The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2011)
11, Number 3, pp. 327–344

Logistic quantile regression in Stata

Nicola Orsini
Unit of Biostatistics
and
Unit of Nutritional Epidemiology
Institute of Environmental Medicine, Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

Matteo Bottai
Division of Biostatistics
University of South Carolina
Columbia, SC
and
Unit of Biostatistics
Institute of Environmental Medicine, Karolinska Institutet
Stockholm, Sweden
matteo.bottai@ki.se

Abstract. We present a set of Stata commands for the estimation, prediction, and graphical representation of logistic quantile regression described by Bottai, Cai, and McKeown (2010, *Statistics in Medicine* 29: 309–317). Logistic quantile regression models the quantiles of outcome variables that take on values within a bounded, known interval, such as proportions (or percentages) within 0 and 1, school grades between 0 and 100 points, and visual analog scales between 0 and 10 cm. We describe the syntax of the new commands and illustrate their use with data from a large cohort of Swedish men on lower urinary tract symptoms measured on the international prostate symptom score, a widely accepted score bounded between 0 and 35.

Keywords: st0231, lqreg, lqregpred, lqregplot, logistic quantile regression, robust regression, bounded outcomes

1 Introduction

Some variables take on values within a bounded, known interval. Examples of these intervals include proportions (or percentages) within 0 and 1, school grades between 0 and 100 points, visual analog scales between 0 and 10 cm, quality of life index between 1 and 10, and international prostate symptom scores (IPSS) between 0 and 35.

In this article, we describe a set of Stata commands for the estimation of logistic quantile regression, a method described by Bottai, Cai, and McKeown (2010) modeling quantiles (for example, median) of bounded outcomes. Koenker and Bassett (1978) introduced quantile regression over three decades ago, and the popularity of this method has grown ever since; Koenker (2005) gives an extensive description of quantile regression. The traditional linear regression models the conditional expectation of an outcome variable given a set of covariates. Quantile regression models its conditional quantile instead and can be estimated with the Stata commands `qreg`, `iqreg`, `sqreg`, and `bsqreg`. Quantile regression is a powerful tool for comparing, more thoroughly than the mean alone, various aspects (location, scale, and shape) of any kind of distribution of the outcome across different covariate patterns.

When research interest lies in the mean of bounded response variables, beta regression and fractional logit models are useful methods. Beta regression (Smithson and Verkuilen 2006) is implemented in Stata as the `betafit` package, available from the SSC archive (Buis, Cox, and Jenkins 2011). Beta regression assumes that the regression residual follows a beta distribution and can be used to investigate how the conditional mean and standard deviation depend on explanatory variables (Smithson and Verkuilen 2006). The fractional logit model (Papke and Wooldridge 1996) can be estimated using Stata's `glm` command (see [R] **glm**) (Baum 2008), and it is fully robust and relatively efficient under the generalized linear model assumption.

2 Logistic quantile regression

In this section, we follow the description provided by Bottai, Cai, and McKeown (2010). Suppose we have a sample of n observations on some continuous outcome y_i, $i = 1, \ldots, n$, and an s-dimensional vector of covariates $x_i = \{x_{1,i}, \ldots, x_{s,i}\}^T$. The quantile regression model is

$$y_i = x_i^T \beta_p + \varepsilon_i$$

where the $\beta_p = \{\beta_{p1}, \ldots, \beta_{ps}\}^T$ indicate the unknown regression parameters. For any given $p \in (0,1)$, we assume that $P(\varepsilon_i \leq 0 | x_i) = p$ or, equivalently, that $P(y_i \leq x_i^T \beta_p | x_i) = p$. The p quantile of the conditional distribution of y_i given x_i is defined as

$$Q_y(p) = x_i^T \beta_p \tag{1}$$

If $p = 0.5$, then $Q_y(0.5)$ is the conditional median, the value that splits the conditional distribution of the response variable into two parts with equal probability. No other assumptions are required on the distribution of the regression residual ε_i.

Quantile regression has several desirable properties. For example, its estimation, contrary to the regression on the mean, is equivariant to monotonic transformations of the outcome; that is, $Q_{h(y)}(p) = h\{Q_y(p)\}$ for any nondecreasing function h, while $E\{h(y)\} \neq h\{E(y)\}$ where $E(y)$ denotes the mean of y. Bottai, Cai, and McKeown (2010) exploit this property and define the logistic quantile regression to model continuous outcomes that are bounded within a known interval as

$$y_i \in (y_{\min}, y_{\max}) \tag{2}$$

where $y_{\min}$ and $y_{\max}$ do not denote the smallest and largest observed sample values but the limits of the feasible interval of the outcome variable.

To accommodate the constraint (2), we assume that for any quantile p there exists a fixed set of parameters β_p and a known nondecreasing function h from the interval $(y_{\min}, y_{\max})$ to the real line (a function often referred to as *link*), such that

$$h\{Q_y(p)\} = x_i^T \beta_p$$

Because a continuous outcome bounded within the unit interval resembles a probability, or a propensity, among a variety of suitable choices for the link function h, Bottai, Cai, and McKeown (2010) opt for the logistic transformation

$$h(y_i) = \log\left(\frac{y_i - y_{\min}}{y_{\max} - y_i}\right) = \mathrm{logit}(y_i)$$

The inverse transform is

$$Q_y(p) = \frac{\exp{(x_i^T \beta_p)} y_{\max} + y_{\min}}{1 + \exp{(x_i^T \beta_p)}}$$

Regression coefficients can be estimated using quantile regression by regressing the transformed outcome $h(y_i)$ on x using (1):

$$Q_{h(y_i)}(p) = Q_{\mathrm{logit}(y_i)}(p) = x_i^T \beta_p$$

This is analogous to logistic regression, which applies the same transform to model a probability. Transforming has an identical goal in both models: to facilitate modeling while constraining inference about the outcome within the feasible range, $(0, 1)$ for a probability and $(y_{\min}, y_{\max})$ for the continuous bounded outcome. When the sample data take on the lower limit, $y_{\min}$, or the upper limit, $y_{\max}$, a small quantity can be added to $y_{\max}$ and subtracted from $y_{\min}$.

