Lab 5: Transformations and nonlinearity

Data: mri.csv (see the mri.pdf file for data dictionary/useful information).

Practical objective: To gain familiarity with implementation and interpretation of models involving transformations and basis expansions.

Scientific objectives: To explore predictive models for FEV in adults.

Noteworthy commands: Below is a list of Stata commands and options that will be helpful for this lab.

- `mkspline`
- `eform`

Exercises: Below is a set of exercises that we will go through individually, in small groups, and/or together as appropriate and as time permits.

Setup: Consider the following four candidate models for developing FEV reference ranges among adults (I will denote a natural cubic spline on x with K knots at the default four quantiles selected by Stata by $s_K(x)$; further, I will use "$\times$" as shorthand notation for interaction terms).

$$\log(\text{FEV}) = \beta_0 + s_3(\text{age}) \times s_3(\text{pack years}) + \varepsilon \quad [1]$$

$$\text{FEV} = \beta_0 + \beta_1 \, \text{age} + \varepsilon \quad [2]$$

$$\log(\text{FEV}) = \beta_0 + \beta_1 \, \text{male} + s_3(\text{age}) + s_3(\text{pack years}) + \varepsilon \quad [3]$$

$$\text{FEV} = \beta_0 + s_4(\text{age}) + \varepsilon \quad [4]$$

Exercise 1: Rank these models in order of the least to the most degrees of freedom used.

Exercise 2: Consider the following two subgroups:

Group 1: 90 year-old women with a smoking history of 100 pack-years.
Group 2: 75 year-old men with a smoking history of 25 pack-years.

Before fitting the models, explore the data to hypothesize patterns you might expect in comparing the width of 95% prediction intervals across models and between groups. Justify your hypotheses.

Exercise 3: Confirm (or refute) your hypotheses from Exercise 2 by forming each prediction interval, the code for which can be found on the following page.

Exercise 4: Perform diagnostics to evaluate the degree to which you trust Models [1] through [4] to produce reliable prediction intervals for Groups 1 and 2.

