The Stata Journal (2011)
11, Number 1, pp. 1–29

A procedure to tabulate and plot results after flexible modeling of a quantitative covariate

Nicola Orsini
Division of Nutritional Epidemiology
National Institute of Environmental Medicine
Karolinska Institutet
Stockholm, Sweden
nicola.orsini@ki.se

Sander Greenland
Departments of Epidemiology and Statistics
University of California–Los Angeles
Los Angeles, CA

Abstract. The use of flexible models for the relationship between a quantitative covariate and the response variable can be limited by the difficulty in interpreting the regression coefficients. In this article, we present a new postestimation command, `xblc`, that facilitates tabular and graphical presentation of these relationships. Cubic splines are given special emphasis. We illustrate the command through several worked examples using data from a large study of Swedish men on the relation between physical activity and the occurrence of lower urinary tract symptoms.

Keywords: st0215, xblc, cubic spline, modeling strategies, logistic regression

1 Introduction

In many studies, it is important to identify, present, and discuss the estimated relationship between a quantitative or continuous covariate (also called predictor, independent, or explanatory variable) and the response variable. In health sciences, the covariate is usually an exposure measurement or a clinical measurement. Regression models are widely used for contrasting responses at different values of the covariate. Their simplest forms assume a linear relationship between the quantitative covariate and some transformation of the response variable. The linearity assumption makes the regression coefficient easy to interpret (constant change of the predicted response per unit change of the covariate), but there is no reason to expect this assumption to hold in most applications.

Modeling nonlinear relationships through categorization of the covariate or adding a quadratic term may have limitations and rely on unrealistic assumptions, leading to distortions in inferences (see Royston, Altman, and Sauerbrei [2006] and Greenland [1995a,c,d]). Flexible alternatives involving more flexible, smooth transformations of the original covariate, such as fractional polynomials and regression splines (linear,

quadratic, or cubic), have been introduced (see Steenland and Deddens [2004]; Royston, Ambler, and Sauerbrei [1999]; Marrie, Dawson, and Garland [2009]; Harrell, Lee, and Pollock [1988]; and Greenland [2008; 1995b]) and are available in Stata (see [R] **mfp** and [R] **mkspline**). Nonetheless, these transformations complicate the contrast of the expected response at different values of the covariate and may discourage their use.

The aim of this article is to introduce the new postestimation command `xblc`, which aids in the interpretation and presentation of a nonlinear relationship in tabular and graphical form. We illustrate the procedure with data from a large cohort of Swedish men. The data examine the relationship between physical activity and the occurrence of lower urinary tract symptoms (LUTS) (Orsini et al. 2006). We focus on cubic-spline logistic regression for predicting the occurrence of a binary response. Nonetheless, the `xblc` command works similarly after any estimation command and regardless of the strategy used to model the quantitative covariate.

The rest of this article is organized as follows: section 2 provides an introduction to different types of cubic splines (not necessarily restricted); section 3 shows how to obtain point and interval estimates of measures of association between the covariate and the response; section 4 describes the syntax of the postestimation command `xblc`; section 5 presents several worked examples showing how to use the `xblc` command after the estimation of different types of cubic-spline models and how to provide intervals for the predicted response rather than differences between predicted responses; and section 6 compares other approaches (categories, linear splines, and fractional polynomials) with model nonlinearity, which can also use the `xblc` command.

2 Cubic splines

Cubic splines are generally defined as piecewise-polynomial line segments whose function values and first and second derivatives agree at the boundaries where they join. The boundaries of these segments are called knots, and the fitted curve is continuous and smooth at the knot boundaries (Smith 1979).

To avoid instability of the fitted curve at the extremes of the covariate, a common strategy is to constrain the curve to be a straight line before the first knot or after the last knot. The `mkspline` command can make both linear and restricted cubic splines since Stata 10.0 (see [R] **mkspline**). In some situations, restricting splines to be linear in both tails is not a warranted assumption. Therefore, we next show how to specify a linear predictor for a quantitative covariate X with neither tail restricted, only the left tail restricted, only the right tail restricted, or both tails restricted.

A common strategy for including a nonlinear effect of a covariate X is to replace it with some function of $X, g(X)$. For example, $g(X)$ could be $b_1X + b_2X^2$ or $b_1 \ln(X)$. For the (unrestricted) cubic-spline model, $g(X)$ is a function of the knot values k_i, $i = 1, \ldots, n$, as follows:

$$g(X) = b_0 + b_1X + b_2X^2 + b_3X^3 + \sum_{i=1}^{n} b_{3+i}\max(X - k_i, 0)^3$$

where the math function $\max(X-k_i, 0)$, known as the "positive part" function $(X-k_i)_+$, returns the maximum value of $X - k_i$ and 0. A model with only the left tail restricted to be linear implies that $b_1 = b_2 = 0$, so we drop X^2 and X^3:

$$g(X) = b_0 + b_1X + \sum_{i=1}^{n} b_{1+i}\max(X - k_i, 0)^3$$

A model with the right tail restricted to be linear is equal to the left-tail restricted model based on $-X$ with knots in reversed order and with the opposite sign of the ones based on the original X, which simplifies to

$$g(X) = b_0 + b_1(-X) + \sum_{i=1}^{n} b_{1+i}\max(k_i - X, 0)^3$$

A model with both tails restricted has $n - 1$ coefficients for transformations of the original exposure variable X,

$$g(X) = b_0 + b_1X_1 + b_2X_2 + \cdots + b_{n-1}X_{n-1}$$

where the first spline term, X_1, is equal to the original exposure variable X, whereas the remaining spline terms, $X_2, \ldots, X_{n-1}$, are functions of the original exposure X, the number of knots, and the spacing between knots, defined as follows:

$$\begin{aligned} u_i =& \max(X - k_i, 0)^3 \quad \text{with} \quad i = 1, \ldots, n \\ X_i =& \{u_{i-1} - u_{n-1}(k_n - k_{i-1})/(k_n - k_{n-1}) + u_n(k_{n-1} - k_{i-1})/(k_n - k_{n-1})\}/ \\ & (k_n - k_1)^2 \\ & \text{with} \quad i = 2, \ldots, n-1 \end{aligned}$$

More detailed descriptions of splines can be found elsewhere (see Greenland [2008]; Smith [1979]; Durrleman and Simon [1989]; Harrell [2001]; and Wegman and Wright [1983]).

3 Measures of association, p-values, and interval estimation

Modeling a quantitative covariate using splines or other flexible tools does not modify the way measures of covariate–response associations are defined.

An estimate of a measure of association between two variables usually ends up being a comparison of the predicted (fitted) value of a response variable (or some function of it) across different groups represented by the covariate. For example, the estimated association between gender and urinary tract symptoms compares the predicted urinary tract symptoms for men with the expected urinary tract symptoms for women. Such a comparison can take the form of computing the difference between the predicted values but can also take the form of computing the ratio.

For a quantitative covariate, such as age in years or pack-years of smoking, there can be a great many groups because each unique value of that covariate represents, in principle, its own group. We can display those using a graph, or we can create a table of a smaller number of comparisons between "representative" groups to summarize the relationship between the variables.

Contrasting predicted responses in the presence of nonlinearity is more elaborate because it involves transformations of the covariate. We illustrate the point using the restricted cubic-spline model; similar considerations apply to other types of covariate transformations. The linear predictor at the covariate values z_1 and z_2 is given by

$$g(X = z_1) = b_0 + b_1 X_1(z_1) + b_2 X_2(z_1) + \cdots + b_{n-1} X_{n-1}(z_1)$$
$$g(X = z_2) = b_0 + b_1 X_1(z_2) + b_2 X_2(z_2) + \cdots + b_{n-1} X_{n-1}(z_2)$$

so that

$$g(X = z_1) - g(X = z_2) =$$
$$b_1 \{X_1(z_1) - X_1(z_2)\} + b_2 \{X_2(z_1) - X_2(z_2)\} + \cdots + b_{n-1} \{X_{n-1}(z_1) - X_{n-1}(z_2)\}$$

The interpretation of the quantity $g(X = z_1) - g(X = z_2)$ depends on the model for the response. For example, within the family of generalized linear models, the quantity $g(X = z_1) - g(X = z_2)$ represents the difference between two mean values of a continuous response in a linear model (see [R] **regress**); the difference between two log odds (the log odds-ratio [OR]) of a binary response in a logistic model (see [R] **logit**); or the difference between two log rates (the log rate-ratio) of a count response in a log-linear Poisson model with the log of time over which the count was observed as an offset variable (see [R] **poisson**).