Regarding inference about β_p, it has been shown in quantile regression that bootstrap standard errors outperform asymptotic standard errors (Rogers 1992; Gould 1992; Buchinsky 1995). Therefore, the present Stata commands use bootstrap as the default method for estimating standard errors. When multiple quantiles are estimated, for each bootstrap sample, regression coefficients are estimated for each quantile of interest. Thus one can also test and construct confidence intervals comparing regression coefficients across quantiles of the response.

3 Stata syntax

Inference about the logistic quantile regression model above can be carried out with the new Stata commands `lqreg`, `lqregpred`, and `lqregplot`. We describe their syntax in this section and illustrate their use in section 4.

3.1 lqreg

`lqreg` estimates logistic quantile regression for bounded outcomes. It produces the same coefficients as `qreg` or `sqreg` (see [R] **qreg**) for each quantile of a logistic transformation of *depvar*. `lqreg` estimates the variance–covariance matrix of the coefficients by using either bootstrap (default) or closed formulas.

`lqreg` *depvar* [*indepvars*] [*if*] [*in*] [`,` `quantiles(`*numlist*`)` `reps(`#`)` `seed(`#`)` `ase` `cluster(`*varlist*`)` `ymin(`#`)` `ymax(`#`)` `generate(`*varname*`)` `level(`#`)` `nodots`]

After `lqreg` estimation, `qreg postestimation` (see [R] **qreg postestimation**) is available. In addition, `lqreg` has two specific postestimation commands described in sections 3.2 and 3.3.

Options

`quantiles(`*numlist*`)` specifies the quantiles to be estimated and should contain numbers between 0 and 1, exclusive. Numbers greater than 1 are interpreted as percentages. The default, `quantiles(0.5)`, corresponds to the median.

`reps(`#`)` specifies the number of bootstrap replications to be used to obtain an estimate of the variance–covariance matrix of the estimators (standard errors). For example, the default, `reps(100)`, would perform 100 bootstrap replications.

`seed(`#`)` sets the random-number seed. Because bootstrapping is a random process, this option is important to reproduce the results (see [R] **set seed**).

`ase` specifies the asymptotic standard errors as implemented in `qreg` (see [R] **qreg**).

`cluster(`*varlist*`)` specifies the variables identifying resampling clusters. If `cluster()` is specified, the sample drawn during each replication is a bootstrap sample of clusters. `cluster()` works only if `reps()` is also specified.

`ymin(`#`)` sets the lower bound of *depvar* to be used in the logistic transformation. The default is the minimum value of *depvar* minus half of the minimal increment of *depvar*.

`ymax(`#`)` sets the upper bound of *depvar* to be used in the logistic transformation. The default is the maximum value of *depvar* plus half of the minimal increment of *depvar*.

`generate(`*varname*`)` creates a new variable containing the logistic transformation of *depvar*.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`.

`nodots` suppresses display of the replication dots when using bootstrap.

Saved results

`lqreg` saves the following in `e()`:

Scalars			
`e(N)`	number of observations	`e(n_q)`	number of quantiles requested
`e(df_r)`	residual degrees of freedom	`e(q#)`	the quantiles requested
`e(ymin)`	lower bound for *depvar*	`e(rank)`	rank of `e(V)`
`e(ymax)`	upper bound for *depvar*	`e(convcode)`	`0` if converged; otherwise, return code for why nonconvergence
Macros			
`e(cmd)`	`lqreg`	`e(eqnames)`	names of equations
`e(cmdline)`	command as typed	`e(properties)`	`b V`
`e(depvar)`	name of dependent variable	`e(predict)`	program used to implement `predict`
Matrices			
`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix of the estimators
Functions			
`e(sample)`	marks estimation sample		

3.2 lqregplot

The postestimation command `lqregplot` plots any regression coefficient with confidence bands against a dense set of quantiles.

`lqregplot` *varname* [`,` `quantiles(`*numlist*`)` `level(`*#*`)` `reps(`*#*`)` `ase` `seed(`*#*`)` `nosmooth` `loptions(`*string*`)` `generate(`*varname1 varname2 varname3 varname4*`)`]

Options

`quantiles(`*numlist*`)` specifies the quantiles to be estimated and should contain numbers between 0 and 1, exclusive. Numbers greater than 1 are interpreted as percentages. The default, `quantiles(0.5)`, corresponds to the median.

`level(`*#*`)` specifies the confidence level, as a percentage, for confidence intervals. The default is `level(95)` or as set by `set level`.

`reps(`*#*`)` specifies the number of bootstrap replications to be used to obtain an estimate of the variance–covariance matrix of the estimators.

`ase` specifies the asymptotic standard errors as implemented in `qreg` (see [R] **qreg**).

