Splines with parameters that can be explained in words to non-mathematicians

Roger Newson (King's College, London, UK)

`roger.newson@kcl.ac.uk`

- What splines are, and their uses.
- Including splines in generalized linear models
- Choosing a basis: plus-functions, Schoenberg B-splines, and reference splines
- The programs `bspline` and `frencurv`
- A demonstration (using the `auto` data)

What splines are

- A k**th-degree spline** is a function from the x-axis to the y-axis, defined using an ascending sequence of **knots** $s_0 < s_1 < \ldots < s_q$ on the x-axis.

- (Typically, the sequence of knots is assumed to be part of an extended sequence of form $\ldots s_{-1} < s_0 < \ldots < s_q < s_{q+1} \ldots$ extending outwards to $\pm\infty$.)

- In each interval $s_j \leq x < s_{j+1}$ between two successive knots, the spline is equal to a kth degree polynomial.

- At each knot s_j, the first $k-1$ derivatives of the spline are continuous.

Therefore, a spline of degree 0 is a step function, a spline of degree 1 is linearly interpolated between the knots, and splines of degree 2, 3 and higher are interpolated as curves.

Typical uses of splines

- Splines are used *mostly* for adjusting for "uninteresting" trends while measuring "interesting" trends.

- For instance, we once observed 297 asthmatic patients for the 821 days from 1 April 1997 to 30 June 1999, and recorded whether or not each patient visited a doctor on each day on account of respiratory problems. We also measured pollution levels on each day.

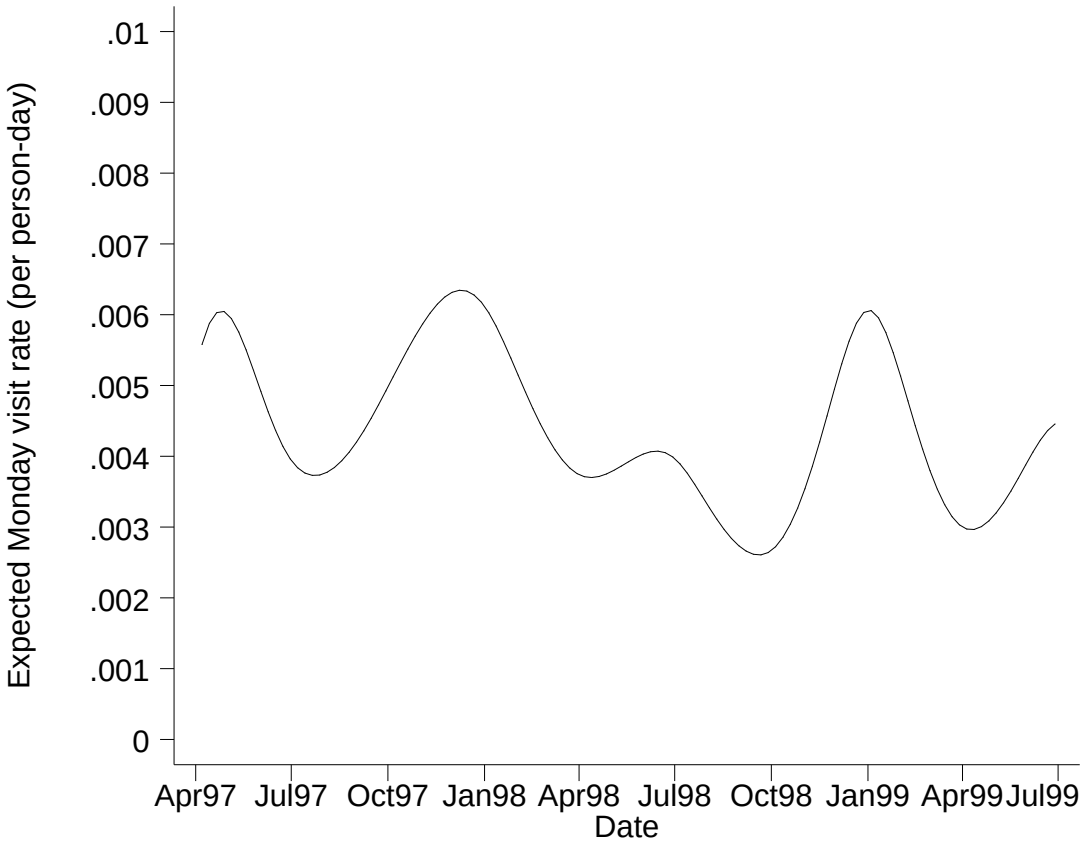
- We measured the effects of pollution using a logit model, including a cubic spline (with one knot per quarter year) to model seasonal and longer-term trends, and odds ratios for days of the week and for elevated pollution levels.

