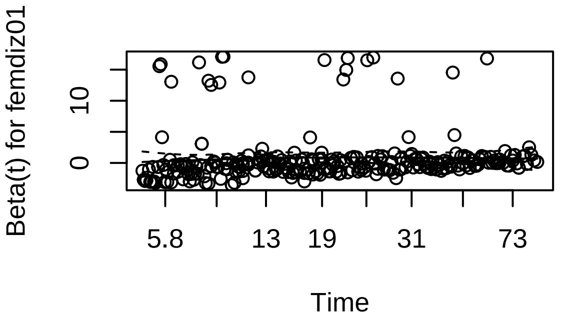
Question 1d: Test for the Proportional Hazard Assumption

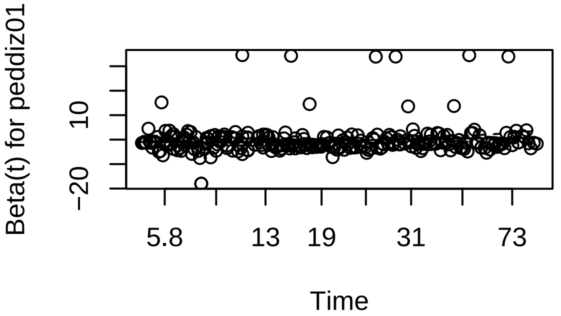
```
## Test proportional hazards assumption for model1
ph_test <- cox.zph(model1)
ph_test
```

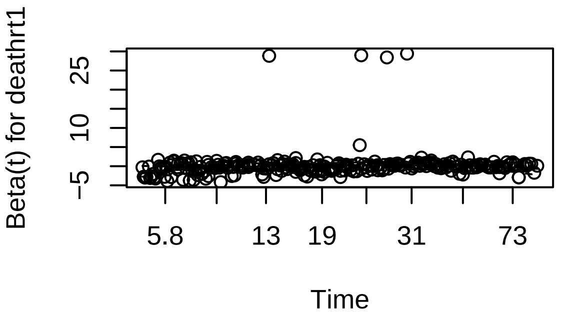
```
          chisq df       p
mandiz01  0.854  1 0.35555
femdiz01  0.550  1 0.45831
peddiz01  0.515  1 0.47301
deathrt1 11.480  1 0.00070
lethal    1.514  1 0.21858
hhosleng  4.381  1 0.03635
hosp01    0.116  1 0.73337
stafcder 12.046  1 0.00052
wpnoavg3  2.085  1 0.14874
GLOBAL   33.718  9 0.00010
```

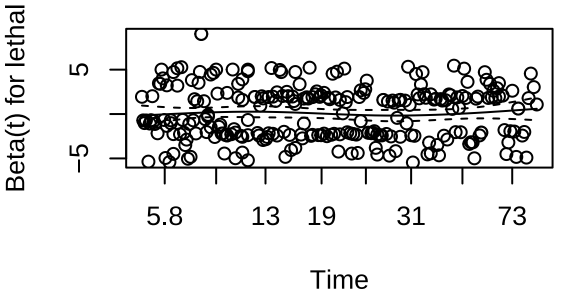
```
par(mfrow = c(2, 2))
plot(ph_test)
```