`seed(`*#*`)` sets the random-number seed.

`nosmooth` specifies not to smooth the plot of the regression coefficients. Smoothing is the default.

`loptions(`*string*`)` specifies `lowess` (see [R] **lowess**) options (for instance, `bwidth()` or `mean`) when smoothing the regression coefficients versus the set of specified quantiles.

`generate(`*varname1 varname2 varname3 varname4*`)` saves the variables required to reproduce the plot: quantile, point estimate, lower bound, and upper bound of the regression coefficient, to be saved in *varname1*, *varname2*, *varname3*, and *varname4*, respectively. This option is useful if one wants to customize the plot using `graph twoway` (see [G-2] **graph twoway**).

3.3 lqregpred

The postestimation command `lqregpred` creates new variables containing the untransformed predicted quantiles of *depvar* and optionally plots them versus a covariate of interest.

`lqregpred` *stubname* [*if*] [*in*] [`,` `for(`*varlist*`)` `at(`*var* `=` *#* [*var* `=` *#* [...]]`)`
`plotvs(`*varname*`)`]

Options

`for(`*varlist*`)` specifies the covariate, modeled using one or more transformations `for()`, for which to compute the (partial) predicted values, evaluating the remaining covariates at the value of 0 unless specified differently with the `at()` option.

`at(`*var* `=` *#* [*var* `=` *#* [...]]`)` specifies the values of the covariates not specified in the `for()` option. `at()` works only if the `for()` option is also specified.

`plotvs(`*varname*`)` creates plots of the untransformed predicted quantiles versus a quantitative covariate.

4 Example

Lower urinary tract symptoms are a common problem in aging men. The severity of these symptoms is frequently measured by the IPSS, whose values are bounded between 0 and 35. Severity of the symptoms is generally interpreted as follows: 0–7 is mild, 8–19 is moderate, and 20–35 is severe. The distribution of the values of IPSS is often markedly skewed to the right, and in most studies the score is dichotomized as mild or no symptoms (0–7) and moderate to severe symptoms (8–35). The binary outcome is then modeled with logistic regression. The cutoff value 7 is clearly arbitrary, albeit widely accepted, and analyzing the resulting binary outcome may be inefficient and potentially misleading.

In this section, we analyze IPSS as a bounded score with logistic quantile regression on a sample of 30,377 men in central Sweden aged 45–79 years; this sample is described in more detail by Orsini et al. (2006). The main covariate of interest is a total physical activity score (variable `tpa`), a combination of intensity and duration for a combination of daily activities, expressed in metabolic equivalents (MET; kcal/kg/hour).

The sample distribution of response variable IPSS is highly skewed with a concentration of observed values in a narrow range close to zero (figure 1). About 50% of the subjects had IPSS less than 3. Figure 2 shows the distribution of IPSS by categories of physical activity. The median and interquartile ranges of IPSS decrease with increasing levels of physical activity. Comparing the lowest (≤ 30) to the largest (> 54) category of physical activity, the 0.25 quantile of IPSS decreases by 1 unit and the median decreases by 4 units.

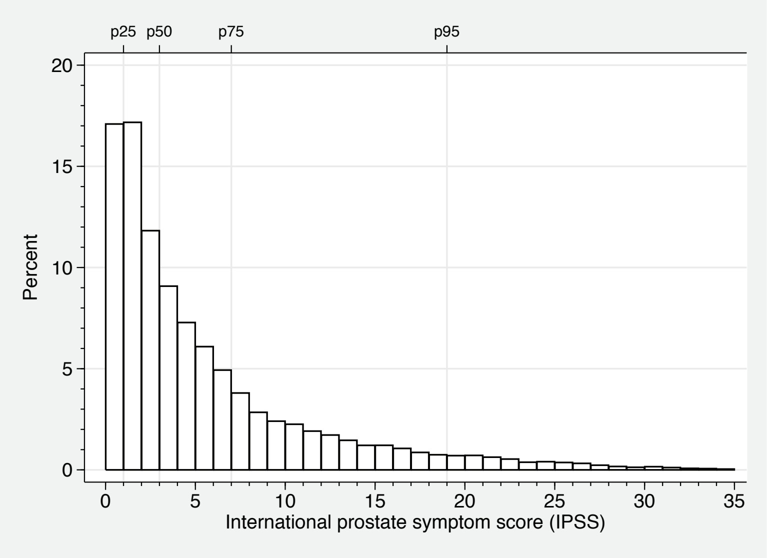

Figure 1. Distribution of the IPSS; the top axis shows the 25th, 50th, 75th, and 95th percentiles

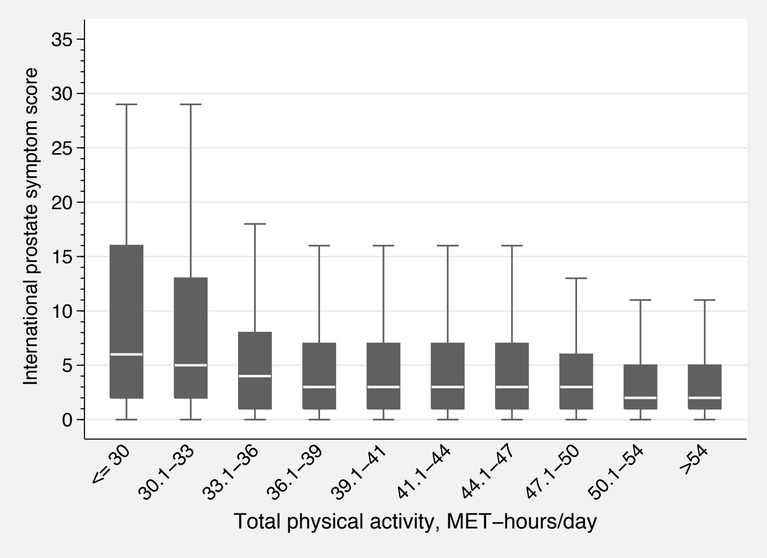

Figure 2. Plot of various observed quantiles of IPSS according to categories of physical activity intervals

