Fundamentals of Statistical Modeling (VT21)

Andrea Discacciati
Karolinska Institutet
Stockholm, Sweden

Lab 3

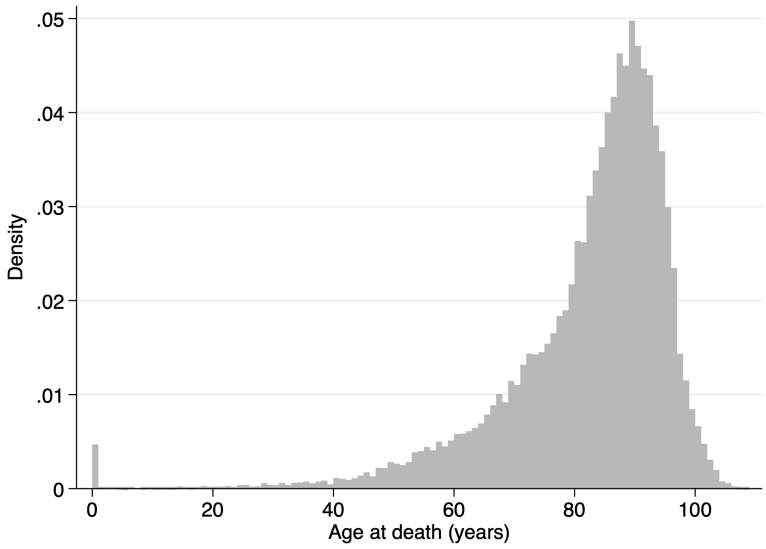
Load the dataset and the `mlci` command

```
. 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
```

Exercise 1

I retrieved data on age at death among females in Switzerland in 2016 from `http://www.mortality.org` (variable `age`) ($n = 33,638$). There are no censored observations (we know the age at death for all individuals). Plot an histogram of age at death. What can we say about the distribution?

```
. hist age, width(1) name(p0, replace)
(bin=109, start=0, width=1)
. graph export p0.png, replace
(file p0.png written in PNG format)
```



Assume that $f(age)$ follows a generalized extreme values distribution. Estimate the parameters μ and σ. Constrain σ to be positive.

Remember: we're assuming that the variable *age* is Standard-Exponential–distributed after we apply the transform $G(y)$. The pdf of a Standard Exponential distribution is $f_{SE}(u) = \exp(-u)$.

```
. local G = "exp((age-{mu})/exp({theta}))"
. local g = "exp((age-{mu})/exp({theta}))/exp({theta})"
. local f = "exp(-`G´)*`g´"
```

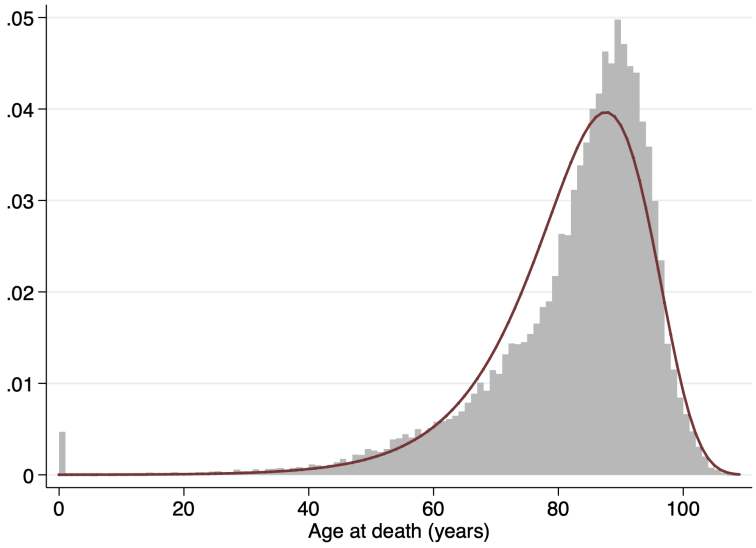
```
. mlexp(ln(`f´))
initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -370018.19
rescale:       log likelihood = -207462.71
rescale eq:    log likelihood =  -137283.6
Iteration 0:   log likelihood =  -137283.6
Iteration 1:   log likelihood = -130159.81
Iteration 2:   log likelihood = -129500.71
Iteration 3:   log likelihood = -129497.57
Iteration 4:   log likelihood = -129497.57
Maximum likelihood estimation
Log likelihood = -129497.57                     Number of obs     =     33,638
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/mu	87.58256	.052907	1655.40	0.000	87.47886	87.68625
/theta	2.227608	.0044218	503.78	0.000	2.218941	2.236275

