

## August 2005 Stata Application Tutorial 2: Parametric Models

---

Data Note: Code makes use of UNFINAL.dta and cabinet.dta. Both data sets are available on the Event History website. Code is based on Stata version 8.

---

**Preliminaries:** Basic Parametrics and the Proportionality Property in the Weibull Model (keep PH results in mind when we study Cox) .

Let's start with basic model with two binary covariates:

### Exponential

```
. streg civil interst, dist(exp) nohr
```

```
      failure _d:  failed
analysis time _t:  duration
```

```
Iteration 0:  log likelihood = -103.03289
Iteration 1:  log likelihood = -90.211473
Iteration 2:  log likelihood = -86.44131
Iteration 3:  log likelihood = -86.354656
Iteration 4:  log likelihood = -86.354481
Iteration 5:  log likelihood = -86.354481
```

Exponential regression -- log relative-hazard form

```
No. of subjects =          54          Number of obs   =          54
No. of failures =          39
Time at risk    =          3994
Log likelihood  = -86.354481          LR chi2(2)       =          33.36
                                          Prob > chi2      =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
civil	1.169344	.3588703	3.26	0.001	.4659714 1.872717
interst	-1.6401	.4954337	-3.31	0.001	-2.611132 -.6690679
_cons	-4.350864	.2132007	-20.41	0.000	-4.76873 -3.932999

Exponentiating the coefficients yields the hazard ratio:

```
. display exp(_b[civil])
3.2198805
```

Inference? Interventions prompted by civil wars are about 3.2 times more likely to fail than when compared to the baseline category of internationalized civil wars. Note that in this parameterization, the coefficients are expressed in terms of the hazard rate. Thus, a positive coefficient implies the risk (or  $h[t]$ ) is increasing (and hence the survival time ( $S[t]$ ) is decreasing).

Re-estimating the exponential as an “accelerated failure time model”—that is a linear model for  $\log(t)$ .

```
. streg civil interst, dist(exp) nohr time
```

```
      failure _d: failed
analysis time _t: duration
```

```
Iteration 0:  log likelihood = -103.03289
Iteration 1:  log likelihood = -90.211473
Iteration 2:  log likelihood = -86.44131
Iteration 3:  log likelihood = -86.354656
Iteration 4:  log likelihood = -86.354481
Iteration 5:  log likelihood = -86.354481
```

Exponential regression -- accelerated failure-time form

```
No. of subjects =          54          Number of obs   =          54
No. of failures =          39
Time at risk    =          3994
Log likelihood  = -86.354481          LR chi2(2)       =          33.36
                                          Prob > chi2      =          0.0000
```

_____ _t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
civil	-1.169344	.3588703	-3.26	0.001	-1.872717 - .4659714
interst	1.6401	.4954337	3.31	0.001	.6690679 2.611132
_____ _cons	4.350864	.2132007	20.41	0.000	3.932999 4.76873

Here, the coefficients are expressed in terms of  $\log(t)$ . Thus, a positive coefficient implies the expected log survival time is increasing with the covariate; a negatively signed coefficient implies the survival time is decreasing. The hazard ratios are obtained as before (though we exponentiate  $-\beta$  instead of  $\beta$  [why?]).

```
. display exp(-_b[civil])
3.2198805
```

As it should be, it makes no difference which parameterization you choose: the hazard ratio is the same (3.22).

Let’s examine the PH property a bit. Below, I compute the estimated hazard rates for each covariate:

Civil Wars:

```
. display exp(-(_b[_cons]+_b[civil]*1))
.04152249
```

Interstate Conflicts:

```
. display exp(-(_b[_cons]+_b[interst]*1))
.00250125
```

ICWs:

```
. display exp(-(_b[_cons]))
.01289566
```

The PH property says that since the increase (or decrease) in the hazard rate is a multiple of the baseline hazard rate, the change in the hazard rate is *proportional* to the baseline hazard. Here, then, are the hazard ratios:

Civil Wars:

```
. display .04152249/.01289566
3.219881
```

Interstate Conflicts:

