fp_select: model selection for univariable fractional polynomials

Patrick Royston
MRC Clinical Trials Unit at UCL
Aviation House
125 Kingsway
London WC2B 6NH
UK.
Phone: +44 (0) 20 7670 4736
Email: j.royston@ucl.ac.uk

20 February 2017
Summary

Since their original publication in 1994, fractional polynomials have steadily gained popularity as a tool for flexible parametric modeling of regression relationships. Here, I present \texttt{fp\_select}, a post-estimation tool for \texttt{fp} which allows the user to select a parsimonious fractional polynomial model according to a closed test procedure known as the fractional polynomial selection procedure or FSP. A brief introduction to fractional polynomial models is given, and examples of the use of \texttt{fp} and \texttt{fp\_select} to select such models with real data are provided.

1 Introduction

Since the original publication in 1994 (Royston & Altman (1994)), fractional polynomials (FPs) have steadily gained popularity as a tool for flexible parametric modeling of regression relationships in both univariable and multivariable settings. A recent enquiry in Google Scholar (17 January 2017) yielded 1289 citations of Royston & Altman (1994) to date. For those unfamiliar with FPs, I provide a very brief introduction below. For a much wider view, please see Royston & Sauerbrei (2008), and the MFP website at http://mfp.imbi.uni-freiburg.de and many other articles cited therein.

Concisely stated, a fractional polynomial (FP) is a special type of polynomial that may include logarithms, noninteger powers, and repeated powers. Each time a power repeats in an FP function of $x$, it is multiplied by another $\ln (x)$. One may write an FP in $x$ as

$$x^{(p_1,p_2,...,p_m)} \beta$$

where the positive integer $m$ is known as the degree or dimension of the FP. For example, an FP in $x$ with powers $(-1,0,0.5,3,3)$ and coefficients $\beta$ has the following form:

$$x^{(-1,0.5,3,3)} \beta = \beta_1 x^{-1} + \beta_2 \ln (x) + \beta_3 x^{0.5} + \beta_4 x^3 + \beta_5 x^3 \ln (x)$$

In the above example, the dimension of the FP is $m = 5$.

Despite their somewhat dry definition, FPs are not just a mathematical abstraction. With a suitable range of powers, it turns out that they provide a considerable range of functional forms in $x$ that are useful in regression models of real data. The default set of powers from which FP powers are selected is $\{-2,-1,-0.5,0,0.5,1,2,3\}$, 0 signifying log (i.e. \texttt{ln}). In practice, even FP1 and FP2 functions (FPs of dimension 1 and 2) offer much more flexibility than polynomials of the same degree, that is, than linear or quadratic functions. See for example Figure 1, which
Figure 1: Examples of some of the functional forms available with FP2 functions with various powers \((p_1, p_2)\).

shows schematically some FP2 functions with various powers \((p_1, p_2)\) and coefficients \((\beta_1, \beta_2)\).

The aim of flexible regression models for a single continuous covariate \(x\) is to provide a succinct and accurate approximation of the relationship between \(x\) and a response \(y\) without resorting to ‘categorization’ (discretization) of the covariate into groups. Further material on FPs, including a discussion of the pitfalls of categorization and the motivation and potential advantages of FPs, with a real example, may be found at http://mfp.imbi.uni-freiburg.de/fp.

Univariable FP regression models have been available in official Stata for two decades, following the release of the `fracpoly` command in Stata 5 (1997). Following a ground-up rewrite, the current official implementation of univariable FPs as `fp` appeared in Stata 12 (2011). Utilising a revised command syntax and FP search algorithm, `fp` extended the types of regression model in which FPs could be fit.

An important concept in flexible regression modeling is parsimony: the need to remove ‘dead wood’ from a model, mainly to avoid overfitting and improve the interpretability of the selected model. An example of dead wood in univariable FP modeling would be the inclusion of high-dimensional FP terms not supported by the data. Such terms would likely produce ‘wiggly’ fitted curves that exhibit uninterpretable local features (recognised as an issue also when fitting standard polynomials of high dimension). For an example of the curve instability that can result
with overfitted spline models (another type of smoother), see Royston & Sauerbrei (2008) Figure 3.3.

With FP modeling, a technique known as the function selection procedure or FSP is available. The aim of the FSP is, if possible, to simplify an FP model to one of lesser complexity by appropriate statistical testing. In this article, I outline how the FSP works, and introduce a new fp post-estimation command, fp-select, that implements the FSP. I illustrate the use of fp-select in an example with real data.