Inference about the conditional mean IPSS through ordinary least-squares regression would have several limitations because the normality and homoskedasticity (constant variance) assumptions of the conditional mean of the outcome are clearly untenable. Two useful parametric regression methods to model the conditional mean IPSS are a beta regression model (`betafit`) and a fractional logit model (`glm`). Both commands require prior transformation of the original bounded outcome to a unit interval. Beta regression requires an additional transformation to avoid the boundaries (exact 0s and 1s), as suggested by Smithson and Verkuilen (2006). Below is the syntax for both models modeling physical activity with indicator variables.

```
. use http://nicolaorsini.altervista.org/data/pa_luts
(Data source: Orsini et al. The Journal of Urology. 2006. 176(6):2546-50)
// Transform the [0,35] bounded outcome in [0,1] bounded outcome
. generate ipssb = (ipss-0)/(35-0)
// Transform the [0,35] bounded outcome in (0,1) bounded outcome
. generate ipssc = [(ipss-0)/(35-0)*(30377-1)+.5]/30377
// Beta regression
. xi: betafit ipssc, mu(i.tpac)
  (output omitted)
// Fractional logit regression
. xi: glm ipssb i.tpac, family(binomial) link(logit) vce(robust)
  (output omitted)
```

The estimates of both models indicate that the mean of the transformed IPSS significantly decreases with increasing values of physical activity. Interpretation of both regression coefficients and predicted values from these models requires transforming back to the original scale of the outcome. The conditional mean alone cannot fully describe the shift in location and shrinkage in spread of the distribution of a bounded outcome evident in figure 2. This can be described directly and with no assumptions about the residuals by modeling quantiles of the bounded outcome with logistic quantile regression.

4.1 Categorical predictor

We now present how to estimate quantiles of IPSS as a function of covariates using the proposed `lqreg` command. The minimum and maximum values used for the logistic transformation of the response can be set directly using the `ymin(#)` and `ymax(#)` options. If `ymin(#)` and `ymax(#)` are not specified, the `lqreg` command computes the logistic transformation of the bounded response variable subtracting a quantity (half of the minimal increment of the outcome) from its minimum value and adding the same quantity to its maximum value to ensure that the transform is defined for all values of the response. Confidence intervals are obtained by using 100 bootstrap samples, and the seed of the pseudorandom-number generator can be specified to reproduce the confidence intervals.

Because IPSS ranges from 0 to 35, half of the smallest increment is $1/2 = 0.5$. Therefore, $\texttt{ymin()} = 0 - 0.5 = -0.5$ and $\texttt{ymax()} = 35 + 0.5 = 35.5$. The default logistic transformation of IPSS is

$$y = \log\{(\texttt{ipss} + 0.5)/(35.5 - \texttt{ipss})\}$$

We start by modeling the median ($p = 0.5$) of the logistic transformation of IPSS (y) as a function of physical activity categorized in 10 intervals (`tpac`), using the lowest category as referent (≤ 30 MET-hours/day). The equation of the model is

$$\begin{aligned} Q_y(p) = {} & \beta_{p0} + \beta_{p1}\texttt{\_Itpac\_2} + \beta_{p2}\texttt{\_Itpac\_3} + \beta_{p3}\texttt{\_Itpac\_4} + \beta_{p4}\texttt{\_Itpac\_5} + \\ & + \beta_{p5}\texttt{\_Itpac\_6} + \beta_{p6}\texttt{\_Itpac\_7} + \beta_{p7}\texttt{\_Itpac\_8} + \beta_{p8}\texttt{\_Itpac\_9} \\ & + \beta_{p9}\texttt{\_Itpac\_10} \end{aligned}$$

```
. xi: lqreg ipss i.tpac, nodots seed(123)
i.tpac            _Itpac_1-10        (naturally coded; _Itpac_1 omitted)

Logistic Quantile Regression                          Number of obs =     30377
Bounded Outcome: ipss(-.5, 35.5)                      Bootstrap(100) SEs

------------------------------------------------------------------------------
             |              Bootstrap
        ipss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q50          |
     _Itpac_2 |  -.3779775    .535728    -0.71   0.480    -1.428027    .6720719
     _Itpac_3 |   -.610909    .551716    -1.11   0.268    -1.692296    .4704776
     _Itpac_4 |  -.8934761   .5248025    -1.70   0.089    -1.922111    .1351589
     _Itpac_5 |  -.8934761   .5248025    -1.70   0.089    -1.922111    .1351589
     _Itpac_6 |  -.8934761   .5248025    -1.70   0.089    -1.922111    .1351589
     _Itpac_7 |  -.8934761    .531642    -1.68   0.093    -1.935517    .1485645
     _Itpac_8 |  -.8934761   .5432536    -1.64   0.100    -1.958276    .1713238
     _Itpac_9 |  -1.260254   .5248025    -2.40   0.016    -2.288889   -.2316185
    _Itpac_10 |  -1.260254   .5248025    -2.40   0.016    -2.288889   -.2316185
        _cons |  -1.335001   .5248025    -2.54   0.011    -2.363636    -.306366
------------------------------------------------------------------------------
```