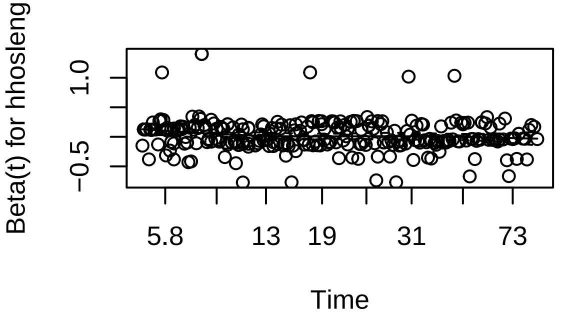

Beta(t) for mandiz01
25
10
−5
5.8
13
19
31
73
Time

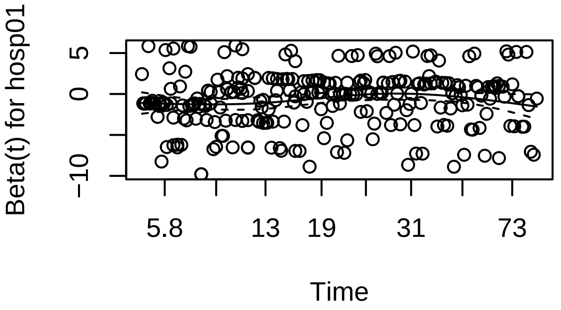

Beta(t) for femdiz01
10
0
5.8
13
19
31
73
Time

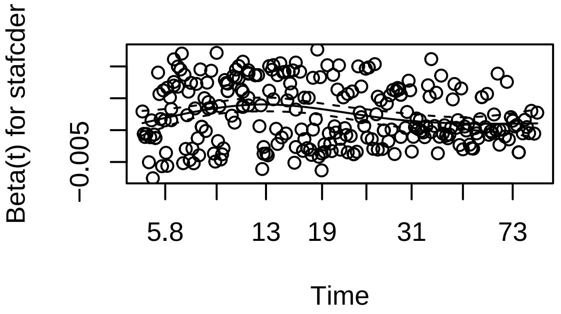

Beta(t) for peddiz01
10
−20
5.8
13
19
31
73
Time

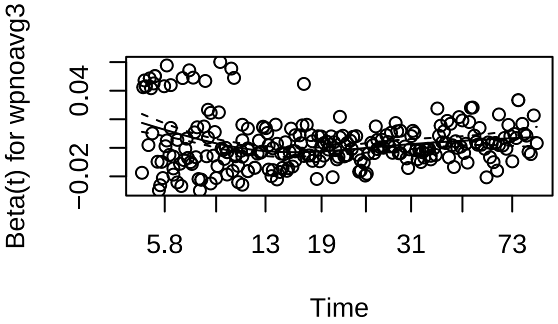

Beta(t) for deathrt1
25
10
−5
5.8
13
19
31
73
Time


Beta(t) for lethal
5
−5
5.8
13
19
31
73
Time


Beta(t) for hhosleng
1.0
−0.5
5.8
13
19
31
73
Time


Beta(t) for hosp01
5
0
−10
5.8
13
19
31
73
Time


Beta(t) for stafcder
−0.005
5.8
13
19
31
73
Time



```r
# Alternative specification with interactions
altmodel <- coxph(
  Surv(acttime, censor) ~ mandiz01 + femdiz01 + peddiz01 +
    deathrt1:acttime + lethal + hhosleng:acttime +
    hosp01 + stafcder:acttime + wpnoavg3,
  data = CarpenterFdaData
)

summary(altmodel)
```

```
Call:
coxph(formula = Surv(acttime, censor) ~ mandiz01 + femdiz01 + 
    peddiz01 + deathrt1:acttime + lethal + hhosleng:acttime + 
    hosp01 + stafcder:acttime + wpnoavg3, data = CarpenterFdaData)

  n= 408, number of events= 262 

                       coef  exp(coef)   se(coef)       z Pr(>|z|)    
mandiz01         -1.393e-01  8.699e-01  3.633e-01  -0.384   0.7013    
femdiz01          2.415e-01  1.273e+00  2.540e-01   0.951   0.3417    
peddiz01          4.299e-01  1.537e+00  3.528e-01   1.218   0.2231    
lethal           -1.264e-01  8.813e-01  1.800e-01  -0.702   0.4826    
hosp01            2.664e-01  1.305e+00  2.213e-01   1.204   0.2287    
wpnoavg3          4.252e-03  1.004e+00  8.766e-04   4.850 1.23e-06 ***
deathrt1:acttime -1.128e-02  9.888e-01  8.730e-03  -1.292   0.1963    
acttime:hhosleng -1.272e-03  9.987e-01  7.414e-04  -1.715   0.0863 .  
acttime:stafcder -2.769e-04  9.997e-01  1.689e-05 -16.388  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                 exp(coef) exp(-coef) lower .95 upper .95
mandiz01            0.8699     1.1495    0.4268    1.7730
femdiz01            1.2732     0.7854    0.7739    2.0946
peddiz01            1.5371     0.6506    0.7698    3.0693
lethal              0.8813     1.1347    0.6192    1.2542
hosp01              1.3052     0.7661    0.8459    2.0141
wpnoavg3            1.0043     0.9958    1.0025    1.0060
deathrt1:acttime    0.9888     1.0113    0.9720    1.0058
acttime:hhosleng    0.9987     1.0013    0.9973    1.0002
acttime:stafcder    0.9997     1.0003    0.9997    0.9998