- The spline therefore represents the log odds of a visit expected on each date, if that date had been a Monday with a low pollution level ("uninteresting epidemics").

- The odds ratios represent the weekly cycle of visit rates, and also the elevated visit rates arising from pollution ("interesting epidemics").

Inverse logit-transformed spline representing expected Monday visit rates

- The outcome variable is the event that a patient visited his/her doctor with a respiratory complaint.
- The baseline model contains a cubic spline over time, together with odds ratios for days of the week other than Monday.
- The graph plots the fitted spline (vertical axis) against time (horizontal axis).
- The spline represents the visit rate per patient-day expected on a particular date, had that date been a Monday.



Including a spline in a generalized linear model

- In a generalized linear model, the conditional mean (or its link function) is equal to the spline plus a sum of "more interesting" terms.
- The spline itself is represented as a linear combination of a set of elementary splines, known as a **basis**.
- The basis of splines is included in the columns of the design matrix as predictor variates. The corresponding fitted parameters define the fitted spline.
- *Unfortunately*, most bases of splines give rise to parameters not easily explained in words to non-mathematicians.
- The standard solution to this problem is to pretend that these parameters do not exist. (This is sometimes done by using jargon such as "non-parametric regression".)

Plus-functions

- The k**th degree plus-function originating at a knot** s is defined as

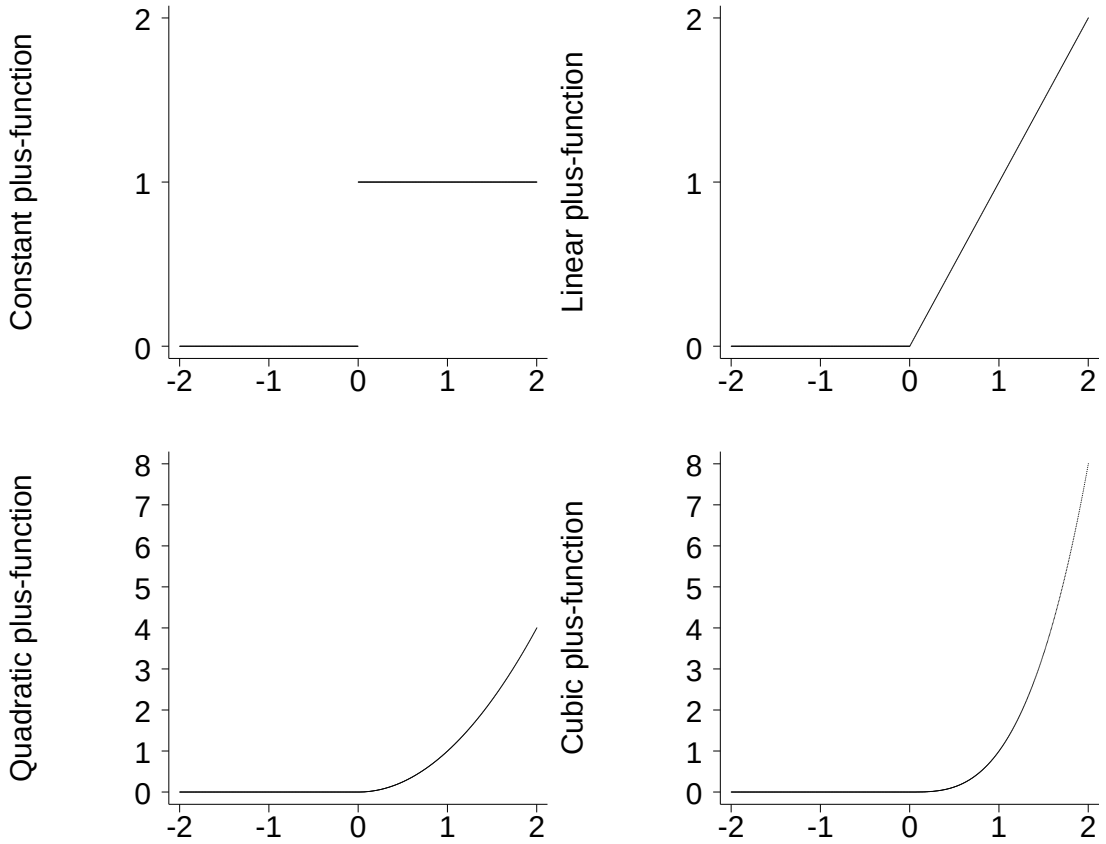
$$P_k(x;s) = \begin{cases} (x-s)^k, & x \geq s, \\ 0, & x < s. \end{cases}$$