```
. display .00250125/.01289566
.1939606
```

ICWs:

```
. display .01289566/.01289566
1
```

Of course we don't have to do this! The PH property implies that the ratio of two hazards is equal to the multiple of the baseline hazard. The "multiple" in this setting is the estimated coefficient. To find the hazard ratios, all we have to do is exponentiate  $-\beta$  (if we're using the AFT model):

Civil Wars:

```
. display exp(-_b[civil])
3.2198805
```

Interstate Conflicts:

```
. display exp(-_b[interst])
.19396062
```

ICWs:

```
. display exp(0)
1
```

...which of course is what we obtained before. Most software programs will compute this for you directly. In Stata I use a predict command to generate the hazard rates and ratios:

```
. predict hazard_rate, hazard
(4 missing values generated)
```

```
. tab hazard_rate
```

predicted hazard	Freq.	Percent	Cum.
.0025013	10	18.52	18.52
.0128957	30	55.56	74.07
.0415225	14	25.93	100.00
Total	54	100.00	

```
. predict hazard_ratios, hr
(4 missing values generated)
```

```
. tab hazard_ratios
```

hazard ratio	Freq.	Percent	Cum.
.1939606	10	18.52	18.52
1	30	55.56	74.07
3.219881	14	25.93	100.00
Total	54	100.00	

And again, it's the same estimates that we computed "by hand."

All of the above holds for the Weibull (the math is a bit different however). Let's consider a Weibull:

```
. streg civil interst, dist(weib) time
```

```
failure _d: failed
analysis time _t: duration
```

Fitting constant-only model:

```
Iteration 0: log likelihood = -103.03289
Iteration 1: log likelihood = -93.501426
Iteration 2: log likelihood = -93.488663
Iteration 3: log likelihood = -93.488663
```

Fitting full model:

```
Iteration 0: log likelihood = -93.488663
Iteration 1: log likelihood = -86.548564
Iteration 2: log likelihood = -84.667898
Iteration 3: log likelihood = -84.655162
Iteration 4: log likelihood = -84.655157
```

Weibull regression -- accelerated failure-time form

```
No. of subjects =          54          Number of obs =          54
No. of failures =          39
Time at risk   =          3994
Log likelihood = -84.655157          LR chi2(2) =          17.67
                                          Prob > chi2 =          0.0001
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
civil	-1.100421	.4457861	-2.47	0.014	-1.974146	-.2266966
interst	1.736832	.6165459	2.82	0.005	.5284242	2.94524
_cons	4.28793	.2652436	16.17	0.000	3.768062	4.807798
/ln_p	-.2145617	.1237889	-1.73	0.083	-.4571834	.02806
p	.806895	.0998846			.6330642	1.028457
1/p	1.239319	.1534138			.97233	1.579619

## Generating the Hazard Rates “the hard way.”

```
. gen lambda_civil=exp(-(_b[_cons]+_b[civil]))  
. gen haz_civil=lambda_civil*e(aux_p)*(lambda_civil*duration)^(e(aux_p)-1)
```

## ...and “the easy way.”

```
. predict hazard_civil, hazard, if civil==1
```

(The “hard way” is “by hand”; the “easy way” is through Stata options [they’re the same numbers!].)