Concordance= 0.95  (se = 0.004 )
Likelihood ratio test= 1009  on 9 df,   p=<2e-16
Wald test            = 300.7  on 9 df,   p=<2e-16
```

```
Score (logrank) test = 311  on 9 df,   p=<2e-16
```

```r
ph_test2 <- cox.zph(altmodel)
print(ph_test2)
```

```
                   chisq df       p
mandiz01          6.3317  1 0.01186
femdiz01          2.6566  1 0.10312
peddiz01          0.8158  1 0.36642
lethal            3.3186  1 0.06850
hosp01            0.5577  1 0.45518
wpnoavg3          0.0146  1 0.90367
deathrt1:acttime  0.9769  1 0.32297
acttime:hhosleng 12.0218  1 0.00053
acttime:stafcder 50.7976  1 1.0e-12
GLOBAL           79.1249  9 2.4e-13
```

```r
# Alternative specification with strata (binary splits)
CarpenterFdaData <- transform(
  CarpenterFdaData,
  deathrt1.re = ifelse(deathrt1 <= 0.08011, 1, 2),
  hhosleng.re = ifelse(hhosleng <= 5.454, 1, 2),
  stafcder.re = ifelse(stafcder <= 1314, 1, 2)
)

altmodel2 <- coxph(
  Surv(acttime, censor) ~ mandiz01 + femdiz01 + peddiz01 +
    strata(deathrt1.re) + lethal +
    strata(hhosleng.re) + hosp01 +
    strata(stafcder.re) + wpnoavg3,
  data = CarpenterFdaData
)

summary(altmodel2)
```

```
Call:
coxph(formula = Surv(acttime, censor) ~ mandiz01 + femdiz01 +
    peddiz01 + strata(deathrt1.re) + lethal + strata(hhosleng.re) +
    hosp01 + strata(stafcder.re) + wpnoavg3, data = CarpenterFdaData)

  n= 408, number of events= 262

               coef  exp(coef)   se(coef)      z Pr(>|z|)
mandiz01  0.7389819  2.0938027  0.3584150  2.062 0.039226 *
femdiz01  0.7932325  2.2105304  0.2631597  3.014 0.002576 **
peddiz01 -0.4232030  0.6549457  0.3677968 -1.151 0.249879
lethal    0.2180611  1.2436630  0.1889195  1.154 0.248396
hosp01   -0.4471474  0.6394496  0.2113949 -2.115 0.034411 *
wpnoavg3  0.0037059  1.0037127  0.0009761  3.797 0.000147 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
mandiz01    2.0938     0.4776    1.0372    4.2269
femdiz01    2.2105     0.4524    1.3198    3.7025
```

```
peddiz01    0.6549     1.5268    0.3185    1.3467
lethal      1.2437     0.8041    0.8588    1.8010
hosp01      0.6394     1.5638    0.4225    0.9677
wpnoavg3    1.0037     0.9963    1.0018    1.0056

Concordance= 0.614  (se = 0.022 )
Likelihood ratio test= 40.57  on 6 df,   p=4e-07
Wald test            = 43.64  on 6 df,   p=9e-08
Score (logrank) test = 46.76  on 6 df,   p=2e-08
```

```
cox.zph(altmodel2)
```

