

# Fundamentals of Statistical Modeling (VT21)

Andrea Discacciati  
Karolinska Institutet  
Stockholm, Sweden

## Lab 4

Load the dataset and the `mlci` command

```
. version 14
. use https://raw.githubusercontent.com/anddis/fsm/master/data/lab4_1.dta, clear
. run https://raw.githubusercontent.com/anddis/fsm/master/do/mlci.do
```

### Exercise 1

We use data from 267 patients diagnosed with oral cancer. We measured time to death ( $y$ ) in subjects with low-grade cancer ( $x = 0$ ) and high-grade cancer ( $x = 1$ ). Some survival times are censored ( $d = 0$ ). First, we plot Kaplan-Meier estimates of the survival functions.

```
. stset y, fail(d)
      failure event:  d != 0 & d < .
obs. time interval:  (0, y]
exit on or before:  failure
```

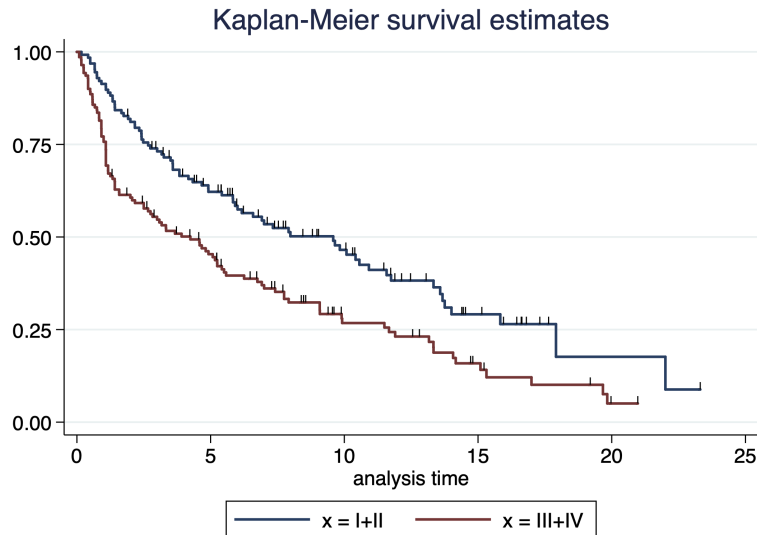
---

267	total observations	
0	exclusions	

---

267	observations remaining, representing	
184	failures in single-record/single-failure data	
1,620.864	total analysis time at risk and under observation	
	at risk from t =	0
	earliest observed entry t =	0
	last observed exit t =	23.258

```
. sts graph, by(x) cens(single) name(km, replace)
      failure _d:  d
      analysis time _t:  y
. graph export km.png, replace
(file km.png written in PNG format)
```



We consider a log-logistic model (AFT) for  $f(y|x)$  (see slides 112). Estimate the model's parameters. Constrain the parameter  $\lambda$  to be positive. Take into account right censoring (see slide 76).

The PDF of a (standard) logistic distribution is:

$$f(y) = \frac{\exp(-y)}{(1 + \exp(-y))^2}$$

while the Survival function is:

$$S(y) = 1 - \frac{1}{1 + \exp(-y)}$$

```
. local lambda = "exp({theta})"
. local G = "(log(y)-({beta0}+{beta1}*x))/`lambda'"
. local g = "1/(`lambda'*y)"
. local f = "exp(-(`G'))/((1+exp(-(`G')))^2)*`g'"
. local S = "1-1/(1+exp(-(`G')))"
. mlexp ((d==1)*ln(`f') + (d==0)*ln(`S'))
initial:      log likelihood = -696.29352
alternative:  log likelihood = -641.95479
rescale:      log likelihood = -641.95479
rescale eq:   log likelihood = -573.91032
Iteration 0:  log likelihood = -573.91032
Iteration 1:  log likelihood = -572.255
Iteration 2:  log likelihood = -572.24452
Iteration 3:  log likelihood = -572.24452
Maximum likelihood estimation
Log likelihood = -572.24452      Number of obs      =      267
```