2 The function selection procedure (FSP)

An important (default) option of fp is compare. The table of FP model comparisons that is presented with compare contains all the elements needed to select a preferred model according to the FSP, an ordered sequence of hypothesis tests. The FSP has the flavor of a closed test procedure (Marcus, Peritz & Gabriel (1976)) that (approximately) protects the ‘familywise’ type 1 error probability for selecting an FP transformation of $x$ at some nominal value, $\alpha$, such as 0.05. For further details of the closed test aspect, see the description of the FSP in Ambler & Royston (2001), there called ‘procedure RA2’. Although fp (and fracpoly) supply the necessary information on which the FSP operates, neither program actually indicates which model the FSP would choose at a given $\alpha$ level.

The FSP starts with an FP model of maximal allowed complexity, defined by its dimension, $m_0$ say. By default in fp, $m_0 = 2$, that is an FP2 (fractional polynomial of dimension or degree 2). The FSP attempts to simplify the model to an FP1 or linear function of $x$ by applying a specific sequence of tests. The sequence of tests for $m_0 = 2$ is described under the heading ‘Methods of FP model selection’ in the Stata manual entry for mfp. See also Royston & Sauerbrei (2008), pp. 82-84.

In general terms, the FSP comprises two parts. The maximum permissible FP degree, $m_0$, is chosen by the analyst a priori and is usually 2. The first part of the FSP is a test for inclusion of an FP-transformed continuous covariate $x$ in the model. Let us call the corresponding significance level $\alpha_{select}$. By convention, if $\alpha_{select} = 1$, no test is done and $x$ (possibly FP-transformed) is included in the model anyway, the final choice of the functional form being determined by the subsequent steps of the FSP. If $\alpha_{select} < 1$, the best-fitting FP$m_0$ model is tested against the model omitting $x$ on $2m_0$ degrees of freedom (d.f.) at significance level $\alpha_{select}$. If the test is significant, the algorithm continues as described below, otherwise $x$ is ‘omitted’ (taken as uninfluential), and the procedure ends.

Let the critical significance level for tests of functional form in the FSP be $\alpha$ (0 < $\alpha$ < 1). Assuming the inclusion test at level $\alpha_{select}$ is ‘passed’, the remaining steps for general $m_0 \geq 1$ are as follows.
1. Test $FP_{m_0}$ against linear (a straight line) in $x$ on $2m_0 - 1$ d.f. at level $\alpha$. If significant, continue; otherwise, stop, with the chosen model for $x$ being a straight line.

2. If $m_0 > 1$, test $FP_{m_0}$ against $FP_1$ on $2(m_0 - 1)$ d.f. at level $\alpha$. If significant, continue; otherwise, stop, with the chosen model for $x$ being $FP_1$.

3. If $m_0 > 2$, test $FP_{m_0}$ against $FP_2$ on $2(m_0 - 2)$ d.f. at level $\alpha$. If significant, continue; otherwise, stop, with the chosen model for $x$ being $FP_2$.

4. Continue in like manner until the test of $FP_{m_0}$ against $FP(m_0 - 1)$ is reached. If significant, the selected model is $FP_{m_0}$; otherwise, it is $FP(m_0 - 1)$. End of procedure.

In some situations, one might have reason to vary the significance levels $\alpha_{\text{select}}$ and $\alpha$, the two ‘tuning’ constants of the FSP. In an observational study, for example, where possible overfitting of the variables in a confounder model is not necessarily a critical issue, one might choose $\alpha_{\text{select}} = 1$ and $\alpha = 0.2$ to select the functional form for a continuous confounder.

3 Example

3.1 Data and preliminary analysis

As an example, I use the IgG data (Isaacs, Altman, Tidmarsh, Valman & Webster (1983)), which may be loaded into Stata by using the command `webuse igg`. The aim is to model $y = \sqrt{\text{igg}}$, the square root of the serum immunoglobulin-G (IgG) concentration in 298 children as a function of $x = \text{age}$, a child’s age in years. The reasons to square-root transform the response are for variance stabilization and normalization of the residuals.