Commands for calculating p-values and predictions are derived using standard techniques available for simpler parametric models (Harrell, Lee, and Pollock 1988). For example, to obtain the p-value for the null hypothesis that there is no association between the covariate X and the response in a restricted cubic-spline model, we test the joint null hypothesis

$$b_1 = b_2 = \cdots = b_{n-1} = 0$$

The linear-response model is nested within the restricted cubic-spline model ($X_1 = X$), and the linear response to X corresponds to the constraint

$$b_2 = \cdots = b_{n-1} = 0$$

The p-value for this hypothesis is thus a test of linear response. Assuming this constraint, one can drop the spline terms $X_2, \ldots, X_{n-1}$, which simplifies the above comparison to

$$g(X = z_1) - g(X = z_2) = b_1 \{X_1(z_1) - X_1(z_2)\}$$

The quantity $b_1 \{X_1(z_1) - X_1(z_2)\}$ is the contrast between two predicted responses associated with a $z_1 - z_2$ unit increase of the covariate X throughout the covariate range (linear-response assumption). Therefore, modeling the covariate response as linear assumes a constant difference in the linear predictor regardless of where we begin the increase (z_2).

Returning to the general case, an approximate confidence interval (CI) for the difference in the linear predictors at the covariate values z_1 and z_2, $g(X = z_1) - g(X = z_2)$, can be calculated from the standard error (SE) for this difference, which is computable from the covariate values z_1 and z_2 and the covariance matrix of the estimated coefficients:

$$\begin{aligned}&[b_1\{X_1(z_1) - X_1(z_2)\} + b_2\{X_2(z_1) - X_2(z_2)\} + \cdots + b_{n-1}\{X_{n-1}(z_1) - X_{n-1}(z_2)\}] \\ &\pm z_{(\alpha/2)} \times \text{SE}[b_1\{X_1(z_1) - X_1(z_2)\} + b_2\{X_2(z_1) - X_2(z_2)\} + \cdots \\ &\quad + b_{n-1}\{X_{n-1}(z_1) - X_{n-1}(z_2)\}]\end{aligned}$$

where $z_{(\alpha/2)}$ denotes the $100(1 - \alpha/2)$ percentile of a standard normal distribution (1.96 for a 95% CI). The postestimation command `xblc` carries out these computations with the `lincom` command (see [R] **lincom**). In health-related fields, the value of the covariate $X = z_2$ is called a reference value, and it is used to compute and interpret a set of comparisons of subpopulations defined by different covariate values.

4 The xblc command

4.1 Syntax

`xblc` *varlist*, `at(`*numlist*`)` `covname(`*varname*`)` [`reference(`#`)` `pr` `eform`
`format(%`*fmt*`)` `level(`#`)` `equation(`*string*`)`
`generate(`*newvar1 newvar2 newvar3 newvar4*`)`]

4.2 Description

`xblc` computes point and interval estimates for predictions or differences in predictions of the response variable evaluated at different values of a quantitative covariate modeled using one or more transformations of the original variable specified in *varlist*. It can be used after any estimation command.

4.3 Options

`at(`*numlist*`)` specifies the values of the covariate specified in `covname()`, at which `xblc` evaluates predictions or differences in predictions. The values need to be in the current dataset. Covariates other than the one specified with the `covname()` option are fixed at zero. This is a required option.

`covname(`*varname*`)` specifies the name of the quantitative covariate. This is a required option.

`reference(`#`)` specifies the reference value for displaying differences in predictions.

`pr` computes and displays predictions (that is, mean response after linear regression, log odds after logistic models, and log rate after Poisson models with person-time as offset) rather than differences in predictions. To use this option, check that the previously fit model estimates the constant `_b[_cons]`.

`eform` displays the exponential value of predictions or differences in predictions.

`format(%`*fmt*`)` specifies the display format for presenting numbers. `format(%3.2f)` is the default; see [D] **format**.

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

`equation(`*string*`)` specifies the name of the equation when you have previously fit a multiple-equation model.

`generate(`*newvar1 newvar2 newvar3 newvar4*`)` specifies that the values of the original covariate, predictions or differences in predictions, and the lower and upper bounds of the CI be saved in *newvar1*, *newvar2*, *newvar3*, and *newvar4*, respectively. This option is very useful for presenting the results in a graphical form.

5 Examples

As an illustrative example, we analyze in a cross-sectional setting a sample of 30,377 men (`pa_luts.dta`) in central Sweden aged 45–79 years who completed a self-administered lifestyle questionnaire that included international prostate symptom score (IPSS) questions and physical activity questions (work/occupation, home/household work, walking/bicycling, exercise, and leisure-time such as watching TV/reading) (Orsini et al. 2006). The range of the response variable, the IPSS score, is 0 to 35. According to the American Urological Association, the IPSS score (variable `ipss2`) is categorized in two levels: mild or no symptoms (scores 0–7) and moderate to severe LUTS (scores 8–35). The main covariate of interest is a total physical activity score (variable `tpa`), which comprises a combination of intensity and duration for a combination of daily activities and is expressed in metabolic equivalents (MET) (kcal/kg/hour).

The proportion of men reporting moderate to severe LUTS is $6905/30377 = 0.23$. The odds in favor of experiencing moderate to severe LUTS are $0.23/(1 - 0.23) = 6905/23472 = 0.29$; this means that on average, for every 100 men with mild or no symptoms, we observed 29 other men with moderate to severe LUTS, written as 29:100 (29 to 100 odds). Examining the variation of the ratio of cases/noncases (odds) of moderate to severe LUTS according to subpopulations of men defined by intervals of total physical activity (variable `tpac`) is our first step in describing the shape of the covariate–response association (figure 1).

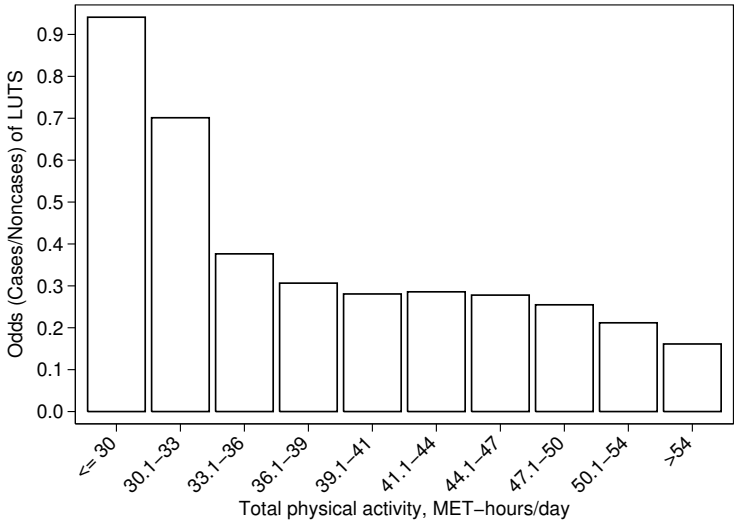

Figure 1. Observed odds (ratio of cases/noncases) of moderate to severe LUTS by categories of total physical activity (MET-hours/day) in a cohort of 30,377 Swedish men.

The occurrence of moderate to severe LUTS decreases more rapidly at the low values of the covariate distribution. There is a strong reduction of the odds of moderate to severe LUTS, going from 94:100 at the minimum total physical activity interval ($\geq$ 30 MET-hours/day) down to 38:100 at the interval 33.1 to 36 MET-hours/day. It follows a more gradual decline in the odds of moderate to severe LUTS to 16:100 in men at the highest total physical activity interval ($>$ 54 MET-hours/day).

Table 1 provides a tabular presentation of the data (total number of men, sum of the cases, range and median value of the covariate) by intervals of total physical activity. About 99% of the participants and 99% of the cases of moderate to severe LUTS are within the range 29 to 55 MET-hours/day. Therefore, results are presented within this range.