```
. mlci exp /theta
 9.277647   95% CI: 9.197589, 9.358402
```

Plot the estimated density $\hat{f}(age)$ over the sample histogram

```
. gen fhat_age = exp(-exp((age-_b[/mu])/exp(_b[/theta])))*exp((age-_b[/mu])/exp(_b[/theta]))/exp(_b[/theta])
. tw (hist age, width(1)) (line fhat_age age, sort), name(p1, replace) legend(off)
. graph export p1.png, replace
(file p1.png written in PNG format)
```



Exercise 2

Inflate the probability of death during the first year of life ($age < 1$), while constraining it to be between 0 and 1. How do you interpret the coefficient η?

Note: we can probably improve the fit of this model by making it more flexible, for example using restricted cubic splines. This is described in the Extra material for Lab 3.

```
. 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
```

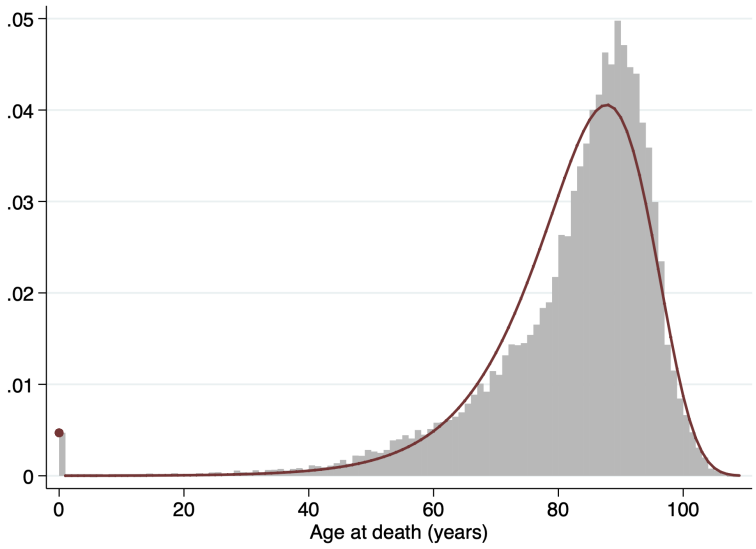
```
rescale:       log likelihood = -374140.51
rescale eq:    log likelihood = -136866.71
Iteration 0:   log likelihood = -136866.71
Iteration 1:   log likelihood = -129626.39
Iteration 2:   log likelihood =  -128639.2
Iteration 3:   log likelihood = -128638.98
Iteration 4:   log likelihood = -128638.98

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
```

Plot the estimated density $\hat{f}(age)$ over the sample histogram

```
. gen fhat_age2 = 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)

. tw (hist age, width(1)) (scatter fhat_age2 age if age<1, sort msiz(small) lc(maroon)) ///
> (line fhat_age2 age if age>=1, sort lc(maroon)), name(p2, replace) legend(off)