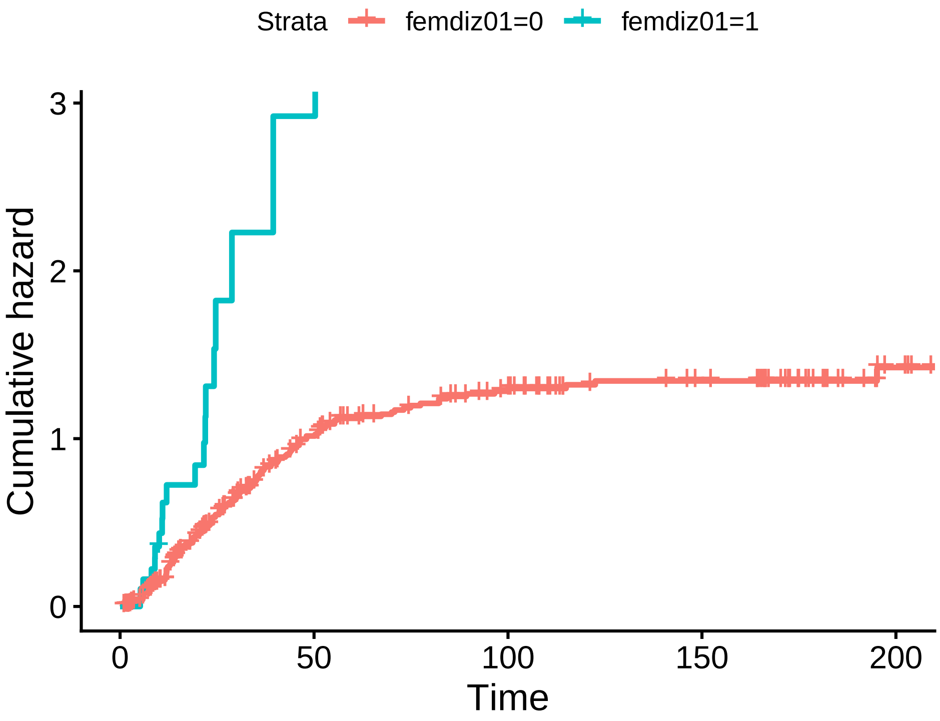
```
         chisq df     p
mandiz01 0.574  1 0.449
femdiz01 0.376  1 0.540
peddiz01 0.363  1 0.547
lethal   0.422  1 0.516
hosp01   2.782  1 0.095
wpnoavg3 2.064  1 0.151
GLOBAL   5.505  6 0.481
```

To test the proportional hazards assumption, I examined the Schoenfeld residuals using cox.zph(model1). The global test is significant, indicating a violation of the proportional hazards assumption in the original model. The individual tests suggest that deathrt1 (p = 0.00070), hhosleng (p = 0.036), and stafcder (p = 0.00052) are the main sources of this violation.

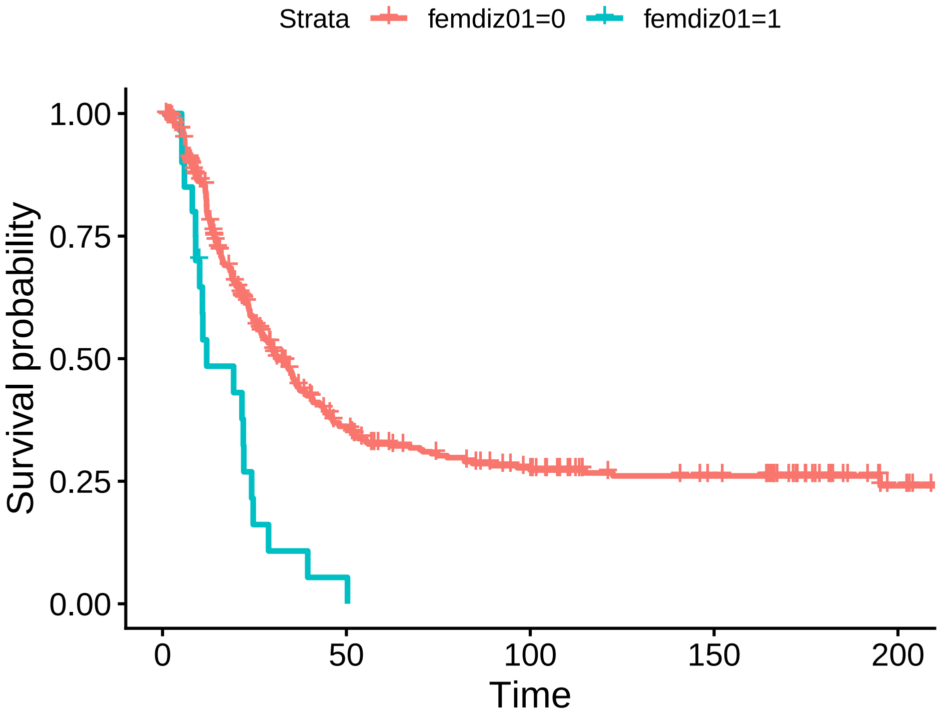
To address this issue, I estimated an alternative Cox model that stratifies these three variables, allowing each to have its own baseline hazard. In the stratified model (altmodel2), the global PH test is no longer significant, indicating that the proportional hazards assumption now holds.

