Final: 12/13 Wednesday 2-5pm WH 100E

Homework #15

1. Suppose that
$X \sim bin(3,p)$,
$C \sim U(\{0.5, 1.5\})$,
$X \perp C$,
$M = X \wedge C$ and $\delta = \mathbf{1}(X \leq C)$.

Generate $n = 100$ RC data (M_i, δ_i)'s, find
the NPMLE's of S_X and f_X,
the MLE's of p, S_X and f_X.

Sol: (1) The PLE of $S(t)$ is

time	*n.risk*	*n.event*	*survival*	*std.err*	*lower95%CI*	*upper95%CI*
0			0.687			
1			0.316			

(2) The MLE of $S(t)$ is $\hat{S}(t) = \begin{cases} 1 & \text{if } t < 0 \\ something & \text{if } t \in [0,1) \\ something & \text{if } t \geq 1 \end{cases}$ as the time is censored after 1.5 ?

(3) Since $\hat{p} = 0.29$, $\hat{f}(x) = \binom{3}{x} 0.29^x 0.71^{3-x}$?

(4) $\hat{S}(x) = \begin{cases} 1 & \text{if } t < 0 \\ 1 - \sum_{i=0}^{x} \binom{3}{i} 0.29^i 0.71^{3-i} & \text{if } 0 \leq x < 3 \end{cases}$?

$\int_0^1 x dx = -\int_1^0 x dx$? $\sum_{i=0}^{3} a_i = -\sum_{i=3}^{0} a_i$? $\binom{0}{0} =$? $\binom{4}{5} = \frac{4!}{4!(-1)!} =$?

(5) $\hat{S}(x) = 1 - \sum_{i=0}^{\lfloor x \rfloor} \binom{3}{i} 0.29^i 0.71^{3-i}$?

(6) $\hat{S}(x) = 1 - \sum_{i=0}^{\lfloor x \rfloor} \binom{3}{i} 0.29^i 0.71^{3-i}$ if $x \leq 3$?

(7) $\hat{S}(x) = 1 - \sum_{i\geq 0}^{\lfloor x \rfloor} \binom{3}{i} 0.29^i 0.71^{3-i}$ if $x \leq 3$?

2. Carry out a simulation on log-logistic regression with RC data. In the simulation, generate $n = 200$ RC data (M_i, δ_i, X_i) from the model $\ln Y = 1 + 4X + \sigma Z$, where $Z \sim$ the logistic with $S(t) = \frac{1}{1+\exp(t)}$, $X \sim U(0,3)$, censoring distribution C is Poisson; the censoring rate should be round 50%, i.e., $\sum(\delta)/n \approx 1/2$.

(a) Draw in the same figure with legend,
the true survival function $S_{Y|X}(\cdot|2)$,
its estimates based on the linear regression, the log-linear regression, Cox' regression, and the MLE's of the survival functions under dist= weibull, normal, lognormal, logistic and loglogistic in survreg().

(b) Estimate $E(Y|X = 2)$ together with the confidence interval (there are 8 of them).

(c) Make comments on these plots and estimates.

Sol. About (c). Before the simulation, what do you expect to see ?

1. The estimates of $S_{Y|X}(\cdot|2)$ should be close to the true values in the cases:
   log-linear,
   log-logistic.
2. The estimate of $E(Y|X = 2)$ should be close to the true value in the case: log-logistic.
3. Why not the others ?

Logistic distribution

Common form $S(t) = \frac{1}{1+exp(t)}$,

With covariate in Splus or R, reparametrization:

$S_Y(y|\mathbf{x}, \beta) = \frac{1}{1+exp(\frac{y-\beta'\mathbf{X}}{\tau})}$,

$Y = \beta'\mathbf{x} + \tau Z$, $Z \sim logistic(0,1)$. $E(Z) = 0$, $\underbrace{\sigma_Z = 1 \text{ and } \tau = \pi/\sqrt{3}}_{change\ to\ \sigma_z = \pi/\sqrt{3}}$.

> e=rlogis(n,0,1) # mean=0, sd=s not 1, as seen in the next few lines

> mean(e)

[1] -0.1285181

> sd(e)

[1] 1.726354

> pi/sqrt(3)

[1] 1.813799

T=rlogis(n,0,1) $\underbrace{<\neq>}_{change\ to\ <=>}$ $T_1, ..., T_n \sim$ logistic(0,1).

