# Stata Vignette for Finite-Tailed CDF-Quantile Distribution Models

Michael Smithson

## Introduction

This document presents demonstrations of the user-defined Stata function cdfquantreg01, which implements regression models for a finite-tailed cdf-quantile family of distributions with support on [0,1] described by Smithson and Shou (2022). The family is an extension of the cdf-quantile distributions first developed in Smithson and Shou (2017). All members of this new family have finite density at 0 and at 1, i.e., they are able to handle cases on the boundaries of the closed unit interval. Smithson and Shou (2022) provide the rationale, derivation, and assessment of this distribution family. The demonstrations herein are based on one of the examples of applications in that paper.

## About the Data

Yoon, Steiner, and Reinhardt (2003) conducted a study of time spent by patients admitted to the emergency department of the University of Alberta Hospital between midnight January 23 and midnight January 29, 1999, for five stages of ED assessment and treatment: Registration, triage assessment, nursing assessment, physician assessment, and disposition decision. While Yoon, et al. analyzed predictors of the total length of stay in the emergency ward, we will follow the analyses in Smithson and Broomell (2022), who examine the proportions of the patients' stays in the various stages.

Smithson and Broomell observed that the data include a substantial number of zeros (e.g., 696 out of 894 patients spending no time in the decision stage). They reduced the zeros by aggregating the decision and physician stages, and aggregating the registration and triage stages. The resulting composition had three parts: Registration-triage, nursing assessment, and physician-decision. We will use that composition here. The Appendix contains a list of the variables with brief descriptions of each of them.

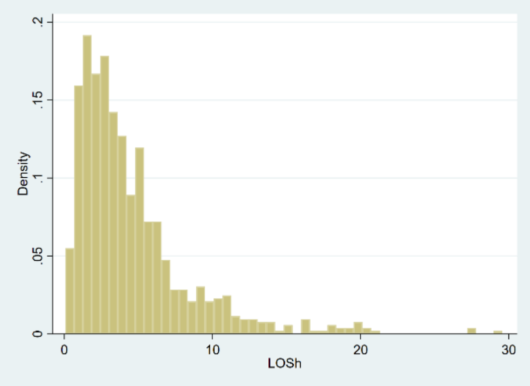
Our example focuses on the proportion of time spent in the registration-triage stage. Patients arriving by ambulance tended to have more life-threatening conditions than those arriving as ``walk-ins", so we expect to find that the ambulance-arrivals spend a smaller proportion of their time in this preliminary and mainly bureaucratic stage because serious cases need to be rushed into treatment. The more serious cases also typically required lengthy nursing and treatment times, so expect that longer length of stay will predict a lower proportion of time spent in the Registration-triage stage.

A quick examination of both relevant variables reveals that the log of the length of stay adequately corrects skew in that variable, and the split between ambulance-arrivals and walk-ins has adequate numbers of cases in both categories (ambulance = 0 are walk-ins and ambulance = 1 are ambulance-arrivals).

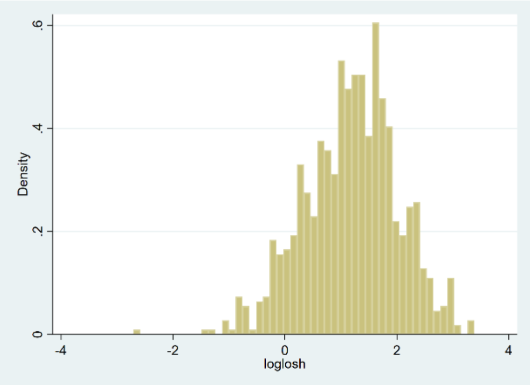
```
. tabulate ambulance

  Ambulance |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        683       76.40       76.40
          1 |        211       23.60      100.00
------------+-----------------------------------
      Total |        894      100.00
```

```
. histogram losh, bin(50)
```




```
. generate loglosh = ln(losh)
. histogram loglosh, bin(50)
```