After the `lqreg` command, it is possible to use all the postestimation commands available for quantile regression. For instance, the `predict` command provides the predicted median of the logistic transformation of IPSS. The untransformed predicted median IPSS can be obtained easily with the postestimation command `lqregpred`.

```
. lqregpred crude
. table tpac, c(mean crude50) f(%2.0f)

--------------------------
tpa       |
intervals | mean(crude50)
----------+---------------
    <= 30 |             7
  30.1-33 |             5
  33.1-36 |             4
  36.1-39 |             3
  39.1-41 |             3
  41.1-44 |             3
  44.1-47 |             3
  47.1-50 |             3
  50.1-54 |             2
      >54 |             2
--------------------------
```

The regression coefficients are differences in medians of logit transform of IPSS, between each physical activity level and the referent. The median IPSS decreases significantly with increasing activity levels. However, men reporting different physical activity levels may be different with respect to sociodemographic, biological, anthropometrical, health, and other lifestyle factors. Age is the strongest predictor of urinary problems. Urinary problems increase with age and occur in most elderly men, while total physical activity decreases with age. Therefore, the estimated decreasing trend in the median IPSS in subpopulations of men reporting higher physical activity levels might be partially or totally explained by differences in the distribution of age.

We include age, centered on the sample mean of 59 years, in the logistic quantile regression model.

```
. generate cage = age-59

. xi: lqreg ipss i.tpac cage, nodots seed(123)
i.tpac            _Itpac_1-10         (naturally coded; _Itpac_1 omitted)

Logistic Quantile Regression                        Number of obs =     30377
Bounded Outcome: ipss(-.5, 35.5)                    Bootstrap(100) SEs

------------------------------------------------------------------------------
             |              Bootstrap
        ipss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q50          |
    _Itpac_2 |  -.1061739   .4390046    -0.24   0.809    -.9666414    .7542936
    _Itpac_3 |  -.4440855   .4310596    -1.03   0.303     -1.28898    .4008094
    _Itpac_4 |  -.5086374     .42238    -1.20   0.229     -1.33652    .3192452
    _Itpac_5 |  -.5839941   .4212758    -1.39   0.166    -1.409712    .2417242
    _Itpac_6 |  -.6493446    .418746    -1.55   0.121    -1.470104    .1714152
    _Itpac_7 |  -.7247013   .4247918    -1.71   0.088    -1.557311    .1079085
    _Itpac_8 |   -.772386   .4293875    -1.80   0.072    -1.614004    .0692316
    _Itpac_9 |  -.8754148   .4365654    -2.01   0.045    -1.731101   -.0197283
   _Itpac_10 |  -.8289242    .442902    -1.87   0.061    -1.697031    .0391823
        cage |   .0376784   .0011155    33.78   0.000      .035492    .0398647
       _cons |  -1.644483   .4237833    -3.88   0.000    -2.475116     -.81385
------------------------------------------------------------------------------
```

As expected, adjustment for age attenuates the magnitude of the association between physical activity and IPSS. We see a statistically significant decreasing trend in age-adjusted median IPSSs with increasing physical activity levels.

Postestimation commands for calculating *p*-values and predictions are the same as those for other Stata regression commands. For example, to obtain the *p*-value for the null hypothesis that there is no association between physical activity and the median IPSS, we test the joint null hypothesis that all the regression coefficients of the indicator variables used to model physical activity are simultaneously equal to zero.

```
. testparm _I*
 ( 1)  [q50]_Itpac_2 = 0
 ( 2)  [q50]_Itpac_3 = 0
 ( 3)  [q50]_Itpac_4 = 0
 ( 4)  [q50]_Itpac_5 = 0
 ( 5)  [q50]_Itpac_6 = 0
 ( 6)  [q50]_Itpac_7 = 0
 ( 7)  [q50]_Itpac_8 = 0
 ( 8)  [q50]_Itpac_9 = 0
 ( 9)  [q50]_Itpac_10 = 0
       F(  9, 30366) =   16.25
            Prob > F =    0.0000
```

The small *p*-value indicates a statistically significant association between physical activity and median IPSS.

4.2 Simultaneous logistic quantile regression

So far, we have examined the association between physical activity and one quantile (median) of IPSS. We now consider four quantiles (0.25, 0.50, 0.75, 0.95) to assess whether the association between physical activity and IPSS varies according to the quantile of IPSS. We can get the age-adjusted predicted quantiles with the `lqregpred` command:

```
. quietly xi: lqreg ipss i.tpac cage, quantiles(25 50 75 95) seed(123) nodots
. lqregpred vpadj, for(_Itpac_2- _Itpac_10) at(cage=0)
```