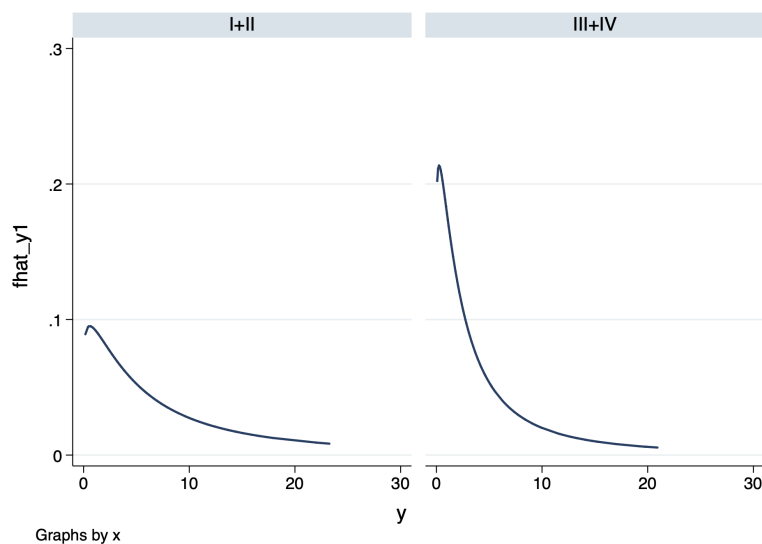
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
/beta0	2.051811	.1455026	14.10	0.000	1.766631	2.336991
/beta1	-.8090195	.1993996	-4.06	0.000	-1.199836	-.4182034
/theta	-.1109087	.060722	-1.83	0.068	-.2299216	.0081042

```
. mlci exp /beta1
.4452944    95% CI: .3012437, .6582283
. mlci exp /theta
.8950204    95% CI: .7945959, 1.008137
```

Plot the estimated densities  $\hat{f}(y|x)$ .

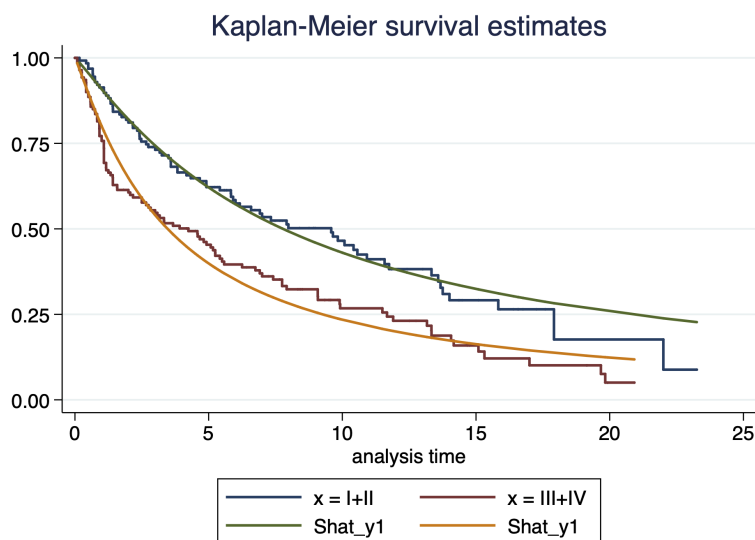
```
. gen fhat_y1 = exp(-((ln(y)-(_b[/beta0]+_b[/beta1]*x))/exp(_b[/theta])))/((1+exp(-ln(y)-(_b[/beta0]+_b[/beta1]*x))/exp
> (_b[/theta]))^2)*1/(exp(_b[/theta])*y)
```

```
. tw (line fhat_y1 y, sort), name(p1, replace) by(x) ylabel(0(0.1)0.3)
. graph export p1.png, replace
(file p1.png written in PNG format)
```



Plot the estimated survival functions  $\hat{S}(y|x)$  together with the Kaplan-Meier estimates.

```
. gen Shat_y1 = 1-1/(1+exp(-((ln(y)-(_b[/beta0]+_b[/beta1]*x))/exp(_b[/theta]))))
. sts graph, by(x) name(km1, replace) addplot((line Shat_y1 y if x == 0, sort) ///
> (line Shat_y1 y if x == 1, sort))
      failure _d: d
      analysis time _t: y
. graph export km1.png, replace
(file km1.png written in PNG format)
```



## Exercise 2

Use a RCS transformation of time to death to make the log-logistic model more flexible (see slide 118).