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

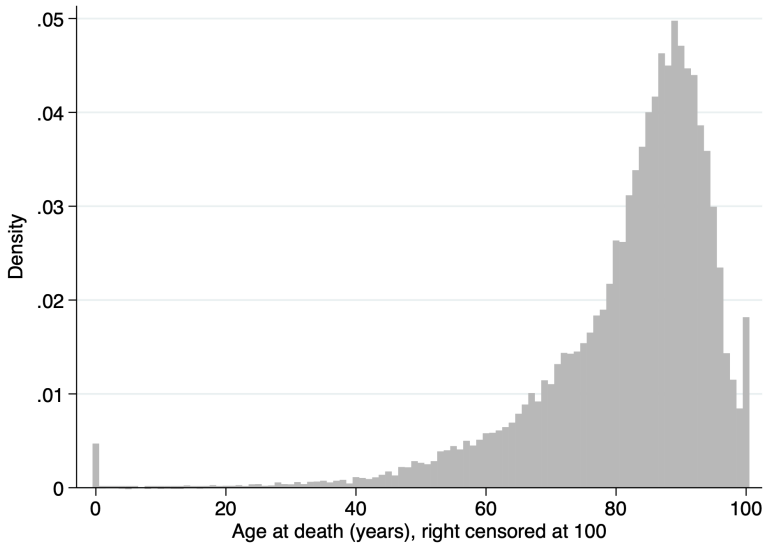


Exercise 3

Assume now that all ages above 100 years were recorded as 100 years (those ages are right-censored at 100 years) (variable `age100`).

Plot an histogram of age at death. Note the spike at $age = 100$ due to the censored observations.

```
. hist age100, discrete name(p00, replace)
(start=0, width=1)

. graph export p00.png, replace
(file p00.png written in PNG format)
```



Assume that $f(age)$ follows a generalized extreme values distribution. Estimate the parameters η, μ and σ. Constrain η to be between 0 and 1. Constrain σ to be positive. Take into account right-censoring in age-at-death. You'll need to generate an event/censoring indicator variable, first.

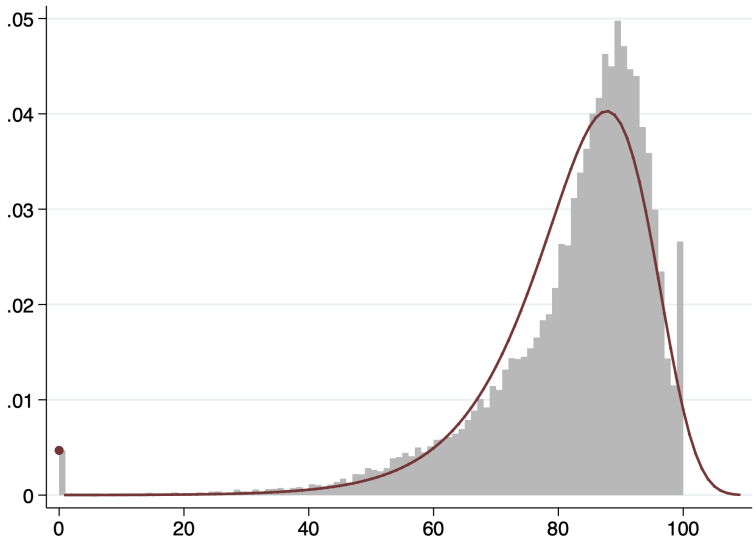
```
. gen d = (age < 100)
. local G = "exp((age100-{mu})/exp({theta1}))"
. local g = "exp((age100-{mu})/exp({theta1}))/exp({theta1})"
. local f = "exp(-`G´)*`g´"
. local S = "exp(-`G´)"
. local eta = "invlogit({theta2})"
. mlexp ((age<1)*ln(`eta´) + (age>=1)*ln((1-`eta´)*((`f´)^(d==1) * (`S´)^(d==0))))
initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -696974.21
rescale:       log likelihood = -371471.36
rescale eq:    log likelihood = -135999.31
Iteration 0:   log likelihood = -135999.31
Iteration 1:   log likelihood = -129310.49
Iteration 2:   log likelihood =  -127781.5
Iteration 3:   log likelihood = -127776.38
Iteration 4:   log likelihood = -127776.38
Maximum likelihood estimation
Log likelihood = -127776.38                     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	87.74612	.0521505	1682.56	0.000	87.6439	87.84833
/theta1	2.207031	.0045158	488.73	0.000	2.19818	2.215881