- Given an infinite sequence of knots $\ldots s_0, s_1, s_2, \ldots$, any kth degree spline based on these knots is a linear combination of the kth degree plus-functions originating at the knots.

- The values of a kth degree spline in the interval $s_j \leq x < s_{j+1}$ between two successive knots depend only on plus-functions originating at s_j and at the k knots *immediately* to the left.

Constant, linear, quadratic and cubic plus-functions originating at zero

- The constant plus-function is a step function, whereas the others tend to infinity.
- A spline with unit knots is a linear combination of plus-functions originating at the unit knots.
- Parameters corresponding to plus-functions are subject to instability problems when the model is fitted.
- Also, they are expressed in y-axis units per k'th power x-axis unit. (Not easy to explain!)



The Schoenberg B-spline basis

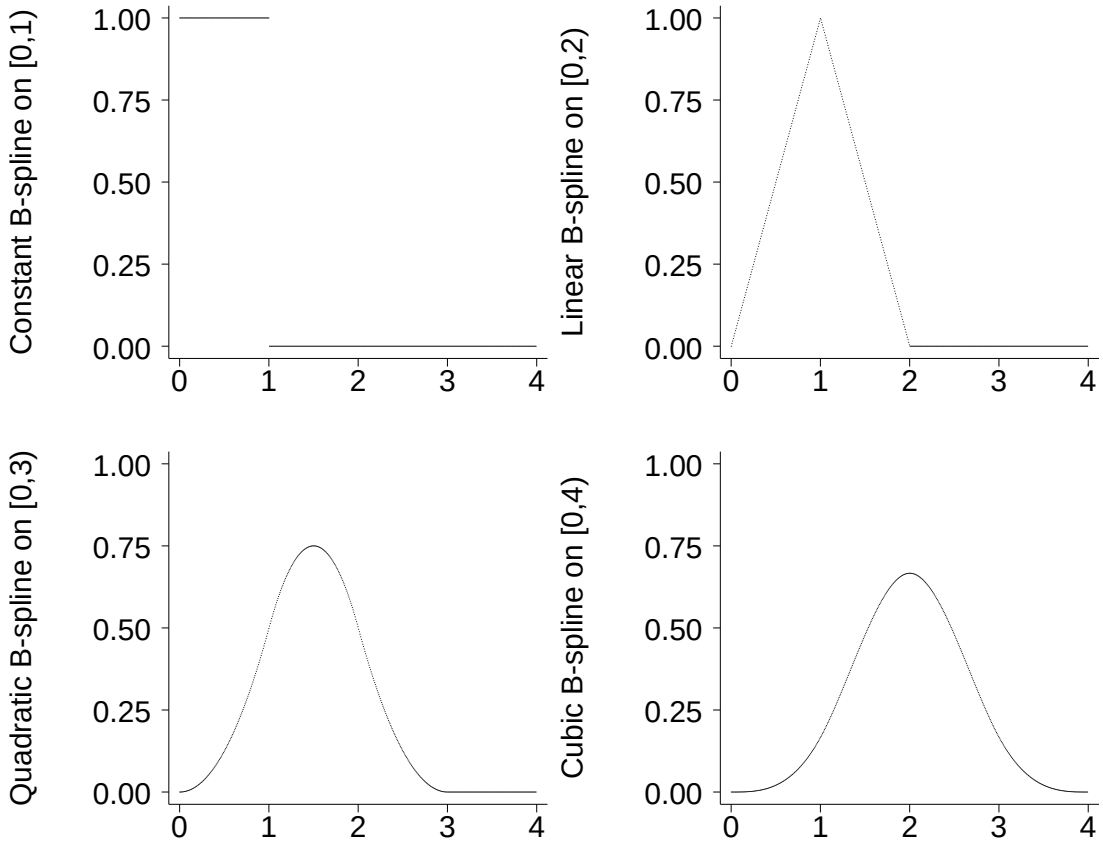
- In the 1960s, I. J. Schoenberg introduced the **B-spline basis** to solve the instability problems associated with plus-functions.