Table 1. Tabular presentation of data, unadjusted and age-adjusted ORs with 95% CI for the association of total physical activity (MET-hours/day) and occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men.

No. of subjects	No. of cases	Exposure range	Exposure median	Unadjusted OR [95% CI]*	Age-adjusted OR [95% CI]*
66	32	≤ 30	29	1.00	1.00
427	176	30.1–33	32	0.71 [0.55, 0.93]	0.85 [0.65, 1.12]
2761	755	33.1–36	35	0.42 [0.30, 0.58]	0.60 [0.43, 0.84]
7524	1765	36.1–39	38	0.31 [0.23, 0.43]	0.47 [0.34, 0.66]
5074	1112	39.1–41	40	0.31 [0.23, 0.43]	0.45 [0.32, 0.62]
5651	1256	41.1–44	43	0.31 [0.23, 0.43]	0.41 [0.30, 0.57]
4782	1040	44.1–47	45	0.30 [0.22, 0.41]	0.40 [0.29, 0.55]
2359	479	47.1–50	48	0.27 [0.20, 0.37]	0.39 [0.28, 0.54]
1373	240	50.1–54	52	0.24 [0.17, 0.33]	0.37 [0.27, 0.52]
360	50	> 54	55	0.21 [0.15, 0.30]	0.36 [0.25, 0.51]

* Total physical activity expressed in MET-hours/day was modeled by right-restricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model. The value of 29 MET-hours/day, as the median value of the lowest reference range of total physical activity, was used to estimate all ORs.

5.1 Unrestricted cubic splines

We first create unrestricted cubic splines with four knots at fixed and equally spaced percentiles (20%, 40%, 60%, and 80%). Varying the location of the knots (for instance, using percentiles 5%, 35%, 65%, and 95% as recommended by Harrell's book [2001]) had negligible influence on the estimates.

```
. generate all = 1
. table all, contents(freq p20 tpa p40 tpa p60 tpa p80 tpa)
```

all	Freq.	p20(tpa)	p40(tpa)	p60(tpa)	p80(tpa)
1	30,377	37.2	39.6	42.3	45.6

```
. generate tpa2 = tpa^2
. generate tpa3 = tpa^3
. generate tpap1 = max(0,tpa-37.2)^3
. generate tpap2 = max(0,tpa-39.6)^3
. generate tpap3 = max(0,tpa-42.3)^3
. generate tpap4 = max(0,tpa-45.6)^3
```

Ideally, the number of knots and their placement will result in categories with reasonably large numbers of both cases and noncases in each category. While there are no simple and foolproof rules, we recommend that each category have at least five and preferably more cases and noncases in each category and that the number of cases and number of noncases each are at least five times the number of model parameters. Further discussion on the choice of location and number of knots can be found in section 2.4.5 of Harrell's book (2001). Harrell also discusses more general aspects of model selection for dose–response (trend) analysis, as do Royston and Sauerbrei (2007).

We first fit a logistic regression model with unrestricted cubic splines for physical activity and no other covariate.

```
. logit ipss2 tpa tpa2 tpa3 tpap1 tpap2 tpap3 tpap4
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood = -16187.593
Iteration 2:   log likelihood = -16185.014
Iteration 3:   log likelihood = -16185.014
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(7)      =     194.46
                                                  Prob > chi2     =     0.0000
Log likelihood = -16185.014                       Pseudo R2       =     0.0060
```

ipss2	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
tpa	9.759424	3.383661	2.88	0.004	3.12757	16.39128
tpa2	-.2985732	.0986663	-3.03	0.002	-.4919556	-.1051909
tpa3	.0029866	.0009553	3.13	0.002	.0011143	.0048589
tpap1	-.009595	.0035502	-2.70	0.007	-.0165532	-.0026368
tpap2	.0094618	.0052404	1.81	0.071	-.0008093	.0197328
tpap3	-.0049394	.0040546	-1.22	0.223	-.0128863	.0030074
tpap4	.0027824	.0019299	1.44	0.149	-.0010001	.0065649
_cons	-104.8292	38.52542	-2.72	0.007	-180.3376	-29.32074

Because the model omits other covariates, it is called uncontrolled or unadjusted analysis, also known as "crude" analysis.

The one-line postestimation command `xblc` is used to tabulate and plot contrasts of covariate values. It allows the user to specify a set of covariate values (here 29, 32, 35, 38, 40, 43, 45, 48, 52, and 55) at which it computes the ORs, using the value of 29 MET-hours/day as a referent.

```
. xblc tpa tpa2 tpa3 tpap1 tpap2 tpap3 tpap4, covname(tpa)
> at(29 32 35 38 40 43 45 48 52 55) reference(29) eform generate(pa or lb ub)
tpa           exp(xb)    (95% CI)
29            1.00       (1.00-1.00)
32            0.71       (0.55-0.93)
35            0.41       (0.30-0.57)
38            0.31       (0.23-0.43)
40            0.31       (0.23-0.42)
43            0.30       (0.22-0.41)
45            0.30       (0.22-0.41)
48            0.28       (0.20-0.38)
52            0.22       (0.16-0.31)
55            0.19       (0.13-0.28)
```

We specify the `eform` option of `xblc` because we are interested in presenting ORs rather than the difference between two log odds of the binary response. For plotting the ORs, a convenient `xblc` option is `generate()`, which saves the above four columns of numbers in the current dataset. The following code produces a standard two-way plot, as shown in figure 2:

```
. twoway (rcap lb ub pa, sort) (scatter or pa, sort), legend(off)
> scheme(s1mono) xlabel(29 32 35 38 40 43 45 48 52 55) ylabel(.2(.2)1.2,
> angle(horiz)  forma t(%2.1fc)) ytitle("Unadjusted Odds Ratios of LUTS")
> xtitle("Total physical activity, MET-hours/day")
```

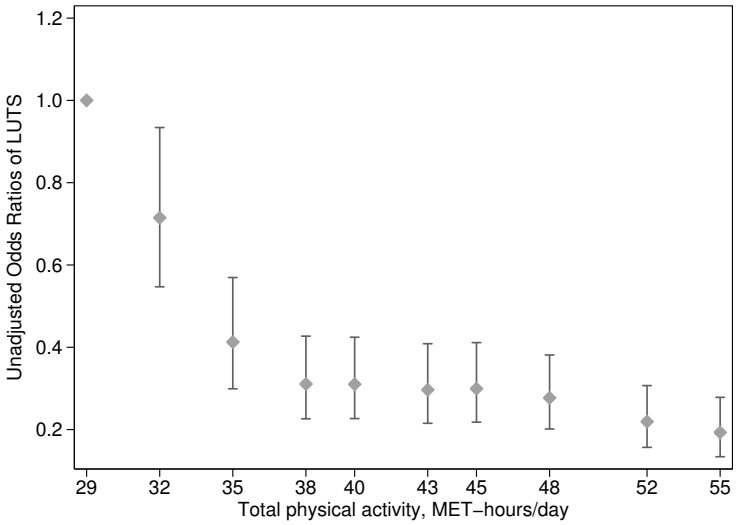

Figure 2. This graph shows unadjusted ORs (dots) with 95% CI (capped spikes) for the relation of total physical activity (MET-hours/day) to the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men. Total physical activity was modeled by unrestricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model. The reference value is 29 MET-hours/day.