Code for Exercise 3:

```
* Import MRI data
import delimited "...mri.csv", clear

* Generate log-transformed FEV
gen logfev = log(fev)

* Flag original observations vs. "prediction" observations
gen flag = 0

set obs `=_N+1'
replace ptid = _N if flag == .
replace flag = 1 if flag == .

set obs `=_N+1'
replace ptid = _N if flag == .
replace flag = 1 if flag == .

* Create observations for two subgroups of interest
replace age = 90 if ptid == _N - 1
replace age = 75 if ptid == _N

replace male = 0 if ptid == _N - 1
replace male = 1 if ptid == _N

replace packyrs = 100 if ptid == _N - 1
replace packyrs = 25 if ptid == _N

* Create natural cubic spline basis expansion
mkspline aspl = age, cubic nknots(3)
mkspline pspl = packyrs, cubic nknots(3)


* Regression Model [1]
regress logfev aspl1-aspl2 pspl1-pspl2 if flag == 0
predict prm1 if ptid >= _N - 1
predict prsd1 if ptid >= _N - 1, stdf
* Mind the degrees of freedom (724 complete observations minus 5 parameters)
local tquantM1 invt(724 - 5, 0.975)
quietly: summarize prm1 if ptid == _N - 1
local fevpr1 = r(mean)
quietly: summarize prsd1 if ptid == _N - 1
local fevsd1 = r(mean)
* Display prediction interval for Group 1 (Model [1])
display "GROUP 1: [" exp(`fevpr1' - `tquantM1' * `fevsd1') ", " exp(`fevpr1' + `tquantM1' * `fevsd1') "]"
quietly: summarize prm1 if ptid == _N
local fevpr1 = r(mean)
quietly: summarize prsd1 if ptid == _N
local fevsd1 = r(mean)
* Display prediction interval for Group 2 (Model [1])
display "GROUP 2: [" exp(`fevpr1' - `tquantM1' * `fevsd1') ", " exp(`fevpr1' + `tquantM1' * `fevsd1') "]"


* Regression Model [2]
regress fev age if flag == 0
predict prm2 if ptid >= _N - 1
predict prsd2 if ptid >= _N - 1, stdf
* Mind the degrees of freedom (725 complete observations minus 2 parameters)
local tquantM2 invt(725 - 2, 0.975)
quietly: summarize prm2 if ptid == _N - 1
local fevpr2 = r(mean)
quietly: summarize prsd2 if ptid == _N - 1
local fevsd2 = r(mean)
* Display prediction interval for Group 1 (Model [2])
display "GROUP 1: [" `fevpr2' - `tquantM2' * `fevsd2' ", "`fevpr2' + `tquantM2' * `fevsd2' "]"
quietly: summarize prm2 if ptid == _N
local fevpr2 = r(mean)
quietly: summarize prsd2 if ptid == _N
local fevsd2 = r(mean)
* Display prediction interval for Group 2 (Model [2])
display "GROUP 2: [" `fevpr2' - `tquantM2' * `fevsd2' ", "`fevpr2' + `tquantM2' * `fevsd2' "]"
```

```
* Regression Model [3]
regress logfev male aspl1-aspl2 pspl1-pspl2 if flag == 0
predict prm3 if ptid >= _N - 1
predict prsd3 if ptid >= _N - 1, stdf
* Mind the degrees of freedom (724 complete observations minus 6 parameters)
local tquantM3 invt(724 - 6, 0.975)
quietly: summarize prm3 if ptid == _N - 1
local fevpr3 = r(mean)
quietly: summarize prsd3 if ptid == _N - 1
local fevsd3 = r(mean)
* Display prediction interval for Group 1 (Model [3])
display "GROUP 1: [" exp(`fevpr3' - `tquantM3' * `fevsd3') ", " exp(`fevpr3' + `tquantM3' * `fevsd3') "]"
quietly: summarize prm3 if ptid == _N
local fevpr3 = r(mean)
quietly: summarize prsd3 if ptid == _N
local fevsd3 = r(mean)
* Display prediction interval for Group 2 (Model [3])
display "GROUP 2: [" exp(`fevpr3' - `tquantM3' * `fevsd3') ", " exp(`fevpr3' + `tquantM3' * `fevsd3') "]"


* Regression Model [4]
regress fev aspl1-aspl2 if flag == 0
predict prm4 if ptid >= _N - 1
predict prsd4 if ptid >= _N - 1, stdf
* Mind the degrees of freedom (725 complete observations minus 3 parameters)
local tquantM4 invt(725 - 3, 0.975)
quietly: summarize prm4 if ptid == _N - 1
local fevpr4 = r(mean)
quietly: summarize prsd4 if ptid == _N - 1
local fevsd4 = r(mean)
* Display prediction interval for Group 1 (Model [4])
display "GROUP 1: [" `fevpr4' - `tquantM4' * `fevsd4' ", "`fevpr4' + `tquantM4' * `fevsd4' "]"
quietly: summarize prm4 if ptid == _N
local fevpr4 = r(mean)
quietly: summarize prsd4 if ptid == _N
local fevsd4 = r(mean)
* Display prediction interval for Group 2 (Model [4])
display "GROUP 2: [" `fevpr4' - `tquantM4' * `fevsd4' ", "`fevpr4' + `tquantM4' * `fevsd4' "]"
```

Code for Exercise 4:

```
drop if flag == 1

* MODEL [1]
regress logfev aspl1-aspl2 pspl1-pspl2
predict stres, rstudent
predict fitted
lowess stres fitted, yline(0)
qnorm stres
drop stres fitted

* MODEL [2]
regress fev age
predict stres, rstudent
predict fitted
lowess stres fitted, yline(0)
qnorm stres
drop stres fitted

* MODEL [3]
regress logfev male aspl1-aspl2 pspl1-pspl2
predict stres, rstudent
predict fitted
lowess stres fitted, yline(0)
qnorm stres
drop stres fitted

* MODEL [4]
regress fev aspl1-aspl2
predict stres, rstudent
predict fitted
lowess stres fitted, yline(0)
qnorm stres
```