- Each **kth degree B-spline** is positive in an interval bounded by $k+2$ successive knots $s_h < s_{h+1} < \ldots < s_{h+k+1}$, and zero elsewhere. It is defined in terms of the plus-functions originating at these $k+2$ successive knots.

- *Therefore*, the values of a kth degree spline in the interval $s_j \leq x < s_{j+1}$ between two successive knots depend only on $k+1$ successive local B-splines. These originate at s_j and at the k knots immediately to the left, and terminate at s_{j+1} and at the k knots immediately to the right.

Constant, linear, quadratic and cubic B-splines originating at zero with unit knots

- A kth degree B-spline is positive in an interval bounded by $k+2$ successive knots, and zero elsewhere.
- The component B-splines are therefore local in their effects on the spline as a whole.
- This solves the instability problems associated with plus-functions.
- *However*, the parameters are still not easy to explain in words to a non-mathematician.



The reference spline (or "French curve") basis

- Given a set of knots, a **reference spline basis** is defined as a linear transformation of the B-spline basis for the same knots.

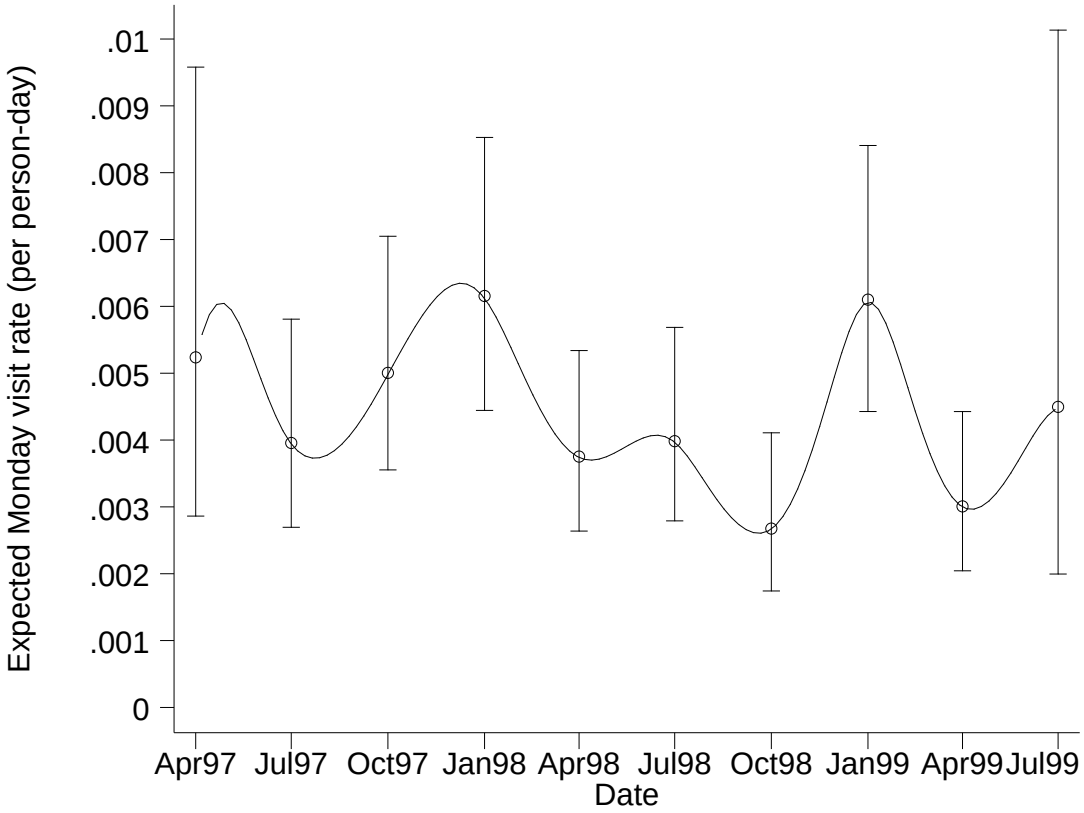
- Each reference spline is centred on a **reference point** on the x-axis.