The substantive results remain similar: femdiz01 and mandiz01 are positively associated with the hazard of approval, hosp01 is associated with a lower hazard, and wpnoavg3 remains a small but significant positive predictor. Thus, the stratified model appropriately resolves the PH violation while preserving the core findings.

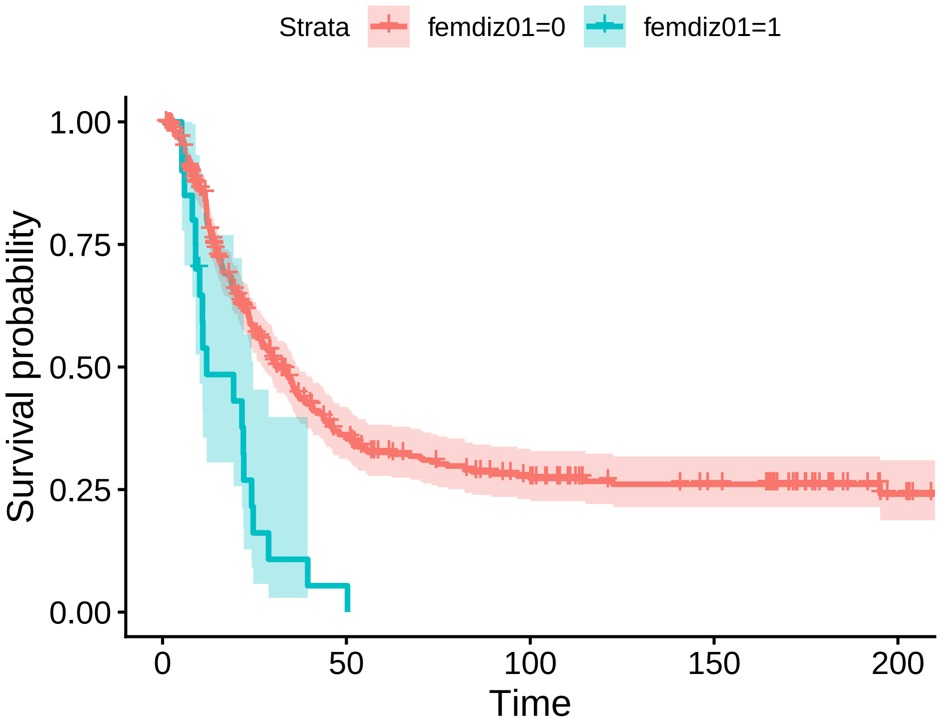
```
# Additional visualization tools
require(survminer)
model4<-survfit(Surv(acttime, censor)~femdiz01, data=CarpenterFdaData)

ggsurvplot(model4, data=CarpenterFdaData, fun="cumhaz")
```