Generate a RCS transform of  $y$  (V2) and its derivative (v2) using `rcsген`. Estimate the model's parameters. Constrain the parameter  $\lambda$  to be positive

```
. rcsген y, ген(V) дген(v) df(2)
```

Variables V1 to V2 and v1 to v2 were created

```
. local lambda = "exp({theta})"
. local G = "(log(y)+{eta1}*y+{eta2}*V2-({beta0}+{beta1}*x))/`lambda'"
. local g = "(1/y+{eta1}+{eta2}*v2)/`lambda'"
. local f = "exp(-(`G'))/((1+exp(-(`G')))^2)*`g'"
. local S = "1-1/(1+exp(-(`G')))"
. mlexp ((d==1)*ln(`f') + (d==0)*ln(`S'))
initial:      log likelihood = -696.29352
final:        log likelihood = -696.29352
rescale:      log likelihood = -696.29352
Iteration 0:  log likelihood = -696.29352 (not concave)
Iteration 1:  log likelihood = -596.84712 (not concave)
Iteration 2:  log likelihood = -573.66527
Iteration 3:  log likelihood = -566.46684
Iteration 4:  log likelihood = -562.42408
Iteration 5:  log likelihood = -562.22973
Iteration 6:  log likelihood = -562.22929
Iteration 7:  log likelihood = -562.22929
```

Maximum likelihood estimation

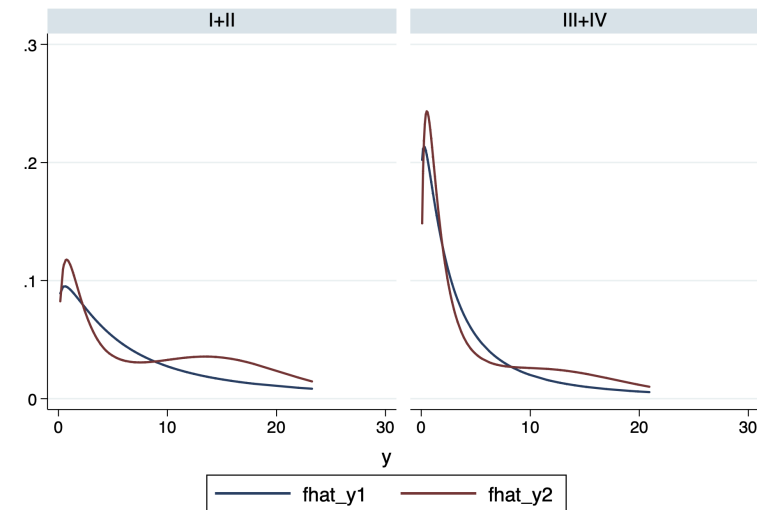
Log likelihood = -562.22929                      Number of obs       =       267

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
/eta1	-.1457699	.0290741	-5.01	0.000	-.2027541	-.0887858
/eta2	-.0008936	.0001706	-5.24	0.000	-.0012279	-.0005592
/beta0	1.334615	.2061882	6.47	0.000	.9304931	1.738736
/beta1	-.5995634	.1623232	-3.69	0.000	-.917711	-.2814157
/theta	-.3866109	.1145084	-3.38	0.001	-.6110432	-.1621786

```
. mlci exp /beta1
.5490513 95% CI: .3994323, .7547145
. mlci exp /theta
.6793554 95% CI: .5427843, .8502893
.
. test [eta1]_cons [eta2]_cons
( 1) [eta1]_cons = 0
( 2) [eta2]_cons = 0
      chi2( 2) = 28.76
      Prob > chi2 = 0.0000
```