```
. mlci exp /theta1
 9.088688   95% CI:   9.0086, 9.169488
. mlci invlogit /theta2
  .004697   95% CI: .0040201, .0054873
```

Plot the estimated density $\hat{f}(age)$ over the sample histogram

```
. gen fhat_age3 = 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)
. tw (hist age100, width(1)) (scatter fhat_age3 age if age<1, sort msize(small) lc(maroon)) ///
> (line fhat_age3 age if age>=1, sort lc(maroon)), name(p20, replace) legend(off)
. graph export p20.png, replace
(file p20.png written in PNG format)
```



Exercise 4

We know that age at death was actually recorded in integer years. The exact age at death is therefore unknown to us. We only know it happened between $\lfloor age \rfloor$ and $\lfloor age + 1 \rfloor$ years.

Estimate the parameters μ and σ. Constrain σ to be positive. Take into account interval-censoring and right-censoring at 100 years.

```
. gen age100_plus_1 = age100 + 1
. local Sy = "exp(-exp((age100-{mu})/exp({theta1})))"
. local Su = "exp(-exp((age100_plus_1-{mu})/exp({theta1})))"
. local eta = "invlogit({theta2})"
. mlexp ((age<1)*ln(`eta´) + (age>=1)*ln((1-`eta´)*(`Sy´-`Su´)^(d==1) * (`Sy´)^(d==0)))
initial:       log likelihood =     -<inf>  (could not be evaluated)
feasible:      log likelihood = -696974.21
rescale:       log likelihood = -371547.51
rescale eq:    log likelihood = -136937.58
Iteration 0:   log likelihood = -136937.58
Iteration 1:   log likelihood = -128566.68
Iteration 2:   log likelihood = -127661.04
Iteration 3:   log likelihood = -127655.04
Iteration 4:   log likelihood = -127655.04
Maximum likelihood estimation
Log likelihood = -127655.04                     Number of obs     =     33,638
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/theta2	-5.356118	.0797434	-67.17	0.000	-5.512412	-5.199824
/mu	88.227	.051902	1699.88	0.000	88.12528	88.32873
/theta1	2.201819	.0045281	486.25	0.000	2.192944	2.210694

```
. mlci exp /theta1
 9.041443   95% CI: 8.961556, 9.122043
. mlci invlogit /theta2
  .004697   95% CI: .0040201, .0054873
```

Exercise 5

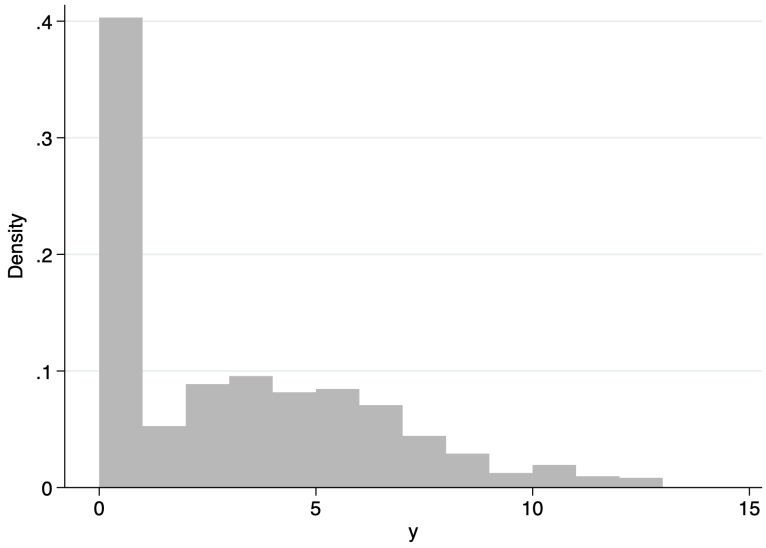
We measured how many times a random sample of 722 subjects were admitted to the hospital in 2016. Plot an histogram of the variable y.

```
. use https://raw.githubusercontent.com/anddis/fsm/master/data/lab3_2.dta, clear
.
```

```
. tab y
```