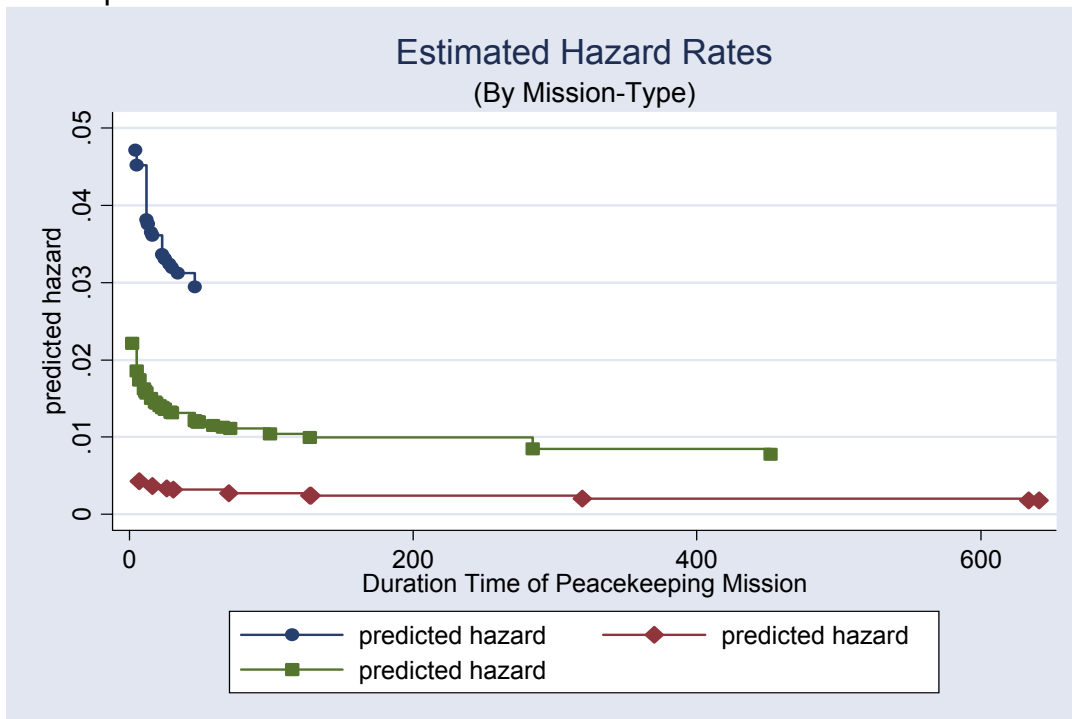
Let’s generate the hazard rates for the other two mission types:

```
. predict hazard_interst, hazard, if interst==1  
(48 missing values generated)  
. predict hazard_icw, hazard, if interst==0 & civil==0  
(28 missing values generated)
```

We could graph them:

```
twoway (scatter hazard_civil _t, connect(s) msymbol(O)) (scatter  
hazard_interst_t, connect(s) msymbol(D)) (scatter hazard_icw_t, connect(s)  
msymbol(S)), xtitle(Duration Time of Peacekeeping Mission) title(Estimated  
Hazard Rates ) subtitle((By Mission-Type)) saving(c:\ehbook\icpsr_unhazrates,  
replace)
```

Which produces:



(The top line denotes civil wars; the middle line denotes internationalized civil wars; the bottom line denotes interstate conflicts). There is a big difference between a plot of the hazard rate from a Weibull and the plot of the hazard rate from the exponential. (What is it?)

We could use the “predict” option to generate hazard ratios:

```
. predict hr_interst, hr, if interst==1
(48 missing values generated)

. predict hr_civil, hr, if civil==1
(44 missing values generated)

. predict hr_icw, hr, if civil==0 & interst==0
(28 missing values generated)

. tab hr_interst
```

hazard ratio	Freq.	Percent	Cum.
.2462419	10	100.00	100.00
Total	10	100.00	

```
. tab hr_civil
```

hazard ratio	Freq.	Percent	Cum.
2.430081	14	100.00	100.00
Total	14	100.00	

```
. tab hr_icw
```

hazard ratio	Freq.	Percent	Cum.
1	30	100.00	100.00
Total	30	100.00	

...or we could do it “by hand.”

```
. display exp(-(_b[interst]))^(e(aux_p))
.24624185

. display exp(-(_b[civil]))^(e(aux_p))
2.4300808

. display exp(-(0))^(e(aux_p))
1
```

It all works in a pretty simple way!