Plot the estimated densities  $\hat{f}(y|x)$

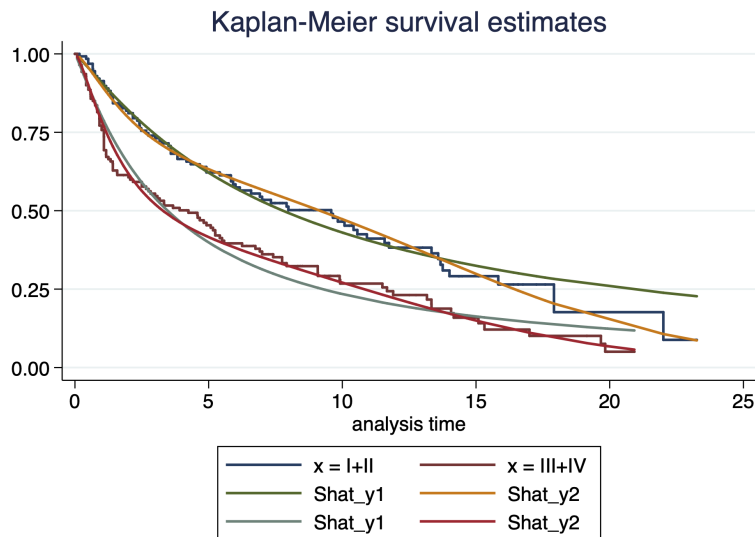
```
. gen fhat_y2 = exp(-((ln(y)+_b[/eta1]*y+_b[/eta2]*V2-(_b[/beta0]+_b[/beta1]*x))/exp(_b[/theta])))/((1+exp(-ln(y)+_b[/eta1]*y+_b[/eta2]*V2-(_b[/beta0]+_b[/beta1]*x))/exp(_b[/theta]))^2)*(1/y+_b[/eta1]+_b[/eta2]*v2)/(exp(_b[/theta]))
. tw (line fhat_y1 fhat_y2 y, sort), name(p3, replace) by(x) ylabel(0(0.1)0.3) legend(rows(1))
. graph export p2.png, replace
(file p2.png written in PNG format)
```



Graphs by x

Plot the estimated survival functions  $\hat{S}(y|x)$  over the Kaplan-Meier estimates.

```
. gen Shat_y2 = 1-1/(1+exp(-(ln(y)+_b[/eta1]*y+_b[/eta2]*V2-( _b[/beta0]+_b[/beta1]*x))/exp(_b[/theta])))
. sts graph, by(x) name(km2, replace) addplot((line Shat_y1 Shat_y2 y if x == 0, sort) ///
> (line Shat_y1 Shat_y2 y if x == 1, sort))
      failure _d: d
      analysis time _t: y
. graph export km2.png, replace
(file km2.png written in PNG format)
```



### Exercise 3

We consider a Weibull model (PH) (see slide 124). Estimate the model's parameters. Constrain the parameter  $k$  to be positive.

But first: we need to extend the likelihood on slide 123 in order to accommodate the censored observations.

The log-likelihood, in the presence of right censoring, is (see slides 76):

$$\log[L(\theta)] = \sum_{i=1}^n I(d_i = 1) \log[f(z_i)] + I(d_i = 0) \log[Sz_i].$$

Knowing that  $f(y) = S(y)h(y)$  and  $S(y) = \exp(-H(y))$  (see slide 121), we can rewrite it as

$$\log[L(\theta)] \doteq \sum_{i=1}^n I(d_i = 1)(\log[\exp(-H(z_i))] + \log[h(z_i)]) + I(d_i = 0) \log[\exp(-H(z_i))] = \sum_{i=1}^n I(d_i = 1) \log[h(z_i)] - H(z_i).$$

This equation justifies the form of the log-likelihood passed to `mlexp` in Exercise 4 and 5.

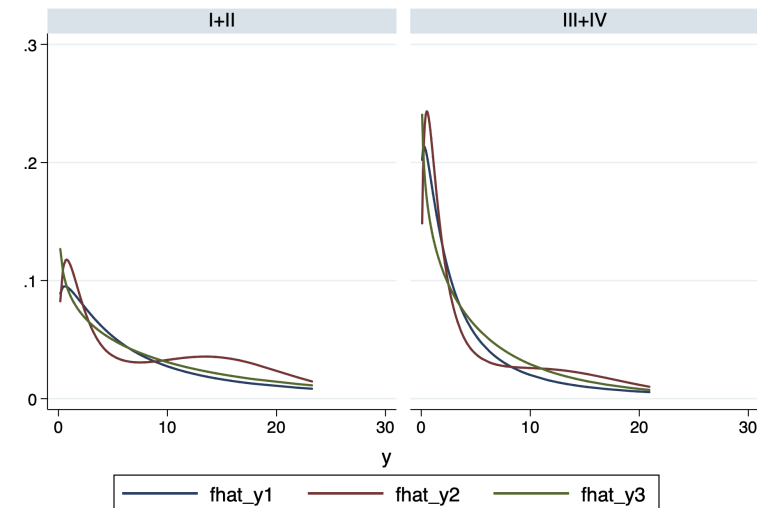
What's the interpretation of  $\beta_1$ ?

```
. local k = "exp({theta})"
. local G = "`k'*log(y)+{beta0}+{beta1}*x"
. local g = "(`k'/y)"
. local H = "exp(`G´)"
. local h = "`H´*`g´"
. mlexp ((d==1)*(ln(`h´)-`H´) + (d==0)*(-`H´))
initial:      log likelihood =  -1620.864
alternative:  log likelihood = -659.98633
rescale:      log likelihood = -659.98633
rescale eq:   log likelihood = -605.31316
Iteration 0:  log likelihood = -605.31316
Iteration 1:  log likelihood = -575.24733
Iteration 2:  log likelihood = -573.66495
Iteration 3:  log likelihood = -573.66061
Iteration 4:  log likelihood = -573.66061
Maximum likelihood estimation
Log likelihood = -573.66061          Number of obs   =          267
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	-.1586235	.0610589	-2.60	0.009	-.2782968 -.0389502
/beta0	-2.141021	.1611149	-13.29	0.000	-2.4568 -1.825241
/beta1	.5370332	.1499405	3.58	0.000	.2431551 .8309113

Plot the estimated densities  $\hat{f}(y|x)$