To get a better idea of the dose–response relation, one can compute the ORs and 95% confidence limits of moderate to severe LUTS for any subpopulation of men defined by a finer grid of values (using, say, a 1 MET-hour/day increment) across the range of interest (figure 3).

```
. capture drop pa or lb ub
. xblc tpa tpa2 tpa3 tpap*, covname(tpa) at(29(1)55) reference(29) eform
> generate(pa or lb ub)
```

(*output omitted*)

```
. twoway (rcap lb ub pa, sort) (scatter or pa, sort), legend(off)
> scheme(s1mono) xlabel(29(2)55) xmtick(29(1)55)
> ylabel(.2(.2)1.2, angle(horiz) format(%2.1fc))
> ytitle("Unadjusted Odds Ratios of LUTS")
> xtitle("Total physical activity, MET-hours/day")
```

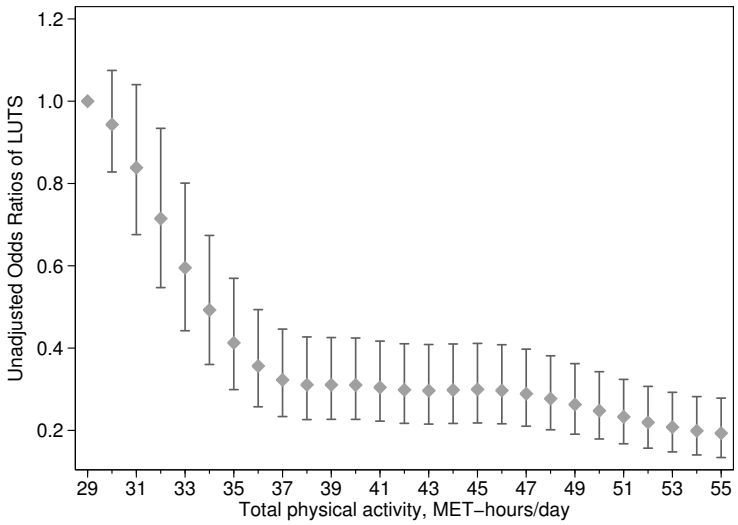

Figure 3. This graph shows unadjusted ORs (dots) with 95% CI (capped spikes) for the relation of total physical activity (MET-hours/day) to the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men. Total physical activity was modeled by unrestricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model. The reference value is 29 MET-hours/day.

To produce a smooth graph of the relation, one can estimate all the differences in the log odds of moderate to severe LUTS corresponding to the 315 distinct observed exposure values, and then control how the point estimates and CIs are to be connected (figure 4).

```
. capture drop pa or lb ub
. quietly levelsof tpa, local(levels)
. quietly xblc tpa tpa2 tpa3 tpap*, covname(tpa) at(`r(levels)´) reference(29)
> eform generate(pa or lb ub)
. twoway (line lb ub pa, sort lc(black black) lp(- -))
> (line or pa, sort lc(black) lp(l)) if inrange(pa,29,55), legend(off)
> scheme(s1mono) xlabel(29(2)55) xmtick(29(1)55)
> ylabel(.2(.2)1.2, angle(horiz)  format(%2.1fc))
> ytitle("Unadjusted Odds Ratios of LUTS")
> xtitle("Total physical activity, MET-hours/day")
```

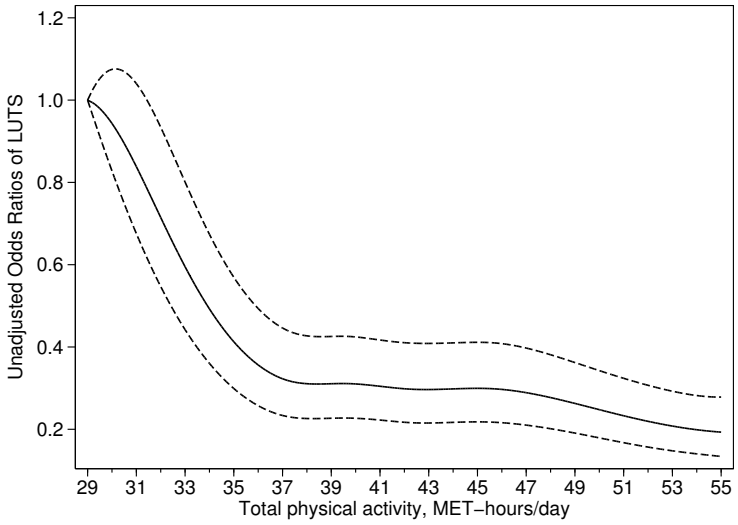

Figure 4. This graph shows unadjusted ORs (solid line) with 95% CI (dashed lines) for the relation of total physical activity (MET-hours/day) to the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men. Total physical activity was modeled by unrestricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model. The reference value is 29 MET-hours/day.

5.2 Cubic splines with only one tail restricted

The observed odds of moderate to severe LUTS decreases more rapidly on the left tail of the physical activity distribution (see figure 1), which suggests that restricting the curve to be linear before the first knot placed at 37.2 MET-hours/day (20th percentile) is probably not a good idea. On the other hand, the right tail of the distribution above 45.6 MET-hours/day (80th percentile) shows a more gradual decline of the odds of moderate to severe LUTS, suggesting that restriction there is not unreasonable.

The left-tail restricted cubic-spline model just drops the quadratic and cubic terms of the previously fit unrestricted model. Given that the model that is left-tail restricted is nested within the unrestricted model, a Wald-type test for nonlinearity beyond the first knot is given by

```
. testparm tpa2 tpa3
 ( 1)  [ipss2]tpa2 = 0
 ( 2)  [ipss2]tpa3 = 0
           chi2(  2) =   18.42
         Prob > chi2 =    0.0001
```

The small p-value of the Wald-type test with two degrees of freedom indicates nonlinearity beyond the first knot. We show how to fit the model and then present the results:

```
. logit ipss2 tpa tpap1 tpap2 tpap3 tpap4
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood = -16195.263
Iteration 2:   log likelihood = -16194.212
Iteration 3:   log likelihood = -16194.212
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(5)      =     176.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -16194.212                       Pseudo R2       =     0.0054

------------------------------------------------------------------------------
       ipss2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tpa |  -.0989318   .0101559    -9.74   0.000     -.118837   -.0790267
       tpap1 |   .0042758   .0009338     4.58   0.000     .0024457    .0061059
       tpap2 |  -.0095806   .0027518    -3.48   0.000     -.014974   -.0041873
       tpap3 |   .0060958    .003147     1.94   0.053    -.0000722    .0122638
       tpap4 |  -.0004933   .0017884    -0.28   0.783    -.0039985     .003012
       _cons |   2.549523   .3753997     6.79   0.000     1.813753    3.285293
------------------------------------------------------------------------------
```

Similarly to what we did after the estimation of the unrestricted cubic-spline model, we use the postestimation command `xblc` to present a set of ORs with 95% confidence limits. The only difference in the syntax of this `xblc` command is the list of transformations used to model physical activity.

```
. xblc tpa tpap*, covname(tpa) at(29 32 35 38 40 43 45 48 52 55) reference(29)
> eform
tpa             exp(xb)     (95% CI)
29              1.00        (1.00-1.00)
32              0.74        (0.70-0.79)
35              0.55        (0.49-0.62)
38              0.41        (0.34-0.49)
40              0.37        (0.31-0.45)
43              0.40        (0.34-0.47)
45              0.39        (0.33-0.46)
48              0.35        (0.29-0.42)
52              0.29        (0.24-0.35)
55              0.25        (0.20-0.32)
```

When assuming linearity only in the right tail of the covariate distribution, as explained in section 2, we first generate the cubic splines based on the negative of the original exposure. We then fit the model:

```
. generate tpan = -tpa
. generate tpapn1 = max(0,45.6-tpa)^3
. generate tpapn2 = max(0,42.3-tpa)^3
. generate tpapn3 = max(0,39.6-tpa)^3
. generate tpapn4 = max(0,37.2-tpa)^3
. logit ipss2 tpan tpapn*
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood = -16189.534
Iteration 2:   log likelihood = -16187.088
Iteration 3:   log likelihood = -16187.087
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(5)      =     190.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -16187.087                       Pseudo R2       =     0.0058
```

ipss2	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
tpan	.0325003	.0074057	4.39	0.000	.0179853	.0470153
tpapn1	-.0007537	.0005094	-1.48	0.139	-.0017521	.0002446
tpapn2	.0018099	.0023017	0.79	0.432	-.0027014	.0063212
tpapn3	.0029923	.0041652	0.72	0.473	-.0051714	.0111561
tpapn4	-.006665	.0032405	-2.06	0.040	-.0130163	-.0003136
_cons	.1675475	.3437625	0.49	0.626	-.5062146	.8413095

Once again, the postestimation command `xblc` facilitates the presentation, interpretation, and comparison of the results arising from different models.

```
. xblc tpan tpapn*, covname(tpa) at(29 32 35 38 40 43 45 48 52 55)
> reference(29) eform
tpa           exp(xb)    (95% CI)
29            1.00       (1.00-1.00)
32            0.71       (0.55-0.93)
35            0.42       (0.30-0.58)
38            0.31       (0.23-0.43)
40            0.31       (0.23-0.43)
43            0.31       (0.23-0.43)
45            0.30       (0.22-0.41)
48            0.27       (0.20-0.37)
52            0.24       (0.17-0.33)
55            0.21       (0.15-0.30)
```

The right-restricted cubic-spline model provides very similar ORs to the unrestricted model, but uses fewer coefficients.