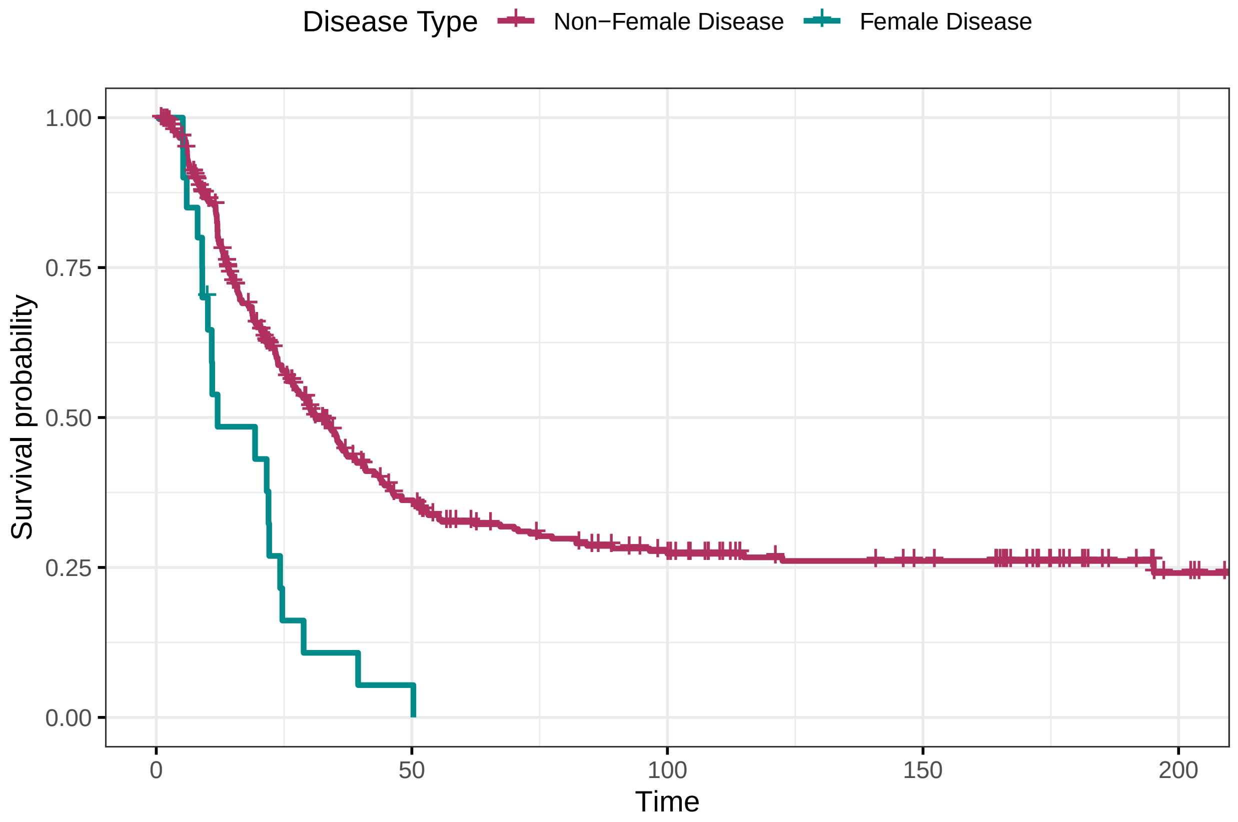
```r
ggsurvplot(model4, data=CarpenterFdaData)
```