## Two-parameter model

We begin with two-parameter distributions ($\theta$, the location and skew parameter, and $\sigma$, the dispersion parameter). We will use the Cauchit-ArcSinh outer-W distribution for this demonstration. Fitting a model using this distribution identifies significant effects of both ambulance arrival and log of length of stay in the expected directions for the $\theta$ submodel (eq1). Note that the coefficients are positive for Ambulance and loglosh, because $\theta$ tracks skew and therefore a positive coefficient predicts a decrease in the median proportion of time spent in the registration-triage stage.

```
. cdfquantreg01 pregptriage i.ambulance loglosh , cdf(cauchit) quantile(asinh)
pos(outer) func(w) twothree(2) zvarlist(i.ambulance loglosh)

initial:       log likelihood =  629.02215
rescale:       log likelihood =  629.02215
rescale eq:    log likelihood =  629.02215
Iteration 0:   log likelihood =  629.02215
Iteration 1:   log likelihood =  855.13443
Iteration 2:   log likelihood =  929.76623
Iteration 3:   log likelihood =  935.77038
Iteration 4:   log likelihood =   935.8292
Iteration 5:   log likelihood =  935.82938
Iteration 6:   log likelihood =  935.82938

                                                Number of obs     =        894
                                                Wald chi2(2)      =      35.08
Log likelihood =  935.82938                     Prob > chi2       =     0.0000
```

```
------------------------------------------------------------------------------
 pregptriage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
 1.ambulance |    1.44701   .4390561     3.30   0.001     .5864763    2.307544
     loglosh |    .602078   .1306889     4.61   0.000     .3459325    .8582236
       _cons |   1.362088   .1370775     9.94   0.000     1.093421    1.630755
-------------+----------------------------------------------------------------
eq2          |
 1.ambulance |  -.1100352   .4265914    -0.26   0.796     -.946139    .7260686
     loglosh |   .2427018   .1257848     1.93   0.054    -.0038319    .4892356
       _cons |  -.4175588   .1285837    -3.25   0.001    -.6695782   -.1655394
------------------------------------------------------------------------------
. estimates store A
```

There is a marginally non-significant effect of loglosh in the $\sigma$ (dispersion) submodel (eq2). Nonetheless, it turns out that a model without the dispersion submodel effects suffers a significant decline in goodness-of-fit. However, a model with interaction-effect terms does not significantly improve fit over the main-effects model (neither of these runs are shown here, but the reader may readily verify these claims by running the additional models). So our final model is one that includes main-effects terms for loglosh and ambulance in both submodels.

An examination of the parameter estimate correlation matrix reveals two correlations whose magnitudes are above 0.85, but the model appears stable and converges to the same solution from alternative starting-values.

```
. estat vce, correlation

Correlation matrix of coefficients of ml model
             | eq1                           | eq2
             |        1.                     |        1.
        e(V) | ambula~e   loglosh     _cons  | ambula~e   loglosh     _cons
-------------+-------------------------------+------------------------------
eq1          |                               |
 1.ambulance |   1.0000                      |
     loglosh |  -0.0893    1.0000            |
       _cons |  -0.1133   -0.6434    1.0000  |
-------------+-------------------------------+------------------------------
eq2          |                               |
 1.ambulance |  -0.9621    0.0688    0.1087  |   1.0000
     loglosh |   0.0597   -0.8908    0.5097  |  -0.0686    1.0000
       _cons |   0.1254    0.5267   -0.7868  |  -0.1334   -0.6391    1.0000
```

The margins command operates as usual in Stata, but the cdfquantreg01_mf program adds functionality by producing marginal predictions of quantiles across categories of categorical predictors. The example below shows this being done for the predicted median by setting the pctle option to 0.5. The predicted marginal median proportion of time spent in the registration-triage state for walk-ins is 0.125 whereas for ambulance-arrivals it is only 0.036.

```
. cdfquantreg01_mf ambulance, pctle(0.5)
Predictive margins                              Number of obs     =        894
Model VCE    : OIM

Expression   : Linear prediction, predict(equation(#1))
------------------------------------------------------------------------------
             |            Delta-method
```

```
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ambulance |
          0  |   2.074836   .1242395    16.70   0.000     1.831331    2.318341
          1  |   3.521847   .4271211     8.25   0.000     2.684705    4.358988
------------------------------------------------------------------------------
(results modresults are active now)

Predictive margins                              Number of obs     =        894
Model VCE    : OIM

Expression   : Linear prediction, predict(equation(#2))
------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ambulance |
          0  |  -.1302452   .1193036    -1.09   0.275     -.364076    .1035856
          1  |  -.2402805   .4157706    -0.58   0.563    -1.055176    .5746149
------------------------------------------------------------------------------
(results modresults are active now)

ambulance
.5 quantile   factor level
--------------------------
.12464288     0bn.ambulance
.03619026     1.ambulance
```