5.3 Cubic splines with both tails restricted

To create a cubic spline that is restricted to being linear in both tails is more complicated, but the `mkspline` command facilitates this task.

```
. mkspline tpas = tpa, knots(37.2 39.6 42.3 45.6) cubic
```

The above line creates the restricted cubic splines, automatically named `tpas1`, `tpas2`, and `tpas3` using the defined knots. We then fit a logistic regression model that includes the three spline terms.

```
. logit ipss2 tpas1 tpas2 tpas3
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood = -16195.572
Iteration 2:   log likelihood = -16194.592
Iteration 3:   log likelihood = -16194.592
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(3)      =     175.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -16194.592                       Pseudo R2       =     0.0054
------------------------------------------------------------------------------
       ipss2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       tpas1 |  -.1009415   .0098873   -10.21   0.000    -.1203203   -.0815627
       tpas2 |    .337423   .0515938     6.54   0.000      .236301    .4385449
       tpas3 |  -.8010405    .130892    -6.12   0.000    -1.057584   -.5444968
       _cons |   2.620533   .3663134     7.15   0.000     1.902572    3.338494
------------------------------------------------------------------------------
```

To translate the estimated linear predictor into a set of ORs, we use the `xblc` command, as follows:

```
. xblc tpas*, covname(tpa) at(29 32 35 38 40 43 45 48 52 55) reference(29)
> eform
tpa           exp(xb)    (95% CI)
29            1.00       (1.00-1.00)
32            0.74       (0.70-0.78)
35            0.55       (0.49-0.61)
38            0.40       (0.34-0.48)
40            0.37       (0.30-0.44)
43            0.40       (0.34-0.47)
45            0.38       (0.32-0.45)
48            0.34       (0.29-0.40)
52            0.29       (0.24-0.35)
55            0.26       (0.21-0.32)
```

Figure 5 shows a comparison of the four different types of cubic splines. Given the same number and location of knots, the greatest impact on the curve is given by the inappropriate linear constraint before the first knot. Using Akaike's information criterion (a summary measure that combines fit and complexity), we found that the unrestricted and right-restricted cubic-spline models have a better fit (smaller Akaike's information criterion) compared with the left- and both-tail restricted cubic-spline models.

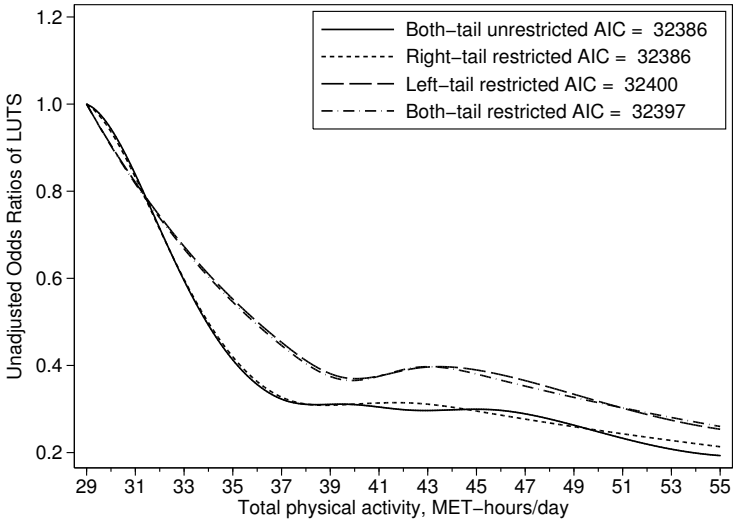

Figure 5. This graph compares unadjusted ORs for the relation of total physical activity (MET-hours/day) with the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men. Total physical activity was modeled by both-tail unrestricted, left-tail restricted, right-tail restricted, and both-tail restricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model. The reference value is 29 MET-hours/day.

The right-restricted model has a smaller number of regression coefficients than does the unrestricted model. Hence, we use the right-restricted model for further illustration of the `xblc` command with adjustment for other covariates and for the presentation of adjusted trends and confidence bands for the predicted occurrence of the binary response.

5.4 Adjusting for other covariates

Men reporting different physical activity levels may differ with respect to sociodemographic, biological, anthropometrical, health, and other lifestyle factors, so the crude estimates given above are unlikely to accurately reflect the causal effects of physical activity on the outcome. We now show that adjusting for such variables (known as potential confounders) does not change how the postestimation command `xblc` works.

Consider age, the strongest predictor of urinary problems. Moderate to severe LUTS increases with age and occurs in most elderly men, while total physical activity decreases with age. Therefore, the estimated decreasing odds of moderate to severe LUTS in subpopulations of men reporting higher physical activity levels might be explained by differences in the distribution of age. Thus we include age, centered on the sample mean

of 59 years, in the right-tail restricted cubic-spline model. For simplicity, we assume a linear relation of age to the log odds of moderate to severe LUTS. We could also use splines for age, but it has negligible influence on the main covariate–disease association in our example.

```
. quietly summarize age
. generate agec = age - r(mean)
. logit ipss2 tpan tpapn* agec
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood = -15533.528
Iteration 2:   log likelihood = -15517.532
Iteration 3:   log likelihood = -15517.526
Iteration 4:   log likelihood = -15517.526
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(6)      =    1529.44
                                                  Prob > chi2     =     0.0000
Log likelihood = -15517.526                       Pseudo R2       =     0.0470
------------------------------------------------------------------------------
       ipss2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        tpan |   .0104343   .0076532     1.36   0.173    -.0045657    .0254343
      tpapn1 |   .0004587   .0005237     0.88   0.381    -.0005676    .0014851
      tpapn2 |  -.0015097   .0023617    -0.64   0.523    -.0061385     .003119
      tpapn3 |   .0040955   .0042703     0.96   0.338    -.0042742    .0124651
      tpapn4 |  -.0048478   .0033233    -1.46   0.145    -.0113615    .0016658
        agec |   .0552749   .0015376    35.95   0.000     .0522612    .0582885
       _cons |  -.9478404   .3556108    -2.67   0.008    -1.644825   -.2508561
------------------------------------------------------------------------------
```

The syntax of the `xblc` command in the presence of another covariate is the same as that used for the unadjusted analysis.

```
. xblc tpan tpapn*, covname(tpa) at(29 32 35 38 40 43 45 48 52 55)
> reference(29) eform
tpa           exp(xb)    (95% CI)
29            1.00       (1.00-1.00)
32            0.85       (0.65-1.12)
35            0.60       (0.43-0.84)
38            0.47       (0.34-0.66)
40            0.45       (0.32-0.62)
43            0.41       (0.30-0.57)
45            0.40       (0.29-0.55)
48            0.39       (0.28-0.54)
52            0.37       (0.27-0.52)
55            0.36       (0.25-0.51)
```

As expected, the age-adjusted ORs of moderate to severe LUTS are generally lower compared with the crude ORs. Thus the association between physical activity and the outcome was partly explained by differences in age (table 1). Entering more covariates in the model does not change the `xblc` postestimation command. To obtain figure 6, the code is as follows:

```
. capture drop pa or lb ub
. quietly levelsof tpa, local(levels)
. quietly xblc tpan tpapn*, covname(tpa) at(`r(levels)´) reference(29) eform
> generate(pa or lb ub)
. twoway (line lb ub pa, sort lc(black black) lp(- -))
> (line or pa, sort lc(black) lp(l)) if inrange(pa,29,55), legend(off)
> scheme(s1mono) xlabel(29(2)55) xmtick(29(1)55)
> ylabel(.2(.2)1.2, angle(horiz) format(%2.1fc))
> ytitle("Age-adjusted Odds Ratios of LUTS")
> xtitle("Total physical activity, MET-hours/day")
```