Figure 3 shows that the predicted 0.25, 0.50, 0.75, and 0.95 age-adjusted quantiles are decreasing with increasing physical activity levels.

```
. twoway
> (line vpadj25 tpa, lp(dash) lc(black) sort c(J))
> (line vpadj50 tpa, lp(longdash_dot) lc(black) sort c(J))
> (line vpadj75 tpa, lp(longdash) lc(black) sort c(J))
> (line vpadj95 tpa, lp(l) lc(black) sort c(J))
> if inrange(tpa,30,60),
> ytitle("International prostate symptom score")
> ylabel(0(5)35, angle(horiz)) ymtick(0(1)35)
> xtitle("Total physical activity, MET-hours/day")
> xlabel(30(5)60)  xmtick(30(1)60)
> legend(label(1 "0.25") label(2 "0.50") label(3 "0.75")
> label(4 "0.95") textfirst ring(0) pos(1) col(1)
> subtitle("Age-Adjusted" "Quantiles") order(4 3 2 1))
> scheme(sj)
```

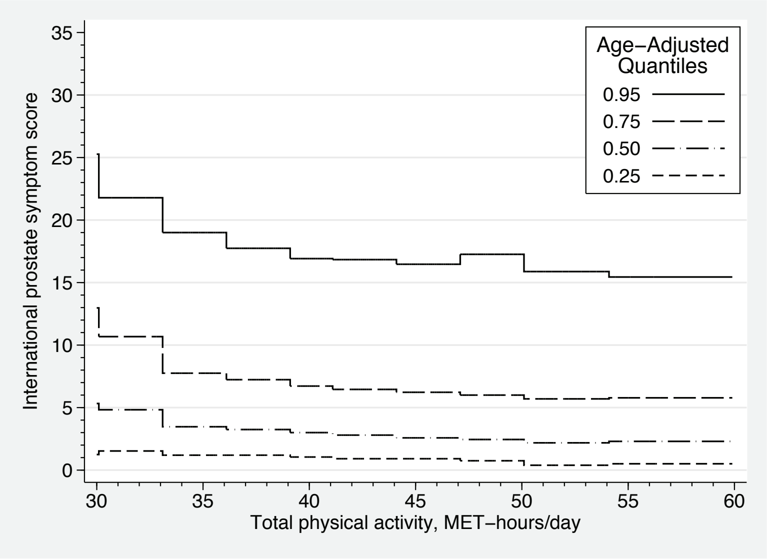

Figure 3. Age-adjusted quantiles of IPSS as a function of physical activity estimated with `lqreg`

One can also perform statistical tests about differences in associations across quantiles of IPSS. For example, is the change in the 0.50 quantile significantly different from the change in the 0.95 quantile, comparing the highest versus the lowest physical activity level?

```
. test [q95]_Itpac_10 = [q50]_Itpac_10
 ( 1)  - [q50]_Itpac_10 + [q95]_Itpac_10 = 0
       F(  1, 30366) =    0.36
            Prob > F =    0.5488
```

The large p-value of the test indicates no evidence for differences between the two age-adjusted quantiles.

4.3 Continuous predictor

We now model physical activity as a continuous covariate, assuming a linear relationship between physical activity and each quantile of the logit transformation of IPSS. The model is

$$Q_y(p) = \beta_{p0} + \beta_{p1}\texttt{tpa} + \beta_{p2}\texttt{cage}$$

```
. lqreg ipss tpa cage, quantiles(25 50 75 95) nodots seed(123)
Logistic Quantile Regression                       Number of obs =     30377
Bounded Outcome: ipss(-.5, 35.5)                   Bootstrap(100) SEs
------------------------------------------------------------------------------
             |              Bootstrap
        ipss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q25          |
         tpa |  -.0290973   .0039163    -7.43   0.000    -.0367735   -.0214211
        cage |    .032492    .002633    12.34   0.000     .0273312    .0376528
       _cons |  -1.931836   .1505163   -12.83   0.000    -2.226854   -1.636818
-------------+----------------------------------------------------------------
q50          |
         tpa |  -.0268447   .0021013   -12.78   0.000    -.0309634   -.0227261
        cage |   .0378809   .0010609    35.71   0.000     .0358016    .0399603
       _cons |   -1.14027   .0842382   -13.54   0.000     -1.30538   -.9751592
-------------+----------------------------------------------------------------
q75          |
         tpa |  -.0233426   .0018965   -12.31   0.000    -.0270599   -.0196253
        cage |   .0417262   .0011054    37.75   0.000     .0395597    .0438928
       _cons |  -.4093629    .078281    -5.23   0.000     -.562797   -.2559288
-------------+----------------------------------------------------------------
q95          |
         tpa |  -.0178706   .0032533    -5.49   0.000    -.0242471   -.0114941
        cage |   .0422159   .0018143    23.27   0.000     .0386598    .0457721
       _cons |   .7177373   .1313472     5.46   0.000     .4602914    .9751833
------------------------------------------------------------------------------
```

With only the exception of the 0.95 quantile, there are decreasing trends of similar magnitudes for the age-adjusted percentiles. Every 1 MET-hours/day increase in total physical activity is associated with a statistically significant reduction in the 0.25, 0.50, 0.75, and 0.95 quantiles of IPSS.

Figure 4 is a twoway plot overlaying a scatterplot and a line plot of predicted responses for each quantile. The figure is obtained directly with the `lqregpred` postestimation command with the `for()` and `plotvs()` options.

```
. by ipss tpa, sort: generate flag = _n==1
. lqregpred adjl if flag == 1, for(tpa) plotvs(tpa)
```

One can customize the twoway plot by using the saved predicted quantiles.