y	Freq.	Percent	Cum.
0	291	40.30	40.30
1	38	5.26	45.57
2	64	8.86	54.43
3	69	9.56	63.99
4	59	8.17	72.16
5	61	8.45	80.61
6	51	7.06	87.67
7	32	4.43	92.11
8	21	2.91	95.01
9	9	1.25	96.26
10	14	1.94	98.20
11	7	0.97	99.17
12	2	0.28	99.45
13	4	0.55	100.00
Total	722	100.00	

```
. hist y, width(1) name(p000, replace)
(bin=13, start=0, width=1)
. graph export p000.png, replace
(file p000.png written in PNG format)
```



Assume that $f(y)$ follows a Bernoulli-Poisson Mixture model. It's similar to the Bernoulli-Negative-Binomial Mixture model, but the density is:

$$(\beta + (1-\beta) * f_{\text{Poi}}(0))^{I(y=0)} \times ((1-\beta) * f_{\text{Poi}}(y))^{I(y>0)},$$

where $f_{\text{Poi}}(y)$ is the pmf of a Poisson distribution (`https://en.wikipedia.org/wiki/Poisson_distribution`) (see Stata's `poissonp()` function). Estimate the model's parameters. Remember to constrain the parameters to their parameter space.

```
. local beta = "invlogit({theta1})"
. local lambda = "exp({theta2})"
. local f = "(y==0)*ln(`beta´+(1-`beta´)*poissonp(`lambda´,0))+(y>0)*ln((1-`beta´)*poissonp(`lambda´,y))"
. mlexp (`f´)
initial:       log likelihood = -2898.2492
alternative:   log likelihood =  -2301.236
rescale:       log likelihood = -1893.7307
rescale eq:    log likelihood = -1701.2662
Iteration 0:   log likelihood = -1701.2662
Iteration 1:   log likelihood = -1533.4708
Iteration 2:   log likelihood = -1492.1875
Iteration 3:   log likelihood =  -1492.173
```

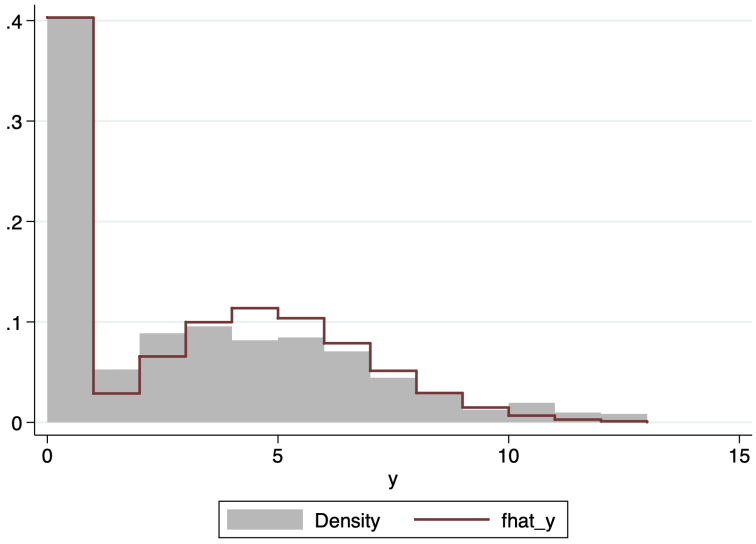
```
Iteration 4:   log likelihood =  -1492.173
Maximum likelihood estimation
Log likelihood =  -1492.173                     Number of obs     =        722
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/theta1	-.419094	.0771304	-5.43	0.000	-.5702669	-.2679211
/theta2	1.517249	.0230008	65.97	0.000	1.472168	1.56233