The pale line in Figure 2 is a smoothed scatter plot of $y$ against $x$. [FIGURE 2 NEAR HERE] The solid line is a local polynomial fit created by Stata’s `lpolyci` graph subcommand with a relatively narrow bandwidth of 0.2, hence the rather ‘wiggly’ curve. Nevertheless, a visual indication of nonlinearity is present. The dashed line is the best-fitting $FP_2$ curve, computed by the `fpfit` subcommand. The commands which created the figure are as follows:

```stata
webuse igg, clear
set scheme sj
graph twoway (lpolyci sqrtigg age, bwidth(0.2)) (fpfit sqrtigg age) ///
(scatter sqrtigg age, msymbol(o) msize(*0.75))
```

A biological argument suggests that since IgG is a blood protein reflecting the maturity of the immune system from birth on, the underlying curve should be monotone increasing. The fitted $FP_2$ curve is in fact monotone. It indicates a rapid rise in IgG in the youngest children followed by a gentler rate of increase subsequently. By contrast, the ‘nonparametric’ local
Figure 2: IgG data with local polynomial and fractional polynomial smoothing.
polynomial fit is nonmonotone, with local features that evidently are present in the data but are unlikely to be real in the population.

3.2 FP model selection

I now consider FP model selection for the IgG dataset. Below I show the output from running \texttt{fp} with the default \texttt{dimension(2)} setting.

\begin{verbatim}
. fp <age> : regress sqrtigg <age>
(fitting 44 models)
(...10%...20%...30%...40%...50%...60%...70%...80%...90%...100%)

Fractional polynomial comparisons:

\begin{verbatim}
<table>
<thead>
<tr>
<th>age</th>
<th>df</th>
<th>Deviance</th>
<th>Res. s.d.</th>
<th>Dev. dif.</th>
<th>P(*)</th>
<th>Powers</th>
</tr>
</thead>
<tbody>
<tr>
<td>omitted</td>
<td>0</td>
<td>427.539</td>
<td>0.497</td>
<td>108.090</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>linear</td>
<td>1</td>
<td>337.561</td>
<td>0.428</td>
<td>18.113</td>
<td>0.000</td>
<td>1</td>
</tr>
<tr>
<td>m = 1</td>
<td>2</td>
<td>327.436</td>
<td>0.421</td>
<td>7.987</td>
<td>0.020</td>
<td>0</td>
</tr>
<tr>
<td>m = 2</td>
<td>4</td>
<td>319.448</td>
<td>0.416</td>
<td>0.000</td>
<td>--</td>
<td>-2 2</td>
</tr>
</tbody>
</table>
\end{verbatim}

(*) P = sig. level of model with m = 2 based on F with 293 denominator df.

\begin{verbatim}
Source | SS    | df | MS     | Number of obs = 298
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>22.2846976</td>
<td>2</td>
<td>11.1423488</td>
<td>F(2, 295) = 64.49</td>
</tr>
<tr>
<td>Residual</td>
<td>50.9676492</td>
<td>295</td>
<td>.172771692</td>
<td>R-squared = 0.3042</td>
</tr>
<tr>
<td>Total</td>
<td>73.2523469</td>
<td>297</td>
<td>.246640898</td>
<td>Root MSE = .41566</td>
</tr>
</tbody>
</table>

sqrtigg | Coef. | Std. Err. | t | P>|t| | [95%, Conf. Interval] |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>age_1</td>
<td>-.1562156</td>
<td>.027416</td>
<td>-5.70</td>
<td>0.000</td>
<td>-.2101713</td>
</tr>
<tr>
<td>age_2</td>
<td>.0148405</td>
<td>.0027767</td>
<td>5.34</td>
<td>0.000</td>
<td>.0093757</td>
</tr>
<tr>
<td>_cons</td>
<td>2.189242</td>
<td>.0473835</td>
<td>46.20</td>
<td>0.000</td>
<td>2.095989</td>
</tr>
</tbody>
</table>
\end{verbatim}

As seen for the entry for m = 2 in the table headed \textit{Fractional polynomial comparisons}, the best-fitting FP2 powers of age are (−2, 2). This FP2 transformation of age is represented
by the two variables \texttt{age\_1} and \texttt{age\_2} that appear in the table of regression estimates.

Although the results are suggestive, the output is not explicit as to whether an FP2 model is really needed, or whether, at significance level 0.05, a simpler model (FP1 or linear) would suffice. Using \texttt{fp\_select} (described in section 4) with \( \alpha_{\text{select}} = \alpha = 0.05 \) immediately after \texttt{fp}, the following result is obtained:

\begin{verbatim}
. fp\_select, alpha(0.05) select(0.05)
\end{verbatim}

selected FP model: powers = (-2 2), df = 4

The output confirms that when \( m_0 = 2 \), an FP2 model is selected at the 0.05 significance level. The selected model can be fit as follows using results (best FP powers) stored by \texttt{fp\_select} in \texttt{r(powers)}:

\begin{verbatim}
. fp <age>, fp('r(powers)') replace: regress sqrtigg <age>
\rightarrow regress sqrtigg age\_1 age\_2
\end{verbatim}