Loglogistic distribution

ln$Y = \beta'\mathbf{x} + \tau Z$, $Z \sim logistic(0,1)$. $E(Z) = 0$ and $\sigma_Z = \pi/\sqrt{3}$.

Remark. About ln$Y = \beta X + W$ (or $Y = e^{\beta X} e^W$).

If $f_{Y|X}(t|x) = e^{-\beta x} f_o(te^{-\beta x})$, where f_o is a df.

Then $e^W = Y/e^{\beta x}$ has d.f. f_o.

Also $S_{Y|X}(t) = S_o(te^{-\beta x})$.

This is due to u-substitution.

$$\int_{-\infty}^{y} e^{-\beta x} f_o(t/e^{\beta x}) dt = \int_{-\infty}^{y/e^{\beta x}} f_o(u) du, \qquad \text{where } u = t/e^{\beta x}.$$

The true survival function of $Y|(X = x)$:

$S_{Y|X}(t|x) = P(Y > t|X = x)$

$= P(lnY > \text{ln}t|X = x)$

$= P(\beta X + W > \text{ln}t|X = x)$

$= P(W > \text{ln}t - \beta' x|X = x)$

$= P(Z > \frac{\text{ln}t - \beta' x}{\tau}|X = x)$

$= \frac{1}{1+e^{\frac{\text{ln}t-\beta' x}{\tau}}}$, $t > 0$ assuming $W \perp X$.

The true survival function:

$$S_{Y|X}(t|2;\beta) = S(t;2,\beta) = \begin{cases} 1(t \le 0) + \frac{\mathbf{1}_{(t>0)}}{1+e^{\frac{\text{ln}t-\beta' x}{\tau}}} & \text{if } \tau \neq 1 \\ 1(t \le 0) + \frac{\mathbf{1}_{(t>0)}}{1+t/e^{\beta' x}} & \text{if } \tau = 1 \end{cases}$$

with $\beta'\mathbf{x} = \beta_1 + \beta_2 x = 1 + 4 \times 2$

Its MLE is $S(t; 2, \hat{\beta})$.

```
b=survreg(Surv(m,d)~x, dist="loglogis")
a=b$coef
s=b$s
t=sort(m)
X=2
plot(t,1/(1+exp((log(t)-(1+4*X)/τ))) # true S(t), τ is given.
lines(t,1/(1+exp((log(t)-(a[1]+a[2]*X)/s))) # MLE
lines(t,1-plogis((log(t)-a[1]-a[2]*X)/s)) # Is it the same ?
lines(t ,1-plogis(log(t),a[1]+X*a[2],s)) # How about this ?
```

Under the linear regression model:

$Y = \beta' X + W$.

$W = Y - \beta' X$.

$M = Y \wedge C$,

$(M_i - \beta' \mathbf{X}_i, \delta_i)$s are RC W_is.

$$\tilde{S}_{Y|X}(t|2) = \tilde{P}(W + 1 + \beta' 2 > t) = \tilde{S}(t),$$

where $\tilde{S}$ is the PLE based on $R_i = \underbrace{m_i - \hat{\beta}_1 - \hat{\beta}_2 X_i}_{RC-residuals} + \hat{\beta}_1 + \hat{\beta}_2 \times X$, $i = 1, ..., n$,

$$\tilde{S}(t) = \prod_{R_{(i)} > t} (1 - \frac{\delta_{(i)}}{n - i + 1}) \text{ where } R_i = m_i - \hat{\beta}' \mathbf{X}_i + \hat{\beta}' \mathbf{x}.$$

$> $ a=bj(Surv(m,d)~x, link="identity", control=list(iter.max=50))$coef

$> $ zz=$\underbrace{\text{m-a[1]-a[2]*x}}_{RC\ residual}$+a[1]+a[2]*X

$> $ plot(survfit(Surv(zz,d)~ 1)

Under the AL model:

```
M=log(m)
a=bj(Surv(m,d)~x, control=list(iter.max=50))$coef
zz=M-a[1]-a[2]*x+a[1]+a[2]*2
plot(survfit(Surv(exp(zz),d)~ 1)
```

Under the Cox model:

```
zz=coxph(Surv(m,d)~x)
plot(survfit(zz,newdata=data.frame(x=2), conf.type="none"))
```

$$\check{S}(t) = (\check{S}_o(t))^{e^{X \times \hat{\beta}}}.$$

Semi-parametric estimate of $E(Y|X)$:
Under the Cox model:

zz=coxph(Surv(m,d)~x)