- The fitted parameter corresponding to each reference spline is simply the value of the spline at its appropriate reference point.

- The spline is therefore like a line drawn using a French curve, interpolated between the values of the spline at the reference points.

Inverse logit-transformed spline representing expected Monday visit rates

- The reference points on the time axis are the first days of each quarter year.
- The parameters of the spline are the visit rates expected on the first day of each quarter year, had that day been a Monday.
- The values of the spline in between reference points are interpolated as if we had used a French curve.



How to choose the reference points?

There are many ways of choosing knots and/or reference points, none of which is uniquely "right". However, a "sensible default" is for the reference points to be:

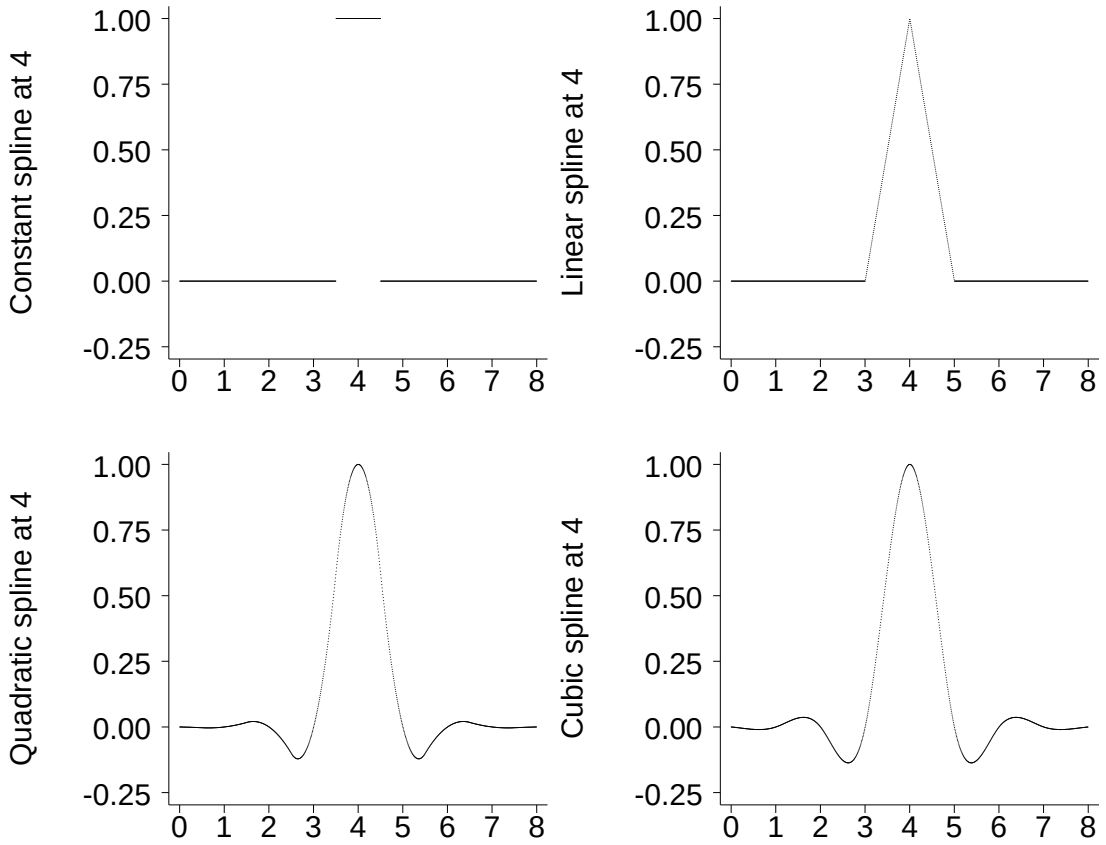
- A subset of the knots themselves, in the case of an odd-degree spline (eg linear, cubic, or quintic);

- A subset of the midpoints between successive knots, in the case of an even-degree spline (eg constant, quadratic, or quartic).

This default has the consequence that, if we fit a quadratic or cubic spline to an interval on the x-axis bounded by an ascending set of reference points $r_0 < r_1 < \ldots < r_q$, then there will be a need for two "extra" reference points $r_{-1} < r_0$ and $r_{q+1} > r_q$. The parameters corresponding to these reference points represent the spline "behaving badly off the edge of the paper".

