

Parametric Analysis

Jibei Zheng jz3425

Data Prep

```
lung_df = lung %>%  
  mutate_at(c(1, 3, 5, 6), .funs = ~as.factor(.))  
lung_df2 = lung_df %>% na.omit()
```

Model Checking

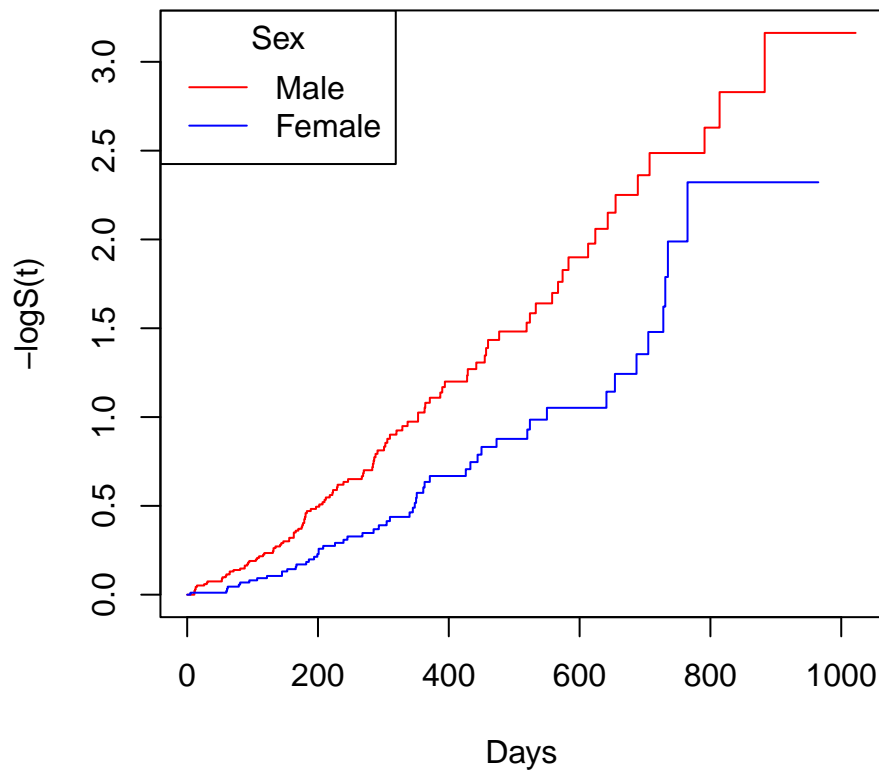
From the KM plot, we assume the survival data follow the Weibull distribution, with non-constant hazard rate.

Use plots to check if the lung data, by sex, follows the exponential distribution or the Weibull distribution.

Plot $-\log\hat{S}(t)$

```
lung_fit = survfit(Surv(time, status == 2) ~ sex,  
  data = lung_df)  
  
plot(lung_fit, col = c("red", "blue"),  
  fun = "cumhaz", xlab = "Days", ylab = "-logS(t)",  
  main = "Negative Log of Estimated Survival Functions")  
legend("topleft", legend = c("Male", "Female"),  
  title = "Sex", col = c("red", "blue"), lty = 1)
```