```
. mlci invlogit /theta1
 .3967336   95% CI: .3611752, .4334175
. mlci exp /theta2
 4.559664   95% CI: 4.358676, 4.769921
```

Plot the estimated density $\hat{f}(y)$ over the sample histogram

```
. gen fhat_y = exp((y==0)*ln(invlogit(_b[/theta1])+(1-invlogit(_b[/theta1]))*poissonp(exp(_b[/theta2]),0))+ ///
> (y>0)*ln((1-invlogit(_b[/theta1]))*poissonp(exp(_b[/theta2]),y)))
. tw (hist y, width(1)) (line fhat_y y, sort connect(J)), name(p3, replace)
. graph export p3.png, replace
(file p3.png written in PNG format)
```



Exercise 6

We consider Y the interval-censored version of a latent (unobserved) variable Y^*. Assume that Y^* follows a gamma distribution. Estimate its parameters.

```
. local beta = "invlogit({theta1})"
. local a = "exp({theta2})"
. local b = "exp({theta3})"
. local f = "(y==0)*ln(`beta´+(1-`beta´)*gammap(`a´,1/`b´))+(y>0)*ln((1-`beta´)*(gammap(`a´,(y+1)/`b´)-gammap(`a´,y/`b´)
> ))"
. mlexp (`f´)
initial:       log likelihood = -2541.5861
alternative:   log likelihood =  -1748.638
rescale:       log likelihood = -1726.2325
rescale eq:    log likelihood = -1499.4998
Iteration 0:   log likelihood = -1499.4998
Iteration 1:   log likelihood = -1485.5559
Iteration 2:   log likelihood = -1469.2134
Iteration 3:   log likelihood = -1468.9548
Iteration 4:   log likelihood = -1468.9545
Iteration 5:   log likelihood = -1468.9545
Maximum likelihood estimation
```

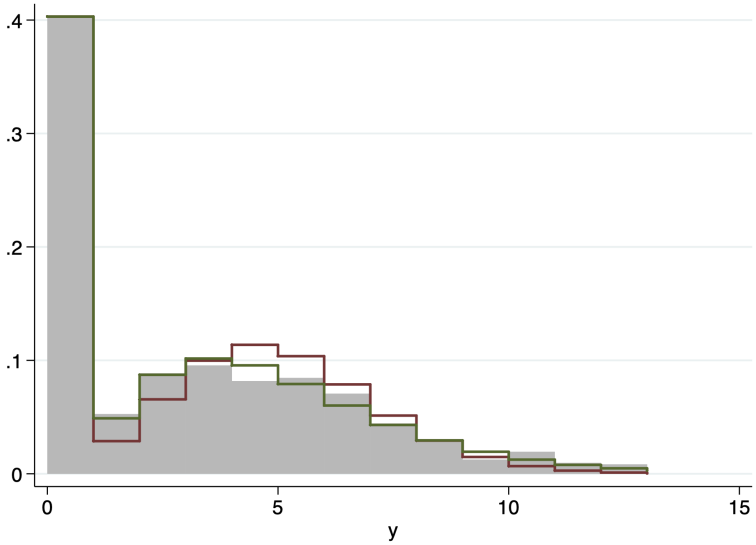
```
Log likelihood = -1468.9545                     Number of obs     =        722
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
/theta1	-.4280537	.0781964	-5.47	0.000	-.5813159	-.2747915
/theta2	1.254735	.0814365	15.41	0.000	1.095122	1.414347
/theta3	.363448	.0819923	4.43	0.000	.2027462	.5241499

```
. mlci invlogit /theta1
 .3945912   95% CI: .3586299, .4317312
. mlci exp /theta2
 3.506908   95% CI: 2.989548,   4.1138
. mlci exp /theta3
  1.43828   95% CI: 1.224762, 1.689022
```