$\sum_t tf(t)$.

w=survfit(zz,newdata=data.frame(x=2), conf.type="none")

summary(w)

time	*n.risk*	*n.event*	*survival*	*std.err*
0.000698	200	1	0.9913	0.00862
...				
1.020888	18	1	0.0434	0.02047

Remark. In the final exam, if $\hat{f}(t)$ is the NPMLE of f_X and $\hat{f}(\infty) = \infty$, for uniqueness, define

$$\hat{E}(X) = \sum_{t<\infty} t\hat{f}(t) + (\underbrace{M_{(n)}}_{=?} + 1)\hat{f}(\infty)$$

Modify the codes as follows.

```
(S = zz$surv[1:7])
(F = 1-S)
F=c(0,F,1)
(f = F[-1]-F[-length(F)])
(f = round(c(f,1-sum(f)),3))
f[1:8]
x=c(zz$time[1:7],1+zz$time[7])
```

Under the LR model and AL model:

The methods are similar.

MLE of $E(Y|X)$:

weibull: $E(Y|X) = \tau\Gamma(1 + 1/\gamma)$, $S(t) = \exp(-(\frac{t}{\tau})^\gamma)$, $t > 0$, $\tau = \exp(\beta'\mathbf{x})$.

s=survreg(Surv(m,d)~x,dist=...)$s

MLE is $exp(a[1] + 2 * a[2]) * gamma(1 + s)$ (note that $s = 1/\gamma$).

lognormal: $E(Y|X) = \exp(\mu + (\tau^2/2))$, $\mu = \beta'\mathbf{x}$ and $\tau =$ Scale.

normal or logistic: $E(Y|X) = \beta'\mathbf{x}$ or predict(zz,newdata=data.frame(x=1))

loglogistic: If we assume $lnY = \beta X + \sigma Z$, where $\mu_Z = 0$, $\sigma = 1$ and $\sigma_Z = \pi/\sqrt{3}$.

Then $E(Y|X) = e^{\beta X}E(e^W)$ is a function of β.

The MLE of $E(Y|X)$ is $e^{\hat{\beta}X}E(e^W)$ ($e^{\beta X} = e^{1+4\cdot 2}$), where $E(e^W) \approx mean(exp(rlogist(n)))$.

```
exp(9)*mean(exp(rlogis(1000)))
exp(9)*mean(exp(rlogis(10000)))
exp(9)*mean(exp(rlogis(100000000)))
```

What do you expect ?

Outputs:

[1] 38,784.22
[1] 72,372.55
[1] 205,273.6

What is wrong ?

Does $E(Y|X)$ exist ?

$F_W(t) = \frac{1}{1+e^{-t}}$ and $S_W(t) = \frac{1}{1+e^{t}}$.
$E(e^W) = \int e^w f_W(w)dw = \int t f_Z(t)dt$, $Z = e^W$.
$S_Z(t) = P(e^W > t) = P(W > \ln t) = \frac{1}{1+e^{\ln t}} = \frac{1}{1+t}$, $t > 0$.
$f_Z(t) = -S'_Z(t) = \frac{1}{(1+t)^2}$, $t > 0$,
$E(e^W) = \int_0^\infty \frac{t}{(1+t)^2}dt = \int_1^\infty \frac{1}{u} - \frac{1}{u^2}du = \infty$.

Remark. The SE can be obtained by the bootstrap method.

But the boostrapped mean is not the MLE, though the difference is small.

Remark. In non-parametric approach or semi-parametric approach with RC data with $\hat{f}(y)$. Define $\hat{E}(Y) = \sum_{t<\infty} t\hat{f}(t) + \hat{f}(\infty)(M_{(n)} + 1)$, as $P(\delta_{(n)} = 0) > 0$ even if $\tau_C = \infty$.

Remark. Let

$$F_X(t) = \begin{cases} t & \text{if } t \in [0, 0.5) \\ 0.5 & \text{if } t \in [0.5, 1) \\ 1 - 0.5e^{-t+1} & \text{if } t \geq 1 \end{cases} \quad (1)$$

The median can be defined to be any point m in $[0.5, 1]$, as $P(X \leq m) = 0.5$.
But for uniqueness, define

$F^{-1}(p) = \inf\{t : F(t) \geq p\}$ or $\sup\{t : F(t) < p\}$.