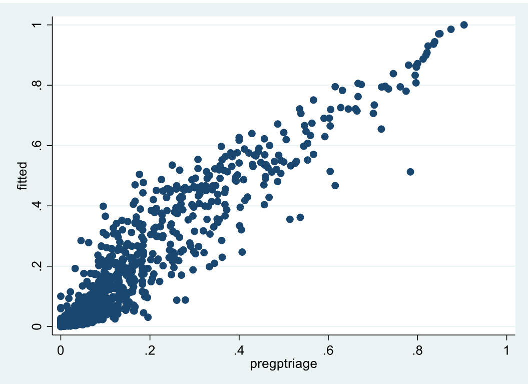
The program cdfquantreg01_p provides post-estimation within- and out-of-sample predictions. The predict command operates in a somewhat un-Stata-like fashion because it adds data to memory. However, this has been permitted in order to allow users to estimate different quantiles. To begin, we can simply obtain fitted values by using the predict command with just the qtile option. As show below, the fitted values' rank-order correlation with the dependent variable is quite high and the scatterplot suggests that the residuals are well-behaved.

```
. predict newvar, qtile
. spearman pregptriage fitted

 Number of obs =     894
Spearman's rho =       0.9344

Test of Ho: pregptriage and fitted are independent
    Prob > |t| =       0.0000

. twoway (scatter fitted pregptriage)
```



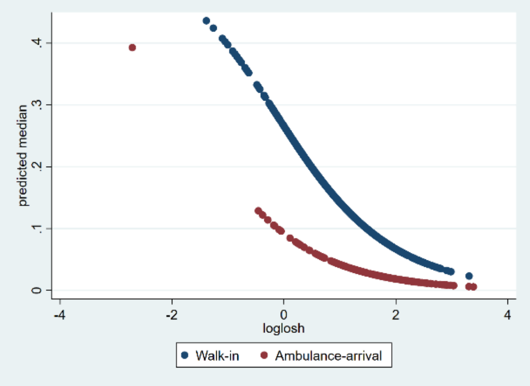
```
. drop xb xd fitted residuals
```

An alternative usage of the predict command with the pctle option, which specifies the quantile being predicted. The two graphs below shows the predicted median and predicted 75th percentile as a function of loglosh, tracked for the walk-ins versus the ambulance-arrivals. This graph effectively displays both main-effects from length of stay and mode of arrival at the emergency ward.

```
. predict newvar, qtile pctle(0.5)
. separate fitted, by(ambulance)

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
fitted0         float   %9.0g                 fitted, ambulance == 0
fitted1         float   %9.0g                 fitted, ambulance == 1

. twoway (scatter fitted0 loglosh, sort) (scatter fitted1 loglosh, sort),
ytitle(predicted median) legend(order(1 "Walk-in" 2 "Ambulance-arrival"))
```

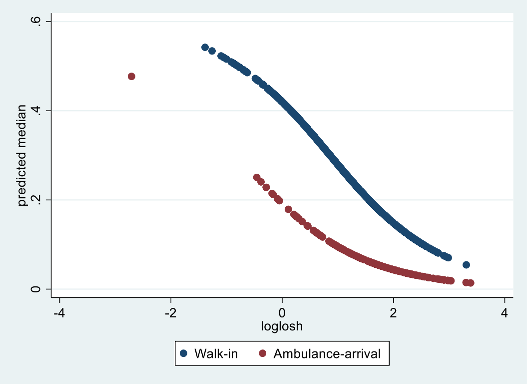



```
. drop xb xd fitted fitted0 fitted1

. predict newvar, qtile pctle(0.75)
. separate fitted, by(ambulance)

              storage   display    value
variable name   type    format     label      variable label
----------------------------------------------------------------------------
fitted0         float   %9.0g                 fitted, ambulance == 0
fitted1         float   %9.0g                 fitted, ambulance == 1

. twoway (scatter fitted0 loglosh, sort) (scatter fitted1 loglosh, sort),
ytitle(predicted median) leg
> end(order(1 "Walk-in" 2 "Ambulance-arrival"))
```