Negative Log of Estimated Survival Functions



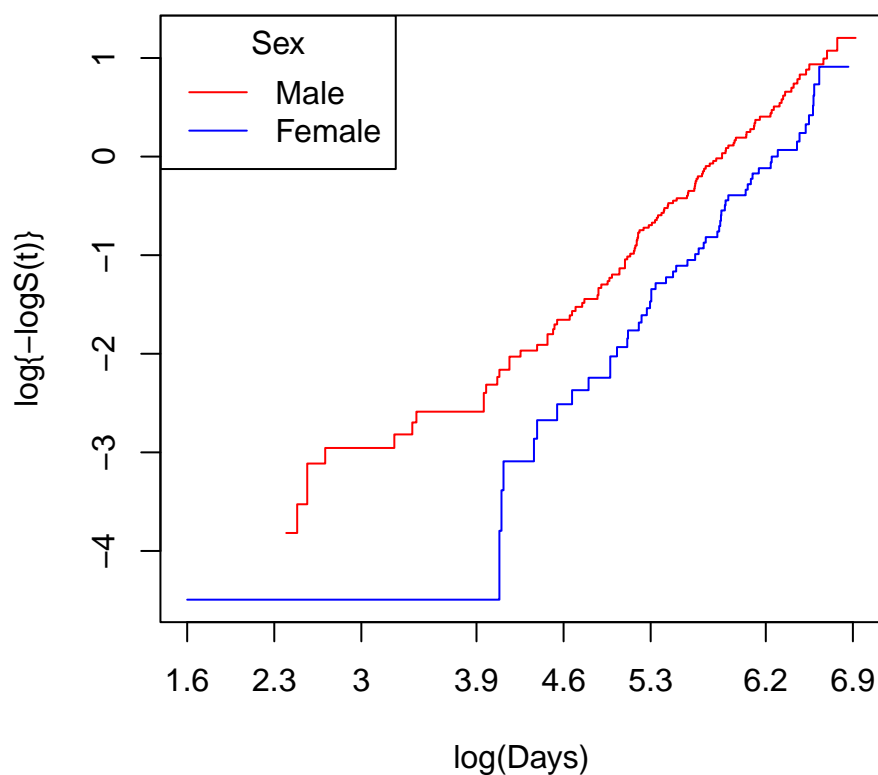
The curve for males is close to a straight line, while the curve for females is obviously non-linear, indicating a better choice of the Weibull distribution.

Plot $\log(-\log S(t))$

```
x_tick = c(5, 10, 20, 50, 100, 200, 500, 1000)
log_x = round(log(x_tick), 1)

plot(lung_fit, col = c("red", "blue"), fun = "cloglog",
     xlab = "log(Days)", ylab = "log{-logS(t)}",
     xaxt = "n", main = "Log of Negative Log of Estimated Survival Functions")
axis(1, at = x_tick, labels = log_x)
legend("topleft", legend = c("Male", "Female"),
     title = "Sex", col = c("red", "blue"), lty = 1)
```

Log of Negative Log of Estimated Survival Function



The slope of the male curve is close to 1, while the slope of the female curve is larger than 1, also indicating a Weibull distribution.