Constant, linear, quadratic and cubic reference splines at 4 with unit reference points

- A kth degree reference spline is centred on a reference point on the x-axis (in this case 4).
- It is equal to 1 at its own reference point, and 0 at all the other reference points.
- *Therefore*, the corresponding parameter of the fitted model is simply the value of the *total* spline at the reference point (in y-axis units).



The programs `bspline` and `frencurv` (insert sg151 in STB-57)

- `bspline` generates a basis of Schoenberg B-splines corresponding to a sequence of knots on the x-axis.

- `frencurv` generates a basis of reference splines corresponding to a sequence of reference points on the x-axis.

- The user can then use a regression program, usually with the `noconst` option, including the generated spline basis in the list of predictor variates.

- Both programs have a "sensible default" method for adding "extra" knots and/or reference points. However, the advanced user can choose to override these defaults "intelligently".

Example: Weight and fuel consumption in US and non-US cars (1)

In the `auto` data set, we use `frencurv` to generate a basis of cubic reference splines in the x-axis variable `weight`, as follows:

```
. frencurv,xvar(weight) refpts(1760 3300 4840) gene(cs) power(3)
. describe cs*
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
cs1             float  %8.4f                  Spline at 220 (INCOMPLETE)
cs2             float  %8.4f                  Spline at 1,760
cs3             float  %8.4f                  Spline at 3,300
cs4             float  %8.4f                  Spline at 4,840
cs5             float  %8.4f                  Spline at 6,380 (INCOMPLETE)
```

By default, `frencurv` has added two extra reference points (at 220 and 6380 pounds) outside the interval bounded by the original `refpts` supplied by the user. The splines for these reference points have variable labels specifying that the spline is "incomplete" at these reference points.

Example: Weight and fuel consumption in US and non-US cars (2)

We then fit a regression model, using the `noconst` option, using the spline to model expected mileage of US cars as a function of weight, together with a difference in mileage for non-US cars of the same weight:

```
. regress mpg cs* foreign,noconst robust
Regression with robust standard errors                Number of obs =      74
                                                      F(  6,    68) =  814.30
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.9792
                                                      Root MSE      =  3.3148
------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         cs1 |   37.23059   12.34003     3.02   0.004     12.60644    61.85475
         cs2 |   32.33655   1.694586    19.08   0.000     28.95506    35.71805
         cs3 |   18.96442   .5237823    36.21   0.000     17.91923    20.00961
         cs4 |   12.08867   .6771901    17.85   0.000     10.73735    13.43998
         cs5 |  -2.428254   10.94945    -0.22   0.825    -24.27755    19.42104
     foreign |  -2.222764   .9946763    -2.23   0.029    -4.207609   -.2379179
------------------------------------------------------------------------------
```

Example: Weight and fuel consumption in US and non-US cars (3)

Using the program `parmest` (available by typing `webseek parmest`), we can list the regression parameters as follows:

```
. list parm label estimate min95 max95 p,noobs
```

parm	label	estimate	min95	max95	p
cs1	Spline at 220 (INCOMPLETE)	37.23	12.61	61.85	0.0036
cs2	Spline at 1,760	32.34	28.96	35.72	0.0000
cs3	Spline at 3,300	18.96	17.92	20.01	0.0000
cs4	Spline at 4,840	12.09	10.74	13.44	0.0000
cs5	Spline at 6,380 (INCOMPLETE)	-2.43	-24.28	19.42	0.8252
foreign	Car type	-2.22	-4.21	-0.24	0.0287

The parameters `cs1` and `cs5` have "crazy" confidence intervals, corresponding to values of the spline "off the edge of the paper". However, the parameters `cs2`, `cs3` and `cs4` have "sensible" values, equal to the expected mileage of US-made cars weighing 1760, 3300 and 4840 pounds. The bottom line states that "foreign" cars typically travel 0.24 to 4.21 *fewer* miles per gallon than US-made cars of the same weight.

Further reading

de Boor C. 1978. *A practical guide to splines*. New York: Springer Verlag.

Newson, R. 2000. sg151: *B*-splines and splines parameterized by their values at reference points on the x-axis. *Stata Technical Bulletin* 57: 20-27.