```
. gen fhat_y3 = exp(ln(exp(exp(_b[/theta])*log(y)+_b[/beta0]+_b[/beta1]*x)*(exp(_b[/theta])/y))-exp(exp(_b[/theta])*log(
> y)+_b[/beta0]+_b[/beta1]*x))
. tw (line fhat_y1 fhat_y2 fhat_y3 y, sort), name(p3, replace) by(x) ylabel(0(0.1)0.3) legend(rows(1))
. graph export p3.png, replace
(file p3.png written in PNG format)
```



Graphs by x

## Exercise 4

We now use a RCS transformation of time to death to make the Weibull model more flexible (see slide 128).

```
. local k = "exp({theta})"
. local G = "`k'*log(y)+{eta1}*y+{eta2}*V2+{beta0}+{beta1}*x"
. local g = "(`k`/y+{eta1}+{eta2}*v2)"
. local H = "exp(`G`)"
. local h = "`H`*`g`"
. mlexp ((d==1)*(ln(`h`)-`H`) + (d==0)*(-`H`))
initial:      log likelihood = -1620.864
final:        log likelihood = -1620.864
rescale:      log likelihood = -1620.864
Iteration 0:   log likelihood = -1620.864
Iteration 1:   log likelihood = -582.7029 (not concave)
Iteration 2:   log likelihood = -568.651
Iteration 3:   log likelihood = -567.19767
Iteration 4:   log likelihood = -565.12405
Iteration 5:   log likelihood = -565.06182
Iteration 6:   log likelihood = -565.06137
Iteration 7:   log likelihood = -565.06137
Maximum likelihood estimation
Log likelihood = -565.06137                Number of obs   =        267
```

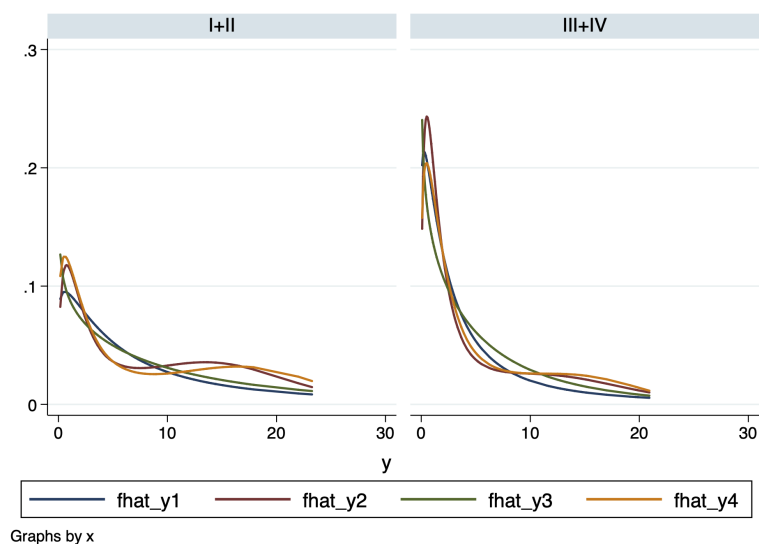
	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
/theta	.2395006	.1123007	2.13	0.033	.0193952 .4596059
/eta1	-.1837425	.0466373	-3.94	0.000	-.2751499 -.092335
/eta2	-.0008603	.000215	-4.00	0.000	-.0012817 -.000439
/beta0	-1.909837	.1633405	-11.69	0.000	-2.229978 -1.589695
/beta1	.5297763	.1499195	3.53	0.000	.2359395 .823613