Let's now include a covariate other than a binary covariate:

```
. streg civil interst borders, dist(weib) time nolog
```

```
      failure _d: failed
analysis time _t: duration
```

```
Weibull regression -- accelerated failure-time form
```

```
No. of subjects =          46                Number of obs =          46
No. of failures =          36
Time at risk   =          3840
Log likelihood = -76.493097                LR chi2(3) =          18.45
                                                Prob > chi2 =          0.0004
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
civil	-1.380352	.4921063	-2.80	0.005	-2.344862	-.4158411
interst	1.806995	.6347777	2.85	0.004	.5628534	3.051136
borders	-.1368689	.0972727	-1.41	0.159	-.3275199	.053782
_cons	4.800974	.4777848	10.05	0.000	3.864533	5.737415
/ln_p	-.2278767	.1328443	-1.72	0.086	-.4882467	.0324932
p	.7962224	.1057736			.6137014	1.033027
1/p	1.255931	.1668432			.968029	1.629457

Let's look at the PH property for this covariate.

```
. gen hazratio_borders=exp(-_b[borders]*borders)^e(aux_p)
```

We'll tabulate these ratios. Note that the ratio is increasing as the number of borders increases.

```
. table hazratio_borders borders
```

hazratio_ borders	borders									
	1	2	3	4	5	6	8	9	13	
1.115138	10									
1.243533		7								
1.38671			6							
1.546373				12						
1.72442					8					
1.922966						3				
2.391271							2			
2.666597								1		
4.123554									1	

BUT, the PH property still must hold. Take the ratio of any adjacent pair:

```
. display 1.546373/1.38671  
1.115138
```

Note that this is equivalent to:

```
. display exp(-_b[borders])^e(aux_p)  
1.1151379
```

...which is the hazard ratio for the “baseline case” (i.e. borders=1; look in the table above!).

---

## Extended Illustration: All Sorts of Parametrics

### Weibull

```
. streg invest polar numst format postelec caretakr, dist(weib) time nolog
```

```
      failure _d:  censor
analysis time _t:  durat
```

```
Weibull regression -- accelerated failure-time form
```

```
No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood   = -414.07496                LR chi2(6)        =          171.94
                                                Prob > chi2       =           0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
invest	-.2958188	.1059024	-2.79	0.005	-.5033838	-.0882538
polar	-.017943	.0042784	-4.19	0.000	-.0263285	-.0095575
numst	.4648894	.1005815	4.62	0.000	.2677533	.6620255
format	-.1023747	.0335853	-3.05	0.002	-.1682006	-.0365487
postelec	.6796125	.104382	6.51	0.000	.4750276	.8841974
caretakr	-1.33401	.2017528	-6.61	0.000	-1.729438	-.9385818
_cons	2.985428	.1281146	23.30	0.000	2.734328	3.236528
/ln_p	.257624	.0500578	5.15	0.000	.1595126	.3557353
p	1.293852	.0647673			1.172939	1.42723
1/p	.7728858	.0386889			.700658	.8525593

### Exponential

```
. streg invest polar numst format postelec caretakr, dist(exp) time nolog
```

```
      failure _d:  censor
analysis time _t:  durat
```

```
Exponential regression -- accelerated failure-time form
```

```
No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood   = -425.90641                LR chi2(6)        =          148.53
                                                Prob > chi2       =           0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
invest	-.3322088	.1376729	-2.41	0.016	-.6020426	-.0623749
polar	-.0193017	.0055465	-3.48	0.001	-.0301725	-.0084308
numst	.515435	.1291486	3.99	0.000	.2623084	.7685616
format	-.1079432	.0435233	-2.48	0.013	-.1932474	-.022639
postelec	.7403427	.134558	5.50	0.000	.4766138	1.004072
caretakr	-1.319272	.2595422	-5.08	0.000	-1.827965	-.8105783
_cons	2.944518	.1663401	17.70	0.000	2.618498	3.270539

## Log-logistic

```
. streg invest polar numst format postelec caretakr, dist(loglog) time nolog
```