\begin{verbatim}
Source | SS df MS Number of obs = 298
-------------+---------------------------------- F(2, 295) = 64.49
Model | 22.2846976 2 11.1423488 Prob > F = 0.0000
Residual | 50.9676492 295 .172771692 R-squared = 0.3042
-------------+---------------------------------- Adj R-squared = 0.2995
Total | 73.2523469 297 .246640898 Root MSE = .41566

sqrtigg | Coef. Std. Err. t P>|t| [95% Conf. Interval]
-------------+----------------------------------------------------------------
age\_1 | -.1562156 .027416 -5.70 0.000 -.2101713 -.10226
age\_2 | .0148405 .0027767 5.34 0.000 .0093757 .0203052
_cons | 2.189242 .0473835 46.20 0.000 2.095989 2.282495
\end{verbatim}

Since in this case the FP2 model was not simplified by \texttt{fp\_select}, the result is the same as that reported by \texttt{fp} for the default \( m_0 = 2 \) model.

### 3.3 Impact of complexity on model selection

Let us see what happens if a more complex model with \( m_0 = 4 \) is taken as the starting point for model selection:

\begin{verbatim}
. fp <age>, dimension(4) replace : regress sqrtigg <age>
\end{verbatim}
(fitting 494 models)
(....10%....20%....30%....40%....50%....60%....70%....80%....90%....100%)

Fractional polynomial comparisons:

<table>
<thead>
<tr>
<th>age</th>
<th>df</th>
<th>Deviance</th>
<th>Res. s.d.</th>
<th>Dev. dif.</th>
<th>P(*)</th>
<th>Powers</th>
</tr>
</thead>
<tbody>
<tr>
<td>omitted</td>
<td>0</td>
<td>427.539</td>
<td>0.497</td>
<td>109.795</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td>linear</td>
<td>1</td>
<td>337.561</td>
<td>0.428</td>
<td>19.818</td>
<td>0.007</td>
<td>1</td>
</tr>
<tr>
<td>m = 1</td>
<td>2</td>
<td>327.436</td>
<td>0.421</td>
<td>9.692</td>
<td>0.149</td>
<td>0</td>
</tr>
<tr>
<td>m = 2</td>
<td>4</td>
<td>319.448</td>
<td>0.416</td>
<td>1.705</td>
<td>0.798</td>
<td>-2 2</td>
</tr>
<tr>
<td>m = 3</td>
<td>6</td>
<td>319.275</td>
<td>0.416</td>
<td>1.532</td>
<td>0.476</td>
<td>-2 1 1</td>
</tr>
<tr>
<td>m = 4</td>
<td>8</td>
<td>317.744</td>
<td>0.416</td>
<td>0.000</td>
<td>--</td>
<td>0 3 3 3</td>
</tr>
</tbody>
</table>

(*) P = sig. level of model with m = 4 based on F with 289 denominator dof.

```
Source | SS       | df   | MS       | Number of obs = 298
-------|----------|------|----------|-------------------|
Model  | 22.5754541 | 4    | 5.6436353 | F(4, 293) = 32.63 |
Residual | 50.6768927 | 293  | .17295878 | Prob > F = 0.0000 |
Total   | 73.2523469 | 297  | .246640898 | Root MSE = .41588 |
```

```
sqrtigg | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
--------|-------|-----------|---|------|---------------------|
age_1   | .8761824 | .1898721 | 4.61 | 0.000 | .5024962 | 1.249869 |
age_2   | -.1922029 | .0684934 | -2.81 | 0.005 | -.3270044 | -.0574015 |
age_3   | .2043794 | .074947 | 2.73 | 0.007 | .0568767 | .3518821 |
age_4   | -.0560067 | .0212969 | -2.63 | 0.009 | -.097921 | -.0140924 |
_cons   | 2.240866 | .1019331 | 21.98 | 0.000 | 2.040252 | 2.44148 |
```

The $m_0 = 4$ model has powers $(0, 3, 3, 3)$. Next, model selection is applied:

```
. fp_select , alpha(0.05) select(0.05)
```

selected FP model: powers = $(0)$, df = 2

Instead of FP2, the selected model is now an FP1 with power $(0)$, that is $\beta_0 + \beta_1 \ln(x)$. 