```
.
. test [eta1]_cons [eta2]_cons
( 1) [eta1]_cons = 0
( 2) [eta2]_cons = 0
      chi2( 2) =    16.21
      Prob > chi2 =    0.0003
```

Plot the estimated densities  $\hat{f}(y|x)$

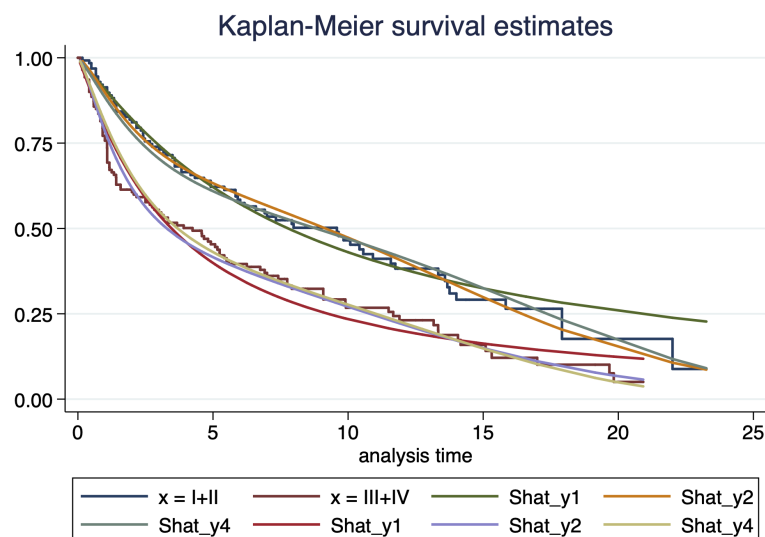
```
. gen fhat_y4 = exp(ln(exp(exp(_b[/theta])*log(y)+_b[/eta1]*y+_b[/eta2]*V2+_b[/beta0]+_b[/beta1]*x)*(exp(_b[/theta])/y+_
> b[/eta1]+_b[/eta2]*v2))-exp(exp(_b[/theta])*log(y)+_b[/eta1]*y+_b[/eta2]*V2+_b[/beta0]+_b[/beta1]*x))
. tw (line fhat_y1 fhat_y2 fhat_y3 fhat_y4 y, sort), name(p4, replace) by(x) ylabel(0(0.1)0.3) legend(rows(1))
```

```
. graph export p4.png, replace
(file p4.png written in PNG format)
```



Plot the estimated survival functions  $\hat{S}(y|x) = \exp(-H(y|x))$  over the Kaplan-Meier estimates.

```
. gen Shat_y4 = exp(-exp(exp(_b[/theta])*log(y)+_b[/eta1]*y+_b[/eta2]*V2+_b[/beta0]+_b[/beta1]*x))
. sts graph, by(x) name(km3, replace) addplot((line Shat_y1 Shat_y2 Shat_y4 y if x == 0, sort) ///
> (line Shat_y1 Shat_y2 Shat_y4 y if x == 1, sort)) legend(cols(4))
      failure _d: d
      analysis time _t: y
. graph export km3.png, replace
(file km3.png written in PNG format)
```



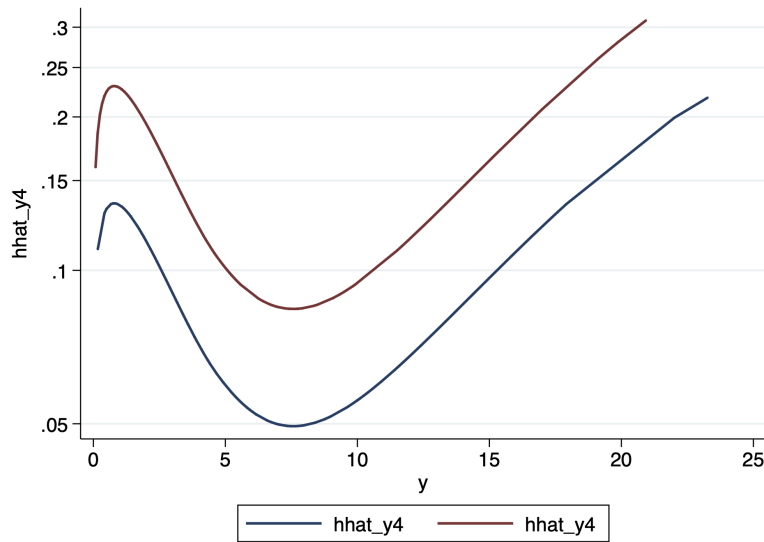
Plot the model-based estimated hazards functions  $\hat{h}(y|x)$ . Use a log scale for the vertical axis to visually check that the model-based hazard functions are actually proportional. **Important:** the hazard functions are proportional because we forced them to be so (using a Weibull PH model)!

Now interpret them, if you can :-)