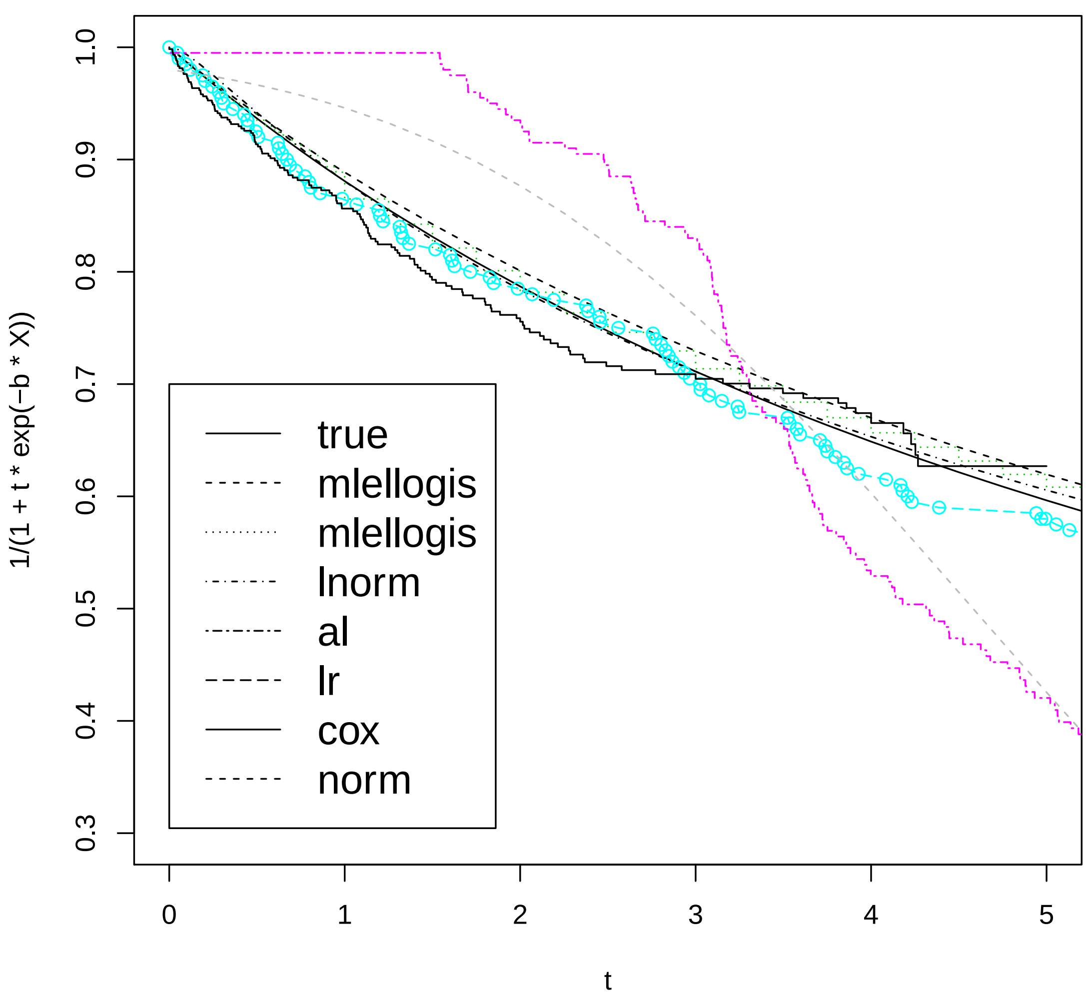
For $X \sim bin(1, 0.5)$, $F^{-1}(0.5) = 0$ by both ways. Notice that $F_X(t) = \begin{cases} 0.5 & \text{if } t \in [0, 1) \\ 1 & \text{if } t \geq 1. \end{cases}$

We can defined median as $t \in [0, 1)$, but not all points in $[0, 1]$.