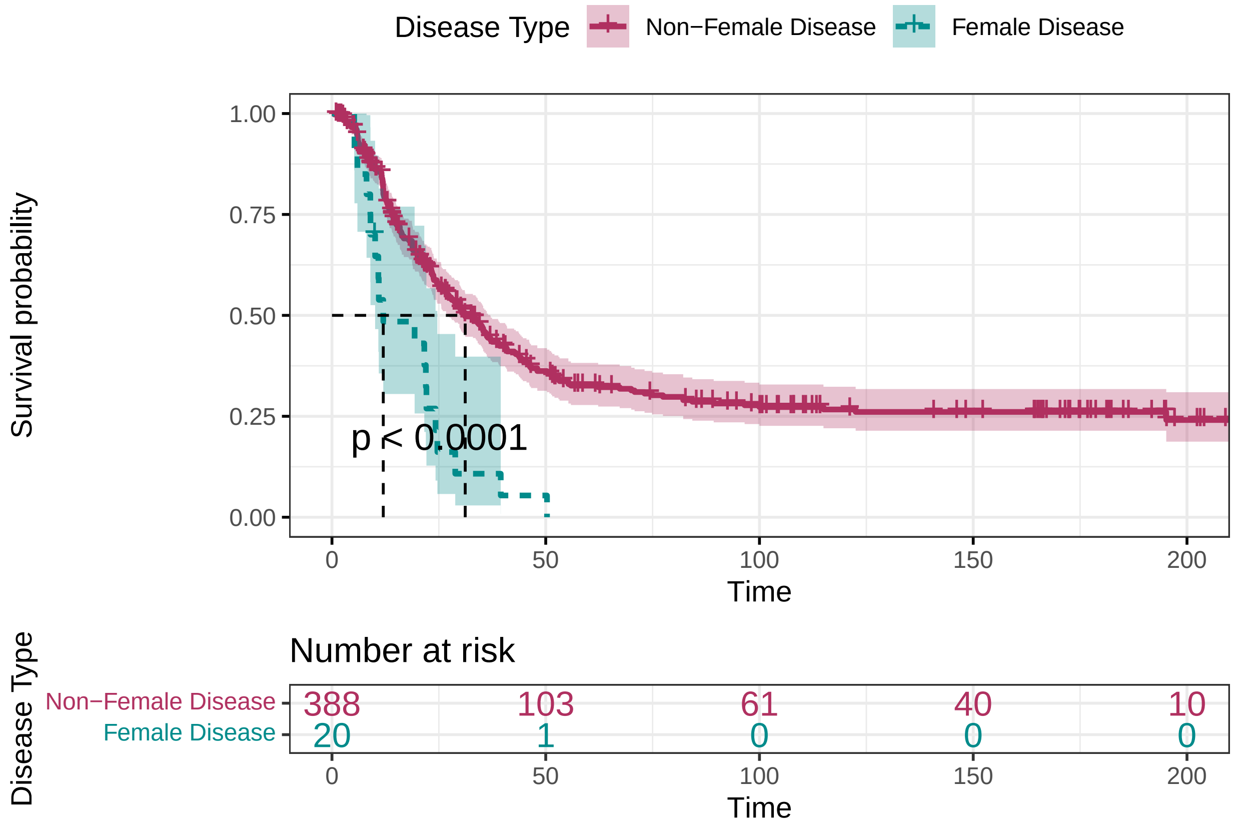
```r
ggsurvplot(model4, data=CarpenterFdaData, conf.int = TRUE)
```



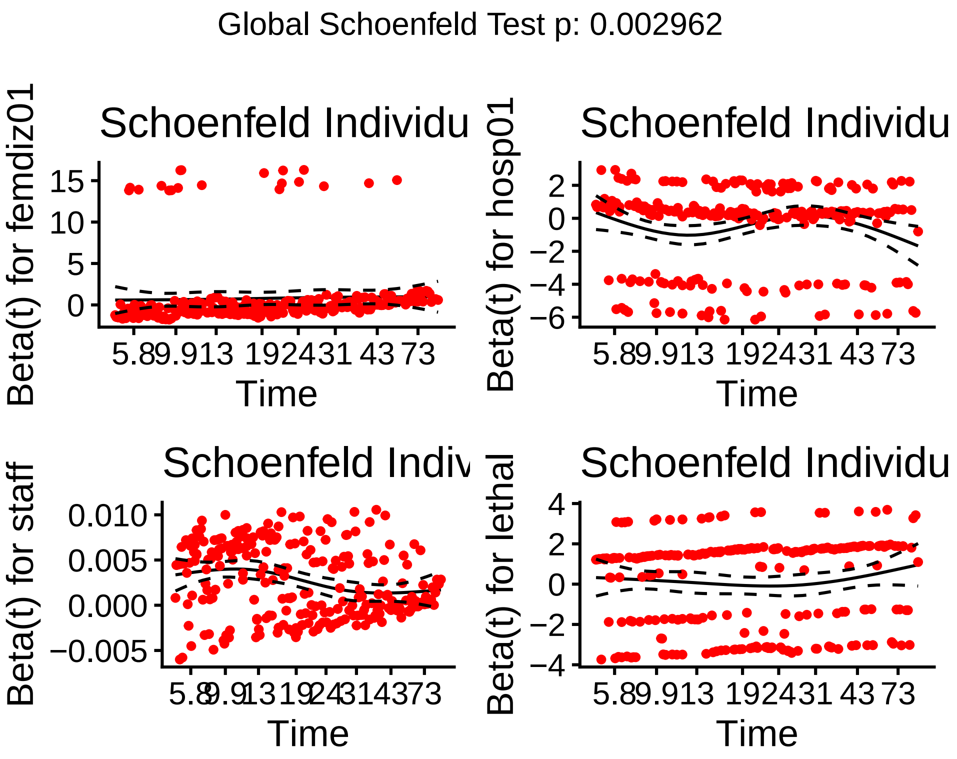
```r
# Survival analysis
# For more details, refer to: https://www.r-bloggers.com/2016/12/survival-analysis-basics/

# Plot 1: KM survival curve with added labels
ggsurvplot(
  model4,
  data = CarpenterFdaData,
  legend.labs = c("Non-Female Disease", "Female Disease"),
  legend.title = "Disease Type",
  palette = c("maroon", "darkcyan"),
  ggtheme = theme_bw()
)
```