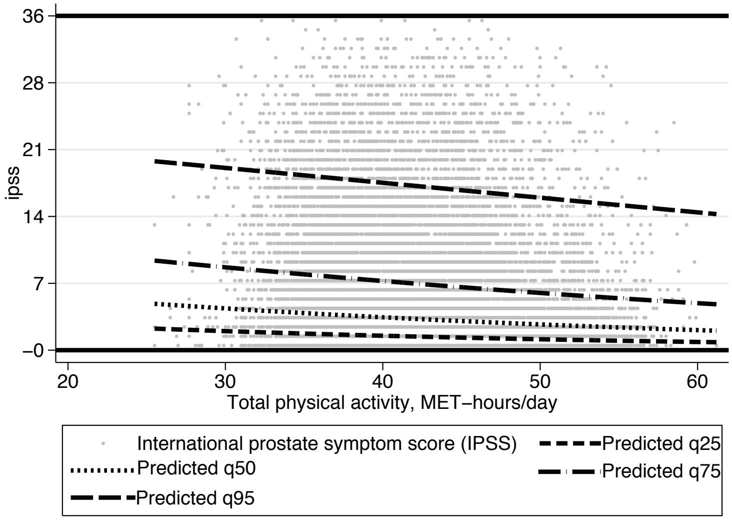

Figure 4. Scatterplot and four age-adjusted quantiles (0.25, 0.50, 0.75, and 0.95 from bottom to top) estimated with `lqreg` assuming linearity for physical activity

A plot of any regression coefficient with its 95% confidence interval for a dense set of quantiles $(0.05, 0.06, \ldots, 0.95)$ can be obtained with the postestimation command `lqregplot`.

```
. lqregplot tpa
Bootstrap(100) 95% Confidence Intervals
```

The coefficient associated with physical activity is highly significant for all quantiles greater than 0.11; its confidence interval does not include 0 (figure 5).

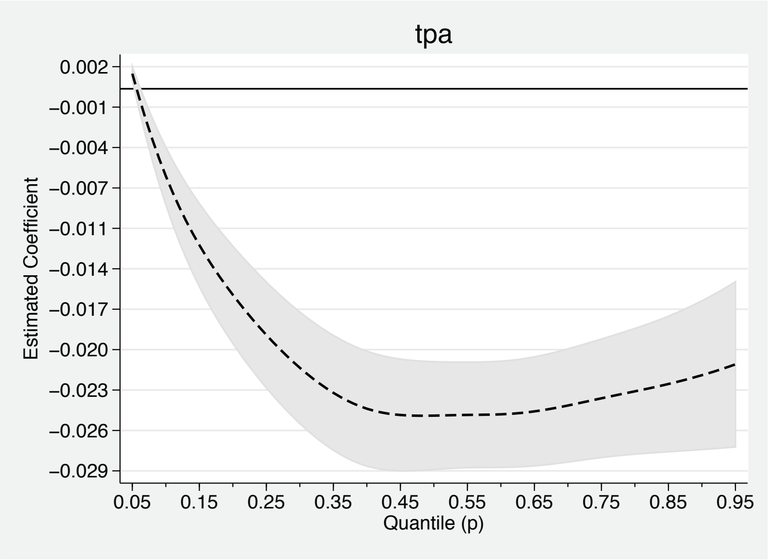

Figure 5. Estimates and 95% confidence bands for the regression coefficient of physical activity for a dense set of quantiles

Similarly to logistic regression for binary outcomes, the linearity assumption between each continuous predictor, either main exposure (physical activity) or confounder (age), and the logit transform of IPSS needs to be assessed. To perform a graphical check of linearity, one can obtain the logistic transformation of the bounded outcome with the `generate()` option of `lqreg`, generate the residuals (observed logit minus predicted logit), and then plot the residuals as a function of the covariate of interest.

A formal p-value for the hypothesis of linearity can be obtained by fitting a logistic quantile regression model with some transformations (polynomials, splines) of the continuous predictors. For instance, we generate restricted cubic spline transformations (3 knots at fixed percentiles 10, 50, and 90 of the distribution) for physical activity and age, and then we fit the logistic quantile regression model with these $3 - 1 = 2$ transformations for each predictor.

```
. mkspline tpas = tpa, nknots(3) cubic
. mkspline cages = cage, nknots(3) cubic
. lqreg ipss tpas1 tpas2 cages1 cages2, quantiles(25 50 75 95) nodots seed(123)
Logistic Quantile Regression                        Number of obs =     30377
Bounded Outcome: ipss(-.5, 35.5)                    Bootstrap(100) SEs
------------------------------------------------------------------------------
             |              Bootstrap
        ipss |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
q25          |
       tpas1 |  -.0442398   .0092667    -4.77   0.000    -.0624029   -.0260766
       tpas2 |   .0179152    .012842     1.40   0.163    -.0072557     .043086
      cages1 |   .0790713   .0057497    13.75   0.000     .0678015     .090341
      cages2 |  -.0566192   .0079938    -7.08   0.000    -.0722874    -.040951
       _cons |  -1.167755     .36403    -3.21   0.001    -1.881269    -.454241
-------------+----------------------------------------------------------------
q50          |
       tpas1 |  -.0403511   .0056028    -7.20   0.000    -.0513328   -.0293694
       tpas2 |   .0193159   .0079029     2.44   0.015      .003826    .0348059
      cages1 |   .0427258   .0030189    14.15   0.000     .0368086    .0486431
      cages2 |   -.007072   .0045496    -1.55   0.120    -.0159894    .0018455
       _cons |  -.5970202   .2177509    -2.74   0.006    -1.023821   -.1702193
-------------+----------------------------------------------------------------
q75          |
       tpas1 |  -.0469142   .0057808    -8.12   0.000    -.0582447   -.0355836
       tpas2 |   .0341496    .008237     4.15   0.000     .0180048    .0502945
      cages1 |    .042501   .0033169    12.81   0.000     .0359997    .0490023
      cages2 |   -.000555   .0048004    -0.12   0.908    -.0099639    .0088539
       _cons |   .4806698   .2190229     2.19   0.028     .0513758    .9099638
-------------+----------------------------------------------------------------
q95          |
       tpas1 |  -.0456865   .0077517    -5.89   0.000    -.0608802   -.0304928
       tpas2 |   .0419179   .0117526     3.57   0.000     .0188823    .0649535
      cages1 |   .0563585   .0058072     9.70   0.000     .0449762    .0677408
      cages2 |  -.0227971   .0077858    -2.93   0.003    -.0380575   -.0075367
       _cons |   1.874565    .286916     6.53   0.000     1.312198    2.436933
------------------------------------------------------------------------------
```