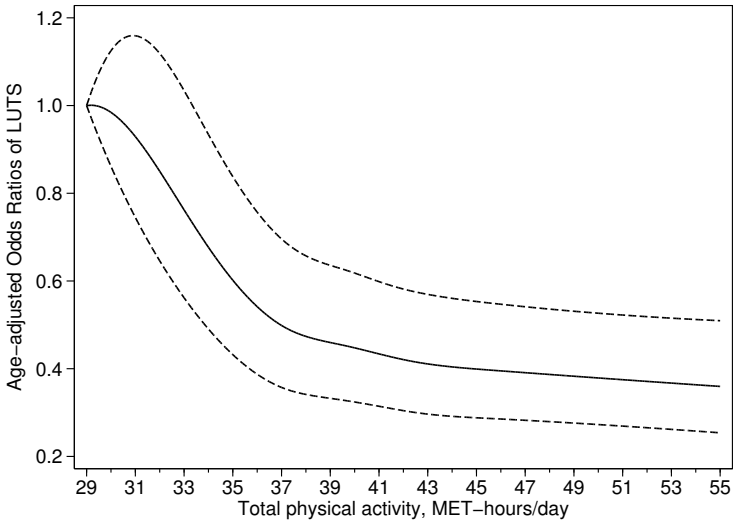

Figure 6. This graph shows age-adjusted ORs (solid line) with 95% CI (dashed lines) for the relation of total physical activity (MET-hours/day) to the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men. Total physical activity was modeled by right-restricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model. The reference value is 29 MET-hours/day.

5.5 Uncertainty for the predicted response

So far we have focused on tabulating and plotting ORs as functions of covariate values. It is important to note that the CIs for the ORs that include the sampling variability of the reference value cannot be used to compare the odds of two nonreference values. The problem arises if one misinterprets the CIs of the OR as representing CIs for the odds. Further discussion of this issue can be found elsewhere (Greenland et al. 1999).

Those readers who wish to visualize uncertainty about the odds of the event rather than the ORs may add the `pr` option (predicted response, log odds in our example) in the previously typed `xblc` command.

```
. capture drop pa
. quietly levelsof tpa, local(levels)
. quietly xblc tpan tpapn*, covname(tpa) at(`r(levels)´) reference(29) eform
> generate(pa rcc lbo ubo) pr
. twoway (line lbo ubo pa, sort lc(black black) lp(- -))
> (line rcc pa, sort lc(black) lp(l)) if inrange(pa,29,55), legend(off)
> scheme(s1mono) xlabel(29(2)55) xmtick(29(1)55) ylabel(.2(.1).8, angle(horiz)
> format(%2.1fc)) ytitle("Age-adjusted Odds (Cases/Noncases) of LUTS")
> xtitle("Total physical activity, MET-hours/day")
```

Figure 7 shows that the CIs around the age-adjusted odds of moderate to severe LUTS widen at the extremes of the graph, properly reflecting sparse data in the tails of the distribution of total physical activity.

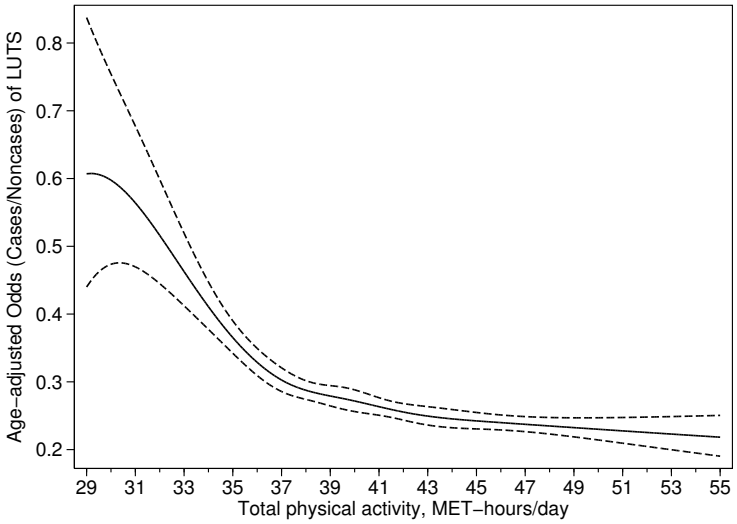

Figure 7. This graph shows age-adjusted odds (ratios of cases/noncases, solid line) with 95% CI (dashed lines) for the relation of total physical activity (MET-hours/day) to the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men. Total physical activity was modeled by right-restricted cubic splines with four knots (37.2, 39.6, 42.3, and 45.6) at percentiles 20%, 40%, 60%, and 80% in a logistic regression model.

6 Use of xblc after other modeling approaches

A valuable feature of the `xblc` command is that its use is independent of the specific approach used to model a quantitative covariate. The command can be used with alternative parametric models such as piecewise-linear splines or fractional polynomials (Steenland and Deddens 2004; Royston, Ambler, and Sauerbrei 1999; Greenland 2008, 1995b). To illustrate, we next show the use of the `xblc` command with different modeling strategies (categorization, linear splines, and fractional polynomials), as shown in figure 8 (in section 6.3).

6.1 Categorical model

We fit a logistic regression model with $10 - 1 = 9$ indicator variables with the lowest interval (≤ 30 MET-hours/day) serving as a referent.

```
. xi:logit ipss2 i.tpac agec, or
i.tpac            _Itpac_1-10         (naturally coded; _Itpac_1 omitted)
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood = -15537.912
Iteration 2:   log likelihood = -15521.805
Iteration 3:   log likelihood = -15521.798
Iteration 4:   log likelihood = -15521.798
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(10)     =    1520.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -15521.798                       Pseudo R2       =     0.0467
------------------------------------------------------------------------------
       ipss2 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Itpac_2 |   .8527221   .2336656    -0.58   0.561     .4983777    1.459004
    _Itpac_3 |    .536358   .1386083    -2.41   0.016     .3232087    .8900749
    _Itpac_4 |   .4730301   .1212153    -2.92   0.003     .2862635    .7816485
    _Itpac_5 |   .4108355   .1055984    -3.46   0.001     .2482452    .6799158
    _Itpac_6 |     .39818   .1022264    -3.59   0.000     .2407393    .6585851
    _Itpac_7 |   .3798165   .0976648    -3.76   0.000     .2294556    .6287081
    _Itpac_8 |   .3534416    .091875    -4.00   0.000     .2123503    .5882776
    _Itpac_9 |   .3658974   .0969315    -3.80   0.000     .2177026    .6149715
   _Itpac_10 |   .2925106   .0871772    -4.12   0.000     .1631012    .5245971
        agec |   1.056869   .0016265    35.94   0.000     1.053686    1.060062
------------------------------------------------------------------------------
```

We estimate the age-adjusted odds of the response with the `xblc` command, as shown in figure 8 (in section 6.3).

```
. quietly levelsof tpa, local(levels)
. quietly xblc _Itpac_2- _Itpac_10, covname(tpa) at(`r(levels)´) eform
> generate(pa oddsc lboc uboc) pr
```

The categorical model implies constant odds (ratio of cases/noncases) of moderate to severe LUTS within intervals of physical activity, with sudden jumps between intervals. The advantages of the categorical model are that it is easy to fit and to present in both tabular and graphical forms. The disadvantages (power loss, distortion of trends, and unrealistic dose–response step functions) of categorizing continuous variables have been pointed out several times (Royston, Altman, and Sauerbrei 2006; Greenland 1995a,b,c,d, 2008).

