Fundamentals of Statistical Modeling (VT20)

Andrea Discacciati
Karolinska Institutet
Stockholm, Sweden

Lab 3 (Extra material on flexible modeling with splines)

Load the dataset and the `mlci` command

```
. version 14
. use https://raw.githubusercontent.com/anddis/fsm/master/data/lab3_1.dta, clear
. run https://raw.githubusercontent.com/anddis/fsm/master/do/mlci.do
```

Assume that $f(age)$ follows a generalized extreme values distribution. Estimate the parameters μ and σ. Constrain σ to be positive. Inflate the probability of death during the first year of life, while constraining it to be between 0 and 1.

```
. local G = "exp((age-{mu})/exp({theta1}))"
. local g = "exp((age-{mu})/exp({theta1}))/exp({theta1})"
. local eta = "invlogit({theta2})"
. local f = "exp(-`G´)*`g´"
. mlexp ((age<1)*ln(`eta´) + (age>=1)*ln((1-`eta´)*`f´))
initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -703081.71
rescale:       log likelihood = -374140.51
rescale eq:    log likelihood = -136866.71
Iteration 0:   log likelihood = -136866.71
Iteration 1:   log likelihood = -129626.39
Iteration 2:   log likelihood =  -128639.2
Iteration 3:   log likelihood = -128638.98
Iteration 4:   log likelihood = -128638.98
Maximum likelihood estimation
Log likelihood = -128638.98                     Number of obs     =     33,638
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/theta2	-5.356114	.0797433	-67.17	0.000	-5.512408	-5.19982
/mu	87.72222	.0516779	1697.48	0.000	87.62094	87.82351
/theta1	2.200033	.0044046	499.49	0.000	2.1914	2.208666

```
. mlci exp /theta1
 9.025309   95% CI:  8.94773,  9.10356
. mlci invlogit /theta2
  .004697   95% CI: .0040201, .0054873
```

Generate the estimated density and the transform $u1 = \hat{F}(y)$ (we'll use it to assess the goodness-of-fit).

Remember: we're assuming that, for $age \geq 1$, the variable age is Standard-Exponential–distributed after we apply the transform $G(y)$. The CDF of a Standard Exponential is $F(y) = 1 - \exp(-y)$

```
. gen fhat_age = invlogit(_b[/theta2])^(age<1) * ///
> ((1-invlogit(_b[/theta2]))* ///
> exp(-exp((age-_b[/mu])/exp(_b[/theta1])))*exp((age-_b[/mu])/exp(_b[/theta1]))/exp(_b[/theta1]))^(age>=1)
.
. gen u1 = invlogit(_b[/theta2]) + (1-invlogit(_b[/theta2])) * (1 - exp(-exp((age-_b[/mu])/exp(_b[/theta1])))) * (age>=1
> )
```

Now, let's include a spline transformation of *age* with 4 degrees of freedom and let's see whether this improves the fit of our generalized extreme values model. Jointly test the 3 parameters η_1, η_2, η_3 to assess whether adding the 3 RCS transforms improves the fit of this model with respect to the "base" model (see above).

We need to help Stata a little by providing reasonable initial values for the model's parameters.

```
. rcsgen age, gen(V) dgen(v) df(4)
Variables V1 to V4 and v1 to v4 were created

. local G = "exp((age+{eta1}*V2+{eta2}*V3+{eta3}*V4-{mu})/exp({theta1}))"

. local g = "exp((age+{eta1}*V2+{eta2}*V3+{eta3}*V4-{mu})/exp({theta1}))*(1+{eta1}*v2+{eta2}*v3+{eta3}*v4)/exp({theta1})
> "

. local eta = "invlogit({theta2})"

. local f = "exp(-`G´)*`g´"

. mlexp ((age<1)*ln(`eta´) + (age>=1)*ln((1-`eta´)*`f´)), from(mu=80 theta1=2 theta2=0 eta1=0 eta2=0 eta3=0)

Iteration 0:   log likelihood = -185382.73  (not concave)
Iteration 1:   log likelihood = -132527.19  (not concave)
Iteration 2:   log likelihood = -129308.19  (not concave)
Iteration 3:   log likelihood = -128997.15
Iteration 4:   log likelihood = -128383.09
Iteration 5:   log likelihood = -127982.99  (not concave)
Iteration 6:   log likelihood = -127979.54
Iteration 7:   log likelihood = -127955.74
Iteration 8:   log likelihood = -127934.54
Iteration 9:   log likelihood = -127934.34
Iteration 10:  log likelihood = -127934.33
Iteration 11:  log likelihood = -127934.33

Maximum likelihood estimation

Log likelihood = -127934.33                     Number of obs     =     33,638
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/theta2	-5.356113	.0797433	-67.17	0.000	-5.512407	-5.199819
/eta1	.0019146	.0000758	25.27	0.000	.0017661	.0020631
/eta2	-.0053266	.000257	-20.72	0.000	-.0058304	-.0048228
/eta3	.0036992	.0002157	17.15	0.000	.0032764	.0041221
/mu	90.98673	1.423035	63.94	0.000	88.19763	93.77583
/theta1	2.449532	.0319658	76.63	0.000	2.38688	2.512183

```
. mlci exp /theta1
 11.58292   95% CI: 10.87949, 12.33183

. mlci invlogit /theta2
  .004697   95% CI: .0040202, .0054873

.
. test [eta1]_b[_cons] [eta2]_b[_cons] [eta3]_b[_cons]

 ( 1)  [eta1]_cons = 0
 ( 2)  [eta2]_cons = 0
 ( 3)  [eta3]_cons = 0

           chi2(  3) =  879.01
         Prob > chi2 =    0.0000
```

Generate the estimated density and the transform $u2 = \hat{F}(y)$.

```
. gen fhat_age1 = invlogit(_b[/theta2])^(age<1) * ///
> ((1-invlogit(_b[/theta2]))* ///
> exp(-exp((age+_b[/eta1]*V2+_b[/eta2]*V3+_b[/eta3]*V4-_b[/mu])/exp(_b[/theta1])))* ///
> exp((age+_b[/eta1]*V2+_b[/eta2]*V3+_b[/eta3]*V4-_b[/mu])/exp(_b[/theta1]))* ///
> ((1+_b[/eta1]*v2+_b[/eta2]*v3+_b[/eta3]*v4)/exp(_b[/theta1])))^(age>=1)

.
. gen u2 = invlogit(_b[/theta2]) + (1-invlogit(_b[/theta2])) * ///
> (1 - exp(-exp((age+_b[/eta1]*V2+_b[/eta2]*V3+_b[/eta3]*V4-_b[/mu])/exp(_b[/theta1])))) * (age>=1)
```

Plot the 2 estimated densities over the sample histogram and the quantile plot.

```
. tw (hist age, discrete) ///
> (scatter fhat_age fhat_age1 age if age<1, sort mcol(navy maroon) msize(small small)) ///
> (line fhat_age fhat_age1 age if age>=1, sort lc(navy maroon)), name(p1, replace) legend(off)

. graph export p1.png, replace
(file p1.png written in PNG format)

.
```

```
. qplot u1 u2, addplot(function y = x) name(p2, replace) legend(off) lc(navy maroon) ///
> msize(vsmall vsmall) msym(O O)

. graph export p2.png, replace
(file p2.png written in PNG format)

.
```