Part of codes for homework 15.

```
n=200
e=rlogis(n,0,1) # mean=0, sd=s not 1, as see next
x=rbinom(n,1,0.5)
b=1
y=exp(b*x+e)
c=rbinom(n,5,0.9)
d=(as.numeric(y<=c))
mean(d)
    [1] 0.69
m=y*d+c*(1-d)
t=c((1:20)/20, 1+ (1:20)/4)
X=2
plot(t,1/(1+t*exp(-b*X)), xlim=c(0,5), ylim=c(0.3,1), type="l", lty=1) # true
w1=survreg(Surv(m,d)~x, dist="loglogis")
(a=w1$coef)
    (Intercept) x
    0.01607511 1.33592149
(s=w1$s)
    [1] 0.9923186
lines(t ,1-plogis((log(t)-a[1]-X*a[2])/s), lty=2) # MLE
lines(t,1/(1+exp((log(t)-a[1]-X*a[2])/s)),type="S", lty=3, col=3) #MLE
w2=survreg(Surv(m,d)~x, dist="lognormal")
a=w2$coef
s=w2$s
lines(t ,1-plnorm(t,a[1]+X*a[2],s), lty=4) # MLE
```

```
w3=bj(Surv(m,d)~x,control=list(iter.max=50))
a=as.numeric(w3$coef)
M=exp(log(m)-a[2]*x+a[2]*X)
lines(survfit(Surv(M,d) 1, conf.type="none"), type="o", lty=5, col=5)
w4=bj(Surv(m, d) ~ x, link="identity",control=list(iter.max=200))
a=as.numeric(w4$coef)
M=m-a[2]*x+a[2]*X
lines(survfit(Surv(M,d) 1, conf.type="none"), type="S", lty=6,col=6)
w5=coxph(Surv(m,d)~x)
lines(survfit(w5,newdata=data.frame(x=2), conf.type="none"), type="S", lty=7, col=1)
w6=survreg(Surv(m,d)~x, dist="gaussian")
a=w6$coef
s=w6$s
lines(t ,1-pnorm(t,a[1]+X*a[2],s), lty=8, col=8) # MLE
legend(0,0.7, + c("true","mlellogis","mlellogis","lnorm","al","lr","cox","norm"),
    lty=c(1,2,3,4,6,5,7,8),cex=1.5)
```



Homework 16. Final: 12/13 Wednesday 2-5pm WH 100E

a. Fit Gehan's data with covariate $X = 1$(treat) to the semi-parametric PH model and the MLE parametric PH models of Weibull, *i.e.*, $S_{Y|X}(t|x) = (S_o(t))^{e^{\beta x}}$, where S_o is Weibull.

b. Obtain the MLEs of $E(Y|X)$ under the previous semi-parametric and parametric assumptions.

Hint: $E(Y|X = x) = \int t f_{Y|X}(t|x)dt$ and $f_{Y|X}(t|x) = -S'_{Y|X}(t|x)$.

c. Check (qqplot and test) whether Gehan's data with covariate treat-control fit the parametric PH models.

d. Is the semi-parametric estimate of $E(Y|X)$ consistent under the PH model ?

Sol. Part a. Parametric model:

> (z=survreg(Surv(time,cens)∼treat,data=gehan))

(Intercept) treatcontrol

3.515687 -1.267335

Scale= 0.7321944

 # What does $X = \mathbf{1}(treat)$ mean ?

> treat[1:3]

[1] control 6-MP control

Levels: 6-MP control # that is (0,1)

> as.numeric(treat)[1:3]

[1] 2 1 2

Hereafter, assume that $X = \mathbf{1}(treat = control)$, and thus it is correct to write "∼ treat".

Continuous semi-parametric Cox's model:

> (a=coxph(Surv(time,cens)∼treat,data=gehan))

	coef	*exp(coef)*	*se(coef)*	*z*	*p*
treatcontrol	1.572	4.817	0.412	3.81	0.00014

Likelihood ratio test=16.4 on 1 df, p=5.26e-05

n= 42, number of events= 30

> b=survfit(a,data.frame(treat="control"), conf.type="none")

> lines(b)

> b=summary(b)

time	*n.risk*	*n.event*	*survival*	*std.err*
1	42	2	0.9227	0.0528
...				
23	7	2	0.0202	0.0244

> (n=length(b$time))

> b$time

[1] 1 2 3 4 5 6 7 8 10 11 12 13 15 16 17 22 23

> b$surv # $(\hat{S}_o(t))^{e^{\hat{\beta}}} = (\hat{S}_o(t))^{4.8}$

[1] 0.92266954 0.84534004 0.80667573 0.72934823 ... 0.05802130 0.02021487

> b=survfit(a,data.frame(treat="6-MP"), conf.type="none")

> plot(b)

> round(b$time,3)

[1] 1 2 3 4 5 6 7 8 10 11 12 13 15 16 17 22 23

> round(b$surv,3) # $\hat{S}_0(t)$

[1] 0.983 0.966 0.956 0.937 0.915 0.880 0.869 0.818 ... 0.554 0.445

Done for part a ?