Fit exponential and Weibull distributions

```
#parametric survival function
fit_exp_m = flexsurvreg(Surv(time, status == 2) ~ 1,
                        data = subset(lung_df, sex == 1), dist = "exp")
fit_exp_f = flexsurvreg(Surv(time, status == 2) ~ 1,
                        data = subset(lung_df, sex == 2), dist = "exp")
fit_weib_m = flexsurvreg(Surv(time, status == 2) ~ 1,
                        data = subset(lung_df, sex == 1), dist = "weibull")
fit_weib_f = flexsurvreg(Surv(time, status == 2) ~ 1,
                        data = subset(lung_df, sex == 2), dist = "weibull")

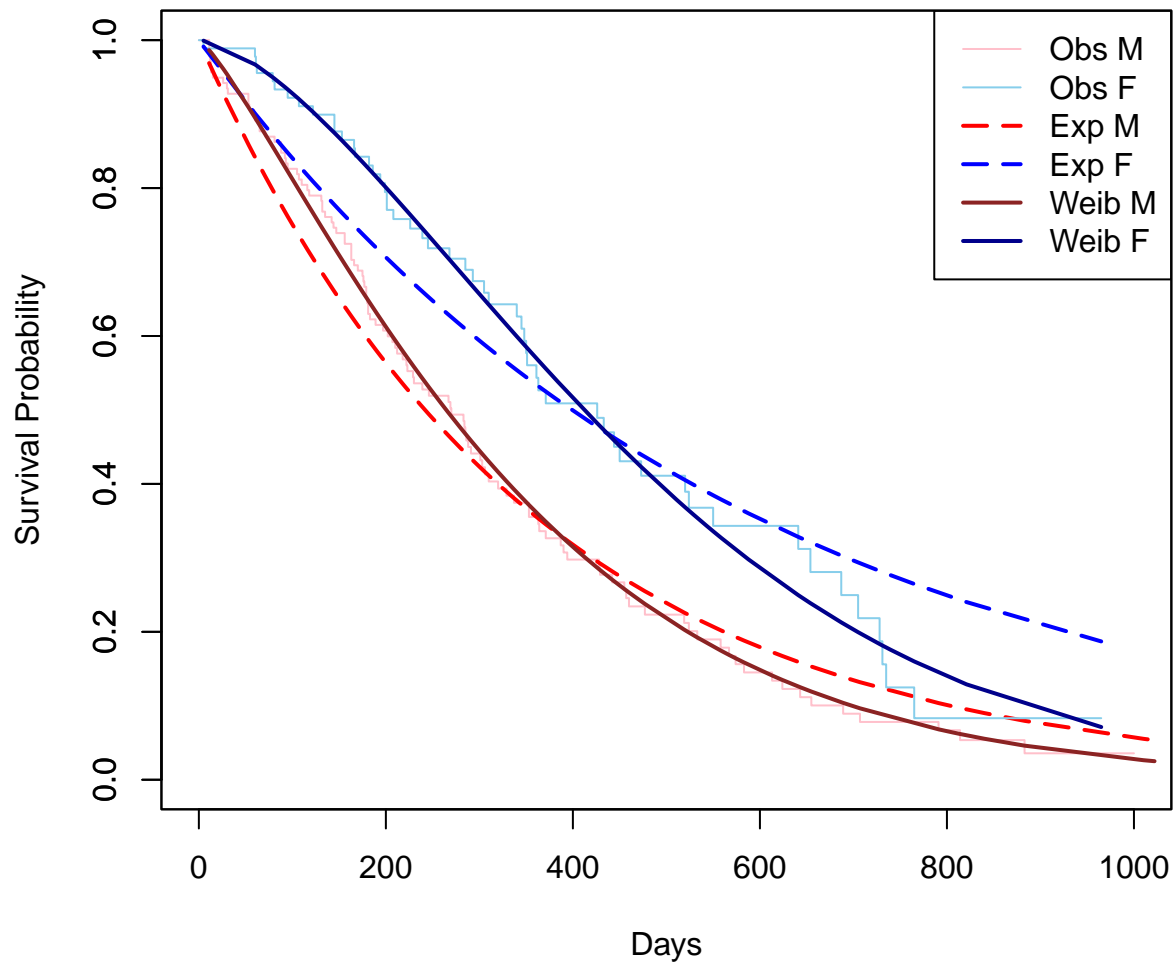
#plot km, exp fitted and weib fitted
plot(fit_exp_m, conf.int = FALSE, ci = FALSE, col = "red", col.obs = "pink",
     lty = "longdash", xlim = c(0,1000),
     xlab = "Days", ylab = "Survival Probability",
     main = "KM and Parametric Est")
par(new = TRUE)
```

```

plot(fit_exp_f, conf.int = FALSE, ci = FALSE, col = "blue", col.obs = "skyblue", lty = "longdash", xlim
plot(fit_weib_m, add = TRUE, ci = FALSE, col = "brown4")
plot(fit_weib_f, add = TRUE, ci = FALSE, col = "blue4")
legend("topright", legend = c("Obs M", "Obs F", "Exp M", "Exp F", "Weib M", "Weib F"),
      col = c("pink", "skyblue", "red", "blue", "brown4", "blue4"),
      lty = c("solid", "solid", "longdash", "longdash", "solid", "solid"),
      lwd = c(1,1,2,2,2,2))

```

KM and Parametric Est



From the plot we can see that fitting a Weibull distribution is actually more precise than an exponential distribution.