```
      failure _d:  censor
analysis time _t:  durat
```

```
Log-logistic regression -- accelerated failure-time form
```

```
No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood  = -424.10921                LR chi2(6)         =          148.72
                                                Prob > chi2        =           0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
invest	-.3367541	.1278083	-2.63	0.008	-.5872538 -.0862544
polar	-.0221958	.0052638	-4.22	0.000	-.0325127 -.0118789
numst	.4830709	.1212506	3.98	0.000	.2454241 .7207177
format	-.1093453	.0419715	-2.61	0.009	-.1916078 -.0270827
postelec	.6408808	.1240329	5.17	0.000	.3977807 .8839808
caretakr	-1.26921	.2310272	-5.49	0.000	-1.722015 -.8164046
_cons	2.728818	.1595866	17.10	0.000	2.416034 3.041602
/ln_gam	-.5657686	.0511353	-11.06	0.000	-.665992 -.4655451
gamma	.5679235	.029041			.5137636 .6277928

## Log-normal

```
. streg invest polar numst format postelec caretakr, dist(lognorm) time nolog
```

```
      failure _d:  censor
analysis time _t:  durat
```

```
Log-normal regression -- accelerated failure-time form
```

```
No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood  = -425.30621                LR chi2(6)         =          150.66
                                                Prob > chi2        =           0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
invest	-.3738013	.1327055	-2.82	0.005	-.6338993 -.1137032
polar	-.021988	.0054825	-4.01	0.000	-.0327336 -.0112424
numst	.5717579	.1232281	4.64	0.000	.3302353 .8132805
format	-.1194982	.0432516	-2.76	0.006	-.2042698 -.0347266
postelec	.6668079	.1292366	5.16	0.000	.4135088 .920107
caretakr	-1.126047	.2576962	-4.37	0.000	-1.631122 -.6209713
_cons	2.632497	.164494	16.00	0.000	2.310095 2.954899
/ln_sig	.0078719	.0439881	0.18	0.858	-.0783432 .0940871
sigma	1.007903	.0443358			.924647 1.098655

## Gompertz (a PH model but not an AFT model)

```
. streg invest polar numst format postelec caretakr, dist(gompertz) nolog

      failure _d:  censor
      analysis time _t:  durat

Gompertz regression -- log relative-hazard form

No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood  = -418.97771                LR chi2(6)        =          159.11
                                                Prob > chi2       =          0.0000
```

_t	Haz. Ratio	Std. Err.	z	P> z	[95% Conf. Interval]
invest	1.442137	.1984577	2.66	0.008	1.101208 1.888615
polar	1.022888	.0057628	4.02	0.000	1.011655 1.034245
numst	.5446288	.0720488	-4.59	0.000	.420238 .7058394
format	1.135794	.04998	2.89	0.004	1.041941 1.238101
postelec	.4104204	.0583874	-6.26	0.000	.3105525 .542404
caretakr	4.382191	1.157914	5.59	0.000	2.610819 7.355392
gamma	.0225632	.005949	3.79	0.000	.0109032 .0342231

## ...and the Generalized Gamma

```
. streg invest polar numst format postelec caretakr, dist(gamma) nolog

      failure _d:  censor
      analysis time _t:  durat

Gamma regression -- accelerated failure-time form

No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood  = -414.00944                LR chi2(6)        =          165.78
                                                Prob > chi2       =          0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
invest	-.3005269	.108745	-2.76	0.006	-.5136633 -.0873906
polar	-.0182998	.0044674	-4.10	0.000	-.0270559 -.0095438
numst	.4692142	.1030895	4.55	0.000	.2671626 .6712659
format	-.1031368	.0342637	-3.01	0.003	-.1702925 -.0359811
postelec	.6807161	.1061356	6.41	0.000	.4726942 .888738
caretakr	-1.328476	.2066422	-6.43	0.000	-1.733487 -.9234647
_cons	2.963114	.1447075	20.48	0.000	2.679492 3.246735
/ln_sig	-.234325	.0802121	-2.92	0.003	-.3915378 -.0771122
/kappa	.9241712	.2065399	4.47	0.000	.5193605 1.328982
sigma	.7911047	.0634561			.6760165 .9257859

From the generalized gamma, we can evaluate the fit of models nested within this encompassing distribution. For these data, model suggests Weibull is preferred to other nest parameterizations.