Answer of (a.1): For Cox's model $S_{Y|X}(t|x) = (S_o(t))^{e^{\beta x}}$, $\hat\beta = 1.57$,

$$\hat S_{Y|X}(t|6-MP) = \hat S_o(t) = \begin{cases} 1 & \text{if } t<1 \\ 0.983 & \text{if } t \in [1,2), \\ \cdots \end{cases}$$

$$\hat S_{Y|X}(t|control) = (\hat S_o(t))^{4.8} = (\hat S_o(t))^{e^{1.57}} = \begin{cases} 1 & \text{if } t<1 \\ 0.923 & \text{if } t \in [1,2) \\ \cdots \end{cases}$$

(a.2) The MLE of Cox's model with the Weibull baseline survival function.

Let $x = \mathbf{1}(X = control)$ and $t > 0$,

$$\begin{aligned} S_{Y|X}(t|x) &= (e^{-\frac{t^\gamma}{\tau}})^{e^{\beta x}} \\ &= e^{-\frac{t^\gamma}{\tau}(e^{\beta x})} \\ &= e^{-\frac{t^\gamma}{e^\alpha}(e^{\beta x})} \\ &= e^{-\frac{t^\gamma}{e^\alpha e^{-\beta x}}} \\ &= \begin{cases} \exp(-\frac{t^\gamma}{e^{\alpha-\beta x}}) \\ \exp(-\left(\frac{t}{e^{\frac{\alpha-\beta x}{\gamma}}}\right)^\gamma) \end{cases} \end{aligned}$$

Which is used in survreg() ?

Example 1. Simulation for understanding the model and output in R:

```
> n=500
> e=rexp(n)
> y=exp(-2+3*log(e))
# y=rweibull(n, scale=exp(-2),shape=1/3)
> zz=survreg(Surv(y)~1)
> summary(zz)
```

$\ln Y = \beta X + \frac{1}{\gamma}\ln Z$

	Value	*Std.Error*	*z*	*p*
(Intercept)	-1.94	0.1536	-12.7	$1.1e-36$
Log(scale)	1.18	0.0354	33.2	$2.2e-242$

Scale= 3.25

$S_Y(t) = P(Y > t)$

$= P(-2 + 3 * \ln Z > \ln t)$ $\qquad \ln Y = \beta X + \frac{1}{\gamma}\ln Z$

$= P(\ln Z > \frac{\ln t+2}{3})$

$= P(Z > \exp(\frac{\ln t+2}{3}))$

$= \exp(-e^{\frac{\ln t+2}{3}})$

$= \exp(-(\frac{t}{e^{-2}})^{1/3})\ if\ t > 0.$

survreg(Surv(m,d)~x) yields: $\begin{pmatrix} & intercept & x & scale \\ & \alpha/\gamma & -\beta/\gamma & 1/\gamma \\ estimates & 3.515687 & -1.267335 & 0.7321944 \\ & \alpha & \beta & \gamma \\ estimates & 4.801576 & 1.730872 & 1.365758 \end{pmatrix}$

Answer to (a.2):

The MLE of $S_{Y|X}(t|x)$ is $(\hat S_o(t))^{e^{1.73x}}$ with

$x = \mathbf{1}(treat = control)$ and $\hat S_o(t) = e^{-t^{1.37}/e^{4.8}}$.

b. $E(Y) = \begin{cases} \beta^{1/\gamma}\Gamma(1+1/\gamma) & \text{if } S_Y(t) = \exp(-t^\gamma/\beta) \\ \tau\Gamma(1+1/\gamma) & \text{if } S_Y(t) = \exp(-(t/\tau)^\gamma). \end{cases}$

```
> a=z$coef

> exp(a[1])*gamma(1+z$s)
   30.78419
> exp(a[1]+a[2])*gamma(1+z$s)
   8.668246
```

Answer to part b.: MLE under parametric model:

$\begin{cases} E(Y|treat = 6-MP) = 30.78419 \\ E(Y|treat = control) = 8.668246 \end{cases}$