Plot the estimated density $\hat{f}(y)$ over the sample histogram

```
. gen fhat_y2 = exp((y==0)*ln(invlogit(_b[/theta1])+(1-invlogit(_b[/theta1]))*gammap(exp(_b[/theta2]),1/exp(_b[/theta3])
> ))+ ///
> (y>0)*ln((1-invlogit(_b[/theta1]))*(gammap(exp(_b[/theta2]),(y+1)/exp(_b[/theta3]))-gammap(exp(_b[/theta2]),y/exp(_b[/
> theta3])))))
. tw (hist y, width(1)) (line fhat_y fhat_y2 y, sort connect(J J)), name(p4, replace) legend(off)
. graph export p4.png, replace
(file p4.png written in PNG format)
```



Which model seems to fit better the data? Tabulate the observed and model-based predicted proportions.

```
. gen N = _N
. bysort y: gen n = _N
. gen obs_p = n / N
. tabstat obs_p fhat_y fhat_y2, by(y) nototal format(%4.3f)
Summary statistics: mean
  by categories of: y
```

y	obs_p	fhat_y	fhat_y2
0	0.403	0.403	0.403
1	0.053	0.029	0.049
2	0.089	0.066	0.087
3	0.096	0.100	0.102
4	0.082	0.114	0.096
5	0.084	0.104	0.079
6	0.071	0.079	0.060
7	0.044	0.051	0.043
8	0.029	0.029	0.029
9	0.012	0.015	0.019
10	0.019	0.007	0.012

```
      11 |      0.010     0.003     0.008
      12 |      0.003     0.001     0.005
      13 |      0.006     0.000     0.003
```

Extra

Let's refit the model in Exercise 3, but this time we use the optimization function `optimize()` (which is the function that `mlexp` calls behind the curtains). `optimize()` is part of Mata, Stata's matrix programming language.

```
. use https://raw.githubusercontent.com/anddis/fsm/master/data/lab3_1.dta, clear
. gen d = (age < 100)
.
. mata
------------------------------------------------- mata (type end to exit) -------------------------------------------------
: mata clear
:
: X = st_data(., ("age", "age100", "d" ))
:
: void model3(todo, beta, ll, S, H) {
> mu = beta[1]
> sigma = exp(beta[2])
> eta = invlogit(beta[3])
>
> external X
> age = X[., 1]
> age100 = X[., 2]
> d = X[., 3]
>
> G = exp((age100 :- mu) :/ sigma)
> g = exp((age100 :- mu) :/ sigma) :/ sigma
> f = exp(-G) :* g
> S = exp(-G)
>
> ll = colsum((age:<1) :* ln(eta) :+ (age:>=1) :* ln((1:-eta) :* ((f):^(d:==1) :* (S):^(d:==0))))
> }
note: argument todo unused
note: argument H unused
:
: S = optimize_init()
: optimize_init_evaluator(S, &model3())
: optimize_init_params(S, (100, log(10), logit(.5)))
: b = optimize(S)
Iteration 0:   f(p) = -168058.62  (not concave)
Iteration 1:   f(p) = -144679.84  (not concave)
Iteration 2:   f(p) = -134653.55  (not concave)
Iteration 3:   f(p) = -130907.79
Iteration 4:   f(p) =  -127911.4
Iteration 5:   f(p) = -127779.65
Iteration 6:   f(p) = -127776.38
Iteration 7:   f(p) = -127776.38
: se = sqrt(diagonal(invsym(-optimize_result_Hessian(S))))
:
: b', se
                  1              2
  1 |    87.74611605    .0521504553
  2 |    2.207030536    .0045158228
  3 |   -5.356118786    .0797434067

: end
-------------------------------------------------------------------------------------------------------------------------------
```