## Application Issues: Interpretation

Suppose we have settled on the Weibull:

```
. streg invest polar numst format postelec caretakr, dist(weib) time nolog
      failure _d:  censor
      analysis time _t:  durat

Weibull regression -- accelerated failure-time form

No. of subjects =          314                Number of obs   =          314
No. of failures =          271
Time at risk    =          5789.5
Log likelihood  = -414.07496                LR chi2(6)          =          171.94
                                                Prob > chi2         =           0.0000
```

_t	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
invest	-.2958188	.1059024	-2.79	0.005	-.5033838	-.0882538
polar	-.017943	.0042784	-4.19	0.000	-.0263285	-.0095575
numst	.4648894	.1005815	4.62	0.000	.2677533	.6620255
format	-.1023747	.0335853	-3.05	0.002	-.1682006	-.0365487
postelec	.6796125	.104382	6.51	0.000	.4750276	.8841974
caretakr	-1.33401	.2017528	-6.61	0.000	-1.729438	-.9385818
_cons	2.985428	.1281146	23.30	0.000	2.734328	3.236528
/ln_p	.257624	.0500578	5.15	0.000	.1595126	.3557353
p	1.293852	.0647673			1.172939	1.42723
1/p	.7728858	.0386889			.700658	.8525593

Reporting expected survival times for various covariate profiles might be useful. The function for the mean is complicated. If we were to generate the function “by hand” we would have a cumbersome statement:

```
gen
expected_S=exp(lngamma(1+1/e(aux_p)))*exp(_b[_cons]+_b[invest]*invest+_b[polar]
*polar + _b[numst]*numst + _b[format]*format + _b[postelec]*postelec +
_b[caretakr]*caretakr)
```

Ugly (though it shows you how to implement the function). Most software programs nowadays have options to generate functions like the mean. Stata does as well.

```
. predict mean_S, mean
```

This will compute the expected survival time for all covariate profiles. It is identical to the “by hand” command issued above. You may be interested in the mean function for particular covariate profiles.

```
. predict mean_S_profile1, mean, if invest==0 & polar==15 & numst==0
```

This gives the expected mean for the case when the investiture and numerical status covariates are set to 0 and the polarization variable is set to its mean (about 15). The remaining covariates are free to vary. The choice of covariate profile you use should be substantively driven (i.e. it’s easy to create a lot of predictions based on nonexistent data points! Be careful!)

The median may also be used to describe the model. Under the Weibull, we could compute the median survival time for all possible covariate profiles:

```
. gen median_S=exp(_b[_cons] + _b[invest]*invest + _b[polar]*polar +  
_b[numst]*numst + _b[format]*format + _b[postelec]*postelec +  
_b[caretakr]*caretakr)*log(2)^(1/e(aux_p))
```

Yes, it’s ugly, but it is the function used to derive the median (note any percentile could be derived here; just substitute 1.33 for the 2 and that will give you the 25<sup>th</sup> percentile; substitute 4 and you’ll get the 75<sup>th</sup> percentile). Of course we could directly compute this using the predict option:

```
. predict med_S, median
```

And we could had “if” statements to generate the median survival time for a specific covariate profile.