In our example, the differences between the categorical model and splines are greater at the low values of the covariate distribution (< 38 MET-hours/day) where the occurrence of moderate to severe LUTS decreases more rapidly (with a steeper slope) compared with the remaining covariate range. Another difference between the two models is the amount of information used in estimating associations. The odds or ratios of odds from the categorical model are only determined by the data contained in the exposure intervals being compared. One must ignore the magnitude and direction of the association in the remaining exposure intervals. For instance, in the categorical model fit to 30,377 men, the age-adjusted OR comparing the interval 30.1–33 MET-hours/day with the reference interval (≤ 30 MET-hours/day) is 0.85 [95% CI $= 0.50, 1.46$]. We would estimate practically the same adjusted OR and 95% CI by restricting the model to 1.6% of the sample (486 men) belonging to the first two categories of total physical activity being compared.

Not surprisingly, the width of the 95% CI around the fitted OR is greater in categorical models compared with restricted cubic-spline models. The fitted OR from a spline model uses the full covariate information for all individuals, and the CI gradually increases with the distance between the covariate values being compared, as it should.

The large sample size and the relatively large number of cases allow us to categorize physical activity in 10 narrow intervals. Therefore, the fitted trend based on the categorical model is overall not that different from the fitted trends based on splines and fractional polynomials (see table 2 on the next page and figure 8 in section 6.3). However, the shape of the covariate–response relationship in categorical models is sensitive to the location and number of cutpoints used to categorize the continuous covariate—potentially more sensitive than fitted curves with the same number of parameters will be to the choice of knots or polynomial terms.

Table 2. Comparison of age-adjusted OR with 95% CI for the association of total physical activity (MET-hours/day) and occurrence of moderate to severe LUTS estimated with different types of models: categorical, linear spline, and fractional polynomial

Exposure range	Exposure median	Categorical model OR [95% CI] *	Linear spline model OR [95% CI] †	Fractional polynomial model OR [95% CI] ‡
≤ 30	29	1.00	1.00	1.00
30.1–33	32	0.85 [0.50, 1.46]	0.76 [0.71, 0.82]	0.69 [0.62, 0.77]
33.1–36	35	0.54 [0.32, 0.89]	0.58 [0.51, 0.67]	0.53 [0.45, 0.63]
36.1–39	38	0.47 [0.29, 0.78]	0.44 [0.36, 0.54]	0.45 [0.36, 0.55]
39.1–41	40	0.41 [0.25, 0.68]	0.43 [0.35, 0.52]	0.41 [0.33, 0.51]
41.1–44	43	0.40 [0.24, 0.66]	0.41 [0.34, 0.49]	0.37 [0.30, 0.47]
44.1–47	45	0.38 [0.23, 0.63]	0.39 [0.33, 0.47]	0.36 [0.29, 0.45]
47.1–50	48	0.35 [0.21, 0.59]	0.37 [0.31, 0.45]	0.35 [0.28, 0.43]
50.1–54	52	0.37 [0.22, 0.61]	0.35 [0.29, 0.41]	0.34 [0.28, 0.41]
> 54	55	0.29 [0.16, 0.52]	0.33 [0.27, 0.39]	0.35 [0.29, 0.42]

* Nine indicator variables.
† One knot at 38 MET-hours/day.
‡ Degree-2 fractional polynomials with powers (0.5, 0.5).

6.2 Linear splines

The slope of the curve (change in the odds of moderate to severe LUTS per 1 MET-hours/day increase in total physical activity) for the age-adjusted association is much steeper below 38 MET-hours/day when compared with higher covariate levels (see figure 7). For example, assume a simple linear trend for total physical activity where we allow the slope to change at 38 MET-hours/day. We then create a linear spline and fit the model, including both the original MET variable and the spline, to obtain a connected, piecewise-linear curve.

```
. generate tpa38p = max(tpa-38, 0)
. logit ipss2 tpa tpa38p agec
Iteration 0:   log likelihood = -16282.244
Iteration 1:   log likelihood =  -15535.79
Iteration 2:   log likelihood = -15520.305
Iteration 3:   log likelihood = -15520.299
Iteration 4:   log likelihood = -15520.299
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(3)      =    1523.89
                                                  Prob > chi2     =     0.0000
Log likelihood = -15520.299                       Pseudo R2       =     0.0468
------------------------------------------------------------------------------
       ipss2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tpa |  -.0902158   .0113921    -7.92   0.000    -.1125439   -.0678877
      tpa38p |    .072256   .0134459     5.37   0.000     .0459026    .0986094
        agec |   .0551684   .0015265    36.14   0.000     .0521766    .0581602
       _cons |   2.154075   .4203814     5.12   0.000     1.330142    2.978007
------------------------------------------------------------------------------
. xblc tpa tpa38p, covname(tpa) at(29 32 35 38 40 43 45 48 52 55) reference(29)
> eform
tpa           exp(xb)    (95% CI)
29            1.00       (1.00-1.00)
32            0.76       (0.71-0.82)
35            0.58       (0.51-0.67)
38            0.44       (0.36-0.54)
40            0.43       (0.35-0.52)
43            0.41       (0.34-0.49)
45            0.39       (0.33-0.47)
48            0.37       (0.31-0.45)
52            0.35       (0.29-0.41)
55            0.33       (0.27-0.39)
```

The above set of age-adjusted ORs computed with the `xblc` command, based on a linear spline model, is very similar to the one estimated with a more complicated right-restricted cubic-spline model (table 2). The advantage of the linear spline in this example is that it captures the most prominent features of the covariate–response association with just two parameters. The disadvantage is that the linear spline can be thrown off very far if the knot selected is poorly placed; that is, for a given number of knots, it is more sensitive to knot placement than to splines with power terms.

To express the linear trend for two-unit increases before and after the knot, we type

```
. lincom tpa*2, eform
 ( 1)  2*[ipss2]tpa = 0
```

ipss2	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.8349098	.0190227	-7.92	0.000	.7984461	.8730387

```
. lincom tpa*2 + tpa38p*2, eform
 ( 1)  2*[ipss2]tpa + 2*[ipss2]tpa38p = 0
```

ipss2	exp(b)	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.9647179	.0073474	-4.72	0.000	.9504242	.9792265

For every 2 MET-hours/day increase in total physical activity, the odds of moderate to severe LUTS significantly decrease by 17% below 38 MET-hours/day and by 4% above 38 MET-hours/day.

6.3 Fractional polynomials

The Stata command `mfp` (see [R] **mfp**) provides a systematic search for the best-fitting (likelihood maximizing) fractional-polynomial function (Royston, Ambler, and Sauerbrei 1999) for the quantitative covariates in the model.

```
. mfp logit ipss2 tpa agec, df(agec:1)
  (output omitted)
Fractional polynomial fitting algorithm converged after 2 cycles.
Transformations of covariates:
-> gen double Itpa__1 = X^-.5-.4908581303 if e(sample)
-> gen double Itpa__2 = X^-.5*ln(X)-.6985894219 if e(sample)
   (where: X = tpa/10)
-> gen double Iagec__1 = agec-1.46506e-07 if e(sample)
Final multivariable fractional polynomial model for ipss2
```

Variable	Initial			Final		
	df	Select	Alpha	Status	df	Powers
tpa	4	1.0000	0.0500	in	4	-.5 -.5
agec	1	1.0000	0.0500	in	1	1

```
Logistic regression                               Number of obs   =      30377
                                                  LR chi2(3)      =    1523.28
                                                  Prob > chi2     =     0.0000
Log likelihood = -15520.604                       Pseudo R2       =     0.0468
```

ipss2	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Itpa__1	-9.545011	3.716881	-2.57	0.010	-16.82996	-2.260058
Itpa__2	-25.4474	6.240732	-4.08	0.000	-37.67901	-13.21579
Iagec__1	.0555331	.0015258	36.40	0.000	.0525426	.0585237
_cons	-1.353425	.0178889	-75.66	0.000	-1.388487	-1.318363

```
Deviance:31041.208.
```

The algorithm found that the best transformation for total physical activity is a degree-2 fractional polynomial with equal powers (0.5, 0.5). To compute the ORs shown in table 2, we type

```
. xblc Itpa__1 Itpa__2, covname(tpa) at(29 32 35 38 40 43 45 48 52 55)
> reference(29) eform
tpa           exp(xb)    (95% CI)
29            1.00       (1.00-1.00)
32            0.69       (0.62-0.77)
35            0.53       (0.45-0.63)
38            0.45       (0.36-0.55)
40            0.41       (0.33-0.51)
43            0.37       (0.30-0.47)
45            0.36       (0.29-0.45)
48            0.35       (0.28-0.43)
52            0.34       (0.28-0.41)
55            0.35       (0.29-0.42)
```

The advantage of using fractional polynomials is that just one or two transformations of the original covariate can accommodate a variety of possible covariate–response relationships. The disadvantage is that the fitted curve can be sensitive to extreme values of the quantitative covariate (Royston, Ambler, and Sauerbrei 1999; Royston and Sauerbrei 2008).