The linear-response model for physical activity is nested within the restricted cubic spline model because `tpas1 = tpa`. Therefore, some departure from linearity can be tested by testing the hypothesis that the regression coefficient of `tpas2` is equal to 0. The restricted cubic spline model indicates some evidence of nonlinearity both for physical activity and for age. Once again, the `lqregpred` command can be useful to depict the estimated percentiles.

Figure 6 shows that the decreasing trends for 0.25, 0.50, 0.75, and 0.95 quantile IPSSs reach the plateau at about 41 MET-hours/day (median physical activity).

```
. lqregpred adjs if flag == 1, for(tpas1 tpas2) plotvs(tpa)
```

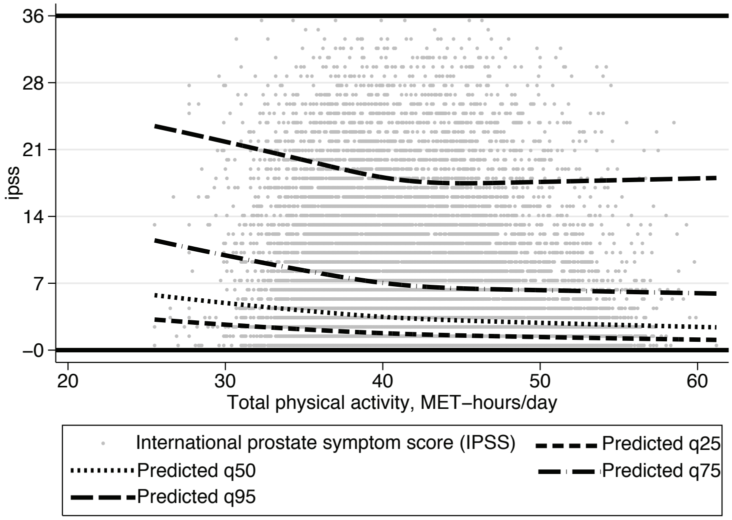

Figure 6. Scatterplot and four age-adjusted quantiles (0.25, 0.50, 0.75, and 0.95 from bottom to top) estimated with `lqreg` modeling physical activity with restricted cubic splines

5 Acknowledgments

Nicola Orsini was partly supported by a Young Scholar Award from the Karolinska Institutet's Strategic Program in Epidemiology. We would like to thank an anonymous reviewer for the comments provided.

6 References

Baum, C. F. 2008. Stata tip 63: Modeling proportions. *Stata Journal* 8: 299–303.

Bottai, M., B. Cai, and R. E. McKeown. 2010. Logistic quantile regression for bounded outcomes. *Statistics in Medicine* 29: 309–317.

Buchinsky, M. 1995. Estimating the asymptotic covariance matrix for quantile regression models: A Monte Carlo study. *Journal of Econometrics* 68: 303–338.

Buis, M. L., N. J. Cox, and S. P. Jenkins. 2011. betafit: Stata module to fit a two-parameter beta distribution. Statistical Software Components S435303, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s435303.html.

Gould, W. 1992. sg11.1: Quantile regression with bootstrapped standard errors. *Stata Technical Bulletin* 9: 19–21. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 137–139. College Station, TX: Stata Press.

Koenker, R. 2005. *Quantile Regression.* New York: Cambridge University Press.

Koenker, R., and G. Bassett, Jr. 1978. Regression quantiles. *Econometrica* 46: 33–50.

Orsini, N., B. RashidKhani, S.-O. Andersson, L. Karlberg, J.-E. Johansson, and A. Wolk. 2006. Long-term physical activity and lower urinary tract symptoms in men. *Journal of Urology* 176: 2546–2550.

Papke, L. E., and J. M. Wooldridge. 1996. Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics* 11: 619–632.

Rogers, W. H. 1992. sg11: Quantile regression standard errors. *Stata Technical Bulletin* 9: 16–19. Reprinted in *Stata Technical Bulletin Reprints*, vol. 2, pp. 133–137. College Station, TX: Stata Press.

Smithson, M., and J. Verkuilen. 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11: 54–71.

About the authors

Nicola Orsini is an associate professor in the Unit of Nutritional Epidemiology and the Unit of Biostatistics at the Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

Matteo Bottai is a professor in the Unit of Biostatistics at the Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.