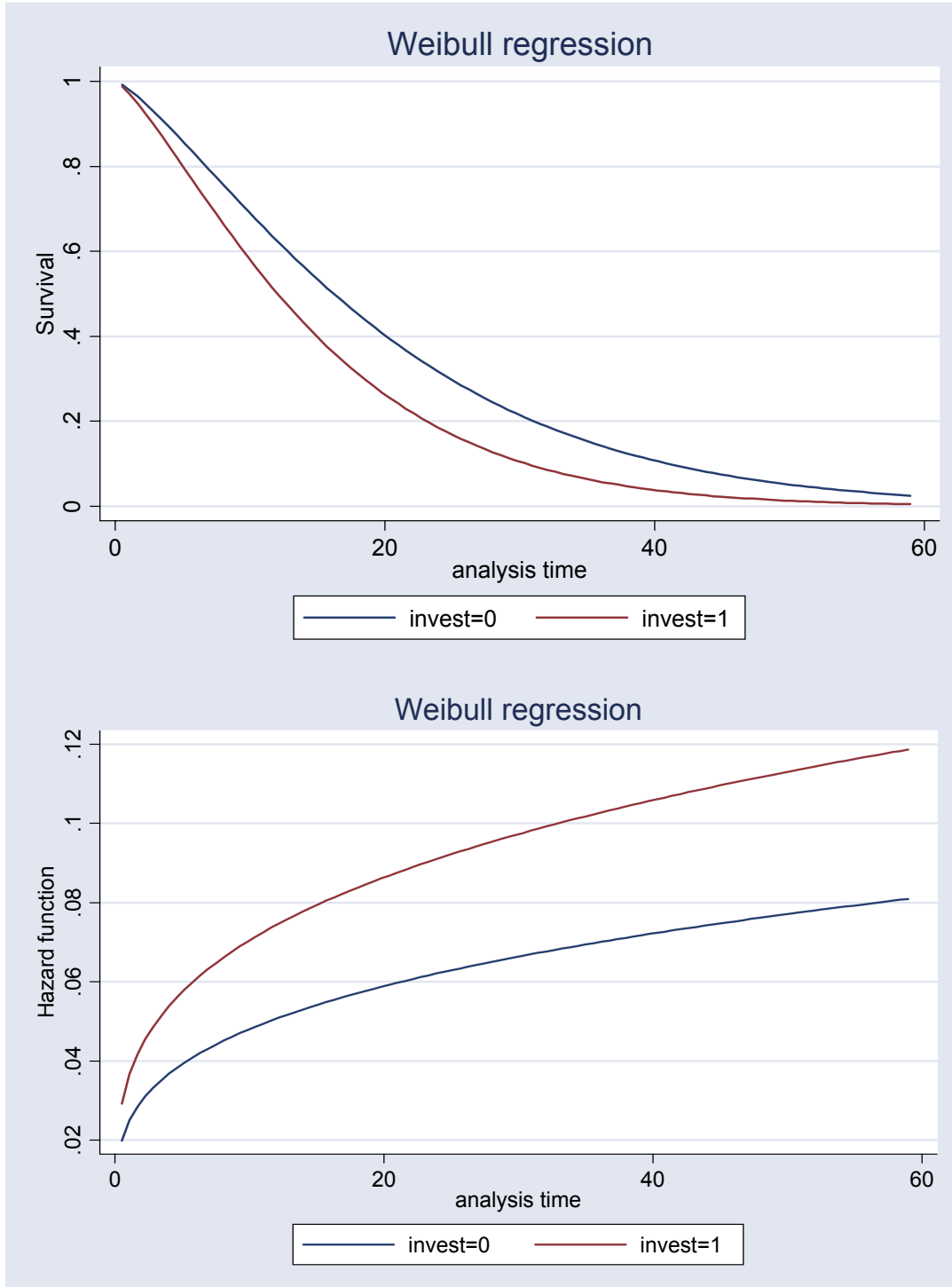
```
. predict med_S_profile1, median, if invest==0 & polar==15 & numst==0
```

This is the same covariate profile as from above (and the same caveats apply).

Graphs of survival functions, hazard rates, and so forth are also useful. In the context of Stata, the stcurv option is very useful. For example, if we wanted the estimated survival function plotted for the case when the investiture requirement was 0 vs. 1, we could use stcurv in the following way:

```
. stcurv, survival at1(invest=0) at2(invest=1)  
. stcurv, hazard at1(invest=0) at2(invest=1)
```

This returns:



(If you don't have a color printer, the top line in the first figure is the case when investiture=0; the top line in the second figure is the case when investiture=1; why are they "flipped"?). Of course these graphs could be improved...but you get the point (I hope!).