Figure 8 provides a graphical comparison of the age-adjusted odds of moderate to severe LUTS obtained with the `xblc` command using the different modeling strategies discussed above.

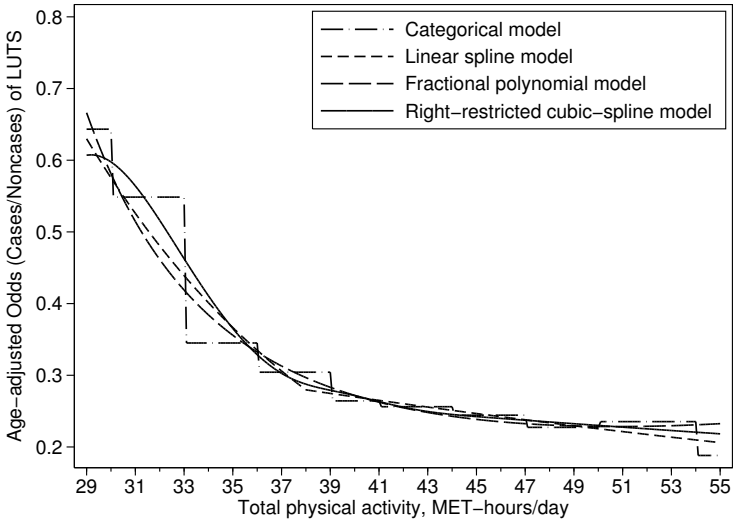

Figure 8. Comparison of covariate models (indicator variables, linear splines with a knot at 38 MET-hours/day, degree-2 fractional polynomial with powers [0.5, 0.5], right-restricted cubic-spline with four knots at percentiles 20%, 40%, 60%, and 80%) for estimating age-adjusted odds for the relation of total physical activity (MET-hours/day) to the occurrence of moderate to severe LUTS in a cohort of 30,377 Swedish men.

7 Conclusion

We have provided a new Stata command, `xblc`, to facilitate the presentation of the association between a quantitative covariate and the response variable. In the context of logistic regression with an emphasis on the use of different type of cubic splines, we illustrated how to present the odds or ORs with 95% confidence limits in tabular and graphical form.

The steps necessary to present the results can be applied to other types of models. The postestimation `xblc` command can be used after the majority of regression analysis (that is, generalized-linear models, quantile regression, survival-time models, longitudinal/panel-data models, meta-regression models) because the way of contrasting predicted responses is similar. The `xblc` command can be used to describe the relation of any quantitative covariate to the outcome using any type of flexible modeling strategy (that is, splines or fractional polynomials). If one is interested in plotting predicted

or marginal effects to a quantitative covariate, one can use the `postrcspline` package (Buis 2008). However, unlike the `xblc` command, the `postrcspline` command works only after fitting a restricted cubic-spline model.

Advantages of flexibly modeling a quantitative covariate include the ability to fit smooth curves efficiently and realistically. The fitted curves still need careful interpretation supported by subject-matter knowledge. Explanations for the observed shape may involve chance, mismeasurement, selection bias, or confounding rather than an effect of the fitted covariate (Orsini et al. 2008; Greenland and Lash 2008). For instance, in our unadjusted analysis, the OR for moderate to severe LUTS is not always decreasing with higher physical activity values. Once we adjust for age, this counterintuitive phenomenon disappears.

This example occurred in a large study in the middle of the exposure distribution where a large number of cases were located. Therefore, the investigator should be aware of the potential problems (instability, limited ability to predict future observations, and increased chance of overinterpretation and overfitting) with methods that can closely fit data (Steenland and Deddens 2004; Greenland 1995b; Royston and Sauerbrei 2007, 2009). Thus, as with any other strategy, subject-matter knowledge is needed when fitting regression models using flexible tools. Other important issues not considered here are how to deal with uncertainty due to model selection, how to assess goodness of fit, and how to handle zero exposure levels (Royston and Sauerbrei 2007, 2008; Greenland and Poole 1995).

In conclusion, the postestimation command `xblc` greatly facilitates the tabular and graphical presentation of results, thus aiding analysis and interpretation of covariate–response relations.

8 References

Buis, M. L. 2008. postrcspline: Stata module containing postestimation commands for models using a restricted cubic spline. Statistical Software Components S456928, Department of Economics, Boston College.
http://ideas.repec.org/c/boc/bocode/s456928.html.

Durrleman, S., and R. Simon. 1989. Flexible regression models with cubic splines. *Statistics in Medicine* 8: 551–561.

Greenland, S. 1995a. Avoiding power loss associated with categorization and ordinal scores in dose–response and trend analysis. *Epidemiology* 6: 450–454.

———. 1995b. Dose–response and trend analysis in epidemiology: Alternatives to categorical analysis. *Epidemiology* 6: 356–365.

———. 1995c. Previous research on power loss associated with categorization in dose–response and trend analysis. *Epidemiology* 6: 641–642.

———. 1995d. Problems in the average-risk interpretation of categorical dose–response analyses. *Epidemiology* 6: 563–565.

———. 2008. Introduction to regression models. In *Modern Epidemiology*, ed. K. J. Rothman, S. Greenland, and T. L. Lash, 3rd ed., 381–417. Philadelphia: Lippincott Williams & Wilkins.

Greenland, S., and T. L. Lash. 2008. Bias analysis. In *Modern Epidemiology*, ed. K. J. Rothman, S. Greenland, and T. L. Lash, 3rd ed., 345–380. Philadelphia: Lippincott Williams & Wilkins.

Greenland, S., K. B. Michels, J. M. Robins, C. Poole, and W. C. Willett. 1999. Presenting statistical uncertainty in trends and dose–response relations. *American Journal of Epidemiology* 149: 1077–1086.

Greenland, S., and C. Poole. 1995. Interpretation and analysis of differential exposure variability and zero-exposure categories for continuous exposures. *Epidemiology* 6: 326–328.

Harrell, F. E., Jr. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Harrell, F. E., Jr., K. L. Lee, and B. G. Pollock. 1988. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute* 80: 1198–1202.

Marrie, R. A., N. V. Dawson, and A. Garland. 2009. Quantile regression and restricted cubic splines are useful for exploring relationships between continuous variables. *Journal of Clinical Epidemiology* 62: 511–517.

Orsini, N., R. Bellocco, M. Bottai, A. Wolk, and S. Greenland. 2008. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *Stata Journal* 8: 29–48.

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.

Royston, P., D. G. Altman, and W. Sauerbrei. 2006. Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine* 25: 127–141.

Royston, P., G. Ambler, and W. Sauerbrei. 1999. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* 28: 964–974.

Royston, P., and W. Sauerbrei. 2007. Multivariable modeling with cubic regression splines: A principled approach. *Stata Journal* 7: 45–70.

———. 2008. *Multivariable Model-building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modelling Continuous Variables*. Chichester, UK: Wiley.

———. 2009. Bootstrap assessment of the stability of multivariable models. *Stata Journal* 9: 547–570.

Smith, P. L. 1979. Splines as a useful and convenient statistical tool. *American Statistician* 33: 57–62.

Steenland, K., and J. A. Deddens. 2004. A practical guide to dose–response analyses and risk assessment in occupational epidemiology. *Epidemiology* 15: 63–70.

Wegman, E. J., and I. W. Wright. 1983. Splines in statistics. *Journal of the American Statistical Association* 78: 351–365.

About the authors

Nicola Orsini is a researcher in the Division of Nutritional Epidemiology at the National Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.

Sander Greenland is a professor of epidemiology at the UCLA School of Public Health and a professor of statistics at the UCLA College of Letters and Science, Los Angeles, CA.