8
Comparisons with Maximum FP complexity, $m_0$

<table>
<thead>
<tr>
<th>FP$m_0$ model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not in model</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Linear</td>
<td>0.002</td>
<td>0.000</td>
<td>0.003</td>
<td>0.007</td>
</tr>
<tr>
<td>$m = 1$</td>
<td>–</td>
<td>0.020</td>
<td>0.092</td>
<td>0.149</td>
</tr>
<tr>
<td>$m = 2$</td>
<td>–</td>
<td>–</td>
<td>0.919</td>
<td>0.798</td>
</tr>
<tr>
<td>$m = 3$</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.476</td>
</tr>
</tbody>
</table>

Selected model $m = 1$  $m = 2$  $m = 1$  $m = 1$

Table 1: $p$-values and selected models arising from FP model comparisons with the IgG data.

Table 1 shows $p$-values from the FSP with increasing maximum complexity. Taking $\alpha_{\text{select}} = \alpha = 0.05$, model comparisons in the FSP pathways for $m_0 = 1, 2, 3, 4$ are presented. [TABLE 1 NEAR HERE]

For all four values of $m_0$, the test of FP$m_0$ against $x$ ‘Not in model’ is highly significant ($p < 0.0005$)—see the third row of Table 1. This confirms that $\text{sqrtigg}$ is associated with $\text{age}$. All tests of FP$m_0$ against linear (fourth row) are also significant, providing evidence that the relationship is nonlinear.

With maximum complexity $m_0 = 2$, the test of FP2 against FP1 is significant at the 5% level, resulting in the selection of an FP2 model (as already seen). This is not the case for $m_0 = 3$ and $m_0 = 4$, where an FP1 model is chosen instead. However, there is no evidence that more complex models with dimension 3 or 4 fit better than FP2. For example, a test of $m = 3$ against $m = 2$ has $p = 0.919$, and a test of $m = 4$ against $m = 2$ has $p = 0.798$ (see Table 1).

The reason why an FP1 function, rather than an FP2 function, is selected when $m_0 > 2$ is presumably an increase in the type 2 error probability (i.e. reduced statistical power) due to redundant parameters being estimated in the models with dimension greater than 2. See Royston & Sauerbrei (2008) section 4.16 for further discussion of the power issue.

4 The fp_select command

4.1 Syntax

The syntax of fp_select is as follows:

\[ \text{fp_select, alpha(#)} [\text{select(#)}] \]

You must run fp to fit FP models before using fp_select.
4.2 Description

Taking the results from the most recent run of `fp`, `fp_select` tries to simplify the most complex reported FP model by applying an ordered sequence of significance tests. The aim is to reduce possible overfitting. The sequence, known as the FSP, approximates a closed test procedure. See the foregoing sections for further details.

4.3 Options

`alpha(#)` (required) defines the significance level for testing less complex models against the most complex FP model that was fitted, FP$_{m_0}$. A typical value of # might be 0.05 or 0.01.

`select(#)` defines the significance level for testing whether the covariate is influential. Specifically, if $m_0$ is the dimension (degree) of the most complex fitted FP model, the test is of FP$_{m_0}$ against the ‘null’ model that omits the covariate. If the covariate is not significant at level # < 1, the procedure terminates. Otherwise, testing continues. Default value of # is 1, meaning that the selection test is not performed and the covariate is automatically included.

4.4 Examples

Fit default FP2 model:

```
. webuse igg
. fp <age>: regress sqrtigg <age>
. fp_select, select(0.05) alpha(0.05)
. display "r(powers)"
```

Fit a more complex FP model:

```
. fp <age>, dimension(4) replace: regress sqrtigg <age>
. fp_select, alpha(0.2)
. display "r(powers)"
```

A multi-equation example:

```
. sysuse auto
. fp <weight>: sureg (price foreign <weight> length) (mpg foreign <weight>)
   > (displ foreign <weight>)
. fp_select, select(0.05) alpha(0.05)
. display "r(powers)"
```


5 Comments

\texttt{fp\_select} fills a gap in the ability of \texttt{fp} to select a parsimonious model. It removes the need to use \texttt{mfp} (searching on one continuous covariate) to select such a model. Note that \texttt{fp} requires that a model return a log likelihood, whereas \texttt{mfp} can fit some additional models (see help on \texttt{mfp}).

Acknowledgments

I thank a reviewer and Dr Tim Morris whose helpful comments prompted me to widen the scope of the article and improve its presentation.

References


About the author

Patrick Royston is a biostatistician with 40 years of experience, with a strong interest in biostatistical methods and in statistical computing and algorithms. He works largely in methodological issues in the design and analysis of clinical trials and observational studies. He is currently focusing on alternative outcome measures in trials with a time-to-event outcome; on problems of model building and validation with survival data, including prognostic factor studies and treatment-covariate interactions; on parametric modeling of survival data; and on novel clinical trial designs.