Parametric Regression Models

Parametric PH Models

```
#backward selection, significance level = 0.05
fit_ph1 = phreg(Surv(time, status == 2) ~ age + sex + ph.ecog + ph.karno +
               pat.karno + meal.cal + wt.loss,
               data = lung_df2, dist = "weibull")
summary(fit_ph1)
#remove meal.cal

fit_ph2 = phreg(Surv(time, status == 2) ~ age + sex + ph.ecog + ph.karno + pat.karno + wt.loss,
               data = lung_df2, dist = "weibull")
summary(fit_ph2)
#remove age

fit_ph3 = phreg(Surv(time, status == 2) ~ sex + ph.ecog + ph.karno + pat.karno + wt.loss,
               data = lung_df2, dist = "weibull")
summary(fit_ph3)
#remove pat.karno

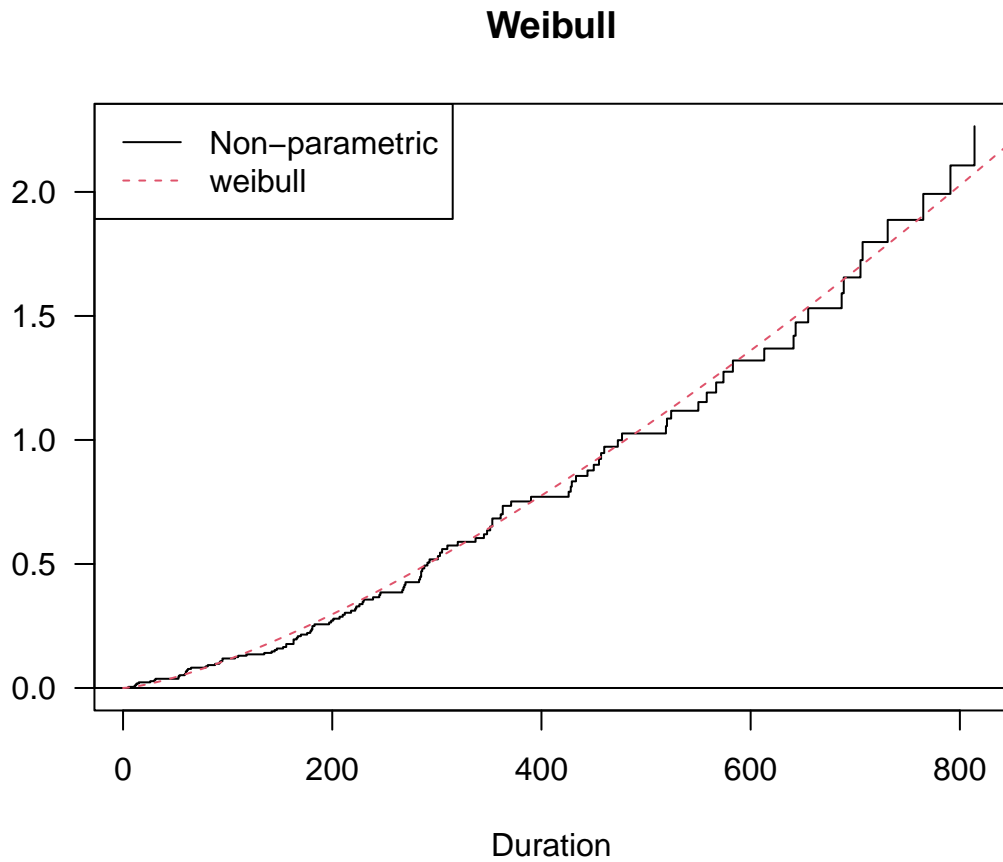
fit_ph4 = phreg(Surv(time, status == 2) ~ sex + ph.ecog + ph.karno + wt.loss,
               data = lung_df2, dist = "weibull")
summary(fit_ph4)
#remove wt.loss

fit_ph5 = phreg(Surv(time, status == 2) ~ sex + ph.ecog + ph.karno,
               data = lung_df2, dist = "weibull")
summary(fit_ph5)
#remove ph.karno

#final model
fit_ph6 = phreg(Surv(time, status == 2) ~ sex + ph.ecog,
               data = lung_df2, dist = "weibull")
summary(fit_ph6)
```

## Covariate		Mean	Coef	Rel.Risk	S.E.	LR p
## sex						0.0087
##	1	0.579	0	1 (reference)		
##	2	0.421	-0.504	0.604	0.197	
## ph.ecog						0.0023
##	0	0.309	0	1 (reference)		
##	1	0.520	0.290	1.337	0.233	
##	2	0.169	0.925	2.523	0.260	
##	3	0.002	1.944	6.986	1.028	
##						
## Events		120				
## Total time at risk		51759				
## Max. log. likelihood		-830.69				
## LR test statistic		20.75				
## Degrees of freedom		4				
## Overall p-value		0.000354567				

```
#compare the estimated baseline hazards with a non-parametric ph model  
fit_cox = coxreg(Surv(time, status == 2) ~ sex + ph.ecog, data = lung_df2)  
check.dist(fit_ph6, fit_cox)
```



The fit of the Weibull baseline function is very close to the non-parametric one.

For a Weibull distribution, the AFT model is also a PH model.