Estimate under the semi-parametric Cox's model: $E(Y|X) = \sum_t t\hat{f}(t) = ...$

c. Since $S_{Y|X}$ satisfies both the Cox's and AL model,

$\ln Y = \beta'\mathbf{X} + \frac{1}{\gamma}\ln Z,$

$\ln Y - \beta'\mathbf{X} = \frac{1}{\gamma}\ln Z,$

RC residuals: $\ln M_i - \hat{\beta}'\mathbf{X}_i$

#weibull $\exp((\ln M_i - \hat{\beta}'\mathbf{X}_i) * \hat{\gamma})$ verses qexp(),

```
> tr=ifelse(treat=="6-MP", 0,1)
> zz=log(time)-a[1]-a[2]*tr
> zz=exp(zz) # Why do this ?
> n=length(zz)
> w0=survfit(Surv(zz,cens) ~ 1)
> z=survreg(Surv(zz,cens)~1,data=gehan)
> a=z$coef
> summary(w0)
   Call: survfit(formula = Surv(zz, cens) ~ 1)
```

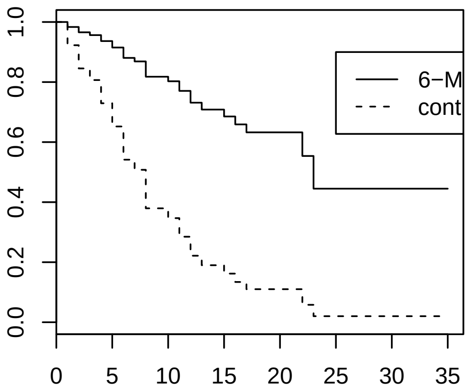
time	*n.risk*	*n.event*	*survival*	*std.err*	*lower95%CI*	*upper95%CI*
0.106	42	2	0.9524	0.0329	0.89011	1.000
...						
2.323	2	1	0.0496	0.0476	0.00758	0.325
2.428	1	1	0.0000	*NaN*	*NA*	*NA*

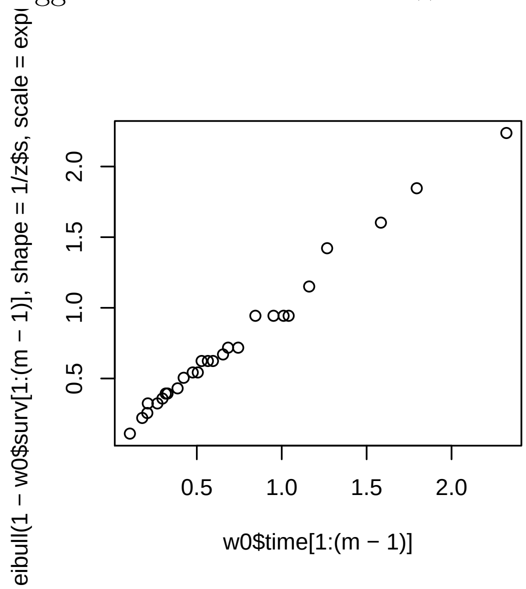
```
> m=length(w0$time)
> plot(w0$time[1:(m-1)],qweibull(1-w0$surv[1:(m-1)], shape=1/z$s,scale=exp(a[1])))
   # Why m-1 ?
> X=c(rep(1,n),rep(0,n))
> d=c(cens,rep(1,n))
> M=c(zz, rweibull(n, shape=1/z$s,scale=exp(a[1])))
> plot(survfit(Surv(M,d) ~X))
> survdiff(Surv(M,d)~X)
```

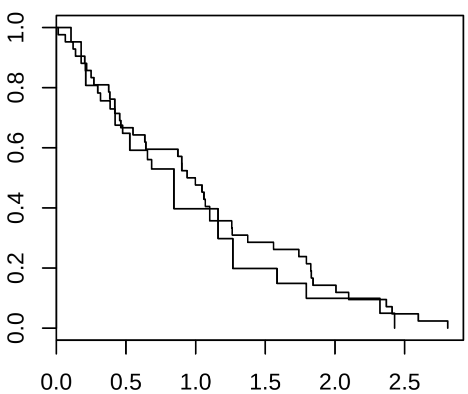
	N	*Observed*	*Expected*	$(O-E)^2/E$	$(O-E)^2/V$
$X=0$	42	42	39.7	0.134	0.314
$X=1$	42	30	32.3	0.165	0.314

Chisq= 0.3 on 1 degrees of freedom, p= 0.575

Answer to part c. The qqplot and the test suggest that the data fit the Weibull model.







d. No standard answer.