### Three-parameter model

The output shown below is from a 3-parameter Cauchit-ArcSinh outer-W model. The additional parameter is $\mu$, the location parameter. The $\mu$ submodel coefficients are in eq1, the $\theta$ submodel coefficients are in eq2, and the $\sigma$ submodel coefficients are in eq3.

```
. cdfquantreg01 pregptriage i.ambulance loglosh , cdf(cauchit) quantile(asinh)
pos(outer) func(w) twothree(3) zvarlist(i.ambulance loglosh) wvarlist(i.ambulance
loglosh)

initial:       log likelihood =  648.96614
rescale:       log likelihood =  648.96614
rescale eq:    log likelihood =  648.96614
Iteration 0:   log likelihood =  648.96614  (not concave)
Iteration 1:   log likelihood =  853.51944
Iteration 2:   log likelihood =  926.83997
Iteration 3:   log likelihood =  931.32839
Iteration 4:   log likelihood =  938.95708
Iteration 5:   log likelihood =  938.99915
Iteration 6:   log likelihood =  938.99918

                                                Number of obs     =        894
                                                Wald chi2(2)      =       1.02
Log likelihood =  938.99918                     Prob > chi2       =     0.6010

------------------------------------------------------------------------------
 pregptriage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
eq1          |
 1.ambulance |   .2298821    .474581     0.48   0.628    -.7002796    1.160044
     loglosh |  -.1408897   .1492391    -0.94   0.345    -.4333929    .1516135
       _cons |  -.2159295   .1370239    -1.58   0.115    -.4844914    .0526325
-------------+----------------------------------------------------------------
eq2          |
 1.ambulance |   1.671946   .4823149     3.47   0.001     .7266265    2.617266
     loglosh |   .5581998   .1487688     3.75   0.000     .2666183    .8497812
       _cons |   1.135094    .182286     6.23   0.000     .7778199    1.492368
-------------+----------------------------------------------------------------
eq3          |
 1.ambulance |  -.2546381   .4438452    -0.57   0.566    -1.124559    .6152824
     loglosh |    .316594    .123377     2.57   0.010     .0747796    .5584084
       _cons |  -.3338415   .1288476    -2.59   0.010    -.5863781   -.0813048
------------------------------------------------------------------------------
```

We can see that the ambulance and loglosh effects in the $\theta$ and $\sigma$ submodels are similar to those in the 2-parameter model, while the $\mu$ submodel has no significant effects. Is this model any better than the 2-parameter model? Of course we cannot compare their log-likelihoods because they are not nested models, but we may compare their AIC or BIC values. The information criteria results from the 2-parameter model are shown below.

```
. estat ic

Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
  modresults |        894         .   935.8294       6   -1859.659  -1830.885
-----------------------------------------------------------------------------
```

The 3-parameter model AIC is very similar, whereas the BIC is decidedly greater, suggesting that the 2-parameter model should be preferred on grounds of parsimony.

```
. estat ic

Akaike's information criterion and Bayesian information criterion
-----------------------------------------------------------------------------
       Model |        Obs  ll(null)  ll(model)      df         AIC        BIC
-------------+---------------------------------------------------------------
           . |        894         .   938.9992       9   -1859.998  -1816.837
-----------------------------------------------------------------------------
```

## References

Smithson, M. & Broomell, S.B. (online 31/01/2022). Compositional Data Analysis Tutorial. *Psychological Methods*. http://dx.doi.org/10.1037/met0000464

Smithson, M. & Shou, Y. (2017). CDF-quantile distributions for modeling random variables on the unit interval. *British Journal of Mathematical and Statistical Psychology*, *70*(3), 412-438. doi: 10.1111/bmsp.12091

Smithson, M. & Shou, Y. (accepted 18/11/22). Flexible cdf-quantile distributions on the closed unit interval, with software and applications. *Communications in Statistics – Theory and Methods*.

Yoon, P., Steiner, I. & Reinhardt, G. (2003). Analysis of factors influencing length of stay in the emergency department. *Canadian Journal of Emergency Medicine*, 5, 155–161.

## Appendix: Codebook for the Data

| variable | contents |
|---|---|
| id | case identification |
| Day | day of the week ( 0 = Sunday) |
| Ambulance | 0 = walk-in; 1 = ambulance-arrival |
| Triage | triage level |
| Triage1 | 1 = triage level 1 |
| Triage2 | 1 = triage level 2 |
| Triage3 | 1 = triage level 3 |
| Triage4 | 1 = triage level 4 |
| Triage5 | 1 = triage level 5 |
| Lab | 1 = laboratory test(s) conducted |
| Xray | 1 = x-ray conducted |
| Other | 1 = other intervention |
| LOS | length of stay in minutes |
| LOSh | length of stay in hours |
| preg | proportion of time in registration stage |
| ptriage | proportion of time in triage stage |
| pnurse | proportion of time in nursing care stage |
| pphysician | proportion of time in consultation with physician(s) |
| pdecis | proportion of time in decisional stage |
| pregptriage | preg + ptriage |
| pphysdecis | pphysician + pdecis |
| prnurse | pnurse/(pnurse + pregptriage) |
| prphysdec | pphysdecis /(pphysdecis + pregptriage) |