```
. gen hhat_y4 = exp(exp(_b[/theta])*log(y)+_b[/eta1]*y+_b[/eta2]*V2+_b[/beta0]+_b[/beta1]*x) * ///
> (exp(_b[/theta])/y+_b[/eta1]+_b[/eta2]*v2)
. tw (line hhat_y4 y if x == 0, sort) (line hhat_y4 y if x == 1, sort), ///
```



```
> yscale(log) name(p5, replace)
. graph export p5.png, replace
(file p5.png written in PNG format)
```



## Extra

Can we fit a so-called “flexible parametric survival model”<sup>1</sup> using the tools we’ve learned so far? Of course. To us, it’s just one possible way of modeling  $y$ .

Note that here we apply RCS transforms to  $z = \log(y)$  instead of  $y$ .

(You’ll need to install the command `stpm2`, first)

```
. cap net install stpm2, from(http://fmwww.bc.edu/RePEc/bocode/s)
. stpm2 x, df(3) scale(h) noorthog nolog
Log likelihood = -395.20707                Number of obs   =        267
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
xb						
x	.5302016	.1499414	3.54	0.000	.2363219	.8240813
_rcs1	2.104715	.3198397	6.58	0.000	1.477841	2.73159
_rcs2	.2025602	.0496948	4.08	0.000	.1051602	.2999602
_rcs3	-.2002837	.0547224	-3.66	0.000	-.3075376	-.0930298
_cons	-1.165862	.2565388	-4.54	0.000	-1.668668	-.6630549

```
. di e(ln_bhknobs)
-2.465104022491821 .223943231484774 1.641711472984396 3.091360584567398
.
. gen double z = log(y)
. rcsgen z, gen(z_rcs) dgen(z_d_rcs) knots(-2.465104022491821 .223943231484774 1.641711472984396 3.091360584567398)
Variables z_rcs1 to z_rcs3 and z_d_rcs1 to z_d_rcs3 were created
. local G = "{eta0}*z+{eta1}*z_rcs2+{eta2}*z_rcs3+{beta0}+{beta1}*x"
. local g = "({eta0}+{eta1}*z_d_rcs2+{eta2}*z_d_rcs3)"
. local H = "exp(`G`)"
. local h = "`H'*`g'"
. mlexp ((d==1)*ln(`h`)-`H`), from(eta0=1 eta1=0 eta2=0 beta0=0 beta1=0)
Iteration 0: log likelihood = -1453.6616 (not concave)
Iteration 1: log likelihood = -513.69571
Iteration 2: log likelihood = -440.91209
Iteration 3: log likelihood = -402.04626
```

<sup>1</sup>Royston, P., & Parmar, M. K. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in medicine*, 21(15), 2175-2197.

Iteration 4: log likelihood = -395.46006  
 Iteration 5: log likelihood = -395.20744  
 Iteration 6: log likelihood = -395.20707  
 Iteration 7: log likelihood = -395.20707

Maximum likelihood estimation

Log likelihood = -395.20707                      Number of obs        =            267

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
/eta0	2.104716	.3198539	6.58	0.000	1.477814	2.731618
/eta1	.2025603	.0496982	4.08	0.000	.1051535	.2999671
/eta2	-.2002838	.0547262	-3.66	0.000	-.3075452	-.0930224
/beta0	-1.165861	.2565491	-4.54	0.000	-1.668688	-.6630343
/beta1	.5302016	.1499414	3.54	0.000	.2363219	.8240813