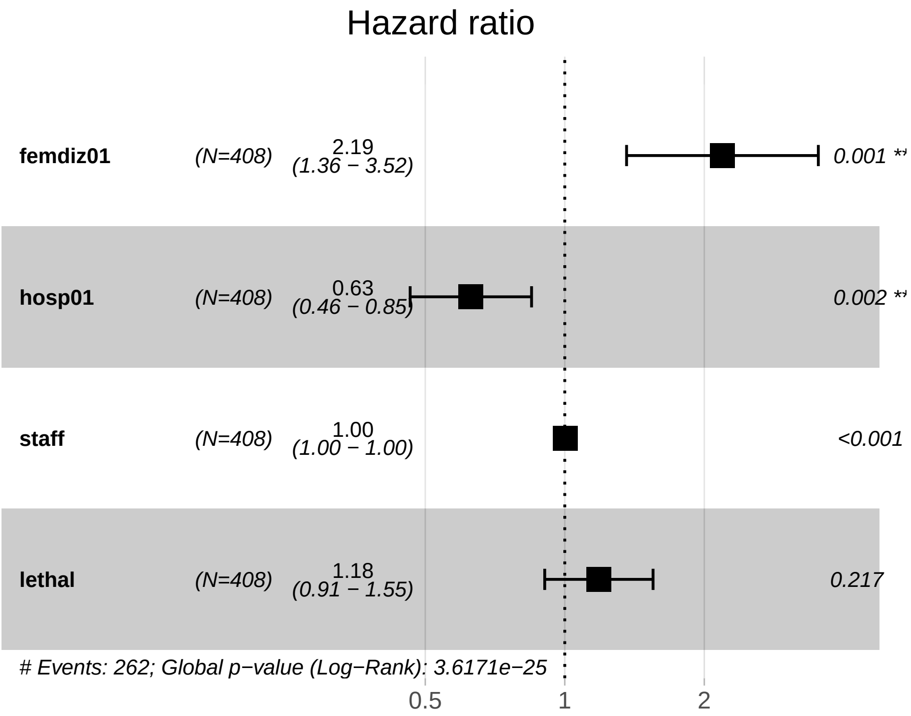
```r
# Plot 2: Most commonly used KM plot
ggsurvplot(
  model4,
  data = CarpenterFdaData,
  pval = TRUE,
  conf.int = TRUE,
  risk.table = TRUE,           # Show risk table
  risk.table.col = "strata",   # Risk table colored by group
  linetype = "strata",         # Different line types by group
  surv.median.line = "hv",     # Horizontal + vertical median lines
  ggtheme = theme_bw(),        # Clean background
  palette = c("maroon", "darkcyan"),  # Custom colors
  legend.labs = c("Non-Female Disease", "Female Disease"),
  legend.title = "Disease Type"
)
```



```r
# Residual analysis
model5<-coxph(Surv(acttime, censor)~femdiz01+hosp01+staff+lethal, data=CarpenterFdaData)
test<-cox.zph(model5)
ggcoxzph(test)
```



```
# Summarizing model results
ggforest(model5)
```


Hazard ratio
femdiz01
(N=408)
(1.36 − 3.52)
0.001 **
hosp01
(N=408)
(0.46 − 0.85)
0.002 **
staff
(N=408)
(1.00 − 1.00)
<0.001
lethal
(N=408)
(0.91 − 1.55)
# Events: 262; Global p−value (Log−Rank): 3.6171e−25